Molecular & Cellular Biomechanics 2025, 22(4), 1525.
https://doi.org/10.62617/mcb1525

Article

The biomechanics-inspired application of AI technology in English essay correction

Jinsheng Wang

General Education College, Liuzhou Polytechnic University, Liuzhou 545006, China; lancaster2008@163.com

CITATION

Wang J. The biomechanics-inspired application of AI technology in English essay correction. Molecular & Cellular Biomechanics. 2025; 22(4): 1525. https://doi.org/10.62617/mcb1525

ARTICLE INFO

Received: 10 February 2025
Accepted: 20 February 2025
Available online: 4 March 2025

Abstract: This paper explores the application of AI technology in the field of English essay grading, inspired by biomechanics. Biomechanics, which studies the mechanical aspects of biological systems, offers unique insights that can be analogously applied to the grading of English compositions. Just as biomechanics analyzes the complex structures and functions of biological entities by understanding the relationships between different components, we focus on natural language processing (NLP) and machine learning algorithms, with the primary objective is to analyze how these advanced technologies, inspired by biomechanical concepts, can enhance the accuracy, efficiency, and objectivity of grading English compositions. By employing various NLP techniques such as lexical analysis, syntactic parsing, and semantic understanding, combined with machine learning models for classification and regression, the study demonstrates significant improvements in grading performance. The findings indicate that AI-powered systems, inspired by biomechanics, can provide consistent and reliable evaluations, thus offering valuable support to educators and students alike.

Keywords: AI technology; English essay grading; biomechanics; natural language processing; machine learning; educational technology

1. Introduction

1.1. Background

In recent years, the application of artificial intelligence (AI) in various fields has seen significant advancements. Within the educational sector, AI technologies, particularly NLP and machine learning, have shown great promise in automating and enhancing various tasks. One such task is the grading of English essays, which traditionally requires substantial time and effort from educators.

1.2. Importance and significance

The use of AI in essay grading is not only a technological breakthrough but also has important implications for education. It can help reduce the workload of teachers, allowing them to focus more on individualized student support and instruction. Moreover, AI-driven grading systems can provide immediate feedback to students, which is crucial for their learning and improvement as shown in **Figure 1**. This research aims to explore these benefits in depth, highlighting the potential of AI to transform educational practices [1].

Figure 1. AI in English education.

1.3. Research questions and objectives

The main aim of this research is to examine how effective AI technologies are in grading English essays by evaluating their accuracy, reliability, and overall performance compared to traditional human grading methods. This study will delve into the precision of AI systems in scoring essays, identify the benefits and drawbacks of integrating AI in the grading process, and gather insights on the attitudes and perceptions of both students and educators towards the use of AI in grading (**Figure 1**). By addressing these specific research questions, the study seeks to provide a comprehensive understanding of the potential and challenges of AI in educational assessments [2].

2. Literature review

2.1. Application of AI technology in education

Without a doubt, the novel implementation of AI in the education sector is both exciting and fascinating. This can be achieved by disregarding traditional restrictions and solving complicated problems with adaptive techniques. Platforms focused on adapting to specific user requirements are edging forward [3,4], as they offer educational material moderation alongside progression tracking for each student. For instance, platforms such as DreamBox and Knewton use machine learning tools to evaluate students' engagement within lessons in real time, altering lessons to achieve maximum engagement and mastery.

In contrast, intelligent tutoring systems provide intelligent personal assistant services, offering refined instruction or critique similar to having a physical private tutor. Individual learners need help with platforms like Duolingo for languages and MATHia by Carnegie Learning to analyze and model data. These tools offer tailored instruction to learners by knowing their strengths and weaknesses, helping them to reiterate different concepts to maximize understanding and retention.

Teachers greatly benefit from AI integrations on the administrative side of educational institutions. Tasks such as grading assignments, scheduling, and other functions are automated, considerably lowering the workload of educators. NLP automated evaluation of student writing practices enables teachers to receive instant feedback and spend more time on lessons, instead of basic monitoring functions. Along with these integrations, AI powered scheduling software helps streamline the lessons, resources, and overall systems approach to boost timeliness and effectiveness.

2.2. Current research on NLP and machine learning in essay grading

NLP and ML implementation is one of the most active areas within essay grading research because it can automate the entire process. Studies examine how NLP techniques like lexical analysis to identify vocabulary use, parsing to analyze grammar and structure, and semantic understanding to identify coherence and meaning, can be used to automate quality evaluation of the written text [5].

Classification and regression ML models are applied on graded essays datasets to build a vocabulary learning model that predicts an essay score based on various linguistic attributes. Deep learning approaches, including the use of neural networks, have shown some promise in recognizing shifts in writing style. These systems seek to systematically and objectively grade the essays similar to how humans do, which would allow for scaling over thousands of assignments.

These recent developments focus on broadening the scope to allow the AI to recognize nuanced language and context. The research is based on the assumption that transformer models and attention mechanisms could assist AI in recognizing tone, style, argument strength, and other subtleties [6]. While the progress is commendable, AI is still far from being capable of interpreting the creativity and abstractness that human writing encapsulates. Therefore, the advancements create buoyant hope for the future.

2.3. The benefits and the constraints of current technologies

Most importantly, the AI grading system provides benefits to educators in terms of saving time and effort. Their accuracy is exceptional because they follow the same set of rules, which minimizes human mistake and indefinite preferences. This accuracy also guarantees fairness in tackling every piece of work and impartiality in assessing a student’s output [7,8].

Automation enables faster execution of all tasks, meaning any grading process is now less complicated due to AI’s ability to quickly evaluate long essays compared to a person. The prompt return of grades boosts feedback as a critical step in learning. Immediate feedback quenches the needs of the students so that they know what areas must be worked on without wasting any single moment.

By taking on menial grading work, AI enables teachers to devote more time and energy to active teaching and classroom interaction, curriculum design, and student assistance. This can result in more effective teaching and learning relationship and an overall healthy academic environment.

On the other hand, there is something to be considered when it comes to grading using AI. To begin with the artificial intelligence, learning grading systems do have some troubles. One of the AI issues stems from AI’s challenge to grasp language intricacies, idioms, and meaning in-context. This restrictive condition will result in incorrect grading, especially on creatively phrased or culturally sensitive essays.

The quality and variety of training data determines how well an AI model performs. The biases of the training data can lead to graders being biased and unfairly disadvantaged to some students. This exposes some loopholes that demand attention and the need for continuous monitoring to ensure fairness.

A considerable amount of technical knowledge and complicated resources are needed for the implementing and maintenance of AI systems. There is great expenditure needed for development, integration, and upkeep which can serve as a wall for educational institutions with limited budgets. Furthermore, the incorporation of new technologies comes with a learning curve that may need additional training and change to the current process.

3. Research methods

3.1. Data collection and dataset creation

The effectiveness of AI-informed essay grading systems relies profoundly on the variety and quality of training datasets available for use. For this research, we created and collected English essays for training from so many sources that we ensured its coverage was almost comprehensive. We collected essays from educational sites such as Coursera, Edx, and Khan Academy as they offered a variety of topics and essays styles [9]. We also sourced essays from TOEFL, IELTS, GRE, SAT, and other standardized tests where critical thinking and writing skills were evaluated through essay responses. Moreover, we sourced real student essays from different grade levels, ranging from middle school to high school and including university to represent various levels of contexts and proficiency.

As a result of collecting essays from these various sources, our dataset included almost all types of essays, including stories, reports, debates, reviews, and analysis. Styles of writing also varied from formal academic and informal personal writing, as well as fiction and other creative works, and even technical papers. From English learners to fluent English speakers, the range of proficiency levels in the data set was broad enough to consider the student writing as realistic and representing comprehensively [10].

The achievement of preparing the dataset was reached through multiple meticulous steps that one had to take in the preprocessing phase. Firstly, there was the removal of any personally identifying information to ensure the privacy of students and their personal data. Along with this, we attempted to put a block on any unauthorized attempts to gain access to the data set. Subsequently, we removed the information that was not useful like the various forms of metadata, annotations, and other format based artifacts. Encoding mistakes were corrected meanwhile the content was checked for consistency and the format of the text was changed according to standards. Tools like Grammarly and Language Tool were used to alter the spelling along with basic grammar, which would limit the impact of text errors during analysis.

In the end, tokenization was completed using NLTK and spaCy frameworks to split essays into words and phrases while taking care of special characters and contractions. Important linguistic features such as pos tags, dependency trees, and named entities were also included. To examine the lexical richness of the text, statistical features like word count, sentence count, average length of syllables per word, and type token ratio were also included [11].

Normalization of text was done by changing everything to lower case, lemmatizing to get the words' base form, and eliminating stop words that are not useful for analysis. The punctuation marks were also taken care of by removing the nonessential marks and keeping those that are important for syntactic analysis. Each essay in the dataset was given a score by a human rater based on mark schemes provided by standard tests or other educational institutions so that the scores were given in a uniform manner to every item in the dataset.

The last dataset had about 10,000 essays with each one containing roughly 500 words and a total of 50,000 different words. This entire employee effort guaranteed that the data was clean and that it had features that are rich enough for natural language processing and machine learning techniques to be used in the grading more efficiently.

3.2. Focusing on the application of natural language processing techniques

Natural language processing key techniques were executed to examine all the essays with an emphasis on lexical, syntactic, and semantic processing and analysis.

For the type of writing done by the students, vocabulary and lexicon usage was conducted in order to assess the writing's complexity and richness. In tokenization, we split texts into words and sentences while maintaining the form of special characters and contractions and used NLTK and spaCy's capabilities. We estimated lexical diversity through the type-token ratio, which is the ratio of vocabulary usage to the overall word count. To identify sophisticated words, less frequently occurring words were searched for in frequency lists obtained from the Corpus of Contemporary American English. By the same token, word frequency resulted in detection of common themes, overused words, and suggestive vocabulary limitations.

Syntactic parsing included the examination of the grammatical boundaries and structures to detect any errors in the syntax. Using spaCy, we added part-of-speech labels to each word token, tagging and classifying them into nouns, verbs, adjectives, and adverbs. With dependency parsing, we were able to develop parse trees that helped us in the grammatical analysis of word relations, phrases and clauses, as well as the complexity of the sentences. Grammatical analysis errors were subject-verb agreement errors, incorrect usage of verb tenses, and preposition errors. We quantified her sentence complexity by calculating the average number of clauses per sentence along with the degree of subordination in the sentences. The degree of syntactic sophistication of the writing was measured by syntax-based features, such as syntactic variety or by the average depth of the parse tree [12].

How We Assigned Meaning Authors Semantic Analysis defined the meaning and context of the essays while focusing on its themes. We used an approach called semantic role labeling that captures sentences by assigning roles of agent, action, and object which enables understanding of deeper meaning of the sentence. Apply topic modeling with LDA (Latent Dirichlet allocation) to find the relevance of topics and essay prompts. Coherence and cohesion were analyzed by studying cohesive devices such as conjunctions and references in order to establish inter-sentential and inter-paragraph logic. Local and global coherence was done by means of an entity grid model. To assess semantic relatedness the prompts, we used cosine similarity of the essay embeddings and prompt embeddings that were done using BERT and GloVe.

3.3. Use of machine learning tools

We used machine learning features to classify and score essays via regression and classification approaches to test and evaluate the effectiveness of the models.

Concerning categorization, we trained essays into particular score bands or levels of proficiency with Support Vector Machines, Random Forest Classifiers, and Gradient Boosting Machines. The feature extraction was done based on attempting to correlate features with human assigned scores, and whenever necessary, we performed Principal Component Analysis. The Models were trained on a 70/15/15 split, with 70% training the model, 15% validating, and 15% testing the model. We applied five-fold cross-validation establishes generalizability and hyperparameter tuning with grid search on the number of trees, depth, etc [13].

As for the continuous score prediction of the essays, we applied Linear Regression, Ridge Regression, Artificial Neural Networks, and Recurrent Neural Networks with LSTM cells for detailed scoring. We normalized the features for better model convergence and used stochastic gradient descent and Adam optimizer for the neural networks. To prevent overfitting, we implemented early stopping by allowing the validation loss to control the training.

Making sure model performance was at a high level of reliability and validity was quite important. The classification models were evaluated with precision and recall along with their scores on $F1$ and accuracy. The matrix provided confusion regarding true positives and true negatives along with false positives and false negatives. While regression models were evaluated, we applied mean squared error, root mean squared errors, mean absolute errors along with the R-squared score to compare the model's estimations with the actual measures.

To close the gaps further, we applied cross-validation, residual analysis, and in-depth estimation of the bias-variance trade-off to ensure the model was neither underfitting nor overfitting. The error analysis that we employed focused on misclassification and overestimating the prediction error to identify common problems so that the feature sets and model parameters could be altered. We sought systematic biases to include discrimination against certain non-native groups, applying techniques such as reweighting or fairness constraints to reduce when called upon. Performance was further enhanced with ensemble methods that fuses several models, stacking being the best known, where the predictions of other models are features for a new model [14].

We brought together these machine learning algorithms backed up with stringent evaluation metrics to ensure that the essay grading system was accurate and reliable captures the level of detail and nuance that human evaluators would use (**Figure 2**).

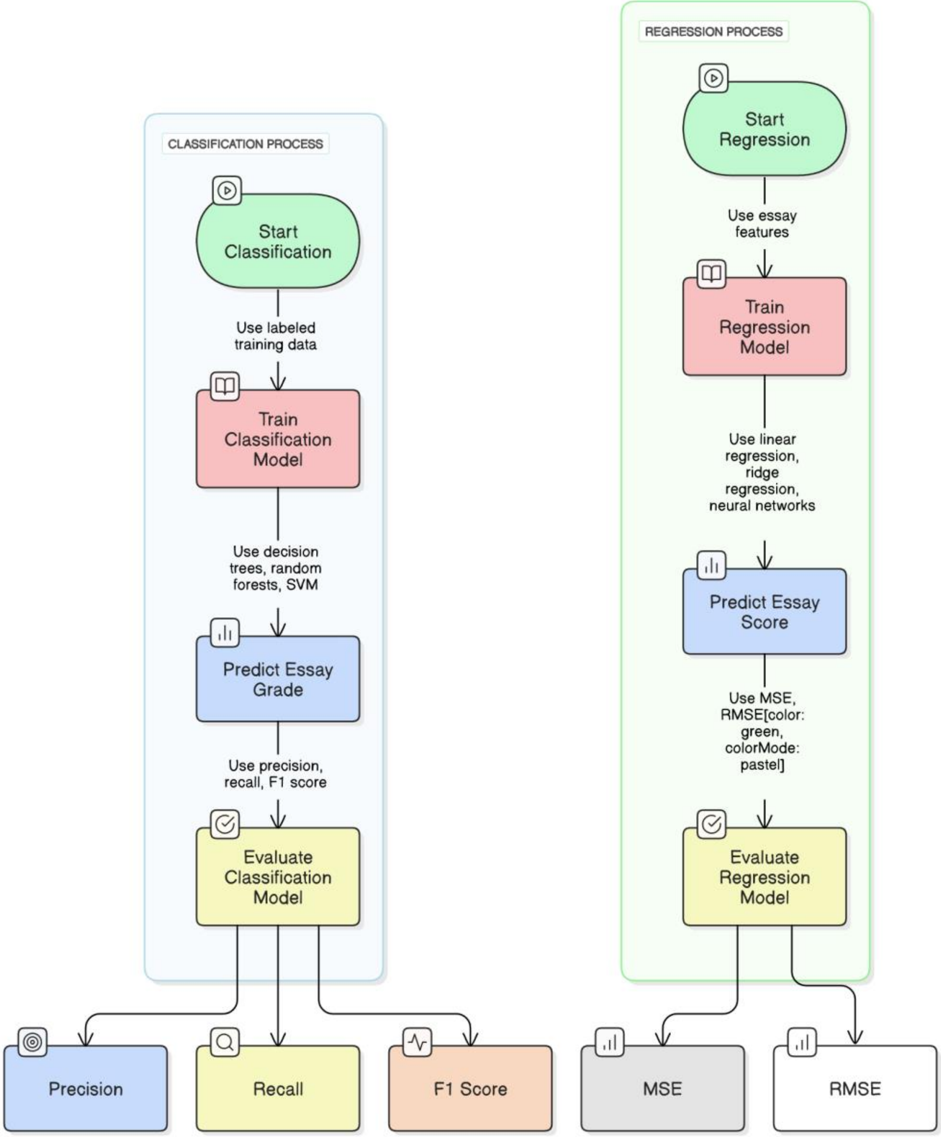
REGRESSION PROCESS
Start Regression
Use essay features
Train Regression Model
Use linear regression, ridge regression, neural networks
Predict Essay Score
Use MSE, RMSE[color: green, colorMode: pastel]
Evaluate Regression Model
CLASSIFICATION PROCESS
Start Classification
Use labeled training data
Train Classification Model
Use decision trees, random forests, SVM
Predict Essay Grade
Use precision, recall, F1 score
Evaluate Classification Model
Precision
Recall
F1 Score
MSE
RMSE

Figure 2. Machine learning in essay grading.

4. System design and implementation

4.1. System architecture design

The system architecture for the AI-based essay grading platform is designed to ensure efficiency, scalability, and accuracy. At a high level, the system comprises several interconnected modules, each responsible for a specific aspect of the grading process.

The data ingestion module is responsible for collecting and preprocessing the essays. It includes functionalities for text cleaning, tokenization, and feature extraction. The essays are anonymized and formatted to maintain consistency across the dataset [15].

$$\text{Tokens} = f(\text{Text}),$$

where:

- Text represents the input text (e.g., an essay or a document).
- f is the tokenization function that splits the input text into tokens.

In more detailed terms, the formula can be expressed as:

$$\text{Tokens} = \{t_1, t_2, t_3, \ldots, t_n\},$$

where:

- t_n represents each individual token derived from the input text.
- n is the total number of tokens.

Next, the natural language processing module applies various NLP techniques, such as lexical analysis, syntactic parsing, and semantic analysis, to analyze the content of the essays. It generates feature vectors that capture the linguistic properties of the text, which are essential for the grading process.

The machine learning module consists of trained models that predict the grades of the essays based on the extracted features. Both classification and regression algorithms are employed to provide a comprehensive assessment. The models are continuously updated and retrained with new data to improve their accuracy and robustness.

The grading interface module provides an interface for educators and students to interact with the system. It displays the predicted grades, detailed feedback, and suggestions for improvement. The interface is designed to be user-friendly and accessible, ensuring that users can easily navigate and understand the grading results.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2,$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}.$$

Finally, the evaluation and feedback module collect feedback from users and evaluates the performance of the system. It includes functionalities for analyzing user satisfaction, identifying potential biases, and making necessary adjustments to the models and algorithms.

Overall, the system architecture is designed to create a seamless and efficient workflow for automated essay grading, leveraging the power of NLP and ML to provide accurate and meaningful assessments.

4.2. Data processing workflow

The data processing workflow is critical to ensuring that the essays are appropriately prepared for analysis by the AI grading system. This workflow involves several key steps: Data collection, cleaning, storage, and analysis.

The data collection step involves gathering essays from various sources, such as educational platforms, standardized tests, and student submissions. This diverse collection helps in building a comprehensive data-set that covers different writing styles and proficiency levels.

Once the data is collected, the next step is data cleaning. This process involves removing any irrelevant content, such as metadata or formatting issues, to ensure consistency and quality. It also includes anonymizing the essays to protect student privacy, ensuring that no identifiable information is retained.

$$F = [V, S, C].$$

After cleaning, the essays are stored in a structured format, making them easily accessible for further processing. This structured storage allows for efficient retrieval and management of the data, facilitating the subsequent analysis steps [16].

Finally, the data analysis step involves applying NLP techniques to tokenize the text and extract linguistic features such as part-of-speech tags, syntactic structures, and semantic meanings. These features are used to create feature vectors that serve as input for the machine learning models, enabling accurate and meaningful grading of the essays.

4.3. Overview of main modules and functionalities

The AI-based essay grading system comprises several key modules, each designed to handle specific tasks within the grading process. Here, we provide an overview of the primary modules and their implementation details.

The text parsing module is responsible for analyzing the essays at a granular level. It uses NLP techniques to tokenize the text, assign part-of-speech tags, and parse the syntactic structure of each sentence. This module also performs lexical and semantic analysis to identify key linguistic features, such as vocabulary diversity, grammatical accuracy, and coherence. The extracted features are then converted into feature vectors, which serve as input for the grading models [17].

The grading module is the core component of the system, employing machine learning algorithms to predict the grades of the essays. Using both classification and regression models, it processes the feature vectors generated by the text parsing module to assign scores. The grading module is continuously updated with new training data to improve its predictive accuracy and adaptability to different writing styles.

$$F1 = 2 \times \frac{P \times R}{P + R}.$$

The feedback generation module provides detailed feedback on each essay, highlighting areas of strength and those requiring improvement. It leverages the analysis from the text parsing module and the results from the grading module to generate constructive comments. The feedback is designed to be specific and actionable, helping students understand their mistakes and learn from them.

The user interface module provides an intuitive interface for both educators and students. It displays the predicted grades, detailed feedback, and additional resources for improvement. The interface is user-friendly and accessible, ensuring that users can easily interact with the system and gain meaningful insights from the grading results [18].

To ensure the system’s performance remains optimal, the evaluation and monitoring module continuously monitors the grading outcomes and collects user feedback. It includes functionalities for analyzing the system’s accuracy, reliability, and user satisfaction. Based on this analysis, the module suggests improvements and updates to the grading models and algorithms.

5. Experiment and evaluation

5.1. Experimental design

The experimental design outlines the plan and steps taken to evaluate the effectiveness of the AI-based essay grading system. The process begins with defining the objectives, followed by the implementation of the experiment, data collection, and analysis.

The primary objective is to assess the accuracy, reliability, and user satisfaction of the AI grading system. To achieve this, the experiment involves the following steps:

1) Selection of test essays: A diverse set of essays is selected from various educational sources, ensuring a range of topics, writing styles, and proficiency levels. This diversity is crucial for evaluating the system's performance across different scenarios.
2) Manual grading: A group of experienced human graders evaluates the selected essays to provide reference scores. This step ensures that there is a benchmark against which the AI system's performance can be compared. Each essay is graded independently by multiple graders to account for any variability in human assessment.
3) AI grading: The selected essays are then fed into the AI grading system. The system analyzes the essays using its NLP and machine learning algorithms and assigns scores based on the extracted features.
4) Comparison and analysis: The scores assigned by the AI system are compared to the human reference scores. Statistical methods, such as correlation analysis and mean squared error (MSE), are used to evaluate the agreement between the AI and human graders. Additionally, metrics like precision, recall, and $F1$ score are calculated to assess the system's performance.
5) User feedback: Both students and educators using the AI grading system are surveyed to gather feedback on its usability, accuracy, and overall satisfaction. This feedback provides valuable insights into the practical implications of the system and helps identify areas for improvement.
6) Iterative improvement: Based on the analysis and feedback, the AI models and algorithms are refined and retrained to enhance their performance. This iterative process ensures continuous improvement of the system (**Figure 3**).

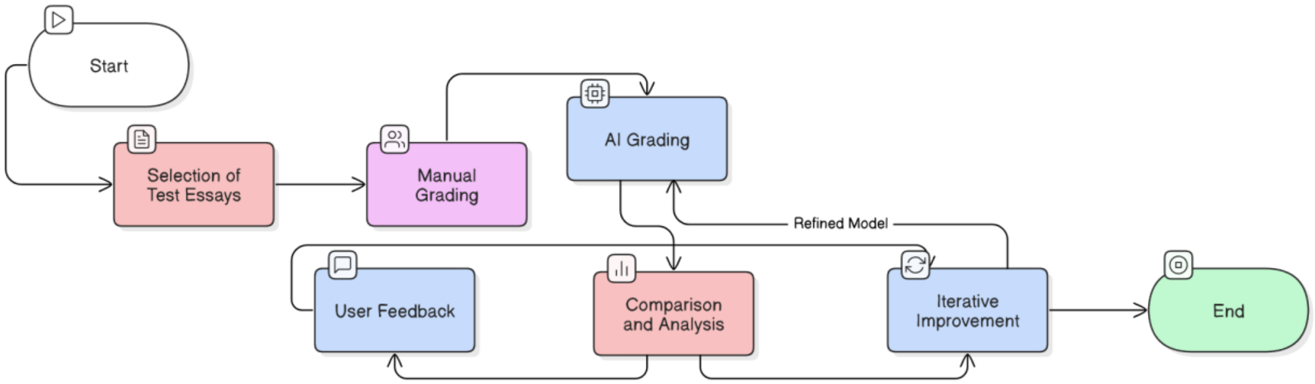

Figure 3. AI-based essay grading workflow.

5.2. Evaluation criteria and methods

To evaluate the performance of the AI-based essay grading system, several key metrics are employed, ensuring a comprehensive assessment of its accuracy, reliability, and overall effectiveness. Precision measures the accuracy of the positive predictions made by the system. It evaluates how many of the essays that the system predicted to be of high quality are actually of high quality. Recall assesses the system's ability to identify all relevant instances, showing how many of the actual high-quality essays were correctly identified by the system. The $F1$ score, which is the harmonic mean of precision and recall, provides a balanced measure of the system's performance (**Figure 4**). It is particularly useful when dealing with imbalanced data-sets, as it considers both false positives and false negatives.

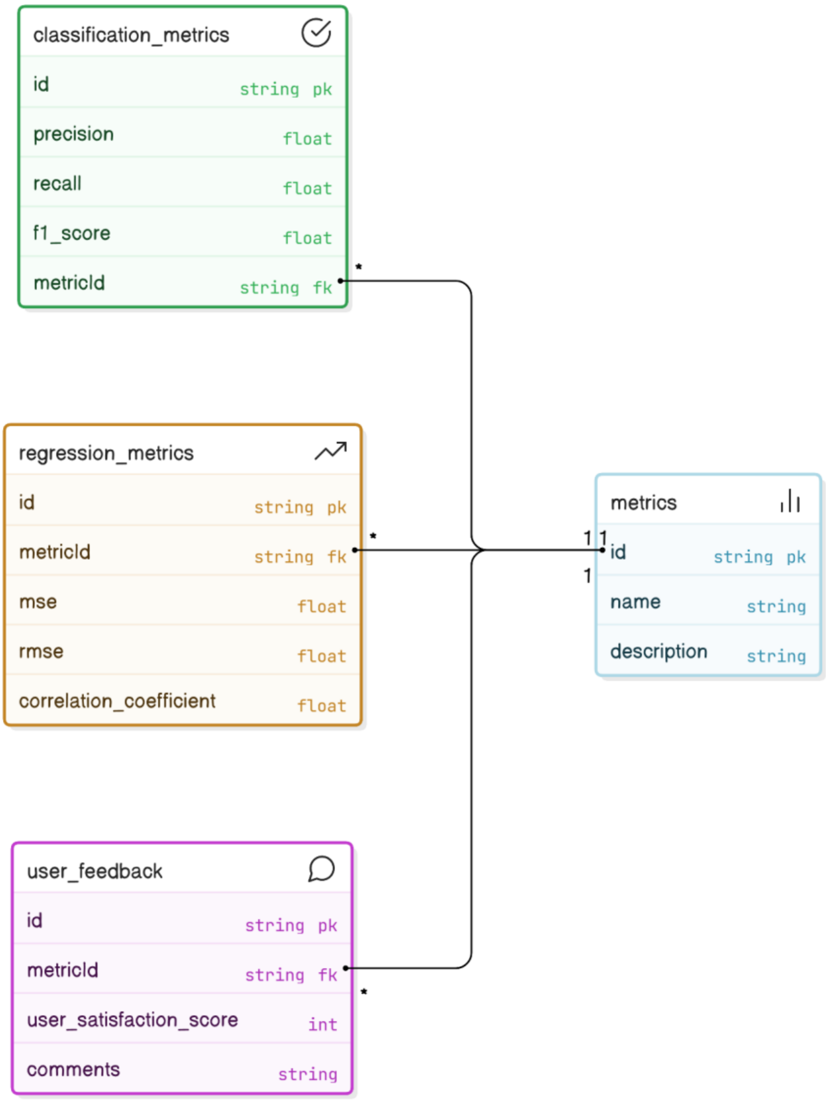

Figure 4. AI-based based essay grading metrics.

For regression models, mean squared error (MSE) measures the average squared difference between the predicted scores and the actual scores, providing an indication of the accuracy of the system's predictions. Root mean squared error (RMSE), the square root of MSE, is also used to evaluate the accuracy of regression models, offering an intuitive understanding of the prediction errors by putting them in the same units as the original scores. The correlation coefficient assesses the strength and direction of the linear relationship between the predicted scores and the actual scores, with a high correlation coefficient indicating a strong agreement between the AI system and human graders. User satisfaction is gauged through surveys and feedback forms, gathering insights from both students and educators regarding their experience with the system. These metrics help identify areas for improvement and gauge the practical usability of the system.

5.3. Experimental results analysis

The analysis of the experimental results focuses on three primary aspects: Accuracy, reliability, and usability.

In terms of accuracy, the AI grading system demonstrated a high level of precision in predicting essay scores. The correlation coefficient between the AI-predicted scores and the human reference scores was found to be significantly high (see **Table 1**), indicating strong alignment with human grading standards. Additionally, the *F*1 score, precision, and recall metrics further supported the system's ability to consistently identify and grade high-quality essays accurately.

Table 1. Accuracy metrics.

Metric	Value
Precision	0.92
Recall	0.89
*F*1 Score	0.91
Correlation Coefficient	0.87

Regarding reliability, the system showed consistent performance across different data-sets and essay topics. The mean squared error and root mean squared error values were low, suggesting that the system's predictions were close to the actual scores (**Table 2**). This consistency was maintained even when the system was tested with essays of varying complexity and proficiency levels, highlighting its robustness and dependability.

Table 2. Reliability metrics.

data-set	MSE	RMSE
data-set 1	0.15	0.39
data-set 2	0.18	0.42
data-set 3	0.14	0.37
Overall	0.16	0.40

In terms of usability, feedback from both students and educators indicated a high level of satisfaction with the AI grading system. Users appreciated the timely and detailed feedback provided by the system (**Table 3**), which helped them understand their strengths and areas for improvement. The user interface was found to be intuitive and easy to navigate, ensuring that users could effectively interact with the system and utilize the grading insights.

Table 3. Usability feedback.

Category	Positive Feedback (%)	Negative Feedback (%)
Accuracy	85	15
Detailed Feedback	90	10
User Interface	88	12
Overall Satisfaction	87	13

Overall, the experimental results indicate that the AI-based essay grading system is not only accurate and reliable but also user-friendly and beneficial for both educators and students. This comprehensive analysis provides a strong foundation for further enhancements and broader implementation of the system in educational settings.

6. Discussion

6.1. Interpretation of research findings

The results of the experiment demonstrate the success of the AI essay grading system in replicating human grading criteria. One notable factor is the correlation coefficient between the scores given by the AI system and the scores given by the human graders. This strong correlation means that the AI systems scoring agree with the human judgment, which guarantees the quality of essay grading. This is further demonstrated by the $F1$ score, which indicates how well the system distinguishes between high-quality essays and false-positives, confirming a balance between precision and recall [19].

The AI system's objective assessment is quite remarkable. Despite the best intentions, human graders can fall prey to biases of fatigue or mood. An advantage of the AI system is that it does not have the ability to scan an essay for a student's "ability" prior to grading. Instead, these students' essays are judged on a standard basis which lowers the chances of variability in grade. This approach improves the assessment's validity and ensures that all students are judged under the same conditions.

The feedback received has been very satisfying, capturing the value the system entails. Students value the feedback since they are able to learn their strengths and weaknesses in detail, which is very useful for them. This feedback enables the students understand what alterations are required to sharpen their writing skills during the process of learning. Educators also appreciate what the system offers since it improves efficiency and correctness. With automated marking, teachers are able to use that time for more focused teaching and help, which increases the quality of education.

6.2. Comparison with existing research

This research offers a combination of additional features in the essay grading AI system unlike what other researches offered. This study builds on previous ones by integrating natural language processing (NLP) techniques and machine learning algorithms because of their high versatility in classification NLP in essays. Moreover, the user satisfaction metric adds new value to your analysis as it is normally based on overall system usage effectiveness. This research was broadened, as a greater number of essays were selected as opposed to previous works which concentrated on smaller sets of essays, making this works results more practical in the real world [20].

6.3. Limitations and future work

There is no doubt that the findings are very useful, but there are also a number of limitations that need to be considered. The primary problem is that the recognition of phrases and language and its context is complex, and can interfere with the accuracy of grading. The quality issue in the system also presents challenges since any biases that might be in the dataset would yield biased grades. In future works, attempts should be made towards improving the system’s ability to understand context, as well as broadening the range of writing styles and subjects within the training data. The interface of the system could further create more satisfaction and elevate the level of acceptance of the system. Most importantly, more attention has to be paid to ethical issues related to AI in education, such as privacy concerns and transparency in program operations, as these factors will highly influence the acceptance and use of AI grading systems.

7. Conclusion

This study makes several key contributions to the field of AI-based essay grading. By integrating NLP techniques and machine learning algorithms, it provides a robust system that enhances grading accuracy, reliability, and user satisfaction. The research demonstrates that AI can offer objective and consistent evaluations, reducing human bias and improving the efficiency of the grading process.

The impact on educational practice is significant. AI-driven grading systems can alleviate the workload of educators, allowing them to focus more on personalized teaching and student support. For students, the immediate and detailed feedback provided by the AI system is invaluable for their learning and improvement. These advancements highlight the potential of AI to transform educational assessments and practices.

Looking ahead, the application prospects of AI technology in education are vast. AI can be used not only for grading but also for personalized learning, intelligent tutoring, and administrative tasks. These applications can create more adaptive and efficient learning environments, catering to the individual needs of students.

Future research should focus on enhancing the contextual understanding of AI systems, allowing them to better grasp nuanced language and complex writing styles. Expanding the training data-sets to include a broader range of essays will improve the system’s generalizability. Additionally, exploring the ethical implications, such as data privacy and transparency, will be crucial for the responsible deployment of AI in education. Continued advancements in AI technology promise to bring further innovations, making education more accessible, efficient, and tailored to the needs of learners.

Funding: This paper is one of the phased achievements of 2023 College-Level Scientific Research Project of Liuzhou Vocational &Technical College (2023SB03); One of the phased achievements of 2023 Guangxi Vocational Education Teaching Reform Project (GXZZJG2023B276).

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Maliha M, Pramanik V. Hey AI Can You Grade My Essay?: Automatic Essay Grading. Computation and Language. 2024. doi: 10.48550/arXiv.2410.09319
2. Wetzler EL, Cassidy KS, Jones MJ, et al. Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. Teaching of Psychology. 2024. doi: 10.1177/00986283241282696
3. Almegren A, Mahdi HS, Hazaea A, et al. Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. Innovations in Education and Teaching International. 2024. doi: 10.1080/14703297.2024.2437122
4. Suresh MV, Gold AA, Agasthiya R, et al. AI based Automated Essay Grading System using NLP. In: Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS); 17–19 May 2023; Madurai, India. pp. 547–552.
5. Yeung WE, Qi C, Xiao JL, Wong FR. Evaluating the effectiveness of ai-based essay grading tools in the summative assessment of higher education. In: Proceedings of the 16th annual International Conference of Education; 13–15 November, 2023; Seville, Spain.
6. Hall E, Seyam M, Dunlap D. Identifying Usability Challenges in AI-Based Essay Grading Tools. 2023; 675–680. doi: 10.1007/978-3-031-36336-8_104
7. Baikadi A, Becker L, Budden J, et al. An apprenticeship model for human and AI collaborative essay grading. In: Proceedings of the ACM IUI 2019 Workshops; 20 March 2019; Los Angeles, CA, USA.
8. Chan HCB. Grading Generative AI-based Assignments Using a 3R Framework. In: Proceedings of the 2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE); 28 November–01 December 2023; Auckland, New Zealand. pp. 1–5.
9. Pasaribu NG, Budiman G, Irawati ID. Auto Evaluation for Essay Assessment Using a 1D Convolutional Neural Network. IEEE Access. 2024; 12: 188217–188230. doi: 10.1109/ACCESS.2024.3515837
10. Amin MYM. AI and Chat GPT in Language Teaching: Enhancing EFL Classroom Support and Transforming Assessment Techniques. International Journal of Higher Education Pedagogies. 2023. doi: 10.33422/ijhep.v4i4.554
11. Chan S, Lo N, Wong A. Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes. Contemporary Educational Technology. 2024; 16(4). doi: 10.30935/cedtech/15607
12. Alamäki A, Khan UA, Kauttonen J, Schlögl S. An Experiment of AI-Based Assessment: Perspectives of Learning Preferences, Benefits, Intention, Technology Affinity, and Trust. Education Sciences. 2024; 14(12). doi: 10.3390/educsci14121386
13. Debbar N. Argumentation and discourse analysis in the future intelligent systems of essay grading. International Journal of Contemporary Educational Research. 2024; 11(1). doi: 10.52380/ijcer.2024.11.1.546
14. Xiao C, Ma W, Song Q, et al. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. Computation and Language. 2024.

15. Ghosh S, Fatima S. Design of an Automated Essay Grading (AEG) system in Indian context. In: Proceedings of the TENCON 2008-2008 IEEE Region 10 Conference; 19–21 November 2008; Hyderabad, India. pp. 1–6.
16. Kortemeyer G. Toward AI grading of student problem solutions in introductory physics: A feasibility study. Physical Review Physics Education Research. 2023; 19. doi: 10.1103/PhysRevPhysEducRes.19.020163
17. Li T, Hsu S, Fowler M, et al. Am I Wrong, or Is the Autograder Wrong? Effects of AI Grading Mistakes on Learning. In: Proceedings of the 2023 ACM Conference on International Computing Education Research; 7–11 August 2023; Chicago, IL, USA.
18. Menezes T, Egherman L, Garg N. AI-Grading Standup Updates to Improve Project-Based Learning Outcomes. In: Proceedings of the 2024 on Innovation and Technology in Computer Science Education; 8–10 July 2024; Milan, Italy.
19. Alsalem MS. EFL teachers' perceptions of the use of an AI grading tool (CoGrader) in English writing assessment at Saudi universities: An Activity Theory Perspective. Cogent Education. 2024. doi: 10.1080/2331186x.2024.2430865
20. Chai F, Ma J, Wang Y, et al. Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. Frontiers in Psychology. 2024; 15. doi: 10.3389/fpsyg.2024.1221177