Online Learning in Educational Research
Volume 5, Issue 2, 319-337.
e_ISSN: 2798-2351
https://www.journal.foundae.com/index.php/oler/index
DOI: 10.58524/oler.v5i2.753

Deep Learning in Automated Essay Scoring for Islamic Education: A Systematic Review

Rokhmatul Khoiro Amin Putri*
UIN Sunan Ampel Surabaya, INDONESIA

Kusaeri Kusaeri
UIN Sunan Ampel Surabaya, INDONESIA

Suparto Suparto
UIN Sunan Ampel Surabaya, INDONESIA

Article Info

Article history:

Received: June 3, 2025
Revised: August 12, 2025
Accepted: October 28, 2025

Keywords:

Assessment;
Automated Essay Scoring;
Deep Learning;
Islamic Education;
Systematic Review.

Abstract

Automated Essay Scoring (AES) is a computer-based scoring system that uses appropriate features to automatically assess or give feedback to students, by combining the power of Artificial Intelligence and natural language processing (NLP) to provide convenience and benefits for evaluators. This study aims to analyze the most effective algorithmic models in evaluating the accuracy and reliability of the Automated Essay Scoring (AES) system, especially in the context of Islamic religious education assessment, as well as examine its advantages and disadvantages in supporting objective and efficient learning evaluation. This study uses the Systematic Literature Review (SLR) approach by following the PRISMA protocol. A total of 31 relevant articles published in the period 2020 to 2025 from the Scopus and Springer databases were analyzed to evaluate the use and effectiveness of algorithms in the development of AES systems. The results show that transformer-based models, specifically BERT, are the most effective algorithms in current AES implementations. BERT excels because of its ability to understand bidirectional context and semantic depth in text. These models generate accurate scores and can provide automated feedback that is close to the quality of human judgment. However, the use of BERT requires large training data and high computing resources. While BERT demands substantial data and computing power, its application in Islamic education highlights the potential of AES to support more objective, consistent, and scalable assessment of students' essays.

To cite this article: Putri, R. K. A., Kusaeri, K., & Suparto, S. (2025). Deep Learning in Automated Essay Scoring for Islamic Education: A Systematic Review. *Online Learning in Educational Research, 5*(2), 319-337. https://doi.org/10.58524/oler.v5i2.753

INTRODUCTION

The shift in the use of computers to assess or score an essay question must be more efficient and accurate than traditional methods (Sevcikova, 2018) In various areas of assessment, essay tests have attracted a lot of attention as a way to measure practical and high-level abilities, such as logical thinking, critical reasoning, and creative thinking (Abosalem, 2015; Hussein et al., 2019; Bernardin et al., 2016; Liu et al., 2014; Rosen & Tager, 2014). Essay tests are widely used to assess higher-order skills such as logical thinking, critical reasoning, and creativity (Abosalem, 2015; Hussein et al., 2019). However, manual grading is often costly, time-consuming, and inconsistent due to assessor subjectivity (Han, 2019; Rahman et al., 2017). These challenges have driven interest in computer-assisted solutions to improve efficiency and reliability.

Automated Essay Scoring (AES) offers a promising alternative by combining the power of *Artificial Intelligence* (AI) and *Natural Language Processing* (NLP). Through the use of key linguistic and structural features, AES offers immediate feedback in digital learning environments. By

* **Corresponding Author**
Rokhmatul Khoiro Amin Putri, UIN Sunan Ampel Surabaya, Indonesia ✉ putriamin01@gmail.com

utilizing text processing, natural language processing, and machine learning algorithms to assess the quality of answers, the system simplifies the grading process that benefits both authors and evaluators. Through the speed of the assessment stage, the speed of providing feedback, and the reduction of inconsistencies of the assessor, the AES system plays a crucial role in improving the effectiveness and time efficiency when conducting evaluations (Chassab et al., 2021; Misgna et al., 2025).

In recent decades, AES approaches have evolved from traditional feature engineering that focuses on essay length, grammar, or vocabulary diversity (Nguyen & Litman, 2018) to automatic feature extraction through algorithms such as Rabin-Karp or Winnowing (Hussein et al., 2019; Lagakis & Demetriadis, 2021). Today, transformer-based models like BERT are widely applied due to their capacity to understand bidirectional context. More recently, the rise of generative AI such as ChatGPT has sparked interest in essay assessment that more closely resembles human evaluation, though issues of data requirements and reliability remain (Bahroun et al., 2023; Ouyang et al., 2023).

Despite these advancements, limited research has focused on the use of AES in Islamic education. Essay assessments are particularly significant in this field, as they evaluate students' ability to interpret, reflect on, and apply religious knowledge, tasks that often require nuanced reasoning beyond rote memorization. Manual scoring in this context is vulnerable to subjectivity and inconsistency, highlighting the potential of AES to support more objective and standardized evaluations of students' written reflections on Islamic concepts.

Given these challenges and opportunities, it is important to conduct a systematic review to better understand the current research landscape. This study applies a Systematic Literature Review (SLR) using the PRISMA protocol to analyze AES models published between 2020 and 2025 in the Scopus and Springer databases. Specifically, it aims to identify the most effective algorithmic models for AES, evaluate their advantages and disadvantages, and highlight their contribution to strengthening objective and efficient assessment in Islamic religious education. By addressing this gap, the study provides valuable insights for exam developers, educators, and policymakers seeking to integrate AI-based evaluation tools into Islamic education.

METHOD

This study uses the Systematic Literature Review (SLR) method using the PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analysis*) protocol. To be able to find out what discussions were discussed in the previous research, several steps were taken in the design process and research strategy, considering that the purpose of this study is to find out the method used in the assessment of the automatic essay exam. The PRISMA protocol helps researchers identify the results of studies that have the right literature according to the objectives of the study through three processes, including *identification*, *screening,* and *eligibility* (Gillath & Karantzas, 2019; Page et al., 2021). The purpose of the PRISMA protocol is to improve the quality of reporting in systematic literature review, to get maximum results in writing, this literature is based on Research Questions (RQ). This Research Question is prepared to focus more on the review of the literature and can make it easier for researchers to find related data. As for the research questions developed to achieve this goal, the following questions become RQ: 1) RQ1: What is the most effective deep learning-based algorithm model for evaluating the accuracy and reliability of Automated Essay Scoring (AES) in the context of Islamic religious education assessment?; 2) RQ2: What are the advantages and disadvantages of using transformer-based models BERT, in Automated Essay Scoring systems for Islamic education assessments?

The literature used in this context consists of journals and articles related to Automated Essay Scoring (AES). This study aims to demonstrate that the automatic assessment of essay exams is one of the most important things in the learning process. The literature preparation process includes four distinct stages: the identification phase, the screening phase, the feasibility phase, and the extraction phase.

Identification Phase

At this stage, the process of selecting articles that are in accordance with the predetermined criteria involves the following steps: First, the selection of sources, there are various databases that can be used in systematic literature review, such as: Web of Science, Springer, Scopus, Science Direct, and others. In this study, the articles used were obtained from the Scopus and Springer databases, because they provide more consistent results, become one of the digital libraries that provide promising results based on titles, abstracts, or keywords, and offer more advanced options. (Lonetti et al., 2023; Naqvi et al., 2023) Second, researchers start the search by utilizing research questions to determine keywords that are important in research related to AES. Then, these keywords are combined by using the Boolean operator and adding quotation marks, replacement characters, or curly brackets to improve the quality of the results. We created a simple set of keywords that highlighted the theme of an Automated Essay Scoring System, which utilizes multiple algorithms to automate essay exam corrections. Searching for articles using keywords that have been defined from search engine results in the following arrangement:

1) "Automated Essay Scoring" AND assessment
2) "Automated essay scoring" AND validity
3) "Automated Essay Scoring" AND algorithm
4) "Automated Essay Scoring" AND evaluation

Screening Phase

Next, the process of filtering titles and abstracts, mapping articles based on titles and abstracts. Titles are filtered for relevance and match the keywords used. Then, the abstract of each article is filtered and scanned according to the predetermined inclusion and exclusion criteria. The authors have established inclusion criteria to assess the relevance of the article to the research objectives and ensure consistent evaluation. The results of the search for data with this criterion are what the author will later use to review the article. Table 1 presents the criteria used by the author in the selection of literature.

Table 1. Exclusion Inclusion Criteria

Criteria	Inclusion	Exclusion
Publication period	2020-2025	Before 2020
Language	English	Other than English
Source type	Articles published in Scopus and Springer	Published articles have not been indexed in Scopus and Springer
Focus study	Focus study on AES evaluated with an algorithmic model	Studies outside the topic of AES
Field study	Social science and education	Health, arts, economics, psychology, Engineering
Types of research	Empirical Research, Literature Review, Qualitative and Quantitative Studies	Studies that are only opinion-only or studies that focus only on the AES theory
Publication stage	Articles in the final stage/ Full text	The article is in the research or registration stage/ not full text

The selection of inclusion and exclusion criteria in this study, the author explains the underlying theoretical reasons, including: First, the search results are limited to that year to ensure the reproducibility of the search results and avoid research that uses concepts or methods that have been used. Second, English is the most widely used language in scientific publications worldwide, making it easier for authors to understand research concepts, and the majority of international research is conducted in English. Third, articles in the field of education focus more on articles with topics to be studied. Fourth, the selection of scientific journal articles ensures that the reviewed research has gone through a peer-review process and has a more guaranteed quality, and avoids research that has not been published or is still in the registration stage. Fifth, to limit the scope of research to be more specific and directed.

Eligibility Phase

Articles that are identified as feasible can then be downloaded with the full text, and the articles that fall under the exclusion criteria can be separated. In this phase, articles that are designated as feasible must be able to map the answers to the questions in the research. The data we extracted from each paper were the title of the paper, the name of the paper's author, the year of publication, the theme raised, the method, and the dataset used. The results of data extraction for each paper are then used to analyze and find out the processes used in AES.

Extraction Phase

This is done after checking the eligibility of the article based on inclusion and exclusion criteria. Articles that qualify for the inclusion criteria will be extracted and analyzed in accordance with the PICO framework. The researcher only focused on articles that discussed AES, which is supported by algorithms to evaluate the accuracy of the system. The search and selection process of literature from the Scopus and Springer databases has been carried out from 01 to 03 April 2025. With a focus on titles, abstracts, and keywords, 1,785 articles were published in 2019-2025 that were relevant to the topic that was the focus of this study, namely AES, which was supported by algorithms. At the identification stage of 2,270 articles were eliminated, 825 were due to duplication and did not meet the inclusion criteria. Furthermore, from the remaining literature, as many as 960 papers were screened to determine studies relevant to the research topic. A total of 607 papers were eliminated for several reasons, such as corrupted files, limited accessibility or missing reports, not appropriate for the research topic, and only in the form of opinions or theories. And 360 articles were at the registration stage and did not use English, and did not match the keyword criteria specified in the study. In the final stage, it is a feasibility assessment of the results of the paper that has screened as many as 31 articles that meet the inclusion criteria and will be included in the data analysis. Figure 1 shows the entire process of selecting this article in the systematic literature review through the PRISMA diagram.

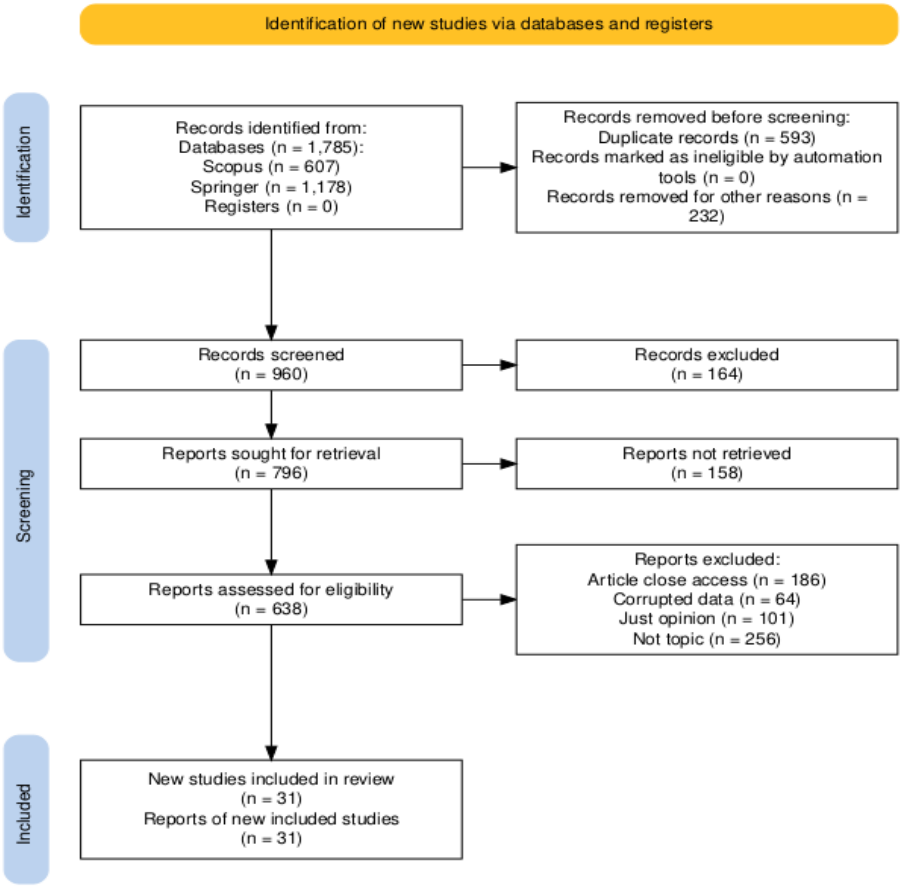

Figure 1. PRISMA Flow Diagram

To facilitate the analysis process, the researcher sorts the research report based on the results, whether the report is related to essay assessment, human assessment, or the essay scoring algorithm. In addition, the researcher analyzes the characteristics of the dataset, such as the methodological research design, the time period in which most of the research was conducted. In summary, review studies are selected through a systematic process with pre-set criteria, so that no bias is intended to be applied to the selected study. We analyze and synthesize the data that has been extracted from each paper in terms of the methods used and the results obtained from each study.

The main purpose of this review is to help readers know what algorithms are used to automatically correct essay answers in the field of education. We also explore the datasets used and the key themes explored. Next, we report the analysis taken by each paper based on the data set taken, the methods used, and the descriptive explanation of the process described in the auto-correcting essay exam. The structure of our report is structured in such a way that it can answer the research questions asked earlier. The findings of our report are presented in the results and discussions.

RESULTS AND DISCUSSION

In this step, we report the analysis of each article based on the data set taken, the methods used, and the descriptive explanation of the AES process. After the identification, screening, and eligibility process, 31 articles were obtained that met the predetermined inclusion criteria. The mapping of these 31 articles will be detailed in various types, including the author's name, year of publication, type of publication, Scopus accreditation, and relevance to research questions (RQ), which are the focus of AES as an evaluation tool. Table 2 presents the articles included in the analysis.

Table 2. Literature Search Results

No	Reference	Explanation
1.	Automated language essay scoring systems: a literature review (Hussein et al., 2019)	AES systems that use manual features and automated features with NLP and machine learning are able to reduce the burden of assessment compared to manual assessment.
2.	Enhanced BERT Approach to Score Arabic Essay's Relevance to the Prompt (Machhout & Zribi, 2024)	Development of the AraBERT model combined with special artificial features, which resulted in a correlation of 0.88 to the human score.
3.	A Transformer-Based Approach for Enhancing Automated Essay Scoring (Chavva et al., 2024)	This study developed an Automated Essay Scoring (AES) system based on the RoBERTa model, as a result, this RoBERTa-based AES system achieved a QWK value of 0.815, which shows a high level of accuracy and reliability
4.	Automatic Essay Scoring: A Review on the Feature Analysis Techniques (Chassab et al., 2021)	Traditional approaches that use only morphological analysis are less able to capture semantic meaning in essays, so more research is needed to overcome these limitations so that AES can be more effective and accurate across multiple domains.
5.	Automated Indonesian Essay Scoring and Holistic Feedback Using Bidirectional Encoder Representations for Transformers (Amalia et al., 2024)	The model developed based on BERT, especially IndoBERT, showed a perfect accuracy of 1.0, *a kappa* score of 0.82, and a QWK score of 0.9 at validation, reflecting a high degree of conformity between the model's prediction and human judgment.
6.	A novel automated essay scoring approach for reliable higher educational assessments (Beseiso et al., 2021)	The model was developed using Bi-LSTM and RoBERTa, resulting in this model being superior in QWK compared to NLP and deep learning

No	Reference	Explanation
7.	Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking (Yang et al., 2020)	This model combines two loss functions (*mean square error loss and batch-wise ListNet loss*) with dynamic weights to significantly improve AES performance, resulting in 3% higher than the typical neural network model.
8.	Neural-Network Architecture Approach: An Automated Essay Scoring Using Bayesian Linear Ridge Regression Algorithm (Catulay et al., 2021)	The model developed is Neural Network with the Bayesian Linear Ridge Regression algorithm, the results of which improve the accuracy and reliability of essay assessment, time efficiency, avoid biased assessment, and provide feedback for students.
9.	Automated Essay Scoring Using Convolutional Neural Network Long Short-Term Memory With Mean of Question-Answer Encoding (Kusumaningrum et al., 2024)	Developed model combines Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) architecture with *the mean of question-answer encoding* approach as an assessment mechanism. The result was able to increase the QWK value by 14.47% or reduce the loss value by 0.23% when compared to the CNN-LSTM model which uses mean over time
10.	More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms (Shin & Gierl, 2020)	The Convolutional Neural Networks (CNNs) model in the AES system provides better performance compared to the traditional model of Support Vector Machines (SVMs) combined with the Coh-Metrix feature.
11.	A Hybrid Automated Essay Scoring Using NLP and Random Forest Regression (Azahar & Ghauth, 2022)	This model combines Natural Language Processing (NLP) and Random Forest Regression to predict essay scores on an ongoing basis. As a result, this model is able to predict essay scores with a lower error rate than Linear Regression and some Deep Learning models, based on MAE, MSE, and RMSE metrics
12.	Integrating Deep Learning into An Automated Feedback Generation System for Automated Essay Scoring (Lu & Cutumisu, 2021)	This study used word-embedding and deep learning models such as CNN, LSTM, and Bi-LSTM, successfully achieving higher scoring accuracy than previous models
13.	A survey on deep learning-based automated essay scoring and feedback generation (Misgna et al., 2025)	Deep learning-based AES models are able to accurately predict essay scores, but there are still limitations in explaining the patterns and features used, making it difficult to provide feedback
14.	A trait-based deep learning automated essay scoring system with adaptive feedback (Hussein et al., 2020)	The Long Short-Term Memory (LSTM) model was able to increase the accuracy of the AES baseline model by 4.6% based on QWK metrics.
15.	Prompt Agnostic Essay Scorer: A Domain Generalization Approach to Cross-prompt Automated Essay Scoring (Ridley et al., 2020)	The Prompt Agnostic Essay Scorer (PAES) successfully overcomes the challenges of cross-topic essay assessment (cross-prompt AES) without the need for target-topic data, both labeled and unlabeled, with a *single-stage* approach
16.	Automated Cross-prompt Scoring of Essay Traits (Ridley et al., 2021)	Automated Cross-prompt Scoring of Essay Traits successfully outperforms both topic-specific and cross-topic assessment methods and enhances AES's ability to deliver assessments
17.	Automatic Scoring of Arabic Essays: A Parameter-Efficient Approach for Grammatical Assessment (Mahmoud et al.,	This study proposes an optimized AraBART pretrained model using various parameter-efficient methods, resulting in this approach providing better performance compared to full fine-tuning

No	Reference	Explanation
	2024)	
18.	Neural Automated Essay Scoring Incorporating Handcrafted Features (Uto et al., 2020)	This study developed the DNN-AES model, a model that is simpler than the RNN but still more accurate in providing essay scores.
19.	Automated essay scoring with SBERT embeddings and LSTM-Attention networks (Nie, 2025)	The integration of Sentence-BERT, BiLSTM, and attention mechanisms in the AES system significantly improves the accuracy of essay assessment, based on the evaluation of *the benchmark dataset*.
20.	On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation (Wang et al., 2022)	The BERT model was able to significantly outperform other deep learning models such as LSTM in the CommonLit Readability Prize dataset
21.	Should You Fine-Tune {BERT} for Automated Essay Scoring?(Mayfield & Black, 2020)	The fine-tuning model of BERT in AES provides similar performance to traditional models but with much higher computing costs.
22.	Automated Essay Scoring Using Transformer Models (Ludwig et al., 2021)	The transformer model is able to produce better performance without the need for hyperparameter tuning compared to the logistic regression model that uses the bag-of-words (BOW) method
23.	Domain-Adaptive Neural Automated Essay Scoring (Cao et al., 2020)	The domain-adaptive framework *working model* was able to achieve the best performance in both in-domain and cross-domain testing using the ASAP dataset
24.	Deep Learning Architecture for Automatic Essay Scoring (Tashu et al., 2022)	This model combines *a multichannel convolutional neural network* (CNN) with *a Bi-gated recurrent unit* (BGRU). The result is a much higher assessment accuracy compared to other deep learning-based AES systems, such as CNN and RNN.
25.	Automated Essay Scoring via Pairwise Contrastive Regression (Xie et al., 2022)	The model developed is called *Neural Pairwise Contrastive Regression* (NPCR), which optimizes regression and rating through a single *loss function*. The results show that this model significantly outperforms the previous method, thus achieving the best performance in AES
26.	Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy (Faseeh et al., 2024)	The developed model incorporates RoBERTa and uses the Lightweight XGBoost (LwXGBoost) algorithm to improve the accuracy of the assessment. The results showed a QWK of 0.941, which signifies a very high level of conformity with human judgment.
27.	Automated Essay Scoring: A Comparative Study of Machine Learning and Deep Learning Approaches (Bansal et al., 2025)	Deep neural network models using Natural Language Processing (NLP) techniques for feature extraction are able to provide essay assessment results that are closest to human judgment compared to conventional machine learning models.
28.	A comprehensive review of automated essay scoring (AES) research and development (Lim et al., 2021)	This study classifies the AES method into content similarity, machine learning, and hybrid, the result is that the hybrid method is superior by being able to combine content analysis and writing style compared to other methods.
29.	FACToGRADE: Automated Essay Scoring System (Das et al., 2022)	Development of an AES system that combines content analysis and structural analysis using LSTM models as well as entity detection techniques. As a result, AES provides automated assessments and feedback quickly, helping students improve the quality of their writing

No	Reference	Explanation
30.	Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of the Bert Model and a Recurrent Neural Network (RNN_Bert_Based) (Eang & Lee, 2024)	The model developed is called RNN_Bert_based, the result of which the combination of RNN and BERT is proven to outperform both traditional and pre-existing text classification models.
31.	Exploring the potential of using an AI language model for automated essay scoring (Mizumoto & Eguchi, 2023)	The use of the GPT-3 *model (text-davinci-003*) for AES on 12,100 essays in the corpus TOEFL11 able to produce a fairly accurate and reliable score when compared to the benchmark scoring standard.

Trend Research Automated Essay Scoring

First, we will pay attention to the publication journey in recent years of studies that fall under the inclusion criteria. Figure 2 shows an article diagram that meets the inclusion criteria.

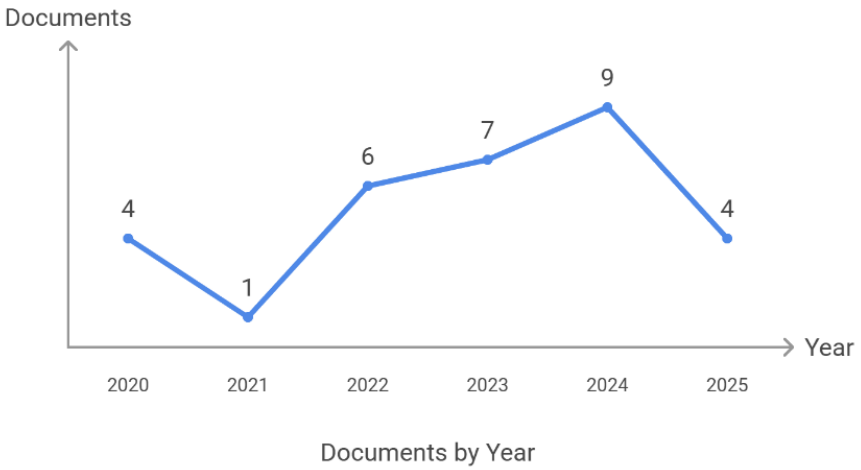

Figure 2. Document by Year

Figure 2 presents the trend of publication of scientific articles on AES in the form of a bar graph. The data visualized covers the period from 2020 to 2025, allowing us to observe the change in the number of publications from year to year. The topic of articles related to AES in deep learning is increasingly in demand, as can be seen from the graph of the numbers that are rising from 2021 to 2025. This is due to the development of technologies and software such as artificial intelligence (AI) and natural language processing (NLP) that allow for the development of more accurate and efficient AES tools. In addition, manual essay grading by teachers is very time-consuming. AES offers a potential solution to reduce teacher workload and provide faster and more consistent assessments. The need for effective and efficient solutions has prompted researchers to study AES-related topics. (Hussein et al., 2019; Ramesh & Sanampudi, 2022)

In 2020, the publication's contribution to the overall total was 12.90%, indicating that research in this area may still be in its early stages or has not received much attention. In 2021, the percentage of publications dropped to only 3.23%, which was the lowest percentage in this study; then in 2022, it increased again, with the percentage of publications to 19.35%. This shows that the topic of AES in deep learning has received attention from researchers. The upward trend will continue until 2024, where in 2023 it will be 22.58%, in 2024 it will reach the highest percentage of 29.03%, and in early 2025 it will be 12.90%. This shows that AES-related research is increasingly in demand by researchers. The increase in AES-related publications between 2020 and 2025, as illustrated in Figure 2, reflects more than just a growing academic interest. It indicates a broader shift in educational assessment practices, driven by the need for scalable, efficient, and reliable evaluation tools. The sharp growth after 2021 aligns with advances in NLP and the increased

reliance on digital education during and after the COVID-19 pandemic. For Islamic Education, this trend signals an urgent demand for automated systems that can reduce teachers' grading burden while maintaining fairness and accuracy in evaluating moral and religious reasoning. The peak in 2024 suggests that AES research is approaching maturity, with a strong emphasis on transformer-based models like BERT, which can better capture the contextual and ethical nuances embedded in Islamic essays compared to earlier machine learning methods. Second, we will present some themes that are interrelated to AES. Figure 3 is a visualization of the grouping of articles related to the corresponding themes.

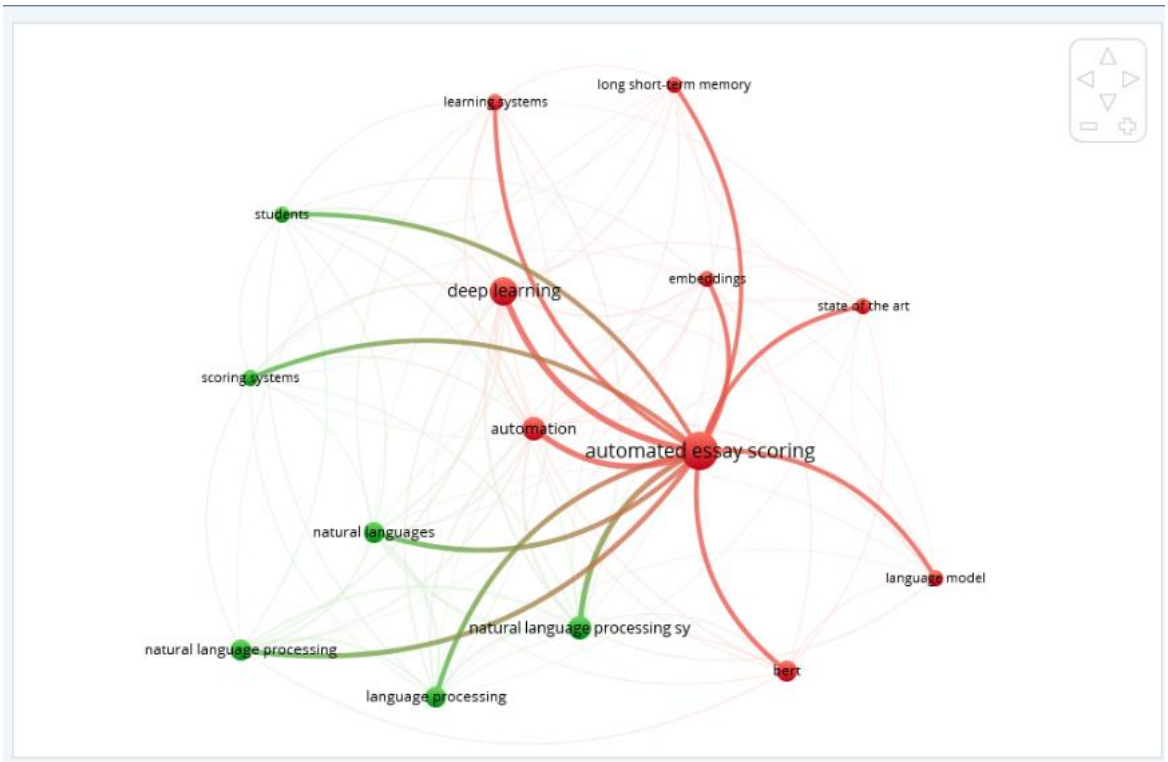

Figure 3. Visualization of VOSviewer

Figure 3 is a visualization of the network of topics or keywords that often appear in AES-related research; each point (*node*) represents a keyword, while the line (*edge*) that connects it shows the existence of a connection or co-emergence in the topic. The size of the node describes the frequency or importance of the keyword in the overall corpus of data, while the color and thickness of the lines indicate the strength of the relationships between concepts. AES is related to several topics, such as "*deep learning*", "*automation*", "*BERT*", and "*language model*". This shows that the modern approach to AES is largely focused on the use of *deep learning technologies* and transformation-based models such as BERT. Figure 3 reveals more than co-occurrence patterns; the clustering of "BERT," "deep learning," and "automation" at the center of the network signifies a paradigm shift from surface-level linguistic analysis toward semantic and context-aware models. The proximity of "students" indicates that AES research increasingly prioritizes feedback mechanisms that directly support learning, not just grading. This emphasis is especially relevant for Islamic Education, where essays are assessed not only for technical accuracy but also for moral reasoning, critical thought, and religious understanding. The dominance of transformer-based models within the keyword clusters suggests that future AES systems will prioritize semantic depth and adaptability, making them particularly valuable for assessing essays that combine faith-based arguments with academic writing. Other nodes, such as "*long short-term memory*" (LSTM), "*state of the art*", and "*learning systems*", indicate the techniques and contexts used to improve the performance of AES systems. (Lagakis & Demetriadis, 2021). On the other hand, the connection between "*students*" and the grading system also reflects that the final focus of AES is to have a direct impact on the assessment of students' essays efficiently and effectively.

The network visualization shown illustrates the semantic relationship and interconnectedness between keywords; in this context, the distance between nodes is not random, but represents the proximity of the relationship of meaning or frequency of co-existence of these words in the article. Keywords that are adjacent to each other indicate that they are often used together in the same context or form closely related topics. For example, "*automated essay scoring*" is very close to "*deep learning*", "*automation*", and "BERT", which indicates that technologies such as deep learning models and transformations such as BERT are often directly associated with AES systems in some studies. Conversely, keywords that are positioned farther away from the center of the keyword indicate that their semantic relevance or use is relatively less frequent. For example, the words "*students*" or "*natural languages*" may have an indirect relationship with technical models such as "*language models*" or "*embeddings*".

This reflects that while these elements are still relevant in the context of AES, the frequency of their simultaneous occurrence may be lower, or the focus is in a different realm. This visualization is particularly useful in identifying the main focuses and relationships between subthemes in the relevant literature. Information about current issues, trends, and methods in the field can be obtained from keywords (Sevcikova, 2018; Hussein et al., 2019) Figure 3 does more than visualize co-occurring keywords; it reveals a clear clustering around “BERT,” “deep learning,” and “automation,” signifying a paradigm shift in AES research from surface-level analysis toward semantic and context-aware models. The proximity of “students” within this network indicates that current AES development is increasingly student-centered, focusing on feedback mechanisms that support learning rather than merely producing scores. This trend is especially relevant in Islamic Education, where assessment is not only about technical accuracy but also about guiding students to express moral reasoning, critical thought, and spiritual reflection. The dominance of transformer-based models in the keyword network suggests that future AES systems will likely prioritize semantic depth and adaptability, making them particularly valuable for evaluating essays that integrate faith-based arguments with academic writing. Keywords are mandatory, provided by the author for each publication, but are also automatically assigned by indexing databases or even extracted as n-grams of the title or abstract. Through thematic maps based on co-occurrence network analysis and grouping using keywords, the authors identified several types of topics (themes) based on density (i.e., level of development) and centrality (i.e., level of relevance). In both emerging and declining topics, such as those related to "BERT" and transformer-based models, they are noteworthy indicators of increasing interest in algorithm-based essay automated grading. (Lim et al., 2021)

The special theme covers several interesting topics, such as embedding-based semantic similarity which is a technique used in AES to conceptually compare the content of a student's essay with a rubric-standard essay (e.g., using Latent Semantic Analysis or BERT to measure the depth of meaning and suitability of content), as well as the creation of automated questions related to AI-based assessment and adaptive learning. The use of algorithms in machine learning and deep learning, such as neural networks, SVM, and LSTM, is becoming the dominant approach that continues to be developed to improve the accuracy of AES systems in mimicking human judgment.

Effectiveness of Algorithmic Models in AES

Automated essay grading systems provide precise and fast feedback that helps in practicing and improving writing skills. One of the advantages of an automated essay grading system is its efficiency and consistency. According to some researchers, there are no biases, prejudices, or stereotypes that are common in human judgment, research conducted by Hussein et al. (2020) and Ke and Ng (2020) discussed an AES model that utilizes handmade features and neural networks, (Chavva et al., 2024; Hussein et al., 2019) in addition, Lagakis and Demetriadis (2021) and Ramesh and Sanampudi (2022) conducted a more in-depth study than the previous research, in this study a discussion was added related to the use of transformer-based AES models (Lagakis & Demetriadis, 2021; Ramesh & Sanampudi, 2022). In addition, Chassab et al. (2021) discuss feature representation methods used in AES tasks. They present a taxonomy of AES feature analysis methods, which are categorized into feature analysis of text structure, frequency, and occurrence of terms, semantic meaning, and morphological aspects. The above research emphasizes the importance of utilizing handmade and auto-generated features to improve assessment

performance and generate direct feedback. There are several AES algorithms used in AES development, including the AES Algorithm with *Machine Learning* (ML), which is increasingly popular in the world of Education. These models assess students' essays and provide direct feedback.

ML models use algorithms to analyze text and extract applicable features such as language style, sentence structure, and vocabulary to find specific trends and patterns. Based on this analysis, ML can automatically score and provide feedback on the parts of the writing that still need improvement. In addition, this model is able to provide personalized feedback and suggestions so that it can help students improve and develop their writing skills. AES and ML systems have the potential to change the traditional way by providing fast, precise, and objective feedback. Some of the most common algorithms used in this context include *Support Vector Machine* (SVM), Random Forest, and XGBoost (Amalia et al., 2024; Beseiso et al., 2021; Yang et al., 2020). *Support Vector Machines* (SVM) can produce accurate scores in assessing Indonesian essays, with an explicit linguistic feature-based approach. This shows that SVM is still relevant and efficient, especially in the context of limited data or light system needs (Catulay et al., 2021; Kusumaningrum et al., 2024).

Shin and Gierl (2021) compared two algorithms: (a) SVM along with Coh-Metrix features as a traditional AES model and (b) *Convolutional Neural Networks* (CNN) approach as a more contemporary deep neural model. They reported that the CNN model performed better, with results more comparable to human raters than traditional models. As witnessed in this example, it has been widely recognized that applying a deep learning approach (i.e., neural network) to AES produces better results than previous approaches. While CNN outperforms SVM in terms of capturing sentence-level patterns, its reliance on local features makes it less effective in handling longer argumentative structures typical in student essays. In Islamic Education, where essays often integrate Qur'anic references, hadith, and moral reasoning across multiple paragraphs, such local focus is insufficient. Transformer-based models like BERT are superior because they capture bidirectional context and semantic depth, enabling them to evaluate both language quality and conceptual coherence. This explains why BERT consistently yields higher QWK scores than CNN or SVM and makes it particularly suited for assessing the depth of understanding in Islamic Education essays. Studies by Azahar et al. (2022) also show that Random Forest is capable of generating essay scoring scores that are close to the accuracy of deep learning models, making them a suitable choice for medium-scale automated assessments. Meanwhile, gradient boosting-based approaches, such as XGBoost, have proven to be superior in terms of accuracy due to their ability to build predictive models gradually and correct prediction errors from previous models.

This is supported by the research of Faseeh et al. (2024), who found that XGBoost can process linguistic features effectively and efficiently that providing a scoring score that is in line with human judgment. However, this machine learning approach has limitations, especially in understanding semantic meanings and complex narrative structures. Therefore, in recent practice, these algorithms are often combined with advanced NLP techniques or deep learning models to improve more accurate assessment results and get closer to human understanding of essays (Lu & Cutumisu, 2021). In the field of artificial intelligence (AI), *deep learning* (DL) is part of an ML technique that works using artificial neural networks to recognize patterns in the data. DL is able to learn from a large amount of big data, both structured and unstructured, and can be applied in various learning paradigms, including *supervised learning*, *unsupervised learning*, and *reinforcement learning* (Lu & Cutumisu, 2021; Misgna et al., 2025).

In the context of AES, the DL model is used to provide a more comprehensive and accurate analysis of students' writing. This model is able to understand complex sentence structures, recognize relevant topics, and detect common errors in essays, resulting in students getting more meaningful and targeted feedback. DL also helps ensure that the AES system is more precise, consistent, and uniform in grading, as its algorithms can be trained to recognize patterns and grade essays fairly and accurately. One of the DL algorithms that is often used in AES is *Long Short-Term Memory* (LSTM). LSTM is a type of neural network that is excellent at processing data in the form of sequences, such as text. LSTM is considered effective because it can remember information in a long and sequential context, making it suitable for understanding the entire content of the essay. Various studies have proven the ability of LSTMs to capture information in the global context of

student writing (Hussein et al., 2020; Mahmoud et al., 2024; Ridley et al., 2020, 2021; Uto et al., 2020).

As technology advances, more advanced transformer-based language models such as BERT (*Bidirectional Encoder Representations from Transformers*) are increasingly used in AES. Compared to traditional machine learning models such as SVM, which rely heavily on handcrafted linguistic features, and CNNs, which primarily capture local text patterns, BERT demonstrates superior performance in AES because of its bidirectional contextual learning. In the context of Islamic Education, where essays often require a nuanced interpretation of Qur'anic verses, hadith, and moral reasoning, the ability to understand both local word dependencies and broader semantic relations is essential. CNNs may effectively detect sentence-level structures but fail to capture long-range dependencies, while SVMs are limited in recognizing meaning beyond surface-level features. By contrast, BERT's transformer architecture allows it to capture subtle semantic differences and logical argumentation across entire essays, making it better suited for evaluating not only language accuracy but also the depth of students' understanding of Islamic values. BERT excels at understanding the context of words in two ways: reading from front and back, which enables it to capture the meaning and relationships between sentences in an essay more accurately. Another study compared the performance of BERT and XLNet on AES standards such as Kaggle ASAP and found that these models consistently produced higher *Quadratic Weighted Kappa* (QWK) scores than conventional approaches, suggesting a more closely conforming level to human grader scores (Nie, 2025; Wang et al., 2022). In addition, research by Mayfield and Black (2021) showed that DistilBERT, a lightweight version of BERT, was able to achieve a very high Cohen's Kappa. This reinforces the finding that transformer models are not only accurate but also consistent in assessment evaluations. While the model even shows scores that are close to *near-perfect* with human scores, other approaches that combine the semantic embedding of BERT with other models, such as LSTM-Attention, have also shown excellent performance.

For example, the SBERT + LSTM-Attention model achieved a QWK of 0.7876, outperforming several other approaches, including XLNet. This confirms that transformer-based models are not only capable of handling the surface aspects of the text but also understanding the context and logical structure of the essay (Ludwig et al., 2021). The deep learning-based approach also allows the development of an AES system that not only assesses scores, but also provides automated feedback regarding the strengths and weaknesses of students' writing, for example, the DeLAES model developed by Tashu et al. (2022), which combines multichannel CNN and bidirectional RNN, shows high performance with a QWK score of up to 0.903 on eight different datasets. This suggests that deep learning models that are architecturally designed to capture various aspects of the linguistic and structural aspects of texts have great potential to mimic or even surpass the consistency of human appraisers in certain contexts.

Yang et al. (2020) showed that the use of BERT directly can improve performance outcomes in screening by expanding its ability to predict essay scores and rankings. Fine-tuning of BERT using the multi-loss function technique, as performed (Yang et al. 2020; Xie et al. 2022; Wang et al. 2022b) and combining BERT with essay-level features (Uto et al. 2020), has yielded results that go beyond neural network methods in AES tasks. In another study conducted by Geetha & Karthika Renuka (2021), various classification models were used to analyze consumer review data, including LSTM, BERT, Naive Bayes Classification, and SVM. The results showed that BERT provided the highest accuracy among all models, reaching 88.48%. Tests that combined the results of the BERT model with the performance of other machine learning algorithms showed that the BERT model outperformed other machine learning algorithms in terms of performance measures. BERT produces better accuracy, outperforming other machine learning techniques in experimental evaluations with high predictions and good accuracy.

The effectiveness of AES is proven to be even higher with the advancement of the algorithms used, especially with the application of machine learning (ML) and deep learning (DL). ML-based models such as Support Vector Machine (SVM), Random Forest, and XGBoost have shown pretty good results in providing consistent scores and approaching human judgment, especially when combined with explicit linguistic features such as sentence structure, language style, and vocabulary (Chassab et al., 2021; Hussein et al., 2019; Lagakis & Demetriadis, 2021). However, this traditional ML model still has limitations in capturing the semantic meaning and complex narrative

structure of essay texts. In contrast, the use of deep learning models such as LSTM and transformer-based models such as BERT further strengthens the effectiveness of AES due to its ability to understand the context of sentences in more depth (Beseiso & Alzahrani, 2020; Eang & Lee, 2024; Faseeh et al., 2024; Mizumoto & Eguchi, 2023). BERT emerged as the most prominent and effective deep learning model in improving the quality of AES systems. BERT also allows the AES system not only to function as a scorer, but also as an automated feedback tool, capable of identifying the strengths and weaknesses of students' writing more comprehensively. Transformer-based models, especially BERT, are currently the most effective approach in AES, offering high accuracy and the ability to understand the context of essays in depth. A hybrid approach that combines artificial linguistic features and semantic representations from deep learning models also shows great potential in improving AES performance.

Advantages and Disadvantages of the AES Algorithm

The most effective algorithm model for evaluating the accuracy and reliability of current AES technology is transformer-based models, specifically BERT. The model has shown superior performance in a wide range of studies, especially in capturing the semantic context and structure of essays in depth. The DL model offers an important strength in automated essay assessment (Chassab et al., 2021; Geetha & Karthika Renuka, 2021). First. This model excels in representation learning, specifically neural networks, which autonomously derive complex representations from textual data, thereby capturing complex patterns and dependencies inherent in essays. Additionally, designers such as *the Recurrent Neural Network* (RNN) or Transformer-based models facilitate the extraction of hierarchical features, which allow for a deeper understanding of context in the essay. As well as the application of transfer learning, which leverages pre-trained language models such as BERT or GPT, allows for effective refinement of essay grading tasks with minimal data, leveraging knowledge gained from a vast textual corpus. In the use of the BERT algorithm, there are several advantages of this system(Bansal et al., 2025; Das et al., 2022; Lim et al., 2021), , including: 1) Deep Context Understanding: BERT uses a two-way attention mechanism, which allows the model to better understand the context of words in sentences. This is especially important in essay assessment, where meaning often depends on a broader context; 2) Powerful Text Representations: BERT generates rich vector representations for each word, which include information about the word in the context of the sentence. It helps in capturing the nuances and complexities of language that are often difficult to measure with traditional methods; 3) Generalization Capabilities: The BERT model can be adapted for a variety of essay assessment tasks without the need for extensive retraining. This makes it flexible and efficient in different applications; 4) Superior Performance: Research shows that BERT can achieve results that are almost on par with other state-of-the-art models in essay grading tasks. For example, in a study conducted by Wang et al. (2022), BERT demonstrated excellent performance in essay assessment by using multi-scale representations that could be studied together, which improved the accuracy of the assessment; 5) Transfer Learning: BERT can leverage transfer learning, where models that have been trained on large datasets can be tailored to specific datasets for essay assessment tasks. This allows the use of existing knowledge to improve model performance in new contexts; 6) Bias Reduction: With proper training, BERT can help reduce bias in essay assessment, as it can learn from a variety of examples and contexts, resulting in a fairer and objective assessment.

However, this approach is not without its drawbacks. DL models often demand substantial annotated data for training, thus posing challenges in scenarios where such data may be scarce, particularly in specialized essay grading tasks. Furthermore, training these models requires significant computing resources and infrastructure, especially for large-scale models such as GPT. Finally, the black box nature of the DL model raises concerns about interpretability, thereby hindering the ability of stakeholders to understand the reasoning behind the scores given for the essay. Eang & Lee (2024) states that although BERT shows superior performance compared to convolutional and repetitive neural networks, it has several limitations: (1) ranking by batch ignores global order information, (2) determining the right weights to balance the loss function for optimization, and (3) BERT only supports sequences of up to 512 tokens, which leads to longer essay cutting.

Automatic Essay Assessment in Islamic Religious Education

AES is an artificial intelligence-based system designed to automatically assess essay writing, taking into account the structure of language, content, and the quality of arguments. In the world of Islamic education, especially in assessing essays that contain religious themes, accuracy in understanding the meaning, context of the evidence, and moral message is important to train students' critical thinking (Mizumoto & Eguchi, 2023). AES has a very important function in the assessment of Islamic Religious Education (PAI) because it can be used to evaluate students' writings with Islamic themes, such as moral themes, *aqidah, fiqh*, Islamic history, or other religious moral issues, AES not only plays a role in providing scores on the technical aspects of writing, but can also be directed to identify how deep students' understanding of the values of Islamic teachings is expressed through essay answers (Alqahtani & Alsaif, 2019). Although AES offers efficiency and consistency in assessing Islamic Education essays, critical challenges remain. Current models may succeed in detecting grammar and structural accuracy, but they struggle with evaluating the depth of the spiritual and moral reasoning component of Islamic learning. For example, distinguishing between a student's superficial mention of Qur'anic verses and a genuine integration of religious argumentation requires advanced semantic understanding. Transformer-based models like BERT offer promise because their bidirectional contextual learning enables more accurate interpretation of nuanced religious expressions. However, limitations such as the "black box" nature of deep learning and the 512-token restriction raise concerns about transparency and completeness in evaluating longer Islamic essays. Thus, while BERT is highly effective, its adoption in Islamic Education must be accompanied by careful adaptation to ensure both fairness and alignment with religious values. AES is a tool to assess and provide feedback on students' essays related to Islamic material, AES not only serves to assess the structure of language and grammar, but also to evaluate students' understanding of Islamic values, for example, students are asked to write essays on topics such as morals in Islam, tolerance between religions, or the history of the Prophet Muhammad PBUH, and the AES system will judge based on the suitability of the content, the logical structure of the argument, and the appropriate use of Islamic terms (Al Awaida et al., 2019).

AES in PAI assessment is to provide assessments quickly, objectively, and consistently, this system can analyze students' writings and provide automatic feedback by assessing the advantages and disadvantages of their writing, including the accuracy of the use of Islamic terms, the conformity of opinions with the postulates of sharia, and the understanding of Islamic teachings (Mahmoud et al., 2024). Thus, AES can encourage students to write more often, develop Islamic critical thinking skills, and acquire reflective learning. To effectively implement AES in PAI assessments, the system must be trained to use a corpus of essays relevant to Islamic themes and to use deep learning models such as BERT or LSTM that have been adapted to better recognize Islamic contexts and values. BERT is present as a transformer-based deep learning model that is trained to read and understand text from two directions (left to right and right to left), so it is very effective in capturing the nuances of meaning and relationships between sentences in students' writing (Mayfield & Black, 2020; Tashu et al., 2022). BERT can also be used to provide automated feedback that is educational in nature, such as showing less powerful parts of an essay in reasoning, or providing suggestions for adding relevant sharia postulates.

This supports the learning process based on reflection and self-improvement in accordance with Islamic values. The long-term use of BERT can improve the quality of formative and summative assessments in PAD, make it easier for teachers to monitor student progress, and encourage the integration of technology with religious values in education (Machhout & Zribi, 2024; Beseiso & Alzahrani, 2020; Gaheen et al., 2021). An automated essay scoring system allows students to submit their essays and obtain feedback in the form of grammar scores and corrected versions of their essays. This system can be further developed to assess various other aspects. For example, it can provide a score based on the suitability of the essay content with the model's answers, and include sample answers as a reference. In addition, the system can also assess the organizational aspects of the writing and provide suggestions on how to organize ideas in a more structured manner. By dividing the essay assessment into components, such as grammar, content, and organization, students can more easily identify weaknesses in their writing and understand the steps that need to be taken to improve the overall quality of the essay.

This study succeeded in identifying the most effective algorithm model, namely BERT, which has proven to be superior in understanding the two-way semantic context and providing assessment results that are close to the quality of human evaluators, especially in assessing students' understanding of Islamic values. In addition, this study maps the latest developments in AES technology and uncovers its advantages and disadvantages, providing a scientific foundation for system developers, educators, and policymakers in designing a fair, effective, and efficient assessment system.

LIMITATIONS

Although this study provides valuable insights into the algorithmic models in Automated Essay Scoring (AES), there are some limitations that must be known. First, the study only covers articles published in the 2020–2025 period and is limited to the Scopus and Springer databases, thus allowing for the omission of relevant studies from other credible sources. Second, although models such as BERT show high effectiveness, this study does not directly test the performance of the model in the context of an Islamic Religious Education essay that is rich in contextual meaning and normative values, so its practical application still requires further adjustment and training. Third, reliance on data available in previous publications also limits in-depth exploration of specific variables that might influence the effectiveness of algorithms in more complex and multicultural educational scenarios.

CONCLUSION

The Automated Essay Scoring (AES) system relies heavily on the development of the algorithmic model used. Studies show that deep learning-based models, especially those that use transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT), are the most prominent and effective approaches today. BERT has proven to be superior in understanding two-way semantic contexts, capturing logical structures in writing, and being able to provide more accurate and consistent assessments than conventional methods such as SVM, Random Forest, and XGBoost. Beyond technical performance, this study emphasizes that in the context of Islamic Education, AES powered by BERT has the potential to evaluate not only grammar and structure but also the depth of students' reasoning when engaging with Qur'anic texts, hadith, and moral arguments. This makes it highly relevant for supporting critical and reflective learning in religious education. From a practical perspective, the integration of AES into Islamic Education can provide significant contributions to teachers, curriculum developers, and policymakers. For teachers, AES can reduce grading workload while offering timely, consistent, and unbiased feedback to students. For curriculum developers, AES insights can highlight common areas of weakness in students' religious writing, guiding the design of more targeted instructional materials. For policymakers, AES offers a scalable solution for large-scale assessments that balances efficiency with the need to uphold fairness and accuracy in evaluating students' understanding of Islamic values.

However, challenges remain, particularly in ensuring that AES systems respect cultural and religious sensitivities, maintain transparency, and are adapted to the specific linguistic features of Islamic texts. Therefore, future research should focus on developing domain-specific AES models trained with Islamic essay corpora, exploring hybrid approaches that combine BERT with interpretable models, and investigating ethical frameworks to ensure fairness in religious-based assessments. Expanding AES beyond technical scoring toward formative feedback aligned with Islamic pedagogical goals will be essential to maximize its role in supporting students' spiritual, moral, and intellectual growth.

AUTHOR CONTRIBUTIONS

RKAP is responsible for designing the research framework, conducting literature searches and selections using the PRISMA protocol, and compiling an initial draft of the article. K and S act as scientific supervisors and reviewers, who provide conceptual input to the methodology, analysis of findings, and ensure the relevance and depth of studies in the context of Islamic Religious

Education. The three contributed to the editing process and have agreed on the final version of the manuscript.

ACKNOWLEDGEMENTS

The author expressed his gratitude to the State Islamic University of Sunan Ampel Surabaya for the academic support and facilities provided during this research process. Thanks are also extended to the anonymous reviewers who have provided valuable input for the improvement of this article. Not to forget, appreciation is given to all parties who have assisted in the process of searching for literature and compiling this article, both directly and indirectly.

REFERENCES

Abosalem, Y. (2015). Assessment techniques and students' higher-order thinking skills. *ICSIT 2018 - 9th International Conference on Society and Information Technologies, Proceedings, 4*(1), 61–66. https://doi.org/10.11648/j.ijsedu.20160401.11

Al Awaida, S. A., Al-Shargabi, B., & Al-Rousan, T. (2019). Automated Arabic essay grading system based on F-score and Arabic WordNet. *Jordanian Journal of Computers and Information Technology, 5*(3), 170–180. https://doi.org/10.5455/jjcit.71-1559909066

Alqahtani, A., & Alsaif, A. (2019). Automatic evaluation for Arabic essays: A rule-based system. *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 1–7. https://doi.org/10.1109/ISSPIT47144.2019.9001802

Amalia, A., Lydia, M. S., Kadir, R. A., Tanjung, F. A. U., Ginting, D. S. B., & Gunawan, D. (2024). Automated Indonesian essay scoring and holistic feedback using bidirectional encoder representations for transformers. *8th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, 96–101. https://doi.org/10.1109/ELTICOM64085.2024.10864959

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 (NAACL-HLT)*, 1, 229–237. https://doi.org/10.18653/v1/n18-1021

Azahar, M., & Ghauth, K. (2022). A hybrid automated essay scoring using NLP and random forest regression (pp. 448–457). *Atlantis Press*. https://doi.org/10.2991/978-94-6463-094-7_35

Bahroun, Z., Anane, C., Ahmed, V., & Zacca, A. (2023). Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability, 15*(17), 12983. https://doi.org/10.3390/su151712983

Bansal, B., Gupta, J., Singh, M., Rani, R., Jaiswal, G., & Sharma, A. (2025). Automated essay scoring: A comparative study of machine learning and deep learning approaches. *5th International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 1–7. https://doi.org/10.1109/ICAECT63952.2025.10958994

Bernardin, H. J., Thomason, S., Buckley, M. R., & Kane, J. S. (2016). Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management, 55*(2), 321–340. https://doi.org/10.1002/hrm.21676

Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications, 11*(10), 204–210. https://doi.org/10.14569/IJACSA.2020.0111027

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*(3), 727–746. https://doi.org/10.1007/s12528-021-09283-1

Cao, Y., Jin, H., Wan, X., & Yu, Z. (2020). *Domain-Adaptive Neural Automated Essay Scoring*. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '20), pages 1011-1020. Association for Computing Machinery.

Catulay, J. J. J. E., Magsael, M. E., Ancheta, D. O., & Costales, J. A. (2021). Neural-network architecture approach: An automated essay scoring using Bayesian linear ridge regression algorithm. *8th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, 196–200.

https://doi.org/10.1109/ISCMI53840.2021.9654801

Chassab, R. H., Zakaria, L. Q., & Tiun, S. (2021). Automatic essay scoring: A review on the feature analysis techniques. *International Journal of Advanced Computer Science and Applications, 12*(10), 252–264. https://doi.org/10.14569/IJACSA.2021.0121028

Chavva, R. K. R., Muthyam, S. R., Seelam, M. S., & Nalliboina, N. (2024). A transformer-based approach for enhancing automated essay scoring. *1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, 1–6. https://doi.org/10.1109/ACET61898.2024.10730000

Das, L. B., Raghu, C. V., Jagadanand, G., George, R. A. R., Yashasawi, P., Kumaran, N. A. A., & Patnaik, V. K. (2022). FACTOGRADE: Automated essay scoring system. *IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 42–48. https://doi.org/10.1109/IAICT55358.2022.9887447

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017). ReaderBench learns Dutch: Building a comprehensive automated essay scoring system for Dutch language. *Artificial Intelligence in Education: 18th International Conference, AIED 2017* (Vol. 10331, pp. 52–63). Springer. https://doi.org/10.1007/978-3-319-61425-0_5

Eang, C., & Lee, S. (2024). Improving the accuracy and effectiveness of text classification based on the integration of the BERT model and a recurrent neural network (RNN_BERT_Based). *Applied Sciences, 14*(18), 8388. https://doi.org/10.3390/app14188388

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y. I. (2024). Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics, 12*(21), 3416. https://doi.org/10.3390/math12213416

Fiacco, J., Adamson, D., & Rose, C. (2023). Towards extracting and understanding the implicit rubrics of transformer-based automatic essay scoring models. In E. Kochmar et al. (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 232–241). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.20

Fiacco, J., Jiang, S., Adamson, D., & Rosé, C. (2022). Toward automatic discourse parsing of student writing motivated by neural interpretation. In E. Kochmar et al. (Eds.), *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)* (pp. 204–215). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bea-1.25

Gaheen, M. M., ElEraky, R. M., & Ewees, A. A. (2021). Automated students' Arabic essay scoring using trained neural network by e-jaya optimization to support personalized instruction. *Education and Information Technologies, 26*(1), 1165–1181. https://doi.org/10.1007/s10639-020-10300-6

Geetha, M. P., & Renuka, D. K. (2021). Improving the performance of aspect-based sentiment analysis using fine-tuned BERT base uncased model. *International Journal of Intelligent Networks, 2*, 64–69. https://doi.org/10.1016/j.ijin.2021.06.005

Gillath, O., & Karantzas, G. (2019). Attachment security priming: A systematic review. *Current Opinion in Psychology, 25*, 86–95. https://doi.org/10.1016/j.copsyc.2018.03.001

Han, C. (2019). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments. *Measurement: Interdisciplinary Research and Perspectives, 17*(2), 113–116. https://doi.org/10.1080/15366367.2018.1516094

Hua, C., & Wind, S. A. (2019). Exploring the psychometric properties of the mind-map scoring rubric. *Behaviormetrika, 46*(1), 73–99. https://doi.org/10.1007/s41237-018-0062-z

Hussein, M. A., Hassan, H. A., & Nassef, M. (2020a). A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications, 11*(5), 287–293. https://doi.org/10.14569/IJACSA.2020.0110538

Hussein, M. A., Hassan, H., & Nassef, M. (2019a). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208

John Bernardin, H., Thomason, S., Buckley, M. R., & Kane, J. S. (2016). Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management, 55*(2), 321–340. https://doi.org/10.1002/hrm.21678

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 6300–6308. https://doi.org/10.24963/ijcai.2019/879

Kruse, O., Rapp, C., Anson, C. M., Benetos, K., Cotos, E., Devitt, A., & Shibani, A. (2023). *Digital writing technologies in higher education: Theory, research, and practice.* Springer. https://doi.org/10.1007/978-3-031-36033-6

Kusumaningrum, R., Kadarisman, K., Endah, S. N., Sasongko, P. S., Khadijah, K., Sutikno, S., Rismiyati, R., & Afriani, A. (2024). Automated essay scoring using convolutional neural network long short-term memory with mean of question-answer encoding. *ICIC Express Letters, 18*(8), 785–792. https://doi.org/10.24507/icicel.18.08.785

Lagakis, P., & Demetriadis, S. (2021). Automated essay scoring: A review of the field. *International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–6. https://doi.org/10.1109/CITS52676.2021.9618476

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated essay scoring (AES) research and development. *Pertanika Journal of Science and Technology, 29*(3), 1875–1899. https://doi.org/10.47836/pjst.29.3.27

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series, 2014*(1), 1–23. https://doi.org/10.1002/ets2.12009

Lonetti, F., Bertolino, A., & Di Giandomenico, F. (2023). Model-based security testing in IoT systems: A rapid review. *Information and Software Technology, 164*, 107326. https://doi.org/10.1016/j.infsof.2023.107326

Lu, C., & Cutumisu, M. (2021). Integrating deep learning into an automated feedback generation system for automated essay scoring. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)* (pp. 573–579). International Educational Data Mining Society.

Ludwig, S., Mayer, C., Hansen, C., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych, 3*(4), 897–915. https://doi.org/10.3390/psych3040056

Machhout, R. A., & Zribi, C. B. O. (2024). Enhanced BERT approach to score Arabic essay's relevance to the prompt. *IBIMA Business Review,* 2024. https://doi.org/10.5171/2024.176992

Mahmoud, S., Nabil, E., & Torki, M. (2024). Automatic scoring of Arabic essays: A parameter-efficient approach for grammatical assessment. *IEEE Access, 12*, 142555–142568. https://doi.org/10.1109/ACCESS.2024.3470728

Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2020)* (pp. 151–162). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.15

Misgna, H., On, B. W., Lee, I., & Choi, G. S. (2025). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review, 58*(2), 11017. https://doi.org/10.1007/s10462-024-11017-5

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/10.1016/j.rmal.2023.100050

Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S., & Porras, J. (2023). Mitigation strategies against phishing attacks: A systematic literature review. *Computers and Security, 132*, 103387. https://doi.org/10.1016/j.cose.2023.103387

Nguyen, H. V., & Litman, D. J. (2018). Argument mining for improving the automated scoring of persuasive essays. *32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, 5892–5899. https://doi.org/10.1609/aaai.v32i1.12046

Nie, Y. (2025). Automated essay scoring with SBERT embeddings and LSTM-attention networks. *PeerJ Computer Science, 11*, e2634. https://doi.org/10.7717/peerj-cs.2634

Ouyang, F., Wu, M., Zhang, L., Xu, W., Zheng, L., & Cukurova, M. (2023). Making strides towards AI-supported regulation of learning in collaborative knowledge construction. *Computers in Human Behavior, 142,* 107650. https://doi.org/10.1016/j.chb.2023.107650

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A.,

Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery, 88*, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Rahman, A. A., Ahmad, J., Yasin, R. M., & Hanafi, N. M. (2017). Investigating central tendency in competency assessment of design electronic circuit: Analysis using many facet Rasch measurement (MFRM). *International Journal of Information and Education Technology, 7*(7), 525–528. https://doi.org/10.18178/ijiet.2017.7.7.923

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

Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. (2020). Prompt-agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint*. http://arxiv.org/abs/2008.01441

Ridley, R., He, L., Dai, X. Y., Huang, S., & Chen, J. (2021). Automated cross-prompt scoring of essay traits. *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021), 15*, 13745–13753. https://doi.org/10.1609/aaai.v35i15.17620

Rosen, Y., & Tager, M. (2014). Making student thinking visible through a concept map in computer-based assessment of critical thinking. *Journal of Educational Computing Research, 50*(2), 249–270. https://doi.org/10.2190/EC.50.2.f

Sevcikova, B. L. (2018). Human versus Automated Essay Scoring: A Critical Review. *Arab World English Journal, 9*(2), 157-174. https://doi.org/10.24093/awej/vol9no2.11

Shin, J., & Gierl, M. J. (2020). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing, 38*(2), 247–272. https://doi.org/10.1177/0265532220937830

Song, W., Song, Z., Liu, L., & Fu, R. (2020). Hierarchical multi-task learning for organization evaluation of argumentative student essays. *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, 3875–3881. https://doi.org/10.24963/ijcai.2020/536

Tashu, T. M., Maurya, C. K., & Horvath, T. (2022). Deep learning architecture for automatic essay scoring. *arXiv preprint*. http://arxiv.org/abs/2206.08232

Uto, M., & Ueno, M. (2018). Empirical comparison of item response theory models with rater's parameters. *Heliyon, 4*(5), e00622. https://doi.org/10.1016/j.heliyon.2018.e00622

Uto, M., Xie, Y., & Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6077–6088). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.535

Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3416–3425). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.249

Xie, J., Cai, K., Kong, L., Zhou, J., & Qu, W. (2022). Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)* (pp. 2724–2733). Association for Computational Linguistics. https://aclanthology.org/2022.coling-1.240/

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560–1569). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.141

Zawacki-Richter, O., & Jung, I. (2023). *Handbook of open, distance and digital education.* Springer. https://doi.org/10.1007/978-981-19-2080-6

Zupanc, K., & Bosnić, Z. (2018). Increasing accuracy of automated essay grading by grouping similar graders. *ACM International Conference Proceeding Series.* https://doi.org/10.1145/3227609.3227645