Journal of Information Systems Engineering and Management
2025, 10(6s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Automated Grading of Essay Using Natural Language Processing: A Comparative Analysis with Human Raters Across Multiple Essay Types

*Dennis C. Gabon[1], Albert A. Vinluan[2], Jennifer T. Carpio[3]
[1] *Student, IT Faculty, IT Department, College of Information Technology, Central Bicol State University of Agriculture, Camarines Sur, Philippines*
[2] *Professor, IT Faculty, IT Department, College of Computing Studies, Information and Communication Technology, Isabela State University, Isabela, Philippines*
[3] *Professor, IT Faculty, IT Department, College of Computing Studies, San Beda University, Manila, Philippines*
[1]dennis.gabon@cbsua.edu.ph, [2]albert.a.vinluan@isu.edu.ph, [3]jcarpio@sanbeda.edu.ph
ORCID ID number: [1]0009-0008-0839-3909, [2]0009-0001-4570-1391, [3]0009-0000-6146-5088
Corresponding Author: Dennis C. Gabon*

ARTICLE INFO

Received: 15 Oct 2024

Revised: 08 Dec 2024

Accepted: 24 Dec 2024

ABSTRACT

This study explores how well an automated essay grading (AEG) system, built with Natural Language Processing (NLP), aligns with human graders in assessing different types of essays. Using essays from 35 information technology (IT) students manually scored by human raters, the system's performance was evaluated with statistical tools like weighted Cohen's Kappa and the Friedman test. The results showed a moderate to substantial match between the automated essay grading system and human scores across essay types (argumentative, comparison and contrast, descriptive, narrative, and persuasive, suggesting that the system can reliably handle various writing styles. Also, no significant differences in grading reliability were found across essay formats, indicating that the system adapts well to different types of essays. These findings suggest that NLP-based essay grading systems could be valuable in educational settings, especially in large classes with high grading demands. While the system shows strong potential in the academic arena, focusing on assessment and further testing with a broader dataset is recommended to improve its generalizability and address complex writing elements, such as creativity and tone. This study contributes to educational technology by presenting a practical, scalable approach to consistent and objective essay grading, paving the way for the broader use of automated grading tools in education.

Keywords: automated essay grading, natural language processing, grading reliability, essay evaluation, machine learning

INTRODUCTION

Natural Language Processing (NLP) is a type of artificial intelligence (AI) that helps computers understand human language in different forms, such as written text, speech, or even casual notes [1]. As AI technology becomes more common daily, NLP is important for improving how we interact with machines. However, developing these applications is challenging because computers need clear and structured input, while human communication could be more specific and complex [2]. Today, NLP is used in various areas like information extraction, translation, summarizing texts, and grading written responses [3].

Essay types of assessments are widely used in education to assess students' understanding, critical thinking, and ability to connect ideas [4]. However, grading these essays manually can be slow and subjective, leading to inconsistencies. This concern has become more evident as many educational institutions,

like the Central Bicol State University of Agriculture (CBSUA) in the Philippines, have moved and implemented flexible learning modalities. Traditional methods, which rely on human graders, are time-consuming and can vary in accuracy [5].

With advancements in NLP, there is a growing interest in developing automated grading systems that provide quick, fair, and accurate feedback. For example, systems like the Intelligent Essay Assessor (IEA) use techniques such as Latent Semantic Analysis (LSA) to evaluate essays, showing a strong correlation with human grading [6]. Yet, evaluating open-ended responses remains complex, requiring understanding beyond matching keywords [7].

This study aims to develop an automated essay grading system using NLP. The objective is to design a tool that can accurately and consistently evaluate essays based on content and structure. The system will be tested against human grading to ensure its effectiveness and reliability. This project is not meant to replace teachers but to help them by making the grading process more efficient and reducing the time needed for assessment.

By introducing such a system, this study hopes to support educators and improve the quality of student feedback. It offers a solution to the challenges of manual essay grading, thereby contributing to a better learning environment for teachers and students.

METHODOLOGY

This study discusses the key components, focusing on data collection and the system overview. The study aims to develop an automated essay grading system using NLP techniques. The system processes student essay submissions and compares them against preloaded model answers and grading rubrics to produce scores and feedback. The research involved gathering relevant data for system training and evaluation to achieve this.

A. Data Collection

We collected essays from 35 IT students, each bringing their responses from what they remembered and comprehended regarding each essay question. From persuasive arguments to creative narratives, their work gave us plenty to test our system with.

Human raters manually graded each essay submission using a structured rubric focused on content quality, coherence, structure, and grammatical accuracy. These human-graded essays served as a reference dataset for training the automated grading system. The grading rubrics and model answers were stored in a MySQL database, which allowed the machine learning model to learn from and base its evaluations.

This dataset was essential for training the system's machine-learning model and conducting performance comparisons between the system's automated scores and human grades. By looking at a broad range of essay types and varying writing quality, the collected data ensured that the system was exposed to different levels of complexity, allowing for a reliable model. The model answers, and grading rubrics provided by the teachers formed the grading criteria' core, aligning human and machine evaluations with a common standard.

After the developmental part, this study tests the system's reliability by determining the consistency of the generated score to the score given by the human rater. This study used the Statistical Package for Social Science (SPSS) version 27.0 software to analyze data with the help of the following statistical tool: Weighted Cohen’s Kappa in determining the reliability of the system and Friedman test in finding if there is a significant difference on the reliability of the system across different types of essays.

B. System Overview

The system works in two main stages: evaluating the essays and training the grading model, as shown in Figure 1.

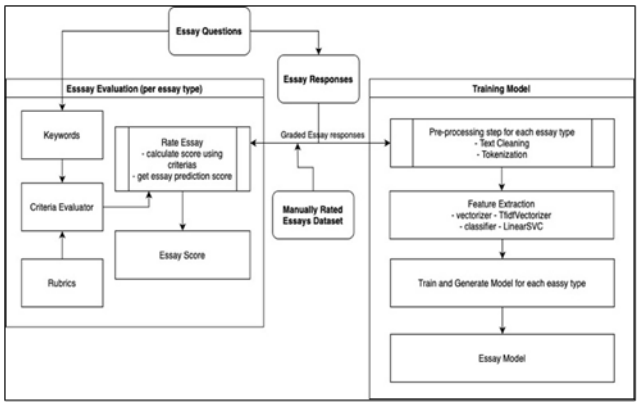

Figure 1. System Overview

In the essay evaluation stage, the system starts by analyzing the essay for keywords to which important terms that reflect the content are considered. It then compares these keywords to a rubric, a set of rules, or criteria for grading. A criteria evaluator evaluates how well the essay meets these criteria and assigns a score. This score is based on specific factors such as how well the essay is structured, how clear the ideas are, and the use of language, depending on the type of essay (persuasive, descriptive, etc.).

In the training model phase, the system learns to grade by studying essays humans have already graded. Here, the system takes these manually graded essays and uses them to learn how to assign scores in the future. It first cleans up the essays by removing unnecessary words, then picks out the most important terms to focus on. These terms are used to train a machine learning model (in this case, a classifier) that will predict scores for future essays based on patterns it finds.

Once the training is complete, the system can automatically grade new essays based on what it has learned. All this information is stored in the database, so students and teachers can access it whenever they need to use it using their login credentials.

Finally, statistical analysis compares its performance to human grading to ensure the system grades accurately and fairly. Using Weighted Cohen's Kappa and the Friedman test, the system's consistency with human raters is measured across different types of essays, ensuring its grades are reliable, objective, and accurate.

RESULTS AND DISCUSSION

A. Preprocessing steps on information extraction

One of the first things the researchers had to do to get the automated essay grading system working was to prepare the text in a way the system could understand. Sorting was done through, and what was written was organized. This is basically what we did with the essays. We used tokenization to break down the essay into smaller pieces like individual words or sentences. This made it easier for the system to pick up on what the essay was talking about.

Text preprocessing is a crucial step in natural language processing (NLP) and information extraction tasks, as it involves cleaning and filtering textual data to remove noise and irrelevant information. This process can significantly impact the efficiency and accuracy of subsequent NLP stages [8], [9]. Studies show that effective preprocessing can improve system performance by up to 30% in tasks like text classification and sentiment analysis [10].

The most common preprocessing techniques are tokenization, stop word removal, and lemmatization. Tokenization breaks down the text into manageable units, such as words or sentences, allowing the system to identify and process linguistic patterns easily. Stop word removal further refines the text by discarding frequently used but contextually insignificant words, enabling the model to concentrate on more meaningful terms and thus enhancing processing efficiency [8]—[10]. Together, these steps help streamline the data, setting a solid foundation for accurate NLP interpretation.

B. Collecting a diverse dataset for training and testing

Another part of making the grading system work was having a good variety of essays to train it on. The researchers gathered essay responses from 35 information technology (IT) students, and they covered different essay types: persuasive, argumentative, descriptive, comparison and contrast, and narrative. This was important because each type of essay has its unique style and purpose. For example, a persuasive essay needs strong, convincing arguments, while a descriptive essay is about painting a picture with words.

The challenge was ensuring we had a good mix of different types and quality levels. If the system only saw well-written essays, it would get used to grading everything as if it were at that high level, which is unrealistic. We included a range of essays, from polished ones to those that needed more work. This way, the system learned how to deal with different levels of writing ability, just like a human grader would.

We split this collection of essays into two groups: one for training the system and one for testing it afterward. This was a way to check how well the system learned to grade essays it had not seen before. It was a good way to ensure the system could handle new, unseen essays fairly and accurately. But there's always room to improve. We could add more essays from different backgrounds and topics to improve the system.

C. Reliability of the automated grading system

Table 1 discusses the reliability of the automated grading system by analyzing its consistency with human rater scores. The collected data were analyzed using SPSS version 27, utilizing weighted Cohen's Kappa as statistical treatment.

Weighted Cohen's Kappa was run to determine if there was an agreement between the score of the automated grading system and human rater across different types of essays. Data revealed that there was moderate agreement between the two along Persuasive (κ =0.512 {95% CI, 0.220 to 0.804}, p < .01), Descriptive (κ =. 0.528 {95% CI, 0.228 to 0.829}, p < .01), and Narrative type of essay (κ = 0.542 {95% CI, 0.154 to 0.930}, p < .001). Moreover, there was substantial agreement between the two raters along Argumentative (κ = 0.767 {95% CI, 0.457 to 0.924}, p < .001) and Comparison and Contrast type of essay (κ = 0.612 {95% CI, 0.457 to 0.920}, p < .001).

Table 1

Reliability of the automated grading system along different types of essay

Type of Essay	Cohen's Kappa (κ)	p-value	95% Confidence interval	
			Upper	Lower
Persuasive	0.512	0.002	0.804	0.220
Argumentative	0.767	0.000	0.924	0.457
Descriptive	0.528	0.001	0.829	0.228
Comparison and Contrast	0.612	0.000	0.920	0.304
Narrative	0.542	0.000	0.930	0.154

Legend:

Cohen's Kappa(k)	*Interpretation*
k < 0.20	Slight
0.21 ≥ k ≤0.40	Fair
0.41 ≥ k ≤0.60	Moderate
0.61 ≥ k ≤0.80	Substantial
k >0.80	Almost Perfect

These findings suggest that the reliability of the automated grading system ranged from moderate to substantial, highlighting all the p-values that were less than 0.01, implying the significance of the reliability of the grading system from 0. Automated assessment systems have shown promising results in various educational contexts. For instance, one study implemented a hybrid system combining static and dynamic analysis to evaluate programming assignments. This approach effectively reduced the grading workload for instructors while maintaining high accuracy and consistency in assessments, even in large-scale online learning environments [11]. Such systems have proven reliable and beneficial, particularly in scenarios where traditional manual grading is challenging due to high student numbers or remote learning conditions. Similarly, our NLP-based grading system demonstrated moderate to substantial agreement with human raters across multiple essay types, indicating its potential to support educators in managing the grading process more efficiently.

D. Differences in the reliability of the system across different types of essays

This section examines the system's reliability differences across different types of essays. The Friedman test was used to analyze data utilizing SPSS version 27. Table II presents the system's reliability differences across different types of essays. As shown in the table, the Chi-square value of 6.122 and p-value of 0.190 suggest that there is no significant difference in the system's reliability across different types of essays.

Table 2

Friedman Test Result

N	5
Chi-Square	6.122
df	4
Asymp. Sig.	.190

These findings suggest that the developed system performed well in checking this study's different types of essays. Previous studies have also explored the performance of automated assessment systems across various contexts. This result aligns with recent research on AES reliability and adaptability. Studies such as [12] support this consistency by noting that AES systems can perform reliably across essay types when calibrated to specific writing attributes. Similarly, research from [5] acknowledges challenges in evaluating nuanced elements like content relevance and cohesion but affirms that AES systems can still yield reliable results across genres with careful design. Additionally, findings from [13] show that AES systems tend to be more reliable for non-struggling students, suggesting that while the system in this study performed well across essay types, further research could explore reliability across different student proficiency levels. Finally, research in [14] highlights potential demographic biases in AES models, underscoring the need for further testing to ensure fairness across diverse student groups.

This suggests that automated systems, when properly calibrated, can provide consistent and fair evaluations regardless of the content or format of the student submissions. Similarly, our study found no significant difference in the system's reliability across different essay types, reinforcing the potential of automated grading tools to handle diverse assessment tasks effectively. By maintaining consistency in evaluation, such systems can significantly reduce biases and improve the scalability of educational assessments.

CONCLUSIONS

This study demonstrated that an automated grading of essay systems utilizing Natural Language Processing (NLP) can be a reliable tool for evaluating diverse essay types with accuracy comparable to human raters. The system showed moderate to substantial alignment with human graders, particularly in structured and rubric-based evaluations through analysis of essay types, including persuasive, descriptive, narrative, argumentative, and comparison and contrast. As reflected in the weighted Cohen's Kappa values and the Friedman test results, statistical analysis confirmed the system's consistency across essay types, suggesting its potential to provide scalable, fair, and objective feedback to students.

While promising, the study is limited by its dataset, which consists of essays from 35 IT students. This constraint affects the model's generalizability, as a more diverse sample could yield broader insights into the system's applicability. Additionally, while the system successfully evaluates structure and language, it may require further refinement to assess more nuanced qualities, such as creativity and tone, which are often better captured through human oversight. Future research should expand the dataset to include a broader range of essay types and demographics and consider implementing advanced NLP techniques to enhance the system's interpretive capabilities.

Overall, this study contributes a practical and efficient approach to essay grading in educational contexts. It demonstrates that with further development, this automated system can serve as a valuable support tool for educators, helping streamline assessment tasks and maintain consistent standards in grading.

ACKNOWLEDGMENT

We are grateful to the College of Information Technology of Central Bicol State University of Agriculture for providing essential tools and facilities, such as the computer laboratory, for the students' testing area. We also extend our sincere thanks to Mrs. Gilda J. Taupa for approving our request. The authors did not receive financing for the development of this research.

We also thank Mr. Nicanor Galang, Jr., Mr. Rey Añonuevo, and Mr. Earl Jay Ruz for their expertise in system design, statistical analysis, and technical support, respectively. Special thanks go to Sir Nikki, Sir Arce, Mam Jocelle, and Mam Dada for their input and suggestions, which greatly enhanced the rigor of our study.

REFERENCES

[1] https://www.coursera.org/articles/natural-language-processing.
[2] Daniel Jurafsky and James H. Martin, Speech and Language Processing, 2e, Pearson Education, 2009
[3] Application of Natural Language Processing - Information Extraction available at: https://www.lifewire.com/applications-of-natural-language- processing-technology-2495544
[4] S. Drolia, P. Agarwal, S. Rupani and A. Singh, "Automated Essay Rater using Natural Language Processing," International Journal of Computer Applications (0975-8887). Vol. 163 – No 10, April 2017.
[5] Ramesh, D., Sanampudi, S.K. An automated essay scoring systems: a systematic literature review. Artif Intell Rev 55, 2495–2527 (2022). https://doi.org/10.1007/s10462-021-10068-2
[6] J. Burstein, K. Kukich, S. Wolff, C. Lu, and M. Chodorow, "The Intelligent Essay Assessor: Applications to Educational Technology," AI Magazine, vol. 19, no. 1, pp. 27-36, 1998.
[7] A. Rokade, B. Patil, S. Rajani, S. Revandkar and R. Shedge, "Automated Grading System Using Natural Language Processing," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1123-1127, doi: 10.1109/ICICCT.2018.8473170.
[8] Tabassum, A., & Patil, R. R. (2020). A survey on text preprocessing & feature extraction techniques in natural language processing. International Research Journal of Engineering and Technology (IRJET), 7(06), 4864-4867.
[9] Sharma, A., & Parmar, M. (2021). A survey on text preprocessing and feature extraction techniques for sentiment analysis of Twitter data. International Research Journal of Computer Science.
[10] Vel, S. S. (2021, March). Preprocessing techniques of text mining using computational linguistics and Python libraries. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 879-884). IEEE.
[11] Zougari, S., Tanana, M., & Lyhyaoui, A. (2022). Validity of a graph-based au-tomatic assessment system for programming assignments: Human versus automatic grading. International Journal of Electrical and Computer Engineering, 12(3), 2867.
[12] Srivastava, K., Dhanda, N., Shrivastava, A., scholar, D.A., & Kalaam (2020). An Analysis of Automated Essay Grading Systems. International Journal of Recent Technology and Engineering.
[13] Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. American Educational Research Journal, 59(6), 1122-1156.
[14] Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021, June). A fairness evaluation of automated methods for scoring text evidence usage in writing. In International Conference on Artificial Intelligence in Education (pp. 255-267). Cham: Springer International Publishing.