Feature Based Classification for Automated Essay Scoring

Muhaimin Hading[a], Muhammad Ikhwan Burhan[b], Andi Nurfadillah Ali[c] ,
A. Inayah Auliyah[d], Mardhiyyah Rafrin[e], Arliyanti Nurdin[f], Wiwit Melayu[g],
[a,b,c,d,e]Institut Teknologi Bacharuddin Jusuf Habibie
[f]Universitas Hasanuddin
[g]Universitas Pendidikan Indonesia
Corresponding Author:
[a]hading.muhaimin@ith.ac.id

ABSTRACT
Essays play a crucial role in traditional assessments, but evaluating them accurately, efficiently, and fairly poses a major challenge for educators. Automated Essay Scoring (AES) aims to address this issue by leveraging computational techniques to support teachers in the grading process. This study explores a classification model to classify the score based on the feature we created. We incorporate additional features aligned with the ASAP 2.0 scoring rubric, such as Lexical Sophistication, Source Adherence, Novelty and Relevance, and a Semantic Disruption feature. These features are used to construct a distributed representation of essays, which is then input into a Support Vector Machine (SVM) model for holistic score prediction. The proposed model achieved a Quadratic Weighted Kappa (QWK) score of 0.8397, indicating a high level of agreement with human raters. The results demonstrate the effectiveness of combining rubric-informed features with a non-linear classifier. The findings can be implemented for educational settings, where the model can be utilized to provide scalable and consistent scoring support, reduce grading workload for instructors, and deliver timely feedback to students. By aligning with rubric-based criteria, the approach can also foster more transparent and constructive learning processes, helping students identify specific areas for improvement in their writing. While the model exhibits strong predictive performance, it also presents limitations related to interpretability and generalizability, especially across diverse writing prompts and domains.

Keywords : Automated Essay Scoring (AES), ASAP 2.0, Support Vector Machine (SVM)

INTRODUCTION

Writing skill is one of the important indicators in evaluating English language proficiency (Wilson & Shermis, 2024). However, the current evaluation system has relied on human ratings. As the number of students continues to grow then the ratio between teacher and student becomes imbalanced, making manual assessment has become impractical. The manual evaluation of student essays is time consuming, lacks subjectivity, inconsistencies indicators and has a bias potential (Ramesh & Sanampudi, 2022). To address these challenges, Automated Essay Scoring (AES) has become a solution. AES is an application of Natural Language Processing (NLP) that automatically assigns a holistic score to a student's essay based on its overall quality (Bai et al., 2022). It offers several advantages, such as a quick review, consistency, objectivity, and scalability.

Research in Automated Essay Scoring (AES) is divided into two categories. The first type is closed-ended essays, which are typically in the form of short answers or question-response formats (Asto Buditjahjanto et al., 2022). Scoring in this context involves measuring the semantic similarity between the student response and the expected answer. Common approaches include semantic algorithms (Abdul Salam et al., 2022; Ayaan & Ng, 2025), word embedding techniques (Lubis et al., 2021) and neural network models (Asto Buditjahjanto et al., 2022; Lubis et al., 2021). These methods perform well if the possible answers are limited and highly structured.

The second category is open-ended essays, which include persuasive, narrative, or argumentative text (Zhang & Litman, 2020). These essays require students to construct original content involving critical thinking and rhetorical strategies. Recent studies in this area have investigated deep learning models (Faseeh et al., 2024; Mesgar & Strube, 2018; Nadeem et al., 2019), trait-based neural networks (He et al., 2022), transformer-based models (Ludwig et al., 2021; Reddy Chavva et al., 2024), and large language models (LLMs) (Dini et al., 2025; Shen et al., 2021; Yancey et al., 2023; Yang et al., 2020). While these models show strong performance, they often suffer from limited interpretability, high computational demands, and a lack of explicit alignment with human scoring rubrics.

To address these limitations, we propose a feature-based classification approach for evaluating English open-ended essays in ASAP dataset. Unlike LLMs, which often operate as black boxes, feature-based methods explicitly capture measurable linguistic, syntactic, and discourse-level characteristics that can be directly mapped to scoring rubrics. This not only enhances interpretability but also provides educators and researchers with actionable insights into why a certain score was assigned. Furthermore, feature-based systems are computationally less demanding, making them more accessible for large-scale deployment in educational settings where resources may be limited. Most importantly, they allow for the integration of domain-specific scoring dimensions (e.g., grammar, coherence, or argument strength), ensuring fairness and consistency in assessment, which remains a critical challenge for purely LLM-driven approaches. Additionally, it aligns with the ASAP 2.0 scoring rubric to ensure consistency with human scoring standards. Through this approach, we aim to advance the understanding of how feature-based methods can classify the essay into level of score.

LITERATURE REVIEW

Automated Essay Scoring (AES) has various methodologies being developed to predict scores. These range from traditional machine learning techniques to deep learning approaches and the large language models (LLMs). Large Language Models (LLMs) have capabilities in understanding and generating text. Recent studies have investigated the use of models such as GPT or T5 for evaluating coherence, relevance, semantic similarity and overall quality of essays (Do et al., 2024; Pack et al., 2024; Yancey et al., 2023). But LLMs have limitations, they require substantial computational resources for training and deploying via APIs is quite expensive.

On the other hand, Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) have been widely used for AES due to their ability to model sequential data and capture the dependencies in text. LSTM models can analyze the flow of information across sentences

and paragraphs, making them suitable for assessing coherence and syntactic structure (Liang et al., 2018; Ramesh & Sanampudi, 2022). However, they often struggle to retain long-range dependencies. Transformer-based architectures like BERT have been used to model contextual and semantic relationships in text. BERT has been shown to perform well in capturing semantic similarity, coherence, and argument strength (Wang et al., 2022). Nevertheless, BERT and similar transformer models are difficult to explain the basis of their scoring decisions.
Another approach is hybrid method (Faseeh et al., 2024; Li & Ng, 2024) that combines various neural networks with handcrafted features like lexical diversity and grammatical or syntactical error. For instance, Fashee et al. proposed a model that integrates vector-based handcrafted features (including lexical, syntactic, and readability features) with deep neural network representations. We adopt a similar approach in this work, using handcrafted features in the Support Vector Machine (SVM) model. These include some features from Fashee's work and we add additional features like Lexical Sophistication, Source Adherence, Novelty and Relevance, and a Semantic Disruption feature. This approach enables the model to capture both surface-level and deeper semantic characteristics of the essay while maintaining interpretability and computational efficiency.

METHODS

Dataset

We conducted our experiment on ASAP 2.0 (Automated Student Assessment Prize) both for training and evaluation. The corpus consists of 24,278 persuasive essays collected from students in grades 6, 8, 9, and 10. All the essays in this corpus are source-based, where students were required to read information from provided source texts into their essays. There were seven prompts in total as we can see the distribution of the data on table 1. Each essay was scored holistically using a standardized rubric from Scholastic Aptitude Test (SAT). The rubric used a 1-6 scale that had interval levels. A score of 6 indicating an effective point of view, outstanding critical thinking, the use of clear examples and reasons, and appropriate evidence from sources. Linguistically, high-scored essays showed strong organization and cohesion as well as the skillful use of language including vocabulary, sentence structure, and grammar and mechanics.

Table 1. Data Distribution

Prompt Name	Score						Total Essays
	1	2	3	4	5	6	
A Cowboy Who Rode the Waves	218	913	827	206	11	0	2175
Car-free cities	139	412	701	539	160	8	1959
Does the electoral college work?	191	513	674	466	170	32	2046
Driverless cars	163	1279	2355	1917	435	21	6170
Exploring Venus	567	1419	1469	808	175	42	4480
Facial action coding system	220	1253	1905	1120	305	80	4883
The Face on Mars	253	1058	1090	497	100	17	3015

Total	**1751**	**6847**	**9021**	**5553**	**1356**	**200**	**24728**

Balancing Dataset

In order to prepare the training and evaluation data, we removed all data associated with copyright-restricted source texts, as the source text information was essential for computing semantic similarity between the student essays and source text. As a result, the dataset was reduced from 24.278 essays to 19.395. The second step, we filtered essays based on length, retaining only those between 150 and 550 words length. This was done to ensure a more generalized dataset by excluding essays that were either too short or excessively long. As a result, the dataset was further reduced from 19.395 to 17.728 essays.

After applying the previous filtering steps, we observed that only five essays remained with a score of 6. Since this small number of examples would not provide sufficient representation for effective training and could potentially introduce noise or class imbalance, we chose to remove them. This led to a final dataset consisting of 17.723 essays, with scores ranging from 1 to 5. Then our new data distribution was as follows: score 1 had 1.493 samples, score 2 had 5.531, score 3 had 6.817, score 4 had 3.569, and score 5 had only 313 samples.

Since our dataset displayed a significant class imbalance between the classes, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to augment the minority classes (Mualfah et al., 2022). After augmentation, the new class distribution became more balanced, particularly for classes 1 and 5, which increased to 3.400 and 3.800 samples respectively, while the sample sizes for the other classes remained unchanged.

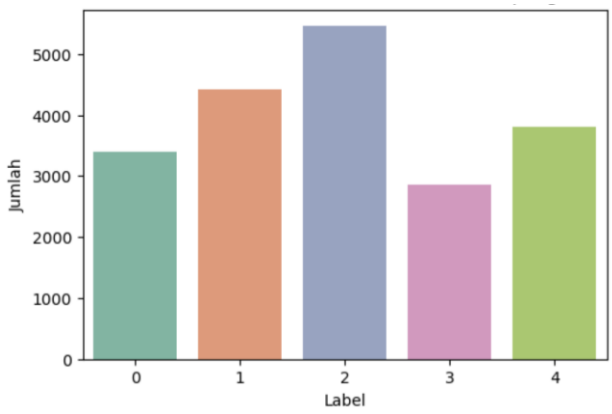

Figure 1. Dataset distribution after SMOTE

Feature Extraction

We extract several features to assess essay quality. First, we incorporate some linguistic features. The features include length based features and syntactic features (Faseeh et al., 2024). We further designed a set of additional features grounded in the assessment rubric of the ASAP 2.0 dataset. First, the lexical sophistication feature aims to capture the writer's lexical maturity. Second, the source adherence feature evaluates students' essays aligns with the content and information provided by the source text. Third, the novelty and relevance feature assesses the uniqueness of an essay relative to other responses to the same prompt, while ensuring that the content remains relevant to the assigned task. Lastly, the semantic disruption feature quantifies

the degree of meaning distortion at the sentence level caused by resulting from grammatical or syntactic errors.

For the features adapted from Uto's work, we utilized the NLTK and spaCy libraries to perform text tokenization, lemmatization, part-of-speech tagging, and stop word identification. For the rubric-based features, lexical sophistication was assessed by first calculating the number of words in the student essay that also appear in the source text, excluding stop words. Then the next step, we identified advanced vocabulary among the content words, where advanced vocabulary is defined as words not included in the Oxford 3000 list. To measure source adherence, we calculated the cosine similarity between the essay and the source text. Both texts were converted into fixed-length vector representations using the Sentence-BERT model (all-MiniLM-L6-v2) provided by Hugging Face. The cosine similarity score, computed using Equation 1, quantifies the degree of semantic alignment (S) between the essay $(\underline{A})$and the source text $(\underline{B})$. The process for computing source adherence is illustrated in Figure 2.

$$S = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

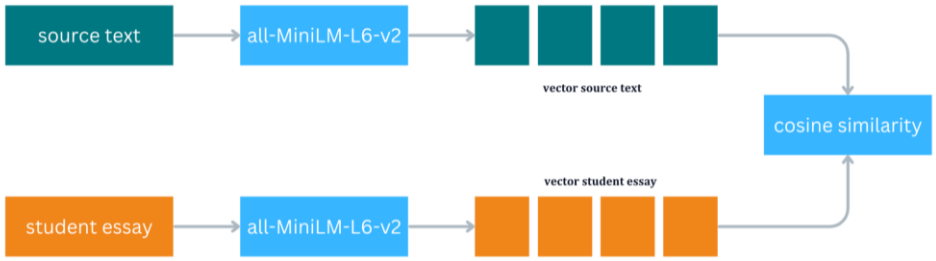

Figure 2. The flow to compute the similarity between source text and essay

For the novelty and relevance feature, we calculated the mean semantic distance (μ_x) of each essay $(x_i)$from the centroid embedding of all essays (n) responding to the same prompt.

$$\mu_x = \frac{\sum_{i=1}^{n} x_i}{n}$$

Additionally, we computed the cosine similarity between the essay and the assignment prompt to ensure that the content remains relevant to the given task.

For the semantic disruption feature, we first identified grammatical and syntactic errors using the LanguageTool Python library. The original essay $(\underline{A})$ and its corrected version $(\underline{B})$ were then converted into vector representations. Cosine similarity (S) was calculated between the two vectors to assess the semantic shift introduced by the errors. If the similarity score (S) was below 0.90, the instance was considered to exhibit semantic disruption due to grammatical or syntactic errors. The illustrated process of computing the semantic disruption feature is presented in Figure 3.

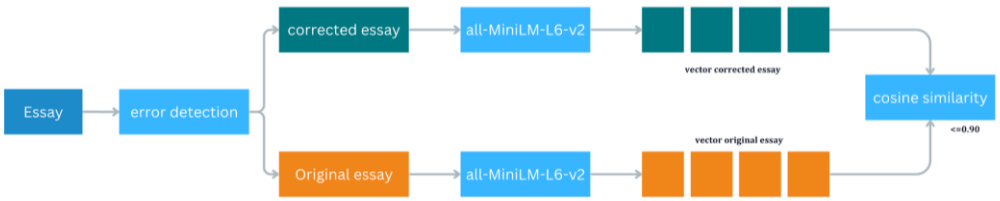

Figure 3. The flow to compute semantic disruption feature

The details of the feature are listed in Table 2. In total, fourteen features were developed for training the model.

Table 2. Feature List

Source Feature	Feature Name	Detail
Uto Feature (Fashee et .al, 2024)	Length based feature	Number of words
		Number of sentence
		Number of lemmas
	Syntactic feature	Number of nouns
		Number of verbs
		Number of Adjectives
		Number of adverbs
		Number of conjunctions
Rubric-based Feature	Lexical sophistication feature	Number of content words
		Number of advance vocabulary
	Source adherence feature	Cosine similarity between essay and source text
	Novelty and relevance feature	Number of uniqueness (mean distance)
		Cosine similarity between essay and assignment
	Semantic disruption feature	Cosine similarity before and after grammar and syntax error corrected in the sentences

Model

In this study, we divided the dataset into training and evaluation. The models were trained using fourteen features that capture various aspects of the essay content. Among the models employed, we utilized the Support Vector Machine (SVM), an algorithm that has demonstrated strong performance in classification tasks. SVM works by identifying an optimal hyperplane that maximizes the margin between data points of different classes, thereby enhancing the model's generalization ability. To handle non-linearly separable data, we employed the Radial Basis Function (RBF) kernel. The RBF kernel is a powerful and widely used kernel in SVM that enables the algorithm to map input data into a higher-dimensional feature space. This transformation allows the model to find a linear separating hyperplane in the transformed space, even when the original data is not linearly separable. The RBF kernel computes

similarity between data points $K(X_1, X_2)$ using a Gaussian function, which depends on the squared Euclidean distance between points and a kernel parameter that controls the width of the Gaussian.

$$K(X_1, X_2) = exp\left(-\frac{|X_1 - X_2|^2}{2\sigma^2}\right)$$

RESULTS AND DISCUSSION

To evaluate the performance, we used Quadratic Weighted Kappa (QWK) as the evaluation metric. QWK is particularly suitable for tasks involving ordinal classification, where the labels represent ordered categories. QWK takes into account the degree of disagreement between predicted and the actual scores. It penalizes larger discrepancies more heavily than smaller ones. For example, predicting a score 5 instead of a score 4 is less severe than predicting a score 5 instead of a score 1.

$$QWK = 1 - \frac{\Sigma_{i,j} W_{i,j} O_{i,j}}{\Sigma_{i,j} W_{i,j} E_{i,j}}$$

The Support Vector Machine (SVM) model achieved a Quadratic Weighted Kappa (QWK) score of **0.8397**, indicating a high level of agreement with the human-assigned essay scores. This result demonstrates that SVM is highly effective in modeling the ordinal nature of the scoring task. The strong performance suggests that the features used in the model were informative and allowed SVM to successfully separate the different classes in a high-dimensional space.

When compared with prior studies that employed deep learning approaches such as Long Short-Term Memory (LSTM) networks, Bidirectional Encoder Representations from Transformers (BERT), or hybrid architectures combining neural and feature-based methods, the performance of the SVM model is competitive. For instance, LSTM-based models typically achieve QWK scores in the range of 0.8489(Ramesh & Sanampudi, 2022), reflecting their capacity to capture sequential dependencies in text. Transformer-based models such as BERT and its variants have reported stronger performance, often exceeding 0.781–0.847 (Wang et al., 2022) in certain datasets, but at the expense of substantially higher computational requirements and reduced interpretability. Hybrid methods that integrate handcrafted features with neural embeddings have reached 0.801 (Uto et al., 2020), but they tend to be more complex to implement. Thus, the SVM's performance of 0.8397 highlights its balance between accuracy, efficiency, and interpretability.

The implications of this approach for education are substantial. Automated essay scoring systems based on feature-based models like SVM can assist educators by providing consistent, efficient, and scalable assessments, especially in large classroom settings where manual grading is impractical. By aligning with established scoring rubrics, such systems can deliver timely feedback to students, enabling them to better understand their strengths and weaknesses in writing. Furthermore, feature-based models can offer interpretable feedback by pinpointing specific linguistic or structural elements that influenced the score, thereby serving as a valuable learning tool rather than merely a grading mechanism. This dual role of assessment and

feedback has the potential to enhance writing instruction and foster the development of critical language skills among students.

CONCLUSION

The experimental results demonstrate that the Support Vector Machine (SVM) model, when combined with both linguistic features and rubric-based features, can effectively predict holistic essay scores. Achieving a Quadratic Weighted Kappa (QWK) score of 0.8394, the model shows a high level of agreement with human raters, confirming its capability in handling the ordinal nature of essay scoring. The use of a non-linear kernel (RBF) enabled the model to capture complex relationships in the feature space, while the margin-maximizing property of SVM contributed to its strong generalization and predictive accuracy.

For future work, several directions can be pursued to extend this study. First, the proposed feature-based classification approach can be validated on additional datasets beyond ASAP to examine its robustness across diverse essay prompts, genres, and writing styles. Second, a systematic comparison with state-of-the-art large language models (LLMs) would provide deeper insights into the trade-offs between performance, interpretability, and computational efficiency. Finally, hybrid approaches that integrate explicit linguistic features with contextual embeddings from transformer models may offer an effective compromise, combining the transparency of feature-based systems with the semantic depth of neural architectures.

LIMITATION

Despite the performance, several limitations should be acknowledged. First, while SVM performs well with carefully engineered features, it lacks the capacity to automatically learn hierarchical representations from raw text as deep learning models do. This may limit its adaptability to more complex scoring criteria or unstructured inputs. Second, the model's interpretability is constrained when using non-linear kernels, making it more challenging to provide feedback or explain score assignments to users.

REFERENCES

Abdul Salam, M., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. *PLOS ONE*, *17*(8), e0272269. https://doi.org/10.1371/journal.pone.0272269

Asto Buditjahjanto, I. G. P., Idhom, M., Munoto, M., & Samani, M. (2022). An Automated Essay Scoring Based on Neural Networks to Predict and Classify Competence of Examinees in Community Academy. *TEM Journal*, 1694–1701. https://doi.org/10.18421/TEM114-34

Ayaan, A., & Ng, K.-W. (2025). Automated grading using natural language processing and semantic analysis. *MethodsX*, *14*, 103395. https://doi.org/10.1016/j.mex.2025.103395

Bai, J. Y. H., Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Cefa Sari, B., & Marin, V. I. (2022, September). Automated Essay Scoring (AES) Systems: Opportunities and Challenges for Open and Distance Education. *Tenth Pan-Commonwealth Forum on Open Learning*. https://doi.org/10.56059/pcf10.8339

Dini, L., Brunato, D., Dell'Orletta, F., & Caselli, T. (2025). TEXT-CAKE: Challenging Language Models on Local Text Coherence. In O. Rambow, L. Wanner, M. Apidianaki,

H. Al-Khalifa, B. Di Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 4384–4398). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.296/

Do, H., Kim, Y., & Lee, G. (2024). Autoregressive Score Generation for Multi-trait Essay Scoring. In Y. Graham & M. Purver (Eds.), *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 1659–1666). Association for Computational Linguistics. https://aclanthology.org/2024.findings-eacl.115/

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y.-I. (2024). Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics*, *12*(21), 3416. https://doi.org/10.3390/math12213416

He, Y., Jiang, F., Chu, X., & Li, P. (2022). Automated Chinese Essay Scoring from Multiple Traits. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3007–3016). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.266/

Li, S., & Ng, V. (2024). Automated Essay Scoring: A Reflection on the State of the Art. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17876–17888. https://doi.org/10.18653/v1/2024.emnlp-main.991

Liang, G., On, B.-W., Jeong, D., Kim, H.-C., & Choi, G. S. (2018). Automated Essay Scoring: A Siamese Bidirectional LSTM Neural Network Architecture. *Symmetry*, *10*(12), 682. https://doi.org/10.3390/sym10120682

Lubis, F. F., Mutaqin, M., Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A. A., & Rosmansyah, Y. (2021). Automated Short-Answer Grading using Semantic Similarity based on Word Embedding. *International Journal of Technology*, *12*(3), 571. https://doi.org/10.14716/ijtech.v12i3.4651

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated Essay Scoring Using Transformer Models. *Psych*, *3*(4), 897–915. https://doi.org/10.3390/psych3040056

Mesgar, M., & Strube, M. (2018). A Neural Local Coherence Model for Text Quality Assessment. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4328–4339). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1464

Mualfah, D., Fadila, W., & Firdaus, R. (2022). Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest. *Jurnal CoSciTech (Computer Science and Information Technology)*, *3*(2), 107–113. https://doi.org/10.37859/coscitech.v3i2.3912

Nadeem, F., Nguyen, H., Liu, Y., & Ostendorf, M. (2019). Automated Essay Scoring with Discourse-Aware Neural Models. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, & T. Zesch (Eds.), *Proceedings of the Fourteenth Workshop on*

Innovative Use of NLP for Building Educational Applications (pp. 484–493). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4450

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, *6*, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Ramesh, D., & Sanampudi, S. K. (2022). *An Improved Approach for Automated Essay Scoring with LSTM and Word Embedding* (pp. 35–41). https://doi.org/10.1007/978-981-16-6616-2_4

Reddy Chavva, R. K., Reddy Muthyam, S., Seelam, M. S., & Nalliboina, N. (2024). A Transformer-Based Approach for Enhancing Automated Essay Scoring. *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, 1–6. https://doi.org/10.1109/ACET61898.2024.10730000

Shen, A., Mistica, M., Salehi, B., Li, H., Baldwin, T., & Qi, J. (2021). Evaluating Document Coherence Modeling. *Transactions of the Association for Computational Linguistics*, *9*, 621–640. https://doi.org/10.1162/tacl_a_00388

Uto, M., Xie, Y., & Ueno, M. (2020). Neural Automated Essay Scoring Incorporating Handcrafted Features. *Proceedings of the 28th International Conference on Computational Linguistics*, 6077–6088. https://doi.org/10.18653/v1/2020.coling-main.535

Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3416–3425. https://doi.org/10.18653/v1/2022.naacl-main.249

Wilson, J., & Shermis, M. (2024). *The Routledge International Handbook of Automated Essay Evaluation*. Routledge.

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating Short L2 Essays on the CEFR Scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020). Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560–1569). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.141

Zhang, H., & Litman, D. (2020). Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8569–8584. https://doi.org/10.18653/v1/2020.acl-main.759