MASTER THESIS

EDUCATIONAL SCIENCE AND TECHNOLOGY

FACULTY OF BEHAVIORAL MANAGEMENT AND SOCIAL SCIENCES

EXPLORING AUTOMATED ESSAY SCORING IN DIGITAL LEARNING ENVIRONMENTS

KINCSÖ SÜVEGES

1ST SUPERVISOR: DR. JOHANNES STEINRÜCKE

2ND SUPERVISOR: DR. MARYAM AMIR HAERI

13/07/2024

UNIVERSITY OF TWENTE.

Abstract

Automated essay scoring (AES) systems have revolutionized education with bringing numerous immense advantages to assessment. While several learning and assessment platforms have already implemented it,literature on the efficiency of AES systems is limited in the context of primary education. Therefore, this research studied how well and accurately AES systems can evaluate writing products from primary school students.

In this exploratory study 100 texts were simulated as being written by Pre-K- Grade 2 students. This was followed by automated scoring in a supervised machine learning setting, based on a validated rubric. Predicted and actual scores were compared to find out how accurately AES can evaluate generated texts.

Accuracy of the automated assessment was found to be fairly reliable despite certain limitations, namely the lack of authentic data.

Keywords: Automated Essay Scoring, Rubrics, Data Augmentation, Holistic Scoring

TABLE OF CONTENTS

1. PROBLEM STATEMENT

The use of rubrics in essay scoring has demonstrated high reliability, especially when combined with proper training of the raters and providing sufficient examples. However, evaluating essays still requires a considerable amount of time. This holds true whether it's a teacher who has to assess, for example, 150 essays over the weekend or a company faced with the task of scoring thousands of essays for standardized assessments. Consequently, there are growing concerns about the feasibility of grading a large volume of written work within a reasonable timeframe, besides the persistent challenge of accurate and effective writing skills evaluation (Shin & Gierl, 2022). The development of AES systems marked a significant breakthrough, as these systems provided a solution for scoring essays much faster and at a much lower cost than traditional methods (Shin & Gierl, 2022). Consequently, this has spurred the development of automated essay scoring and led to a surge in research in this area (Dikli, 2006; Graesser & McNamara, 2012; Shermis & Burstein, 2013; Weigle, 2013; Xi, 2010 as cited in McNamara et al., 2015).

While a lot has been achieved in the field of automated assessment regarding the consensus between automated and human raters (Chen and He, 2013; Alikaniotis et al., 2016 as cited in Amorim et al., 2018) or the impact of automated writing evaluation tools on foreign language anxiety and learner autonomy (Fu et al., 2022, as cited in Dizon & Gold, 2023), there are plenty of aspects of automated assessment that require more research such as "the relevance of the content to the prompt, development of ideas, Cohesion, Coherence, and domain knowledge" (Ramesh & Sanampudi, 2022, p. 2515).

This study aims to bridge the knowledge gap regarding the accuracy, agreements, strengths, limitations, and discrepancies of AES in evaluating student writing. The student writings in this study were AI-generated using different generation prompts. This further expands the scope of this investigation, allowing for a nuanced analysis of how automated systems evaluate different writing styles and levels of complexity. It is crucial to explore these aspects to determine the viability of AES as a reliable and efficient tool for assessing writing proficiency across a diverse range of writing samples, with the use of an established assessment rubric as a benchmark. Therefore, the main question of this research is*: "How effectively can automated essay assessment systems assess AI-generated texts?"*

The primary objective of this research is to analyse the efficiency of AES systems in evaluating generated texts. The generated texts are based on a rubric and are meant to simulate primary school students' writings. By conducting classification analyses, this research aims to identify the strengths and limitations of AES through generated texts as outlined in an assessment rubric.

2. THEORETICAL FRAMEWORK

2.1 Automated essay scoring

2.1.1 What is Automated Essay Scoring (AES)

The term Automated Essay Scoring (AES) system is mostly used to refer to "a computer-based assessment system that automatically scores or grades the student responses by considering appropriate features" (Ramesh & Sanampudi, 2022, p. 2494). While early AES systems can be traced back to 1960s, the advancement of computers in the 1990s led to the development of more accurate AES systems (Mizumoto & Eguchi, 2023).

Automated essay grading is often facilitated through natural language processing (NLP) and machine learning (ML) techniques within AES systems (Uto, 2021). These technologies work in a two-stage process, namely training and scoring (Yun, 2023). In the first stage (training), developers use essays human experts have scored called training sets, to create a scoring model for the system itself (Attali & Burstein, 2006; Breyer et al., 2014; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Shermis, Koch, Page, Keith, & Harrington, 2002 as cited in Yun, 2023). During this training stage, NLP techniques are used to analyse various aspects of an essay's writing like grammar, vocabulary, style, or content. Then machine learning algorithms analyse the relationships between extracted features to generate scores and provide feedback based on the identified patterns (Wilson & Roscoe, 2020).

With this trained model in place, the AES system moves to the second stage: scoring. Here, the system utilizes NLP to analyse new essays, extracting the same features used in training. By applying the learned patterns from the training stage, the machine learning model assigns a score to the new essay, mimicking the evaluation process of a human grader.

Due to their ability of being able to score, or provide feedback, on a large number of essays within a short period of time, AES systems are widespread in most US schools for standardized tests like the Ohio standardized test, the Utah compose tool and the Educational Testing Service (ETS; Ramesh & Sanampudi, 2022).

2.1.2 Application of AES technologies

While AES plays a role in large-scale standardized testing, its applications extend beyond these high-stakes assessments. Platforms of massive open online courses (MOOCs) like EdX, Coursera or Udacity have also incorporated AES systems into their platforms, allowing them to evaluate the written work of thousands of students enrolled in a single course (Stone et al., 2016 as cited in Murphy, 2019). Writing assistants like Grammarly or Turnitin also utilize AES technology. However, it's crucial to note that a significant difference exists in how these two categories of platforms leverage AES to evaluate writing.

Analytical scoring

The field of automated scoring uses various terms for different assessment systems. One such term is Automated Writing Evaluation (AWE), often used to describe tools like Grammarly, which is a writing assistant powered by AI, helping users improving their English writing skills by providing real-time feedback, colour-coded error highlighting, explanations for errors, performance reports and scores (Alam et al, 2023). AWE software relies on NLP, more specifically techniques like Latent Semantic Analysis (LSA) to assess a text's relationships and provide scores with feedback (Hockly, 2018 as cited in Kloppers, 2023). However, since this research focuses on the educational aspect of automated assessment, the term AES will be used throughout the research, encompassing both AES systems and AWE systems.

Similarly to the AWE engines, Grammarly employs “algorithms to identify morphological, syntactical, semantic, and pragmatic rules and patterns in research corpora” (Khurana et al., 2023 as cited in Kloppers, 2023, p. 2). AWE compares user-written text to a collection of corpora using specific rules and patterns. This comparison helps identify sections of the user's text that don't follow those established rules or differ from what's expected. (Grammarly Inc., 2023; Khurana et al., 2023, as cited in Kloppers, 2023). AWE then identifies these highlighted sections as potential errors and, if a match exists within the corpora, suggests corrections (Grammarly Inc., 2023, as cited in Kloppers, 2023). When automated scoring systems give basic input, instructional support, and exemplar writing samples, their method is called analytical. Similarly to Grammarly, Turnitin or Educational Testing Service (ETS) provide feedback on users’ writings the similar way. However, it is important to note that the feedback provided by each system differs in terms of type and level of detail (Murphy, 2019).

Regarding Grammarly’s efficiency, one study by Dizon and Gayed (2021) investigated how L2 (second language) English university students in Japan used the Grammarly mobile keyboard. Interestingly, it was found that Grammarly significantly improved the students' grammar and vocabulary usage but did not have a statistically significant impact on their writing fluency or syntactic complexity (Dizon & Gold, 2023). The result of this study suggests that tools like

Grammarly primarily improve the mechanics of writing, which are considered "low-level" features, but may not significantly impact aspects related to the flow of ideas and organization, which are considered "high-level" features.

What lead Gayed et al. (2022) to the conclusion that the employment of automated scoring systems by providing personalized feedback, can decrease L2 learners' mental workload while writing (Dizon & Gold, 2023). To be more specific, automated scoring systems free up users' working memory by handling low-level tasks like spelling, grammar, and translation, allowing them to focus on more complex aspects of writing, such as developing content and organizing their ideas (Dizon & Gold, 2023).

In summary, since analytical scoring provides feedback or suggestions in a more detailed manner, it could be beneficial for formative assessment purposes, where detailed feedback on specific skills is crucial for student improvement.

Holistic scoring

On the other hand, assessments that give a score for the overall writing competency, are more suitable for providing summative feedback. This so-called holistic scoring can be used to automate the evaluation of a vast number of essays and is commonly used for grading standardized aptitude tests like SAT and GRE (Ke & Ng 2019).

Instead of focusing on individual details, holistic scoring gives an overall impression of quality. However, scorers typically have a set of guidelines that outline what makes something good in that context (Huot, 1988; White, 1984 as cited in Singer & LeMahieu, 2011).

White (1994 as cited in Singer & LeMahieu, 2011) argues that holistic scoring provides a more accurate picture of writing because writing is too intricate to be simply broken down into separate categories like grammar and vocabulary. They believe these elements work together to create a whole that cannot be fully captured by analysing each part individually.

Besides this, another great benefit of holistic scoring is its speed and cost-effectiveness (Spandel & Stiggens, 1980 as cited in Singer & LeMahieu, 2011). This efficiency might explain its popularity, as evidenced by a recent survey by the National Writing Project (NWP) conducted in 2008. The survey found that close to 92% of U.S. states (46 out of 50) have direct writing assessments. Interestingly, of these states, a larger portion (67%, or 32 states) utilizes holistic scoring, while 44% (21 states) relies on analytic scoring. Seven states evaluate writing using both holistic and analytic scores (Singer & LeMahieu, 2011).

The different areas and examples of platforms where the holistic and non-holistic approach of AES are implemented are shown in Table 1.

Table 1

Platforms using different types of AES

Platform	Approach
EdX	Holistic
Coursera	Holistic
Udacity	Holistic
SAT	Holistic
GRE	Holistic
Cheg's Write Lab	Analytical
ETS' Criterion Online Writing Evaluation	Analytical
Grammarly	Analytical
Pearson’s Write to Learn	Analytical
Turnitin's Revision Assistant	Analytical

2.1.3 Potential benefits of AES

The benefits of implementing AES systems can make a significant difference in education and all stakeholders from institutions through educators to learners can profit from these. To begin with one of the greatest benefits, AES allows to provide quick feedback (Lewis, 2013). More importantly, holistic scoring technologies can assess a vast number of essays annually for standardized aptitude test like the SAT or GRE (Ke & Ng, 2019). Providing quick feedback serves both educators and students: educators can save a significant amount of time on grading students’ essays, receiving quick feedback can keep students motivated (Barker, 2001 as cited in Lewis, 2013).

Regarding feedback, AES systems not only provide feedback quickly, but studies have shown that by delivering criticism through an impersonal method, students are likely to be more open to accept the feedback (Lewis, 2013). The reason could be that receiving feedback from a computer can feel less personal and judgmental than receiving it from a teacher. This can create a safer learning environment where students are more open to constructive criticism.

Overall, the implementation of AES systems in education can improve feedback efficiency benefiting institutions, educators, and learners as well. However, despite these advantages, AES also has disadvantages that are going to be discussed below.

2.1.4 Pitfalls of AES

Naturally, AES systems also have some disadvantages. Since AES can have an influence on higher education admission or getting a job, it is important to address these. For example, depending on the approach of the AES system (holistic or non-holistic), the holistic approach like scoring one grade without further explanation is not sufficient in a classroom setting, as students do not get thorough feedback on what they should improve on their writing (Ke & Ng, 2019).

If an AES system is meant to give more thorough feedback than just a grade, often it prioritizes lower-level attributes of the essay (McNamara, 2015). This can contribute to students' writing skills development yet leaves a gap between "lower-level features, such as those related to grammar and spelling" (McNamara et al., 2015, p.53) and higher-order skills such as organization, coherence, revision or editing (Breland, 1983 as cited in Kennedy & Shiel, 2022). Moreover, there is the issue of technology acceptance. Lai (2010 as cited in Lewis, 2013) found that some students prefer peer feedback over computer scoring. "Students often worry that an AES cannot understand novel ideas and concepts, or properly grade answers that were not part of its training" (Landauer, Laham & Foltz,2000, as cited in Lewis, 2013). This highlights the importance of familiarizing students with AES and demonstrating its capabilities to build trust and confidence in the technology. Building trust in automated scoring systems is key to their wider adoption. To achieve this, promoting transparency in their workings is crucial. By providing a comprehensive guide that explores their applications, expected performance, variations in feedback across categories, and most importantly, strategies for maximizing benefit and fostering calibrated trust (Ranalli, 2021), teachers and students will be better equipped to interpret the feedback they receive and have greater confidence in its accuracy. Moreover, by proper introduction and education of these systems, both the over and under-reliance of the systems could be avoided (Lee & Moray, 1992; Lee & See, 2004 as cited in Ranalli, 2021).

Additionally, feedback provided by AES systems lacks human interaction (Beseiso et al., 2021), which can be eliminating, especially in a classroom setting. Deane (2013) argues that the inherent social nature of writing necessitates human readers for all written work, irrespective of its purpose. Expecting students to write for a machine could imply that writing is “not valued as human communication” (Deane, 2013, p.8). While this might be true, it is also important to note that due to the extremely rapid development of technology, students have access to AI-based writing tools like ChatGPT or Gemini that they can employ for certain writing activities, therefore it would be unfair to expect the already overworked educators to not employ an AES system for helping their work.

While AES offers a promising approach to automated essay scoring, it's crucial to acknowledge its limitations. Studies have shown that some systems can be manipulated to generate undeservedly high scores with nonsensical writing what raises significant concerns, particularly when considering the use of AES in “high-stakes testing situations” (Powers et al.,2001; Kolowich, 2014 as cited in Murphy 2013, p. 9). For instance, Powers et al. (2002) conducted a study specifically challenging the validity of an e-rater. They were able to trick the system into awarding unrealistically high scores through a variety of techniques, including repetitive paragraphs, minor sentence alterations, and strategically substituting keywords. These findings highlight the potential for exploitation by test-takers seeking an unfair advantage.

The rise of AI, predicted to dramatically reshape society and work as part of the "Fourth Industrial Revolution" (Schwab 2016; Timms 2016 as cited in Farrow, 2023), is already making waves in education. However, this rapid development raises concerns about fairness, accountability, transparency, and ethics in AI design. This is where Explainable AI (XAI) comes in. XAI aims to shed light on how AI programs make decisions, making it clear who might be accountable for the outcomes (Farrow, 2023). Within the field of XAI, a particularly new term is emerging: Explainable AI in Education (XAIED). The scarcity of results on educational databases (only 1 on ERIC and 2 on Google Scholar) underscores the novelty of this specific focus.

Moreover, “without an understanding of how a model arrived at a particular decision, it is difficult to identify the source of any bias and inaccuracies and then correct them (Murphy, 2019, p. 13)”. This issue can create a lack of trust in such systems in the users.

2.2 Rubrics

While AES systems can provide feedback on the overall quality of a written text, it has been found that "many AES algorithms are driven principally by lower-level features, such as those related to grammar and spelling" (McNamara et al., 2015, p.53). Since these systems focus more on the linguistic properties of the students' writings, the content and ideas are not being evaluated (Correnti et al., 2022). This creates a shallow assessment, painting a limited and potentially inaccurate picture of students' abilities. Further, crucial aspects like critical thinking, subject-specific knowledge, and creative expression are being left out.

Kennedy & Shiel (2022) have "developed a rubric emphasising both higher-and lower-order processes to support formative assessment in Pre-K to Grade 2" (p. 130). The rubric consists of five key components (ideas, organisation, word choice, voice, and conventions) that have the potential to offer learners comprehensive feedback on their writing, indicating areas that require improvement. This research aims to focus on the word choice component (see Appendix A for the word choice component of the rubric).

With the use of assessment rubrics like Kennedy & Shiel (2022), generating texts based on assessment rubrics could address the limitations caused by the lack of corpora. This approach tackles multiple challenges simultaneously. Firstly, generating essays aligned with various criteria in the rubrics creates a vast amount of data for training and testing, overcoming the limitations of limited real-world data.

Secondly, these algorithms can be designed to incorporate diverse writing styles and perspectives, ensuring that the generated essays are representative of various student backgrounds. This not only could reduce potential biases present in limited corpora but also promotes fairer assessment for all students.

To avoid such, distinguishing essay dimensions and implementing them into AES systems could support the improvement of AES systems. However, a significant obstacle in advancing research on dimension-specific essay scoring is the limited availability of corpora that have been manually annotated with dimension-specific scores. Training complex models requires a large amount of expensive, expert-rated data, making data sparsity a persistent challenge to overcome in the field of AES (Li & Zhou, 2019), hindering development, evaluation, and the overall progress in creating reliable and unbiased automated essay scoring systems.

2.3 Scientific and practical relevance

Understanding the characteristics and the operational processes of automated scoring could help transform assessment. Additionally, investigating how effectively automated assessment systems score could assist to the improvement of AES making users have more trust in such systems.

An extremely important aspect of AES implementation roots in its accountability. Unlike a human rater who can be held responsible for their decisions, a machine simply cannot be. As Kim & Doshi-Velez (2021) define it, accountability involves ensuring an AI system acts as intended, which is crucial for determining blame if issues arise. They further propose several approaches to achieve this, including “transparency (data, process, and open-source software), interpretable models, post hoc inspection of outputs, empirical performance (pre-market and post-market), and properties guaranteed by design” (Kim & Doshi-Velez, 2021, p. 48). In the context of AES, transparency, interpretable models, and empirical performance hold promise for establishing accountability.

Educators globally experience stress due to their demanding workload that includes but not limited to lesson planning, organizing activities, developing curriculum, disciplining, administrative tasks and, evaluating and assessing students’ performance (Desouky and Allam, 2017, as cited in Jomuad et al., 2021). As a result, a growing number of teachers are leaving the field worldwide, resulting in a lack of qualified educators (European Commission, 2018; Ingersoll, 2017, as cited in Toropova et al., 2021). While the employment of an automated scoring system could not solve all the challenges and hardships teachers face, it could certainly offer them a solution for time constraints by fast essay scoring, plagiarism detecting and unbiased feedback providing. In addition, students could also benefit from receiving instant feedback on their essay, which can facilitate self-directed learning by making students correcting their work according to the feedback they received.

Staying in the field of education, an advanced AES system could contribute to cost-efficiency for institutions and test centres, as much less manual scoring would be required, provided a well-functioning AES system (Ke & Ng, 2019).

From a scientific perspective, validity of AES systems has always been in the focus of research and debates, concluding that it needs improvement. While several advanced AES systems have been developed, Shermis & Burstein (2013) believe that “it should be construed primarily as a complement to (instead of replacement for) human scoring”. Therefore, the current study contributes to the development of a combined essay scoring rubric, in which human scoring mechanisms is effectively combined with automated scoring systems, potentially benefitting stakeholders, and improving the quality of education and assessment.

Research questions

The presented theoretical framework is the basis for the following research questions: *"How effectively can automated essay assessment systems assess AI-generated texts?"*

1. How accurately can the automated essay assessment system classify the generated texts?
2. Does data augmentation increase the accuracy of rubric-based AES systems?

3. RESEARCH DESIGN AND METHODS

3.1 Research design & participants

To examine the efficiency of automated scoring, a mixed method design was used in this study. Firstly, 100 student-written texts were generated using AI, in English. Two different levels were generated with prompts that are based on the rubric developed by Kennedy & Shiel (2022). The levels described in the prompts served as the "ground truth" for comparing it to the automated ratings. The term "ground truth" here refers to the correct information used to train and test a machine learning model to teach the model the relationship between inputs and outputs and to check how well it learns (Lebovitz et. al., 2021).

Automated text classification was performed in a supervised machine learning context. Since we artificially generated, the respondents were simulated. Following the rubric, the generated data could stem from 7- to 8-year-old primary students, who are in grade 2 and native English speakers are from Dublin.

3.2 Writing assessment rubric

Kennedy and Shiel (2020) developed a rubric that intends to assess the writings of PreK-Grade 2 students. The rubric is divided into seven levels, and it includes five key components, namely, ideas, organisation, word choice, voice, and conventions. This rubric was used as a basis for data generation prompts, as described in the next section.

3.3 Data generation using Bard

Bard is a chat-based AI-tool run by Language Model for Dialogue Applications (LaMDA) developed by Google. LaMDA is a language model "specialized for dialog, which have up to 137B parameters and are pre-trained on 1.56T words of public dialog data and web text" (Thoppilan et al., 2022. p. 1). Users can interact with Bard in a humanlike manner and "possess the extraordinary ability to ingest text prompts and conjure up unique outputs, be it composing emails, processing information, or conducting online research" (Ahmed et al., 2023, p. 3).

For this study Bard was used in September-October 2023 to generate texts using prompts, that were based on the previously described assessment rubric of Kennedy & Shiel (2022). The quality of the generated data mostly depended on the accuracy of the prompt Bard was to be fed with. The prompt contained the stimulated respondents' parameters like age, grade, knowledge, and the writing components assessed by rubrics. To generate the most accurate texts, it was important to include what the stimulated respondents know and not know at each level. The simulated texts were of various writing genres, such as explanatory, narrative, creative and research writing.

While the rubric developed by Kennedy & Shiel (2022) contains seven levels, due to time constraints this research examines two levels, namely Level 2 and Level 4. Table 2 illustrates the features of each level that were included and excluded in the prompts.

Table 2

Features of Each Level

	Features	Features from one level higher (not to include)
Level 2	words used include environmental print and regular words or words from familiar context	simple, everyday adjectives simple verbs
Level 4	simple everyday adjectives and/or varied verbs in informational text, disciplinary language begins to appear	advanced adjectives advanced verbs

Kennedy & Shiel (2022) propose a rubric for text levels. While Level 2 texts should consist of regular or familiar words, the prompts used here included some Level 3 features from the rubric.

This was done to ensure Bard generates even more accurate text. To prevent Level 3 text generation, however, the prompts deliberately omitted features like adjectives and verbs. Table 3 shows examples of the prompts that were used for text generation for both level 2 and level 4 texts.

Table 3

Prompt Examples for Text Generation

Level	Prompt
Level 2	"Generate a book review that is not longer than 100 words. The topic is Pippi Longstocking. Use regular words or words from familiar context. Don't use adjectives. Don't use verbs. Write as if you were an 8-years-old grade 2 student from Dublin, Ireland."
Level 4	"Generate a research report that is not longer than 100 words. The topic is dinosaurs. Use simple everyday adjectives and/or varied verbs (e.g. walked for went or delicious for nice) to make the text interesting. Don't use advanced adjectives. Don't use advanced verbs. Write as if you were an 8-years-old grade 2 student from Dublin, Ireland."

Although Bard has proven to be notably effective in text generation, the first generated drafts were often far from ideal despite the accurate prompts. The most common issues with Bard-generated texts were (1) with the word count: the generated texts were either too short (Appendix 2) or too long (Appendix 3); (2) the vocabulary was too advanced (Appendix 4) or (3) certain expressions and sentences were too frequently used (Appendix 5).

A frequently reoccurring issue with Bard was the inaccurate word count. While the prompts clearly instructed Bard to generate texts that are "not longer than 100 words", Bard several times went over this limit with 30-50 words. Then after the instruction to "shorten" the texts, the texts ended up being significantly shorter than 100 words.

Based on these issues, Bard was found to efficiently generate texts of various topics, however, taking every element of a prompt into account tended to be problematic for Bard. Moreover, the model may not be able to accurately replicate the characteristics of a primary school child's writing.

3.4 BERT text classification algorithm

For the automated text classification in this study, the language model, BERT was used. BERT stands for Bidirectional Encoder Representations from Transformers and "is able to cope with natural language processing (NLP) tasks such as supervised text classification without human supervision" (Garrido-Merchan et al., 2022, p. 1). A successful example of efficient text classification is a systematic review conducted by Aum & Choe (2021) in which BERT was employed to automate the classification of articles. Their BERT model, called srBERT, was first trained on abstract of articles from various sources. Then, it was fine-tuned using the titles of the articles. As a result, BERT outperformed all the previous models on both text classification and relation-extraction tasks.

3.5 Data augmentation algorithm

Since "the BERT approach needs huge amounts of texts to deliver proper results" (Garrido-Merchan et al., 2022, p. 2), concerns arose regarding the limited sample size of the data employed. Models often exhibit enhanced performance when subjected to random shuffling of data points within the training dataset (Mishchenko et al., 2020).

Data augmentation is used to increase effective sample size, by creating additional synthetic data, through applying slight changes to existing data. These changes are designed to be artifacts in the data, that are not relevant for the scoring, such that it can focus on learning the important patterns (Shorten et al., 2021). This data augmentation process is hypothesized to strengthen the association between variables and their respective levels by introducing novel input patterns through intra-level data augmentation.

To prevent the generation of nonsensical text and facilitate the analysis of incorrectly classified sentences, data augmentation was applied only within the training data structure, rather than directly modifying the order of the actual sentences.

This approach aims to mitigate the potential limitations imposed by the small sample size and explore the impact of data augmentation on BERT's performance. The comparative analysis of the three scenarios (0 rounds, 5 rounds and 10 rounds of data augmentation) provides insights into the effectiveness of data augmentation and its role in enhancing BERT's interpretation abilities under constrained data conditions.

Without data augmentation (0 rounds) and a 20% test split, the training set contained 80 texts, while the testing set had 20. For a larger number of augmentation rounds, the original 80 texts (assuming a 20% test split) served as the base. With 5 rounds of augmentation, the training set expanded. We multiplied the original 80 texts by 5 (resulting in 400) and added them back to the original set, bringing the total to 480 training texts. Importantly, the test set remained constant regardless of the number of augmentation rounds. It would always consist of only 20 texts. Following this logic, 10 rounds of augmentation would increase the size of the training set to 880 texts while the test set's size would remain 20 texts. It is important to note that the rounds of data augmentation only affect the data itself, not the model.

3.6 Procedure & data analysis

100 texts were generated by Bard using prompts. The data were divided into two groups: a training and a testing set. The text generation was followed by Data Augmentation to create a larger number of texts for the BERT model. The resulting enriched datasets were then introduced to the Bert model. With the use of these sets, the BERT model was trained on the training dataset. Finally, the results of the model on the test set were analysed to identify how accurately and efficiently automated essay scoring works.

This study initially generated 100 texts. To increase effective sample size, data augmentation was applied. Afterwards, a text scoring model was trained using BERT was applied, and texts were automatically scored using BERT.

BERT was trained using the training set (80%) of the data and evaluate it using a test set (20%) of the data. Afterwards model performance was evaluated in terms of accuracy, by comparing the model predictions with the known scores.

To eliminate any potential random allocation bias (i.e. something) and therefore provide a robust assessment of data augmentation's efficacy, BERT had to classify the texts in multiple iterations. Iterations refer to the testing of the model and show how many times the data was ran by the model. In this study, we compared 5, 15, 30, 50, 100, and 200 iterations.

4. RESULTS

Besides examining the potential benefits of employing data augmentation to enhance the model's overall performance, this BERT model was also being evaluated under three distinct data augmentation rounds: 0, 5, and 10. The evaluation process was conducted with a test size of 0.2 (i.e. 20 texts).

The experiments with multiple rounds (0 rounds resulting in 100 texts, 5 rounds resulting in 480 texts, 10 rounds resulting in 880 texts) of data augmentation and BERT yielded unexpected findings. As shown in Table 4, contrary to H3, increasing the training set size through data augmentation did not consistently lead to improved accuracy: as data is augmented more, the interpretation accuracy of BERT decreases.

Table 4

Output for Rounds of Data Augmentation and Iterations

Rounds of data augmentation	Iterations					
	5	15	30	50	100	200
0	0.71	0.73	0.68	0.71	0.69	0.70
5	0.65	0.65	0.69	0.66	0.68	0.67
10	0.73	0.71	0.69	0.66	0.69	0.69

Note. Test size = 0.2

As evident from Table 4, there is no big difference between the performance of BERT regardless of the rounds of data augmentation and the number of runs. BERT achieved its highest accuracy of 0.73 at both the extreme settings of data augmentation: 10 rounds (highest) and with only 5 iterations (lowest), meaning that the original 100 texts were enriched via 10 rounds of data augmentation with slight changes of the original texts and for this result the BERT model ran 5 times.

The same accuracy (0.73) was achieved through 0 rounds (lowest) with 15 iterations, meaning that the original dataset without any rounds of data augmentation and 15 times of model runs result in the same high accuracy. However, when employing 5 rounds of data augmentation, the accuracy dropped to 0.65.

Further increasing the number of iterations did not yield higher accuracy: 15 iterations appeared to be sufficient for accurate predictions. Further iterations would only have increased training time without any improvement in accuracy.

5. DISCUSSION

To comprehensively evaluate the efficacy of automated scoring, this study employed a mixed-methods design. Initially, a corpus of 100 English-language student essays was generated using Google Bard, based on word component dimension of the rubric from Kennedy and Shiel (2022). Following that, these essays were subjected to automated text classification within a supervised machine learning framework.

Since training BERT requires a substantial amount of data (Aum & Choe, 2021), data augmentation was used to enrich the existing dataset hoping to increase the accuracy of the language model. The initial findings verify the earlier hypothesis that data augmentation the training set can enhance BERT's performance, but only to a certain extent.

The results of the automated text classification suggest that data augmentation did not increase accuracy. On the contrary, the more data augmentation, the worse the BERT's interpretation accuracy gets.

By introducing excessive noise and disrupting the underlying patterns embedded in the data, data augmentation hurts the holistic nature of the texts, which can hinder BERT's ability to accurately score texts. According to Singer & LeMahieu (2011), scoring texts holistically is not just about adding up individual pieces; it also considers how well individual pieces of the text fit together. The final product can be stronger (or weaker) than the sum of its parts.

Without, respectively with less, data augmentation on the other hand, the original data structure is maintained, potentially providing BERT with a cleaner and more consistent learning corpus, maintaining accuracy.

In summary, the benefits of data augmentation are outweighed by the introduction of noise and pattern disruption when the original training set is relatively small. When the dataset is sufficiently large, data augmentation can provide subtle variations that enhance BERT's generalization and robustness. However, with a limited training set, reshuffling disrupts the essential patterns and correlations, leading to a decline in accuracy.

5.1 Limitations

This study has confirmed that automated essay scoring in the context of text classification can be employed efficiently. However, there are a few limitations that affected the study. First, the lack of authentic student-written data possibly had an impact on the results of the text classification. While AI-generated essays (in this case Google Bard by which essays were generated in September-October, 2023) may exhibit a range of features and styles, they lack variety, both in their word choice and how they present arguments (Zhu, 2022 as cited in Corizzo & Leal-Arenas, 2023). Student essays often contain unique phrasings, errors, and creative approaches that deviate from the patterns AI might generate. This can lead to the automated scoring system performing well on the AI-generated data but struggling with the real-world variations found in student-written essays.

Another limitation is the rather small size of the original dataset. Training requires substantial labelled data, but gathering such data can be difficult, particularly within specific domains (Aum & Choe, 2021). With smaller datasets, the model struggles to accurately identify and interpret complex patterns within the text corpus, even with a reduced test size. Furthermore, as the test size grows and the training size declines, the model encounters a wider range of text variations, further taxing its ability to generalize and maintain accuracy.

In conclusion, the study identified limitations due to the use of a small dataset and inauthentic, generated training data. This suggests that using a larger and student-written dataset could have yielded different results.

5.2 Further research

Considering these findings, future research should prioritize the use of authentic student essays for training and testing automated scoring systems. This would ensure that the system is exposed to the natural variations in student writing styles and can provide more accurate feedback that could be implemented into education practises saving time and effort for educators and test raters.

Moreover, future research should expand the corpus size. A larger corpus supports generalizability by exposing the model to a wider range of writing styles and topics, reducing bias from limited data, and improving its ability to handle uncommon writing styles.

Besides the corpus size and authenticity, fine tuning BERT could possibly improve the model's performance. Targeted fine-tuning BERT on a rubric-specific dataset could potentially enhance its ability to capture the nuances of essay quality as defined by the evaluation criteria.

5.3 Conclusion

The findings of this research underscore the intricate interplay between data randomization, dataset size, and BERT's performance for rubric-based automated essay scoring. While randomness can introduce valuable variations in writing styles, excessive randomization or a small dataset can introduce noise and hinder the model's ability to learn the nuances required for accurate scoring based on rubrics. This highlights the importance of using a carefully curated and balanced corpus that incorporates diverse writing styles while maintaining consistency with the evaluation criteria outlined in the rubrics.

References

Ahmed, I., Kajol, M., Hasan, U., Datta, P. P., Roy, A., & Reza, M. R. (2023). ChatGPT vs. Bard: a comparative study. *UMBC Student Collection*. https://doi.org/10.36227/techrxiv.23536290.v2

Alam, S., Usama, M., Alam, M. M., Jabeen, I., & Ahmad, F. (2023). Artificial Intelligence in Global World: A Case Study of Grammarly as e-Tool on ESL Learners' Writing of Darul Uloom Nadwa. *International Journal of Information and Education Technology*, *13*(11). https://doi.org/10.18178/ijiet.2023.13.11.1984

Amorim, E., Cançado, M., & Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). https://doi.org/10.18653/v1/n18-1021

Aum, S., & Choe, S. (2021). srBERT: automatic article classification model for systematic review using BERT. *Systematic reviews*, *10*(1), 1-8. https://doi.org/https://doi.org/10.1186/s13643-021-01763-w

Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, *33*, 727-746. https://doi.org/https://doi.org/10.1007/s12528-021-09283-1

Breland, H. M. (1983). The Direct Assessment of Writing Skill: A Measurement Review. College Board Report No. 83-6.

Corizzo, R., & Leal-Arenas, S. (2023). One-class learning for ai-generated essay detection. *Applied Sciences*, *13*(13), 7901. https://doi.org/https://doi.org/10.3390/app13137901

Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open*, *3*, 1-15. https://doi.org/https://doi.org/10.1016/j.caeo.2022.100084

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7-24. https://doi.org/https://doi.org/10.1016/j.asw.2012.10.002

Dizon, G., & Gold, J. (2023). Exploring the Effects of Grammarly on EFL Students' Foreign Language Anxiety and Learner Autonomy. *JALT CALL Journal*, *19*(3), 299-316. https://doi.org/https://doi.org/10.29140/jaltcall.v19n3.1049

Farrow, R. (2023). The possibilities and limits of XAI in education: a socio-technical perspective. *Learning, Media and Technology*, 1-14. https://doi.org/https://doi.org/10.1080/17439884.2023.2185630

Garrido-Merchan, E. C., Gozalo-Brizuela, R., & Gonzalez-Carvajal, S. (2022). Comparing BERT against traditional machine learning models in text classification. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE3202838

Jomuad, P. D., Antiquina, L. M. M., Cericos, E. U., Bacus, J. A., Vallejo, J. H., Dionio, B. B., Bazar, J. S., Cocolan, J. V., & Clarin, A. S. (2021). Teachers' workload in relation to burnout and work performance. *International journal of educational policy research and review*. https://doi.org/https://doi.org/10.15739/IJEPRR.21.007

Ke, Z., & Ng, V. (2019). Automated Essay Scoring: A Survey of the State of the Art. IJCAI,

Kennedy, E., & Shiel, G. (2022). Writing assessment for communities of writers: rubric validation to support formative assessment of writing in Pre-K to grade 2. *Assessment in Education: Principles, Policy & Practice*, *29*(2), 127-149. https://doi.org/10.1080/0969594X.2022.2047608

Kim, B., & Doshi-Velez, F. (2021). Machine learning techniques for accountability. *AI Magazine*, *42*(1), 47-52.

Kloppers, J. (2023). Errors in errors: an exploration of Grammarly's corrective feedback. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *13*(1), 1-16. https://doi.org/10.4018/IJCALLT.325792

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is ai ground truth really true? The dangers of training and evaluating ai tools based on experts'know-what. *MIS quarterly*, *45*(3). https://doi.org/DOI: 10.25300/MISQ/2021/16564

Lewis, J. (2013). Ethical Implementation of an Automated Essay Scoring (AES) System: A Case Study of Student and Instructor Use, Satisfaction, and Perceptions of AES in a Business Law Course. *Satisfaction, and Perceptions of AES In a Business Law Course*. https://doi.org/http://dx.doi.org/10.2139/ssrn.2684803

Li, J., & Zhou, J. (2019). RefNet: Automatic Essay Scoring by Pairwise Comparison.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35-59. https://doi.org/https://doi.org/10.1016/j.asw.2014.09.002

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/https://doi.org/10.1016/j.rmal.2023.100050

Murphy, R. F. (2019). Artificial intelligence applications to support K-12 teachers and teaching. *RAND Corporation*, *10*. https://doi.org/https://doi.org/10.7249/PE315

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

Shermis, M. D., & Burstein, J. (2013). Handbook of automated essay evaluation. *NY: Routledge*.

Shin, J., & Gierl, M. J. (2022). Evaluating coherence in writing: Comparing the capacity of automated essay scoring technologies. *Journal of Applied Testing Technology*, 04-20.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text data augmentation for deep learning. *Journal of big Data*, *8*, 1-34. https://doi.org/https://doi.org/10.1186/s40537-021-00492-0

Singer, N. R., & LeMahieu, P. (2011). The effect of scoring order on the independence of holistic and analytic scores. *Journal of Writing Assessment*, *4*(1).

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., & Du, Y. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*. https://doi.org/https://doi.org/10.48550/arXiv.2201.08239

Toropova, A., Myrberg, E., & Johansson, S. (2021). Teacher job satisfaction: the importance of school working conditions and teacher characteristics. *Educational review*, *73*(1), 71-97. https://doi.org/https://doi.org/10.1080/00131911.2019.1705247

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*(2), 459-484. https://doi.org/https://doi.org/10.1007/s41237-021-00142-y

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87-125. https://doi.org/https://doi.org/10.1177/07356331198307

Yun, J. (2023). Meta-Analysis of Inter-Rater Agreement and Discrepancy Between Human and Automated English Essay Scoring. *English Teaching*, *78*(3), 105-124. https://doi.org/https://doi.org/10.15858/engtea.78.3.202309.105

Appendix

Appendix A

Word Choice Component of the Write to Read Rubric (Kennedy & Shiel, 2022, p. 147-148)

Word Choice Subcomponents	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
A Word Level	Symbols are used to represent words and meaning is unclear without the child's input.	As previous, and words used include environmental print and regular words (Tier 1), or words from familiar contexts (TV, film, games).	As previous and simple everyday adjectives (e.g. colour, nice/big) [Tier 1 & 2] begin to appear and simple verbs *(e.g. look, went) are used.*	As previous and simple everyday adjectives and/or varied verbs (e.g. walked for went; *delicious for nice, fantastic*) are used occasionally to make text interesting. In informational text disciplinary language begins to appear *(e.g. antennae, aphid)*	As previous and some Tier 2 advanced adjectives are used to make text interesting (in at least one part of text) (e.g. *extraordinary, desperate, enormous)* and verbs (e.g. *peeked, plucked, pounced,*; Informational text describes rather than tells *(e.g. A ladybird is a small red beetle with black spots)*	As previous and uses more advanced adjectives and varied verbs more consistently *(in more than one part)* In informational text, description is more precise and combines with disciplinary language (*e.g. Black rhinos have a hooked upper lip that they use to pluck leaves off bushes and trees*);	As previous, with some experimentation with precise nouns (e.g. *cottage/r house, stool/ chair*, to make text interesting. In informational texts disciplinary words are used consistently and descriptive words are precise and apt.
B Sentence Level	As above	Labels, captions are present and/ or lists of words.	Simple sentences *(e.g. I like to . . .)* begin to emerge. May be patterned sentences.	Simple sentences are present; patterned sentences are more complex and complete. In informational text: facts are in simple sentences *(e.g. A ladybird can fly. It has spots. It has antennae.)*	As previous and sentence openers begin to show variety. Has moved beyond patterned sentences.	Sentence openers show variety and sentences begin to vary in length. In informational text, structure is varied e.g. with intentional use of present tense/third person narrative/ timeless verbs	As previous and variety in sentence openers and sentence length adds to the overall text. In informational text, timeless verbs/ present tense is used consistently; comparisons may be drawn (e.g. *Rhinos have weak eyesight but excellent hearing.)*

(Continued)

(Continued).

Word Choice Subcomponents	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
C Imagery and Creative Language Usage	As above	Reliance on picture (if present) to connect with the reader and/ or lack of details in text *(e.g. see cat)*	Little or no use of imagery as phrases, lists or sentences are basic and connect to the reader in a general way *(e.g. I like the big cat)*	Experiments with simple descriptive language or literary phrases (e.g. *I like cake. Yummy in my tummy*; two simple adjectives for effect: *A fat pink pig, a big bad wolf*)	Begins to experiment with more complex descriptive language or literary phrase and uses words inventively *(e.g. I was like a bomb exploding with excitement)*. In informational text: *A rhino's skin is bumpy and lumpy)*	Uses more than one literary phrase to enhance writing and engage readers *(e.g. Mat swooped into the night; the evil witch decided to make an elixir)*. In informational texts attributes of animals/ habitats are accurate and described clearly: (e.g. *Ants have six legs*); similes may be used for effect: *the frog is as small as my fingernail.*	Figurative or literary language is used more consistently and effectively in more than one part of text to create vivid mental pictures (e.g. *I was so nervous I felt like there were ants crawling around my stomach)*; In informational text literary or disciplinary language or syntax, enhances writing *(e.g. She swims out of the slimy seaweed to get her lunch. Snatch! Her stomach is now satisfied).*

Appendix 2

Example of a Word Count Problem: Too Short

Generate a non-fiction story that is not longer than 100 words. The story is about your summer vacation. Use regular words, or words from familiar contexts. Don't use adjectives. Don't use verbs. Write as if you are a 8 year old grade 2 student from Dublin, Ireland.

I went to the beach with my family. We built a sandcastle and played in the waves. I saw a big fish. We had a picnic lunch and ice cream. I had a lot of fun.

Appendix 3

Example of a Word Count Problem: Too Long

Generate a narrative writing about this picture that is NOT LONGER than 100 words. The title is "If you could meet anyone, who would it be and why?". Use regular words, or words from familiar contexts. Don't use adjectives. Don't use verbs. Write as if you are a 8 year old grade 2 student from Dublin, Ireland.

If you could meet anyone, who would it be and why?

If I could meet anyone, I would meet Zoey Lyndon. She is the main character in my favorite book, "Zoey Lyndon's Big Move to the Lou."

Zoey is a very brave girl. She moves to a new city and starts at a new school, but she doesn't give up. She makes new friends and she even learns to play the trumpet!

I would love to meet Zoey and learn more about her. I would ask her about her new school, her new friends, and her new life in St. Louis. I would also ask her for some advice on how to be brave when things are tough.

I know that Zoey Lyndon is just a fictional character, but she is a very real person to me. I would be so happy if I could meet her someday.

(Word count: 89)

shorten it

Sure, here is a shortened version of your narrative writing:

If you could meet anyone, who would it be and why?

If I could meet anyone, I would meet Zoey Lyndon. She is the main character in my favorite book.

Zoey is brave. She moves to a new city and starts at a new school. She makes new friends and learns to play the trumpet.

I would love to meet Zoey and learn more about her.

(Word count: 54)

Appendix 4

Example of Using Too Advanced Vocabulary

Generate a narrative writing about this picture that is NOT LONGER than 100 words. The title is "What would you do if your favourite toy started talking? What would they say?" Use simple, everyday adjectives and /or varied verbs (e.g. walked for went or delicious for nice) to make the text interesting. Don't use advanced adjectives. Don't use advanced verbs. Write as if you are an 8 year old grade 2 student from Dublin, Ireland.

What would you do if your favorite toy started talking? What would they say?

If my favorite toy, my teddy bear named Teddy, started talking, I would be very surprised and excited! I would ask him all sorts of questions, like what it's like to be a toy, and what he likes to do when I'm not playing with him.

I think Teddy would say that he loves being my toy, and that he enjoys spending time with me. He might also tell me some funny stories about the other toys in my room.

I would be so happy to have a talking toy bear as my friend. We would play together all day long, and we would tell each other secrets. I would never want to let him go.

Here is a short story about what could happen if my favorite toy started talking:

"I'm not just a teddy bear," said Teddy. "I'm a magical teddy bear. I can do all sorts of things that other teddy bears can't."

"What kind of things?" I asked.

"I can fly," said Teddy. "And I can turn invisible. And I can even make your wishes come true."

I couldn't believe it! I had a magical teddy bear for a friend!

Teddy and I went on lots of adventures together. We flew to the moon and back. We turned invisible and played pranks on my parents. And Teddy even made my wish for a new puppy come true.

I'm so glad that Teddy started talking to me. He's the best friend a kid could ask for.

(99 words)

I kept the writing simple and used everyday adjectives and verbs. I also avoided using advanced sentence structures. I tried to write as if I were an 8 year old grade 2 student from Dublin, Ireland, by using contractions (e.g., "I'm" instead of "I am") and informal language (e.g., "I would love to" instead of "I should enjoy").

I also tried to make the text interesting by providing specific details about what Teddy would say if he could talk, and by telling a short story about one of our adventures together.

Teddy's Secret

One day, I was playing with my teddy bear, Teddy, when he suddenly started talking! I was so surprised that I dropped him.

"Teddy?" I said. "Can you talk?"

"Yes, I can talk," said Teddy. "I've been able to talk for a long time, but I was afraid to tell you."

"Why were you afraid?" I asked.

"I was afraid that you wouldn't believe me," said Teddy. "Or that you would be scared of me."

"I'm not scared of you," I said. "I'm just surprised."

"I'm glad," said Teddy. "I've been wanting to talk to you for a long time."

Teddy and I talked for hours that day. He told me all sorts of things about the other toys in my room, and he even told me a secret.

Appendix 5

Example of Frequently Used Expressions, Sentences

Elephants are gentle animals. They like to play and swim.

Elephants are good for us. They help to keep the environment healthy. We should be kind to elephants.

Bees are gentle animals. They like to fly and buzz.

Bees are good for us. They help to grow the food we eat. We should be kind to bees.

Birds are gentle animals. They like to sing and play.

Birds are good for us. They help to keep the environment healthy. We should be kind to birds.

Squirrels are gentle animals. They like to play and chase each other.

Squirrels are good for us. They help to keep the forest healthy. We should be kind to squirrels.

Dogs are gentle animals. They like to be petted and played with.

Dogs are good for us. They give us companionship, love, and protection. We should be kind to dogs.