Advance Sustainable Science, Engineering and Technology (ASSET)
Vol. 8, No.4, October 2026, pp. 02604029-01 ~ 02604029-017
ISSN: 2715-4211 DOI: https://doi.org/10.26877/asset.v8i4.2924

Design and Evaluation of a BERT- and T5-Based Automated Essay Scoring and Feedback Framework

Hendi Triatmoko*, Sani Muhamad Isa

Master of Computer Science, School of Computer Science, Bina Nusantara University, Jl. Kebon Jeruk Raya No. 27, Kemanggisan, West Jakarta 11530, Indonesia

*hendi.triatmoko@binus.ac.id

Abstract. Manual essay assessment is often time-consuming, subjective, and difficult to implement consistently in large-scale educational settings. Therefore, this study aims to develop and evaluate an integrated AES framework that combines essay scoring, spelling error detection, semantic relevance analysis, and automated feedback generation. The proposed framework integrates BERT for essay score prediction, pyspellchecker for spelling error detection, SBERT for semantic relevance analysis, and T5 for automated feedback generation. Experiments were conducted using a publicly available dataset consisting of 1,785 essays for model development and 592 essays for testing. The BERT model achieved strong predictive performance with an MAE of 0.2381, RMSE of 0.2847, and R^2 of 0.9513 on the test dataset, indicating high agreement with human scoring. The spelling detection module successfully identified linguistic errors, while the SBERT model produced relevance scores ranging from 0.58 to 0.95. The T5 model generated feedback that was consistent with essay scores, spelling quality, and semantic relevance. These findings demonstrate that integrating multiple NLP components can extend AES functionality beyond score prediction by providing comprehensive writing evaluation and formative feedback. Practically, the proposed framework supports scalable essay assessment and automated feedback generation. However, its generalizability is limited on a single-topic dataset and synthetic feedback data.

Keywords: Automated essay scoring, BERT T5, natural language processing, deep learning, supervised learning.

(Received 2025-10-31, Revised 2026-06-13, Accepted 2026-08-27, Available Online by 2026-10-31)

1. Introduction

Essay assessment remains one of the most widely used forms of evaluation in education because it enables students to demonstrate critical thinking, argumentative skills, and the ability to organize ideas coherently [1]. However, the large number of students across educational institutions presents significant challenges for educators in assessing essays. Manual essay grading is often time-consuming,

© 2026 The Author(s). Advance Sustainable Science, Environmental Engineering and Technology (ASSET) are published under the terms of the Creative Commons Attribution-ShareAlike (CC BY-SA 4.0). Authors retain the copyright to their work. Users may read, copy and distribute the work in any medium provided the authors and the journal are appropriately credited

susceptible to evaluator subjectivity, and difficult to standardize across raters [2]. Furthermore, providing detailed and constructive feedback requires substantial effort, limiting instructors' ability to offer timely responses that support improvements in student learning outcomes [3].

To address these challenges, Automated Essay Scoring (AES) systems have been developed to perform essay evaluation automatically using machine learning and Natural Language Processing (NLP) techniques [4]. AES has received increasing attention due to its ability to improve grading efficiency, reduce evaluator bias, and support large-scale assessment. Transformer-based architectures, particularly Bidirectional Encoder Representations from Transformers (BERT), have demonstrated superior capabilities in understanding contextual relationships among words and sentences, achieving state-of-the-art performance across various NLP tasks such as text classification, question answering, and essay scoring [5]. Consequently, BERT has emerged as one of the most promising approaches for developing more accurate and objective AES systems [6].

Beyond score prediction, the ability to provide feedback is increasingly recognized as an essential component of intelligent educational assessment systems [7]. Feedback helps students identify weaknesses in their writing and improve their communication skills [8]. Nevertheless, most AES studies continue to focus primarily on essay score prediction without integrating automated feedback generation mechanisms [9]. Advances in transformer-based text generation models have also created new opportunities for automated feedback systems. The Text-to-Text Transfer Transformer (T5) has shown excellent performance across a wide range of text generation tasks by reformulating NLP problems into a unified text-to-text framework [10]. Through its encoder-decoder architecture, T5 can generate contextually appropriate responses and recommendations based on the input information. This capability makes T5 particularly suitable for generating personalized feedback based on essay quality indicators [11]. In addition, other NLP techniques, such as spelling error detection and semantic relevance analysis, can provide complementary information regarding writing quality. Spelling error analysis can be used to evaluate linguistic accuracy, while semantic relevance analysis can measure the extent to which essay content aligns with the assigned topic [12]. Integrating these components has the potential to create a more comprehensive writing evaluation system.

Despite the growing adoption of transformer-based models in AES, several research gaps remain. First, most existing studies focus exclusively on essay score prediction without incorporating automated feedback generation. Second, research integrating multiple assessment dimensions—such as essay quality, spelling accuracy, and semantic relevance—within a unified framework remains relatively limited. Third, empirical evidence regarding the effectiveness of combining BERT for essay scoring, Sentence-BERT (SBERT) for semantic relevance analysis, pyspellchecker for spelling error detection, and T5 for automated feedback generation is still scarce, particularly in educational assessment contexts. These gaps highlight the need for a more integrated AES architecture capable of performing both assessment and feedback generation simultaneously.

This study seeks to address these limitations by proposing an Automated Essay Scoring and automated feedback generation framework that integrates four main components: (1) a BERT-based regression model for essay score prediction, (2) a pyspellchecker module for spelling error detection, (3) a Sentence-BERT (SBERT)-based module for semantic relevance analysis with respect to the assigned topic, and (4) a T5 model for automated feedback generation. Unlike conventional AES approaches that focus solely on score estimation, the proposed framework integrates multiple writing quality indicators to provide a more comprehensive evaluation process. This integration is expected to enhance grading consistency while delivering constructive feedback that supports the development of students' writing skills.

Based on these considerations, this study aims to develop and evaluate an automated essay scoring and feedback generation framework based on BERT, SBERT, pyspellchecker, and T5. To achieve this objective, the study addresses the following research questions:

RQ1: How accurately can a fine-tuned BERT model predict essay scores using a publicly available essay dataset?

RQ2: How effectively can a spelling error detection module identify linguistic errors that may affect

essay quality?

RQ3: How effectively can SBERT-based semantic relevance analysis evaluate the alignment between essay content and the assigned topic?

RQ4: To what extent can a T5-based text generation model produce automated feedback aligned with essay scores, spelling quality, and semantic relevance?

The findings of this study are expected to contribute to the advancement of intelligent educational technologies by providing a scalable, consistent, and objective essay assessment framework. From a practical perspective, the proposed system can assist educators in reducing grading workloads while enabling students to receive faster and more informative feedback. From an academic perspective, this research extends the application of transformer-based architectures in Educational Natural Language Processing by demonstrating the feasibility of integrating essay scoring, linguistic analysis, semantic relevance evaluation, and automated feedback generation within a unified framework.

2. Methods

2.1 Research Design

This study employed an experimental research method to develop and evaluate an Automated Essay Scoring (AES) and automated feedback generation system based on Natural Language Processing (NLP). The proposed system integrates four main components: a BERT model for essay score prediction, pyspellchecker for spelling error detection, Sentence-BERT (SBERT) for semantic relevance analysis, and a T5 model for automated feedback generation. The integration of these components aims to create an assessment process that is not only accurate but also capable of providing constructive feedback to essay writers.

In general, the research workflow begins with the collection of an essay dataset that has been previously scored by human raters. The dataset then undergoes a preprocessing stage consisting of text cleaning, normalization, and tokenization. The preprocessed essays are used as input to the BERT model for essay score prediction. In parallel, spelling error analysis is conducted using pyspellchecker, while topic relevance analysis is performed using SBERT. The outputs from these three analyses are subsequently combined as input parameters for the T5 model to generate automated feedback that reflects the overall quality of the essay.

The final stage of the study involves evaluating the system's performance using quantitative metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) for the essay scoring model. In addition, qualitative evaluation is conducted to assess the effectiveness of spelling error detection, semantic relevance analysis, and the quality of the generated feedback. The entire research process is designed to produce an automated essay assessment framework that can be implemented in educational environments in a more efficient, objective, and consistent manner. The complete research workflow is illustrated in Figure 1.

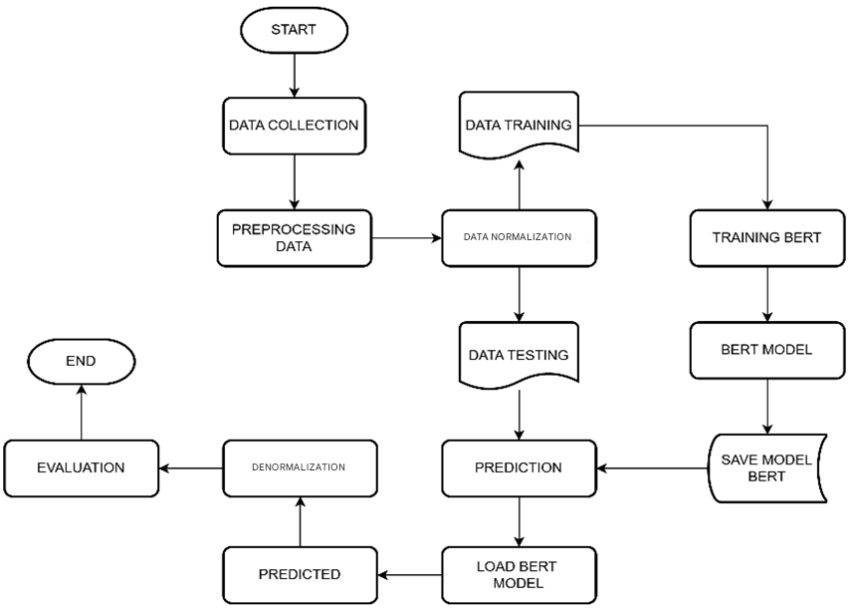


Figure 1. Research Stages

2.2 Data Sources and Characteristics

This study utilized the Automated Essay Scoring Dataset obtained from Kaggle (https://www.kaggle.com/datasets/thevirusx3/automated-essay-scoring-dataset). The dataset was selected because it is one of the most widely used datasets in Automated Essay Scoring (AES) research and provides essay scores assigned by human raters. The availability of these reference scores enables the application of a supervised learning approach to train models for automatically predicting essay quality.

The dataset consists of essays written by eighth-grade students in response to a writing prompt concerning the impact of computer use on society. Each essay was evaluated by human raters and assigned a score ranging from 2 to 12. Overall, the dataset contains 1,785 essays for model development and 592 essays for final evaluation. The average essay length is approximately 350 words, providing substantial variation in sentence structure, vocabulary usage, and argumentative quality, which are essential for effective model training.

Prior to model development, the dataset was examined to ensure the absence of missing or duplicate entries. All essays meeting the completeness criteria were included in the study. The dataset was subsequently divided into training and validation sets using an 80:20 ratio. The training set was used to train the BERT model, while the validation set was employed to monitor model performance during training and prevent overfitting. After the training process was completed, the model's performance was evaluated using the 592 essays provided as an independent test set.

Since the proposed scoring model operates as a regression model, essay scores were normalized into an appropriate numerical range before training. This normalization process was performed to improve learning stability and accelerate convergence during model training. After generating predictions, the predicted scores were transformed back to their original scale through a denormalization process, enabling direct comparison with the scores assigned by human raters.

Although the dataset provides sufficient data for developing an AES system, several limitations should be acknowledged. First, all essays were written in response to a single topic, limiting the ability to assess the model’s generalizability across different writing prompts. Second, the dataset contains only English-language essays; therefore, the performance of the proposed framework in other languages requires further investigation. Third, the relatively homogeneous characteristics of the writers, who belong to the same educational level, may restrict the diversity of writing styles available for model learning. Consequently, the findings of this study should be interpreted in light of these limitations.

Table 1. Research Dataset Characteristics

Characteristic	Value
Dataset Source	Kaggle Automated Essay Scoring Dataset
Essay Type	Persuasive Essay
Education Level	Grade 8
Number of Training Samples	1,785 essays
Number of Test Samples	592 essays
Average Essay Length	Approximately 350 words
Score Range	2–12
Language	English
Reference Scoring Method	Human Raters

2.3 Data Preprocessing

Data preprocessing was conducted to improve the quality of the textual data and ensure that all essays were formatted appropriately for processing by transformer-based models. This stage is essential because the quality of input data directly influences the model's ability to learn linguistic and semantic patterns embedded within the text. Furthermore, preprocessing aims to reduce noise that may interfere with the model training process.

The first step involved text cleaning. During this stage, special characters, irrelevant symbols, and text elements that did not contribute to the essay assessment process were removed. Subsequently, all text was converted to lowercase to maintain consistency in word representation and reduce variations among tokens that share the same meaning but differ in letter casing. This process facilitates more effective learning of word representations by the model.

The next stage was text normalization. Normalization was performed to simplify the textual content and make it more suitable for model processing. In this study, normalization included the removal of unnecessary characters and the adjustment of text formatting to meet the input requirements of the BERT model. The output of the cleaning and normalization processes was a more structured text corpus that was ready for tokenization.

After normalization, tokenization was performed using the BERT Tokenizer provided by the HuggingFace Transformers library. Tokenization is the process of converting text into smaller units (tokens) that can be interpreted by the BERT model. Each token was subsequently transformed into numerical representations consisting of input IDs and attention masks. Input IDs represent tokens in numerical form, whereas attention masks indicate which tokens should be attended to by the model during training. In this study, the maximum sequence length was set to 512 tokens, corresponding to the maximum capacity of the BERT Base model. This configuration allowed most essays to be processed without significant information loss, while padding and truncation mechanisms ensured consistent input dimensions throughout the training process.

To address variations in essay length, padding and truncation mechanisms were applied. Padding was used to add empty tokens to shorter texts so that all inputs shared a uniform length. Conversely, truncation was used to shorten texts that exceeded the maximum input length supported by the model. The implementation of these mechanisms enabled efficient training while maintaining model stability. The final output of the preprocessing stage was a dataset represented in numerical format and ready for use in the BERT model training process. The data preprocessing workflow implemented in this study is illustrated in Figure 2.

2.4 Model Architectures

The proposed framework consists of four main components: BERT-based essay scoring, spelling error detection, semantic relevance analysis, and T5-based feedback generation. These four components are designed to operate in an integrated manner to provide a comprehensive evaluation of essay quality. The system workflow begins with essays that have undergone the preprocessing stage. The essays are then processed in parallel by the essay scoring module, spelling error detection module, and semantic

relevance analysis module. The outputs generated by these three modules are subsequently used as inputs for the T5 model to produce automated feedback tailored to the specific characteristics and quality of each essay.

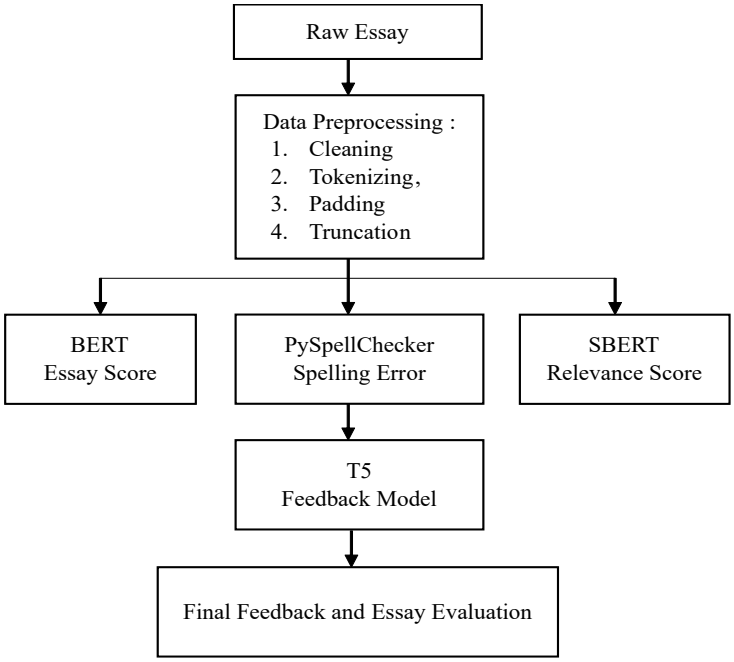


Figure 3. Architecture of the Proposed Automated Essay Scoring Framework

Essay assessment was performed using the Bidirectional Encoder Representations from Transformers (BERT) model with the bert-base-uncased configuration. This model was selected due to its ability to capture bidirectional contextual information, enabling it to learn complex semantic relationships within textual data [13]. During the training stage, the text representations generated by the BERT encoder layers were passed to a regression layer to predict essay scores as continuous values. The model was fine-tuned using essays that had been scored by human raters, allowing it to learn the relationship between textual characteristics and overall writing quality.

In this study, the model output consisted of normalized essay scores during training, which were subsequently transformed back to their original score range during the evaluation stage. The predicted score served as the primary indicator of essay quality and was used as one of the input parameters for the automated feedback generation process.

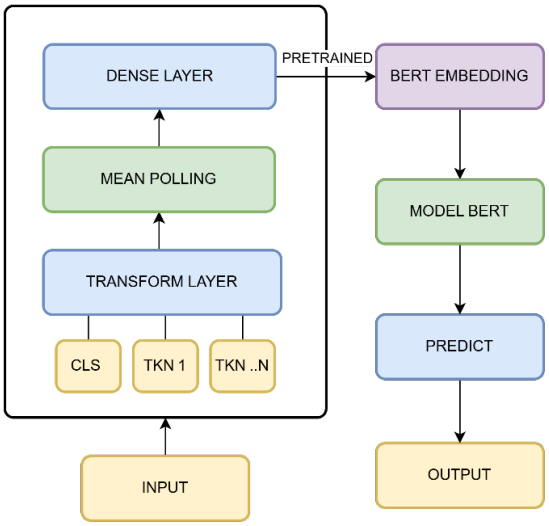


Figure 4. BERT-Based Essay Scoring Architecture

As illustrated in Figure 4, the scoring process begins by feeding the preprocessed essay text into the BERT Tokenizer. The tokenizer converts the text into tokens that can be interpreted by the model and generates input IDs and attention masks as inputs to the BERT encoder layers. Through the self-attention mechanism, BERT learns the relationships among words across the entire text, producing rich contextual representations that capture both linguistic and semantic information.

The representations generated by the encoder layers are subsequently aggregated using a mean pooling strategy. This method is employed to obtain a single vector representation that summarizes the overall content of the essay. The pooled vector is then passed to a fully connected layer that functions as a regression model for predicting essay scores. This approach enables the model to learn the relationship between the linguistic characteristics of essays and the scores assigned by human raters.

The model training process was conducted using the prepared training dataset. Model parameters were optimized using the Adam Optimizer, which is widely recognized for its effectiveness in training deep learning models. Throughout the training process, model performance was monitored using the validation dataset to ensure that the model could generalize effectively to previously unseen data.

To reduce the risk of overfitting, an early stopping mechanism was implemented. This mechanism automatically terminates the training process when no significant improvement is observed on the validation dataset. Such an approach allows the model to achieve optimal performance without necessarily completing all predefined training epochs. In this study, training was automatically halted once the model reached a stable error level and showed no meaningful improvement over several consecutive iterations.

To evaluate linguistic quality, this study employed the pyspellchecker library, which is based on an English-language dictionary. This module identifies words that are not found in the standard dictionary and calculates the number of spelling errors present in each essay. Error detection is performed after basic tokenization by comparing each token against the vocabulary contained in the pyspellchecker dictionary. The number of detected spelling errors is not used to modify the essay score predicted by BERT; instead, it serves as an additional indicator of writing quality. This information is subsequently passed to the T5 module together with the essay score and semantic relevance value to generate more informative and diagnostic feedback.

Semantic relevance analysis was performed using Sentence-BERT (SBERT) with the all-MiniLM-L6-v2 model. SBERT was employed to generate sentence embeddings capable of capturing semantic meaning more effectively than keyword-based approaches. In this study, essay embeddings were compared with reference embeddings using cosine similarity to measure the degree of alignment between essay content and the assigned topic. Cosine similarity values range from 0 to 1, where values closer to

1 indicate a higher level of semantic relevance. The relevance score serves as an indicator of content quality and is used as one of the inputs to the automated feedback generation module. Consequently, the proposed system evaluates not only linguistic quality but also the extent to which essay content addresses the assigned topic.

Automated feedback generation was carried out using the Text-to-Text Transfer Transformer (T5) model. This module receives three primary input parameters: the essay score predicted by BERT, the number of spelling errors detected by pyspellchecker, and the semantic relevance score calculated by SBERT. These parameters are converted into a textual format and used as input for generating feedback that reflects the overall quality of the essay. In this study, each parameter has a different value range according to the characteristics of the corresponding module.

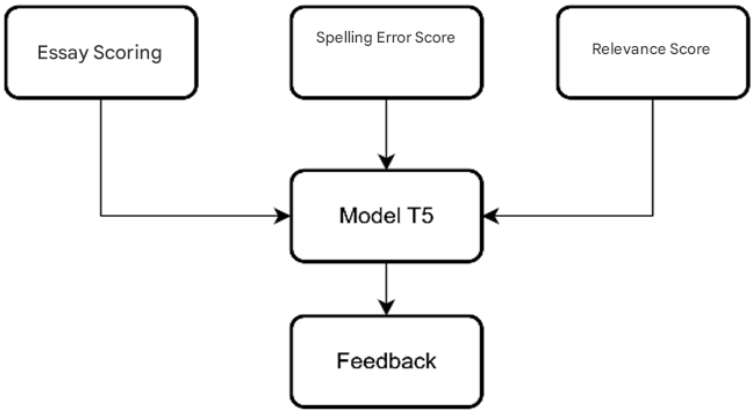


Figure 5. Flow Feedback Generation

2.5 Training Procedures and Hyperparameters

The training process was conducted to optimize the model's ability to predict essay scores based on data labeled by human raters. The BERT model (bert-base-uncased) was employed as the pre-trained backbone and subsequently fine-tuned using the preprocessed essay dataset. During training, the textual representations generated by the BERT encoder were passed to a regression layer to produce continuous essay score predictions. The loss function used was Mean Squared Error (MSE), which is widely adopted in regression tasks because of its effectiveness in measuring the discrepancy between predicted scores and reference scores.

The training data were divided into mini-batches to improve computational efficiency and optimization stability. Model parameters were updated using the Adam (Adaptive Moment Estimation) optimization algorithm with a learning rate specifically adjusted for fine-tuning transformer-based models. Throughout the training process, model performance was monitored using a validation dataset that was separate from the training data. This approach enabled periodic evaluation of the model's generalization capability and helped prevent overfitting.

To further enhance training stability, an early stopping mechanism was implemented. Training was automatically terminated when no improvement in validation performance was observed after a predefined number of epochs. Although the maximum number of training epochs was set to 3,000, the training process stopped earlier once the model reached stable performance. In the conducted experiments, early stopping halted training at epoch 17, as the model had achieved a low error rate and showed no significant improvement in subsequent iterations. This strategy helped reduce the risk of overfitting while also minimizing computational time.

In addition to the evaluation of the BERT model, the spelling error detection and semantic relevance analysis modules did not require additional training because they relied on pre-existing models and libraries. The pyspellchecker module was directly employed to identify spelling errors based on an English-language dictionary, while SBERT (all-MiniLM-L6-v2) was used to generate semantic text

representations and compute cosine similarity scores. Meanwhile, the T5 model was trained using 1,000 synthetic feedback samples constructed from combinations of essay scores, spelling error counts, and semantic relevance levels. Therefore, the feedback generation component in this study is positioned as a feasibility study, demonstrating the potential for integrating automated feedback generation into an Automated Essay Scoring (AES) framework.

Table 2. Training Hyperparameters

Parameter	Value
Base Model	BERT Base Uncased
Task	Essay Score Regression
Tokenizer	BERT Tokenizer
Maximum Sequence Length	512
Optimizer	Adam
Loss Function	Mean Squared Error (MSE)
Maximum Epochs	3000
Early Stopping	Enabled
Validation Strategy	Hold-out Validation (80:20)
SBERT Model	all-MiniLM-L6-v2
Spelling Detection	PySpellChecker
Feedback Model	T5
Synthetic Feedback Dataset	1,000 Instances

2.6 Evaluation Metrics and Experimental Protocol

Evaluation was conducted to assess the performance of each component within the proposed framework, namely BERT-based essay scoring, spelling error detection using pyspellchecker, semantic relevance analysis using SBERT, and feedback generation using T5. All experiments were performed using an essay dataset scored by human raters, which served as the primary reference for evaluating the models' ability to predict writing quality.

For the BERT-based essay scoring module, evaluation was carried out using an independent test set consisting of 592 essays that were not used during either the training or validation phases. Model performance was measured using three regression metrics commonly employed in Automated Essay Scoring (AES) research: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2 Score). MAE was used to measure the average absolute difference between predicted scores and the reference scores assigned by human raters. RMSE was employed to impose a greater penalty on large prediction errors, while the R^2 Score measured the proportion of variance in essay scores that could be explained by the model. Lower MAE and RMSE values indicate better model performance, whereas an R^2 Score closer to 1 indicates stronger predictive capability.

The spelling error detection module was evaluated functionally by calculating the number of words identified as spelling errors using pyspellchecker. Since the dataset did not provide annotated spelling errors as ground truth, quantitative evaluation using metrics such as precision, recall, and F1-score could not be performed. Therefore, the output of this module was utilized as a diagnostic indicator that complements the overall assessment of writing quality. Meanwhile, semantic relevance analysis using SBERT was evaluated based on the cosine similarity scores generated between essay embeddings and reference embeddings. Relevance scores ranged from 0 to 1, where higher values indicated a greater degree of alignment between the essay content and the assigned topic.

Table 3. Evaluation Metrics Used in the Study

Module	Evaluation Metric	Purpose
BERT Essay Scoring	MAE	Measure average prediction error

BERT Essay Scoring	RMSE	Measure prediction error magnitude
BERT Essay Scoring	R^2 Score	Measure explanatory power of the model
Spelling Detection	Number of Detected Errors	Assess linguistic accuracy indicators
Semantic Relevance (SBERT)	Cosine Similarity	Measure topic relevance
Feedback Generation (T5)	Qualitative Assessment	Evaluate consistency between feedback and essay characteristics

3. Results and Discussion

3.1 Performance of BERT-Based Essay Scoring

The performance of the BERT model in automated essay scoring was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Evaluation was conducted on an independent test set that was not used during the training process to assess the model's ability to generalize to unseen essays. The evaluation results are presented in Table 4.

Table 4. Performance of the BERT Model on Training and Test Data

Metric	Training Data	Test Data
MAE	0,2069	0,2381
RMSE	0,2524	0,2847
R^2	0,9720	0,9513
Loss	0,0637	0,0811
Best Epoch	17	17

Based on Table 4, the BERT model demonstrated excellent performance on both the training and test datasets. The MAE value increased only slightly from 0.2069 on the training set to 0.2381 on the test set. Similarly, the RMSE value showed a relatively small increase from 0.2524 to 0.2847. The limited difference between the two datasets indicates that the model possesses strong generalization capability and does not suffer from significant overfitting. This finding is further supported by the relatively stable loss values observed across the training and test datasets.

The R^2 values of 0.9720 for the training data and 0.9513 for the test data indicate that the model is capable of explaining more than 95% of the variation in essay scores assigned by human raters. These results suggest that the contextual representations generated by BERT effectively capture various important characteristics of written essays, including argument quality, coherence between ideas, content relevance, and language usage. Therefore, the model does not merely learn simple statistical patterns from the data but is also capable of understanding more complex semantic relationships within essay texts.

The performance achieved in this study is consistent with the findings of Fianu et al. [14], who reported that BERT-based models can achieve higher scoring accuracy than traditional linguistic feature-based approaches. BERT's self-attention mechanism enables the model to capture relationships among words and sentences more effectively, resulting in predictions that more closely resemble human judgments [15]. The results of the present study reinforce these findings, as the model maintained a low prediction error on both the training and test datasets.

The findings are also consistent with the study by Xue et al. [16], which demonstrated that fine-tuning BERT significantly improves performance in AES tasks compared with earlier generations of deep learning models. The model's ability to leverage linguistic knowledge acquired during the pre-training phase enables it to adapt efficiently to the characteristics of the target essay dataset. A similar

pattern was observed in this study, where the model achieved optimal performance at epoch 17 through the implementation of an early stopping mechanism, resulting in a more efficient training process without compromising predictive accuracy.

Compared with studies employing Long Short-Term Memory (LSTM) architectures, the results of this study suggest superior performance in understanding essay context. Han et al. [17] explained that LSTM models are capable of capturing sequential word dependencies in text; however, they remain limited in modeling long-range contextual relationships within essays containing complex argumentative structures. In contrast, BERT utilizes a bidirectional self-attention mechanism that allows each word to be analyzed within the context of the entire sentence and paragraph. This capability is one of the key factors contributing to the model's more stable predictions and closer alignment with human scoring. Furthermore, the findings of this study support recent developments in the field of Automated Essay Scoring (AES), which demonstrate the superiority of transformer-based architectures over CNN- and RNN-based models.

3.2 Spelling Error Detection Results

The experimental results demonstrate that the proposed system is capable of identifying various types of spelling errors found in essays, including missing letters, extra letters, typographical errors, and the use of words that do not conform to the standard English vocabulary. Based on the analysis of the test data, essays with higher scores tended to contain fewer spelling errors than essays with lower scores. This finding suggests a relationship between linguistic quality and overall writing quality.

Table 5. Examples of Spelling Error Detection Results on the Test Dataset

Essay ID	Essay Score	Number of Words	Spelling Errors
E01	11	378	0
E02	10	341	1
E03	8	322	3
E04	6	298	7
E05	4	271	12

As shown in Table 5, the number of spelling errors generally increases as essay scores decrease. The essay with a score of 11 contained no spelling errors, whereas the essay with a score of 4 contained 12 spelling errors. Although spelling accuracy is not the sole determinant of essay quality, these findings indicate that accurate language use contributes to the effectiveness of written communication and may influence perceptions of overall writing quality.

The findings of this study are consistent with those reported by Abdulkarem and Krivtsun [18], who argued that linguistic features, including spelling and grammatical accuracy, are important indicators in automated essay scoring systems. According to their findings, excessive spelling errors can reduce text readability and hinder the effective communication of ideas. Consequently, many modern AES systems incorporate linguistic features as supporting components in the essay evaluation process.

The results also support the findings of Lim et al. [19] in the review of AES, which demonstrated that linguistic features make a significant contribution to the quality of automated assessment. However, unlike e-rater, which incorporates spelling-related features directly into the scoring process, the present study utilizes spelling error detection as a diagnostic information source for generating more informative feedback to users. As a result, the proposed system is not solely focused on score prediction but also assists users in identifying linguistic weaknesses within their writing.

Despite its ability to detect various forms of spelling errors, this study has certain limitations. The dataset used does not provide annotated spelling errors as ground truth, preventing the evaluation of detection accuracy using quantitative metrics such as precision, recall, and F1-score. Similar limitations have been reported in many AES studies that rely on publicly available datasets, where detailed linguistic error annotations are generally unavailable. Therefore, the spelling error detection results presented in this study should be viewed as a complementary diagnostic analysis rather than a fully

validated error detection benchmark.

Overall, the integration of the spelling error detection module adds significant value to the proposed framework. Most previous AES studies have focused primarily on predicting essay scores based on content and structural features, whereas this study introduces an additional linguistic evaluation dimension that can be leveraged to generate more comprehensive feedback. Consequently, the system is capable not only of automatically assigning essay scores but also of providing information that can help users improve the linguistic quality of their writing.

3.3 Semantic Relevance Analysis

Semantic relevance analysis was conducted to measure the degree of alignment between essay content and the assigned topic. Unlike BERT-based essay scoring, which focuses on predicting an overall essay score, the semantic relevance module aims to evaluate whether the arguments and content presented in the essay remain within the scope of the designated topic. In this study, semantic relevance analysis was performed using SBERT with the all-MiniLM-L6-v2 architecture, which generates vector representations (sentence embeddings) for each essay.

The relevance level was calculated using cosine similarity between the embedding of a target essay and a reference embedding derived from essays within the dataset. Cosine similarity values range from 0 to 1, where values closer to 1 indicate a higher degree of semantic alignment with the assigned topic. The resulting relevance scores were subsequently used as one of the input parameters for the automated feedback generation process, providing information regarding the appropriateness of essay content.

Table 6. Examples of Semantic Relevance Analysis Results on the Test Dataset

Essay ID	Essay Score	Relevance Score
E01	11	0,95
E02	10	0,91
E03	8	0,84
E04	6	0,72
E05	4	0,58

As shown in Table 6, essays with higher scores generally exhibit higher semantic relevance values than lower-scoring essays. The essay with a score of 11 achieved a relevance score of 0.95, indicating that its content was highly aligned with the assigned topic. In contrast, the essay with a score of 4 obtained a relevance score of 0.58, suggesting that portions of its content deviated from the intended topic. These findings indicate that topic relevance is an important factor contributing to overall essay quality.

The results demonstrate that the SBERT model is capable of distinguishing between relevant and less relevant essays based on the semantic representations it generates [20]. This capability is particularly important because essay evaluation depends not only on grammatical quality and argumentative structure but also on the writer's ability to maintain focus on the assigned topic. By utilizing richer semantic representations than keyword-based approaches, SBERT can identify similarities in meaning even when different vocabulary is used.

The findings of this study are consistent with those reported by Riyanto et al. [21], who introduced SBERT as a more efficient approach for measuring semantic similarity between sentences and documents. A similar capability was observed in the present study, where SBERT generated relevance scores that effectively reflected the degree of alignment between essay content and the assigned topic.

The results also support the findings of Chang et al. [22], who demonstrated that transformer-based embedding models outperform traditional keyword-matching approaches such as BM25 and TF-IDF in text retrieval and relevance evaluation tasks. Their study showed that transformer embeddings can learn richer semantic representations and capture conceptual relationships beyond surface-level lexical overlap, enabling more accurate identification of relevant content. This capability explains why SBERT in the present study was able to distinguish essays that maintained a strong topical focus from those that

deviated from the assigned topic, even when different vocabulary or writing styles were used.

Compared with most previous AES studies, the incorporation of semantic relevance analysis in this research provides an additional evaluation dimension that is rarely integrated directly into automated essay assessment systems. Most previous studies have focused primarily on essay score prediction using machine learning or deep learning models, while topic alignment is often not evaluated as a separate component. By integrating SBERT into the framework, this study not only generates essay scores but also provides information regarding the degree of relevance between essay content and the assigned topic.

Although the results demonstrate promising potential, this study has certain limitations because no human-annotated relevance labels were available as ground truth. Consequently, the semantic relevance evaluation remains exploratory in nature and is primarily used as a supporting indicator within the automated feedback generation process. Future research could develop datasets that include topic relevance annotations, enabling a more comprehensive quantitative evaluation of SBERT's performance.

Overall, the integration of SBERT-based semantic relevance analysis makes a significant contribution to the proposed system. This module enables the framework to assess the extent to which essay content aligns with the assigned topic, resulting in feedback that is more informative and not solely focused on the final score. Consequently, the proposed system provides a more comprehensive evaluation than conventional AES approaches that focus exclusively on essay score prediction.

3.4 T5 Feedback Generation Results

In addition to generating assessment scores, the proposed system was designed to provide automated feedback that helps users understand the quality of their writing. To achieve this objective, the study employed the Text-to-Text Transfer Transformer (T5) model to generate feedback based on a combination of essay scores, spelling error counts, and semantic relevance levels. This approach enables the system to produce more informative recommendations than conventional Automated Essay Scoring (AES) systems, which typically provide only numerical scores. [23]

Table 7. Examples of Automated Feedback Generation Results

Essay	Theme	Essay Score	Essay Spelling Mistakes	Essay Relevance	Feedback Essay
Essay 1	Wedding	40%	0	29%	The relevance to the topic is very poor, with many sections straying from the main topic. The spelling is excellent, with no errors whatsoever. The essay is very poor, with numerous errors in structure and organization of ideas.
Essay 2	Computer Influence	72%	8	83%	The relevance is very good, you've managed to present information that is relevant to the topic. The spelling is very poor, with many words not spelled correctly. This essay is quite good, although some sections could be better organized.
Essay 3	Computer Development	58%	1	54%	The relevance is clear, but needs more depth to better focus on the topic. Some errors were found, but they do not affect the overall understanding of the text. The essay is quite good, but some sections still need development to be more organized.

Based on Table 7, the T5 model was able to generate different types of feedback according to the characteristics of each essay. The generated feedback considered not only the final essay score but also linguistic quality and topic relevance. These results indicate that the model successfully integrates information obtained from the BERT, pyspellchecker, and SBERT modules into recommendations that are easily understandable for users.

For Essay 1, the system generated an essay score of 40% with a low relevance level (29%) and no detected spelling errors. The feedback primarily focused on the poor alignment between the essay content and the assigned topic, despite the excellent spelling quality. This finding demonstrates that the system can distinguish between linguistic quality and content quality. In other words, an essay that is free from spelling errors does not necessarily exhibit high overall quality if its content is not relevant to the assigned topic.

For Essay 2, the system generated a score of 72% with a high relevance level (83%), although eight spelling errors were detected. The generated feedback emphasized the strong content quality and high topic relevance while still highlighting weaknesses in linguistic accuracy. This result indicates that the model is capable of identifying both strengths and weaknesses in a balanced manner, allowing users to obtain a more comprehensive understanding of their essay quality.

Meanwhile, Essay 3 received a score of 58%, with a moderate relevance level (54%) and one spelling error. In this case, the feedback focused primarily on the need to further develop ideas and improve topic focus throughout the discussion. This finding suggests that the model evaluates not only language quality but also the depth of discussion and the organization of ideas within the essay.

Overall, the results indicate that higher essay scores and stronger topic relevance tend to produce more positive feedback from the system. Conversely, essays with lower scores and relevance levels generally receive more recommendations related to content improvement, essay organization, and topic focus. This pattern demonstrates that the T5 model is capable of generating feedback that is consistent with the analytical results produced by the preceding modules.

The findings of this study are consistent with the work of Majeed et al. [24], who argued that AES systems possess greater pedagogical value when they provide feedback that helps users understand the strengths and weaknesses of their writing. According to Cutumisu, numerical scores alone are often insufficient to support learning because they do not indicate which aspects require improvement. The results of the present study support this argument, as the T5 model successfully transformed quantitative evaluation results into recommendations that are more meaningful and accessible to users.

The findings are also consistent with Raffel et al. [25], who demonstrated that T5 is highly effective in generating contextually relevant text through a unified text-to-text framework. In the present study, the generated feedback aligned with the combination of essay scores, spelling error counts, and semantic relevance values, suggesting that T5 can function not only as a text generation model but also as a mechanism for translating quantitative evaluation results into meaningful feedback.

Although the results are promising, this study still has several limitations. The T5 model was trained using 1,000 synthetic feedback samples generated according to predefined rules rather than authentic feedback provided by teachers or professional raters. Consequently, the diversity and depth of the generated feedback remain more limited than those of human-generated feedback. Future studies should therefore incorporate evaluations by educational experts and utilize authentic feedback datasets to further improve the quality and pedagogical value of the generated recommendations.

3.5 Implications for Educational Technology

The findings of this study demonstrate that the integration of NLP technologies and transformer-based models has significant potential to enhance the effectiveness of assessment processes in digital learning environments. The proposed system is capable not only of automatically generating essay scores but also of providing spelling error analysis, topic relevance evaluation, and constructive feedback. These capabilities indicate that artificial intelligence can be leveraged to support a more comprehensive evaluation process than conventional automated assessment approaches, which typically focus solely on score generation.

From a practical perspective, the proposed system can help reduce educators' workload in essay assessment, particularly in classes with large numbers of students. The evaluation process, which traditionally requires substantial time and effort, can be performed more rapidly and consistently through the use of BERT and T5 models. In addition to improving efficiency, the system enables students to receive feedback more quickly, allowing them to identify weaknesses in their writing and make timely improvements. The availability of prompt feedback is widely recognized as an important factor in supporting continuous and student-centered learning.

The findings also indicate that the integration of SBERT-based semantic relevance analysis and pyspellchecker-based spelling error detection extends the functionality of digital assessment systems. Information regarding topic alignment and linguistic quality provides a more comprehensive picture of students' writing abilities. Consequently, the system functions not only as an evaluation tool but also as a diagnostic instrument that can assist teachers in identifying students' learning difficulties and designing more targeted instructional strategies.

Nevertheless, the use of automated assessment systems in education is not intended to replace teachers entirely. The proposed framework is more appropriately positioned as a decision support tool that assists the evaluation and learning process. Human assessment remains essential for considering aspects that are difficult to measure automatically, such as creativity, originality of ideas, and depth of argumentation. Therefore, combining the analytical capabilities of artificial intelligence with the professional judgment of educators can lead to assessment processes that are more effective, objective, and focused on improving learning quality. [26]

4. Conclusion

This study successfully developed an Automated Essay Scoring (AES) and automated feedback generation framework that integrates BERT, pyspellchecker, SBERT, and T5 into a unified system to support AI-driven essay evaluation. The results demonstrate that the BERT model can predict essay scores with high accuracy and relatively low prediction error, while the pyspellchecker module effectively identifies spelling errors and the SBERT module accurately analyzes the semantic relevance of essays to the assigned topic. The integration of these three components enables the T5 model to generate automated feedback that is consistent with essay quality, spelling accuracy, and topic relevance. Consequently, the proposed system functions not only as an automated assessment tool but also as a learning support mechanism that provides diagnostic information to users. These findings highlight the significant potential of transformer-based approaches to enhance digital assessment by combining scoring, linguistic analysis, semantic analysis, and feedback generation within a comprehensive framework.

Despite these promising results, the study has several limitations. The dataset used was restricted to a single writing topic and a single language, and the T5 model was trained using synthetic feedback data. As a result, the generalizability of the framework and the quality of the generated feedback require further investigation. Future research should therefore employ more diverse datasets, incorporate evaluations conducted by educators or language experts, and integrate additional dimensions such as grammatical accuracy, argumentation quality, and deeper discourse analysis. Such improvements may further enhance the accuracy, reliability, and pedagogical value of the system in supporting the ongoing transformation of digital education.

Declaration of AI and AI assisted technologies in the writing process

In preparing this manuscript, the author(s) employed ChatGPT and Gemini AI as auxiliary tools to support language polishing, improve sentence coherence, refine academic wording, and enhance the overall readability of the text. All outputs and suggested revisions generated by these tools were carefully assessed and edited by the author(s) to ensure that the scientific content, arguments, and

conclusions remained accurate and fully under the authors' control. The author(s) accept full responsibility for the final manuscript.

Declaration of Competing Interest

The authors declare that they have no financial, professional, institutional, or personal interests that could reasonably be perceived as influencing the research process, interpretation of the findings, or preparation of this manuscript.

References

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

[2] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2495–2527, 2022, doi: https://doi.org/10.1007/s10462-021-10068-2.

[3] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring and feedback generation.," *Artif. Intell. Rev.*, vol. 58, no. 2, p. 36, 2025, doi: https://doi.org/10.1007/s10462-024-11017-5.

[4] A. Bilonikar, H. Chandra, V. Dapolikar, A. Halbe, and N. Mulla, "Automated Evaluation and Quiz Generation using Deep Learning and Transformers," in *2026 IEEE International Conference for Convergence in Computing Technology (I3CTCON)*, 2026, pp. 1–6. doi: https://doi.org/10.1109/I3CTCON68242.2026.11507516.

[5] Zaki and M. Zayyanu, "R EVOLUTIONISING T RANSLATION T ECHNOLOGY : A C OMPARATIVE S TUDY OF V ARIANT T RANSFORMER M ODELS - BERT , GPT AND T5," *Computer Science & Engineering: An International Journal (CSEIJ)*, vol. 14, no. 3, pp. 15–27, 2024, doi: https://doi.org/10.5121/cseij.2024.14302.

[6] A. M. Elmassry, N. Zaki, N. Alsheikh, and M. Mediani, "A Systematic Review of Pretrained Models in Automated Essay Scoring," *IEEE Access*, vol. 13, pp. 121902–121917, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3584784.

[7] W. Xu, R. Mahmud, and W. L. Hoo, "A Systematic Literature Review: Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios?," *IEEE Access*, vol. 12, pp. 77639–77657, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3399163.

[8] S. Gombert *et al.*, "From the Automated Assessment of Student Essay Content to Highly Informative Feedback: a Case Study," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 4, pp. 1378–1416, 2024, doi: https://doi.org/10.1007/s40593-023-00387-6.

[9] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring and feedback generation: H. Misgna et al.," *Artif. Intell. Rev.*, vol. 58, no. 2, p. 36, 2024, doi: https://doi.org/10.1007/s10462-024-11017-5.

[10] C. Myneni, V. M. Devadiga, V. K. K, N. S., and N. Z., "EduCheck: An AI-Driven Framework for Multi-Modal Assessment and Performance Analysis," in *2025 3rd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIHEI)*, 2025, pp. 1–6. doi: https://doi.org/10.1109/IDICAIHEI65991.2025.11378451.

[11] D. Wang and J. Wang, "The impact mechanism of AES on improving English writing achievement," *Sci. Rep.*, vol. 15, no. 1, p. 3928, 2025, doi: https://doi.org/10.1038/s41598-025-88021-4.

[12] F. Li, X. Xi, Z. Cui, D. Li, and W. Zeng, "Automatic Essay Scoring Method Based on Multi-Scale Features," 2023. doi: https://doi.org/10.3390/app13116775.

[13] R. H. Chassab, L. Q. Zakaria, and S. Tiun, "An optimized BERT fine-tuned model using an artificial bee colony algorithm for automatic essay score prediction," *PeerJ Comput. Sci.*, vol. 10, p. e2191, 2024, doi: https://doi.org/10.7717/peerj-cs.2191.

[14] E. Fianu, F. Amankwah-Sarfo, P. Ofori, J. K. Amoako, and H. Sumani, "From Traditional Machine Learning Models to Large Language Models: A Systematic Literature Review of Automated Essay Scoring," *SN Comput. Sci.*, vol. 7, no. 5, p. 406, 2026, doi: https://doi.org/10.1007/s42979-026-05028-y.

[15] T. Miao and D. Xu, "KWM-B: Key-information weighting methods at multiple scale for automated essay scoring with BERT," *Electronics (Basel).*, vol. 14, no. 1, p. 155, 2025, doi: https://doi.org/10.3390/electronics14010155.

[16] J. Xue, X. Tang, and L. Zheng, "A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring," *IEEE Access*, vol. 9, pp. 125403–125415, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3110683.

[17] N. Rasdianah, M. Madania, T. S. Tuloli, W. S. Abdulkadir, H. Ahmad, and T. B. A. Suwandi, "Studi Efek Samping Obat Antituberkulosis (OAT) Pada Pasien TB Paru," *Journal Syifa Sciences and Clinical Research*, vol. 4, no. 3, 2022, doi: https://doi.org/10.37311/jsscr.v4i3.16657.

[18] A. Abdulkarem and A. Krivtsun, "An analysis of automated essay scoring frameworks," *Iraqi Journal for Applied Science*, vol. 1, no. 3, pp. 71–81, 2024, doi: https://doi.org/10.69923/av1gt264.

[19] C. T. Lim, C. H. Bong, W. S. Wong, and N. K. Lee, "A comprehensive review of automated essay scoring (AES) research and development," *Pertanika J. Sci. Technol.*, vol. 29, no. 3, pp. 1875–1899, 2021, doi: https://doi.org/10.47836/pjst.29.3.27.

[20] Y. Nie, "Automated essay scoring with SBERT embeddings and LSTM-Attention networks," *PeerJ Comput. Sci.*, vol. 11, p. e2634, 2025, doi: https://doi.org/10.7717/peerj-cs.2634.

[21] S. Riyanto *et al.*, "Employing SBERT for Essay Assessment," in *2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2024, pp. 255–260. doi: https://doi.org/10.1109/IC3INA64086.2024.10732430.

[22] W.-C. Chang, F. X. Yu, Y.-W. Chang, Y. Yang, and S. Kumar, "Pre-training tasks for embedding-based large-scale retrieval," *arXiv preprint arXiv:2002.03932, 2020, doi: https://doi.org/10.48550/arXiv.2002.03932*.

[23] S. Gombert *et al.*, "From the automated assessment of student essay content to highly informative feedback: A case study," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 4, pp. 1378–1416, 2024, doi: https://doi.org/10.1007/s40593-023-00387-6.

[24] U. Majeed, M. Jaffery, A. Nusrat, and S. R. Tahir, "Investigating The Effect Of Automated Writing Evaluation Through Deep Neural Network And Teachers' Written Evaluation On English Writing Performance," *J. Vis. Exp.*, no. 231, p. e69995, 2026, doi: https://doi.org/10.3791/69995.

[25] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020, doi: https://doi.org/10.48550/arXiv.1910.10683.

[26] M. S. Gunawan and S. Karina, "Human Rights and Constitutional Supremacy in Indonesia: Normative Guarantees and Policy Disharmony in Natural Resource Management, Freedom of Expression, and Public Participation," *Journal of Law and Social Politics*, vol. 4, no. 2, pp. 112–121, 2026, doi: https://doi.org/10.59261/jlsp.v4i2.106.