ENK-Score: A Hybrid Approach to Essay Scoring via Embeddings, Neural Models and Keywords

Estellyn Hoo Weay1, a), Hui-Ngo Goh1, b), Choo-Yee Ting1, c), Albert Quek1, d) and Amy Hui-Lan Lim1, e)

1Faculty of Computing and Informatics, Multimedia University, 63000 Cyberjaya, Selangor Malaysia

*b) Corresponding author: hngoh@mmu.edu.my  
a) esweay33@gmail.com*

*c) cyting@mmu.edu.my*

*d) quek.albert@mmu.edu.my*

*e) amy.lim@mmu.edu.my*

**Abstract.** This study focuses on the task of automatic essay scoring (AES) using the Automated Student Assessment Prize (ASAP) dataset, specifically SET 4. The ASAP dataset is a well-established benchmark in the field of AES, containing a diverse set of essays with associated prompt-specific scoring criteria. Various classifiers like Random Forest, Logistic Regression, XGBoost, Bi-LSTM, and GRU, and various transformer models like BERT, DistilledBERT, Albert, and XLNet are used for scoring, in combination of variety of features extracted from the essays. It is observable that VERB keyword lists give the best combination of achieving 0.86 QWK score.

# INTRODUCTION

AES has become a pivotal tool in educational assessment, offering efficiency and standardization in evaluating student writing. However, existing AES systems can benefit from increased accuracy and reliability, achievable through integrating advanced natural language processing (NLP) techniques. Conventional hand-grading methods take a lot of time and are prone to prejudice. To solve these problems, AES systems were created, offering scalability and objectivity. From then on, various algorithms and frameworks have been proposed to experiment on the dataset. This project seeks to increase essay scoring accuracy and dependability by utilizing the advancements in NLP. In contributing to the ongoing discourse on the role of AES in education, the utilization of state-of-the-art NLP methodologies has the potential to transform the assessment dynamics for both learners and educators alike.

## Problem Statement

The manual evaluation of essays is expensive in terms of labor-intensive, and it also prone to subjectivity and inconsistent scoring standards. This discrepancy affects students’ academic performance. Furthermore, timely essay evaluation is a known practical challenge in large-scale educational settings, which may cause a delay in student progress assessments and feedback. Below are the formulated research objectives:

* RQ1: What NLP and machine learning techniques can be applied in developing an automated essay scoring system?
* RQ 2: What features play a vital role in deciding how effective and precise automated essay scoring systems are?

In such, this study aims to investigate evaluation of student essays reliably and accurately by removing bias and achieving manual grading guidelines using machine learning and language processing.

# related work

Supervised machine learning and neural network architecture have greatly improved AES in recent years to assess and grade written assignments with similar accuracy to human reviewers.

## Supervised Learning

Multi-task learning (MTL) incorporating a linear regression step is used to improve AES results, especially when data is limited. This technique improves traditional methods by incorporating various data sources [1]. A new grading system for short responses which combines text similarity algorithms with ridge regression, achieving impressive accuracy on various standards [2]. [3] focusing on supervised learning in constructing domain-specific ontologies and Natural Language Toolkit (NLTK) to enhance AES. [4] integrated word embeddings with string kernels, utilizing both syntactic and semantic text features to enhance essay scoring accuracy.

Random Forest Classifier in combination of word selection, structure, and sentence flow are used to performance evaluation [5]. Despite the commonly used performance metrics for essay scoring, [6] introduces a grading system called “goodness” to evaluate words and phrases based on style in determining the essay quality. The proposed grading system seems to perform better on features such as essay statistics, punctuation, complexity, language modelling, coherence on traditional and deep learning models. [7] focuses on augmented narrative essays written by junior high students for automated grading essay using XGBoost classifier. Python packages such as NLTK, StanfordCoreNLP, and spaCy are used for extracting needed features. [8] investigated an automated technique for evaluating the thesis statements in persuasive essays. Towards the end, the author shared the marked dataset for future studies. [9] studied the use of AES technology to pinpoint students in danger of failing English language arts exams. The research discovered that screeners scored by AES were more precise and faster than those based on word count.

## Neural Learning Architecture

In 2018, [10] introduced a convolutional recurrent neural network to improve essays scoring by combining linguistic, cognitive, and psychological elements with standard word and sentence embeddings. [11] presented an automated essay evaluation using a Siamese Bidirectional LSTM to assess both the essay and a reference essay, it shows improved precision as compared to traditional neural networks. [12] suggested a system for automated essay grading that utilizes reinforcement learning and a four-layer dilated LSTM to enhance accuracy in scoring. In their study of student essays, [13] introduced a Hierarchical Attention Network with a Bi-directional LSTM to investigate deep seman tic analysis and identify complex language structures. [14] developed a model for grading essays automatically which utilizes LSTM and self-attention mechanisms to assess coherence by capturing word relationships over long distances. [15] highlighted how crucial it is for automated essay scoring systems to consider the persuasiveness of arguments and introduced a neural network model to evaluate and offer feedback on this aspect of student essays. [16] presented BERT, a model for language representation that trains deep bidirectional representations from unlabeled text, greatly improving performance on various natural language processing tasks.

In 2019, [17] investigated a dual-layer Bidirectional LSTM network in combination of word embeddings on Google News Corpus and Kaggle dataset to improve the precision of automated essay grading, showing great improvement in educational technology. [18] contrasts conventional embedding techniques such as Bag of Words and LSTM with context-based embedding such as BERT and XLNet, to tackle different length of texts. [19] introduced a Two-Stage Learning Framework blending feature-engineered and end-to-end AES techniques. This method involves using LSTM for initial evaluation and then combining it with manually created features to enhance accuracy using an XGBoost model. [20] criticized conventional AES systems that require extensive customize features and focus on the needs to move towards neural network models such as LSTM that allow for improved management of complex language features without the need for extensive feature engineering.

In 2020, [21] suggested a method for grading essays by utilizing neural networks and GloVe embeddings, determining LSTM as the most successful network and noting that the ideal GloVe dimensions differ depending on the type of network. [22] presented R2BERT, which improves Automated Essay Scoring (AES) by utilizing a pre-trained BERT model with regression and ranking losses and illustrated the efficacy of self-attention in text analysis. [23] proposed a combination AES approach blending manual essay-level characteristics with a Deep Neural Network, improving scoring precision with reduced parameters and simplicity of use across various models. [24] created the Prompt Agnostic Essay Scorer (PAES) in 2020, utilizing POS embeddings instead of word embeddings for a broad essay representation and incorporating different non-prompt-specific characteristics to assess essay quality. [25] introduced the Shared and Enhanced Deep Neural Network (SEDNN) for cross-prompt AES. This model differentiates between prompt-independent and prompt-dependent characteristics to improve accuracy by incorporating more target prompt-associated information. [26] investigated BERT’s application in AES, acknowledging its better performance compared to conventional models but raising concerns about the extensive computational resources needed for fine-tuning.

In 2021, [27] presented the Cross-prompt Trait Scorer (CTS) for the Automated Cross-prompt Scoring of Essay Traits task. The model employs a multi-task structure and a trait-focused attention mechanism to forecast overall and specific trait scores for essays using new prompts, underscoring the importance of investigating scoring traits across different prompts to offer in-depth feedback in AES. [28] examined the use of transformer-based models in AES, highlighting the drawbacks of prior approaches and the benefits of transformers, including enhanced language comprehension. Their research indicated that models such as Albert, Electra, and Mobile-BERT surpass traditional BERT models in AES by achieving better results with fewer parameters, highlighting their efficiency and efficacy.

In 2023, [29] suggested an automated essay scoring (AES) system that makes use of SentenceBERT for sentence vectorization in conjunction with a DNN model. This approach combines surface-level language characteristics and unique prompts to improve semantic analysis, overcoming drawbacks of conventional AES models and DNN-based models. [30] examined how reliable the GPT-3 model is in AES and explored how linguistic factors impact scoring. Their research, which analyzed 12,100 essays, showed that GPT-3 offers precise and dependable ratings, backing its integration with human assessment. [31] investigated how scores from different AES models can be combined using Item Response Theory (IRT) to address variations in scoring specific to each model. Their method represents a change towards mixed models combining manual and automatic feature extraction, intending to improve the precision and efficiency of AES systems by merging predictions from various models.

# METHODOLOgy

There are 8 groups of essays written by students in Grades 7 to 10 for ASAP dataset and set 4 was chosen due it token length to fit language models such as Bert. Write 4 essays, each with an average of 150 words, to fit within Bert’s 512-token limit. This is done to make sure the models can fully analyze the text by choosing essays with word counts that fit within the token limits. This ensures the automated scoring process remains accurate and unbiased by avoiding any potential errors from incomplete assessments.

## Feature Selection

The quality and readability of student essays are evaluated by systematically analyzing their linguistic and structural elements. Different tools and techniques in NLP are used to extract and measure attributes from the text. These characteristics fall under five types of features. Every category covers various parts of the text, giving complete perspective on the quality of the writing as shown in Table 1 and experimented with various training models.

**TABLE 1**. Features extracted

|  |  |
| --- | --- |
| Feature Type | Features |
| Length-based | Numbers of words, sentences, lemmas, punctuation count, Average lengths of words and sentences, Unique word count, and Count of narrative and dialogue |
| Syntactic | Numbers of nouns, verbs, adverbs, adjectives, and conjunctions |
| Word-based | Numbers of spelling errors, grammar errors and stop-words, Percentage of misspelling and grammar error |
| Semantic | Polarity and Subjective |
| Readability | Automated readability index, Coleman–Liau index, Dale–Chall readability score, Difficult word count, Flesch reading ease, Flesch–Kincaid grade, Gunning fog, Linsear write formula, SMOG index, Syllable count |

## Keyword List Extraction

In this study, the keyword list features are proposed to be integrated with the model. This is especially accurate in tasks such as automated essay scoring, as the capability to evaluate content quality effectively influences the dependability of the outcomes. To improve this ability, incorporating a keyword list into the model is a major advancement. This function includes identifying and evaluating certain terms that signal the quality and pertinence of the content to the scoring standards. The scoring of essays could be impacted via writing styles and usage of keywords. Each essay is tagged with part-of-speech (POS) to identify how words are used in an essay depending on their context and words are sorted based on their POS tags. After that, words are stemmed to obtain their basic forms; this is to combine different grammatical variations. Finally, the top 10 words for each score is produced.

## Large Language Model

This work tests four transformer-based models, BERT (base), XLNet, DistilBERT, and Albert (base). BERT (base) heavily utilizes self-attention mechanisms to create a representation of the data input. It considers contextual relationships between words in an essay bi-directional thru a technique called “Masked Language Model” which words are randomly masked during training, prompting the model to predict the masked words according to their surrounding context.

XLNet is an extension of BERT focusing on utilization of permutation-based strategy. It considers all possible arrangements of word order in a sentence using a method called “Permutation Language Modeling” to understand context in both directions and grasp relationships among every word pair in the input sequence.

DistillBERT is a refined version of BERT that includes knowledge distillation in its training process. It is a smaller model to replicate the actions of an existing or larger model by having fewer parameters, layers, and attention heads, resulting in lower computational needs.

Albert(base) introduces two main changes to BERT’s architecture which are reducing parameters and sharing parameters across layers. The embedding matrix is split into two smaller matrices and sharing parameters across all model layers. This is known as cross-layer parameter sharing. It decrease total memory usage and speed up the training process.

## Learning Model

A combination of BERT embeddings with three learning models of Multinomial Logistic Regression, Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks respectively is used to investigate their abilities in handling language data.

Multinomial Logistic Regression is a complex version of logistic regression, suitable when the outcome variable is categorical and multi-class. This technique is crucial for representing the likelihood distribution of categorical results using predictor variables. The central component of this model is the softmax function, also known as the normalized exponential function, designed for handling multiple categories.

GRU is a subset of recurrent neural networks that have fewer parameters, making them simpler and typically faster to train, yet still able to effectively manage dependencies in sequential data. A GRU’s structure consists of two important elements which are the update gate and the reset gate. The update gate decides the influence of past information on future states by determining how much of the past should be carried forward, while the reset gate is focusing on the amount of old information to delete, assist in controlling information flow and improving the network’s ability to adjust to changes in the input data pattern. That is why GRU is capable to adapt different time dependencies in sequences.

Bi-LSTM examines each training sequence in bi-directions using two different types of recurrent neural networks. Both networks are linked to the identical output layer. The two networks, with one handling the sequence from beginning to end and the other handling it from end to beginning, making the prediction more accurate.

# Experimental results

Quadratic weighted kappa (QWK) score is a measure of agreement between two raters for categorical items. It takes into consideration the potential for random agreement, differentiating between expected and observed agreements. This data uses squared variances to give more importance to differences in ratings, penalizing bigger disagreements between raters more than smaller ones. Ablation studies have been carried out across set of features and learning models.

Ablation studies have been carried out across a set of features and learning models, however, only the top performances are extracted to be shown as in Table 2. Each model was evaluated using QWK to determine alignment with human scoring in providing crucial understanding of how effective and applicable these models are for tasks in natural language processing.

A combination of BERT as word embedding with GRU and Bi-LSTM respectively achieved the same QWK score of 0.84, showing their capability in processing sequential data effectively with additional features like word count and readability indices that capture essay content complexity and style.

BERT as transformer is applied to different sets of features. It achieved a QWK score of 0.84, utilizing a variety of features such as unique word count, total word count, readability metrics such as punctuation count and syllables count, and the best QWK score of 0.86 focusing only on a list of verb keywords with an imbalanced dataset. This implies specific keyword characteristics closely match the human scoring rating, and BERT is capable of distinguishing between different levels of writing styles related using verbs in capturing the essay context.

**TABLE 2.** Comparison of model, features and performance

|  |  |  |
| --- | --- | --- |
| **Model** | **Features** | **QWK** |
| XLNET | Essay, Noun count, Verb count, Adverb count, Adjective count | 0.81 |
| ALBERT | Essay, Percentage of grammar mistakes, Noun count, Verb count, Adverb count, Word count | 0.80 |
| BERT | Essay, Unique wordcount, Word count, Percentage of grammar mistakes, Punctuation count, Syllables count, ARI | 0.84 |
| Logistic Regression | Bert Embedding, Word count, Lemma Count, Unique Words Count, Noun count, Stop Word Count, Syllables Count, SMOG Index, Difficult Word Count, Sentence Count | 0.80 |
| GRU | Bert Embedding, Word count, Lemma count, Unique word count, SMOG Index | 0.84 |
| Bi-LSTM | Bert Embedding, Word count, Lemma count, Unique word count, Punctuation count, ARI | 0.84 |
| BERT | **Essay, Verb Keywords List** | **0.86** |

# conclusion

This work achieved its 2 objectives of (1) to experiment with various ML models that can thoroughly analyze and evaluate student essays, and (2) to analyze different features of essays to measure and compare with manual grading.

It is observable that BERT, supplemented with a verb keyword compilation top in terms of its precision and aligned with human assessment criteria. It shows choosing the right features to be blended with the good model will improve performance results.

For future works, perhaps, we should look into other sets in the ASAP dataset to validate the best feature of verb keywords and other transformation models accordingly.

# References

1. R. Cummins, M. Zhang, and T. Briscoe, “Constrained multi-task learning for automated essay scoring,” Association for Computational Linguistics (2016).
2. M. A. Sultan, C. Salazar, and T. Sumner, “Fast and easy short answer grading with high accuracy,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016), pp. 1070–1075.
3. J. O. Contreras, S. Hilles, and Z. B. Abubakar, “Automated essay scoring with ontology based on text mining and NLTK tools,” in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)* (2018), pp. 1–6.
4. M. Cozma, A. M. Butnaru, and R. T. Ionescu, “Automated essay scoring with string kernels and word embeddings,” *arXiv preprint* arXiv:1804.07954 (2018).
5. S. Mathias and P. Bhattacharyya, “ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018).
6. S. Mathias and P. Bhattacharyya, “Thank ‘goodness’! A way to measure style in student essays,” in *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (2018), pp. 35–41.
7. Y. Salim, V. Stevanus, E. Barlian, A. C. Sari, and D. Suhartono, “Automated English digital essay grader using machine learning,” in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)* (IEEE, 2019), pp. 1–6.
8. Z. Ke, H. Inamdar, H. Lin, and V. Ng, “Give me more feedback II: Annotating thesis strength and related attributes in student essays,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Marquez, Eds. (Association for Computational Linguistics, Florence, Italy, 2019), pp. 3994–4004. Available: https://aclanthology.org/P19-1390
9. J. Wilson and J. Rodrigues, “Classification accuracy and efficiency of writing screening using automated essay scoring,” *Journal of School Psychology* **82**, 123–140 (2020).
10. T. Dasgupta, A. Naskar, L. Dey, and R. Saha, “Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring,” in *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (2018), pp. 93–102.
11. G. Liang, B.-W. On, D. Jeong, H.-C. Kim, and G. S. Choi, “Automated essay scoring: A siamese bidirectional LSTM neural network architecture,” *Symmetry* **10**, 682 (2018).
12. Y. Wang, Z. Wei, Y. Zhou, and X.-J. Huang, “Automatic essay scoring incorporating rating schema via reinforcement learning,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), pp. 791–797.
13. Z. Wang, J. Liu, and R. Dong, “Intelligent auto-grading system,” in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)* (IEEE, 2018), pp. 430–435.
14. X. Li, M. Chen, J. Nie, Z. Liu, Z. Feng, and Y. Cai, “Coherence-based automated essay scoring using self-attention,” in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: Proceedings of the 17th China National Conference on Computational Linguistics (CCL 2018) and the 6th International Symposium, NLP-NABD 2018* (Springer, Changsha, China, 2018), pp. 386–397.
15. Z. Ke, W. Carlile, N. Gurrapadi, and V. Ng, “Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)* (2018), pp. 4130–4136.
16. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint* arXiv:1810.04805 (2018).
17. L. Xia, J. Liu, and Z. Zhang, “Automatic essay scoring model based on two-layer bi-directional long-short term memory network,” in *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence* (2019), pp. 133–137.
18. P. U. Rodriguez, A. Jafari, and C. M. Ormerod, “Language models and automated essay scoring,” *arXiv preprint* arXiv:1909.09482 (2019).
19. J. Liu, Y. Xu, and Y. Zhu, “Automated essay scoring based on two-stage learning,” *arXiv preprint* arXiv:1901.07744 (2019).
20. F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, “Automated essay scoring with discourse-aware neural models,” in *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (2019), pp. 484–493.
21. W. Zhu and Y. Sun, “Automated essay scoring system using multi-model machine learning,” in *CS & IT Conference Proceedings* **10**, 12 (2020).
22. R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, “Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking,” in *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020), pp. 1560–1569.
23. M. Uto, Y. Xie, and M. Ueno, “Neural automated essay scoring incorporating handcrafted features,” in *Proceedings of the 28th International Conference on Computational Linguistics* (2020), pp. 6077–6088.
24. R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, “Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring,” *arXiv preprint* arXiv:2008.01441 (2020).
25. X. Li, M. Chen, and J.-Y. Nie, “SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring,” *Knowledge-Based Systems* **210**, 106491 (2020).
26. E. Mayfield and A. W. Black, “Should you fine-tune BERT for automated essay scoring?” in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (2020), pp. 151–162.
27. R. Ridley, L. He, X.-Y. Dai, S. Huang, and J. Chen, “Automated cross-prompt scoring of essay traits,” in *Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 13745–13753 (2021).
28. C. M. Ormerod, A. Malhotra, and A. Jafari, “Automated essay scoring using efficient transformer-based language models,” *arXiv preprint* arXiv:2102.13136 (2021).
29. F. Li, X. Xi, Z. Cui, D. Li, and W. Zeng, “Automatic essay scoring method based on multi-scale features,” *Applied Sciences* **13**, 6775 (2023).
30. A. Mizumoto and M. Eguchi, “Exploring the potential of using an AI language model for automated essay scoring,” *Research Methods in Applied Linguistics* **2**, 100050 (2023).
31. M. Uto, I. Aomi, E. Tsutsumi, and M. Ueno, “Integration of prediction scores from various automated essay scoring models using item response theory,” *IEEE Transactions on Learning Technologies* (2023).