Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15483
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Evaluation of Machine Learning Algorithm for Automatic Grading of School Students' English Essay

Andi Nurfadillah Ali[1)*], Muhaimin Hading[2)], Andi Sahra Suryabuana[3)]
[1),3)] Information System Department, Bacharuddin Jusuf Habibie Institute of Technology, Parepare, Indonesia
[2)] Robotica AI Department, Bacharuddin Jusuf Habibie Institute of Technology, Parepare, Indonesia
[1)]anurfadillah@ith.ac.id, [2)]hading.muhaimin@ith.ac.id, [3)]andisahrasuryabuana@gmail.com

Submitted :Oct 26, 2025 | **Accepted** : Nov 24, 2025 | **Published** : Jan 02, 2026

Abstract: The manual assessment of essays in English language learning often faces challenges related to objectivity and efficiency, especially on a large scale. With advancements in artificial intelligence technology, machine learning-based approaches have begun to be adopted to automate this process through Automated Essay Scoring (AES) systems. However, most existing AES models tend to rely solely on the final scores from the dataset without considering the structural quality of the writing, such as coherence between paragraphs. This study aims to evaluate the effectiveness of machine learning algorithms in assessing school students' essays by adding coherence features as predictor variables in a regression model. This approach uses linguistic feature representation techniques to explicitly build coherence indicators. The proposed model achieved a QWK improvement from 0.69 to 0.89 using SMOTE and coherence features. Meanwhile, human evaluation results showed that the pair of Rater 1 and Rater 2 achieved a QWK of 0.82, the pair of Rater 1 and Rater 3 scored 0.79, and the pair of Rater 2 and Rater 3 scored 0.81. These values indicate a high level of agreement among raters, suggesting that the assessment instrument used is stable. The main contribution of this study is introducing the coherence feature as an explicit predictor in the AES model, filling the gap not provided by standard datasets and proving that coherence improves model accuracy. This research provides practical benefits such as speeding up the evaluation process, reducing teachers' workload, and improving the objectivity and consistency of assessment in language education and evaluation.

Keywords: Assessment; Artificial Intelligence; Automated Essay Scoring (AES); QWK; SMOTE

INTRODUCTION

Assessment plays a crucial role in evaluating students' understanding and competence, and technological advances have transformed how this process is conducted. Learning evaluation or assessment has a strategic role that can measure the extent to which student understanding and competence develop. Essay assessment has its own advantages compared to other forms of assessment such as multiple choice or short answer.

Essay assessment allows students to express their thoughts in depth and demonstrate their analytical abilities and writing skills. Through essays, educator , can assess critical thinking, conceptual understanding, and the ability to organize and express ideas logically. It can assess critical thinking skills, concept understanding, and the ability to organize and convey ideas clearly and logically. Therefore, learning evaluation with essay questions is crucial in providing a more comprehensive picture of students' competencies. Essay assessment requires in-depth analysis of the written content and is often a challenge for educators as it takes time and effort to give fair scores without being influenced by subjectivity. To overcome this challenge, research in the field of natural language processing (NLP) with machine learning algorithms has led to the development of automated scoring systems that can provide fast and consistent assessment of student essays (Radiatul Kamila and Budiyanto, 2025).

The ability to write essays in English is one of the important indicators in assessing language competence, especially in academic contexts and evaluating critical thinking skills. With the development of natural language processing (NLP) and machine learning technologies, Automated Essay Scoring (AES) approaches have been

*Andi Nurfadillah Ali

e-ISSN : 2541-2019
p-ISSN : 2541-044X

widely used to automatically evaluate essays. AES utilizes linguistic features such as syntactic complexity, vocabulary variety, and semantic structure to build predictive models that generate essay scores (Wang J, 2025).

Most previous studies have shown that AES systems can achieve a high level of accuracy in predicting overall essay scores. However, a major challenge remains in measuring more complex aspects of writing quality, such as coherence between sentences and paragraphs, which substantially affects the readability and flow of arguments in essays. This research does not focus on the development of new applications or systems but instead is directed at testing and evaluating machine learning algorithms in improving the accuracy of essay score prediction based on text coherence representation. Given the limitations in terms of time and access to collect student essay data directly, this research utilizes the Automated Student Assessment Prize (ASAP) dataset that has been widely used in previous studies. This dataset provides a collection of essays that have been scored by human raters, making it possible to experiment with supervised learning approaches.

Coherence is an important feature because it determines the unity of ideas in an essay—how sentences and paragraphs are logically connected. Human raters rely heavily on coherence when assessing writing quality, as it affects the clarity of the argument, flow, and overall comprehension. AES models without coherence only assess the surface, such as text length or vocabulary, and thus have the potential to give high scores to structurally weak essays. By incorporating coherence, the assessment becomes more similar to human evaluation, more comprehensive, and more accurate. The coherence feature is underutilized because AES datasets generally only provide a final score, without aspect scores such as coherence, organization, or argumentation, making it difficult for models to learn these features directly. Coherence is abstract and cannot be captured by simple features (n-grams, text length), requiring more complex NLP techniques.

Previous studies on Automated Essay Scoring (AES) generally focus on surface-level or content-based features, while failing to measure discourse-level coherence explicitly as part of the scoring mechanism. As a result, regression-based AES models rarely integrate coherence representations, leaving a methodological gap in how textual organization contributes to score prediction. To address this limitation, this study introduces explicitly engineered coherence features as predictive variables within a regression framework. Additionally, the model's performance is compared across original and SMOTE-balanced datasets, presenting a novel evaluation perspective in AES research.

LITERATURE REVIEW

Essay Scoring & Educational Assessment

Education is a factor that influences humans to be able to process and interact with their surroundings and is a fundamental matter that needs to be considered because education can shape a person's character. Learning evaluation is conducted as a measure of students' understanding of the material provided by teachers, allowing teachers to assess the abilities of each student. Exams themselves are divided into two different types, namely objective exams with multiple-choice answers and subjective exams with essay answers. Essay exams are often a challenge for teachers and assessors in identifying and giving objective scores to each student's answers (Mubarok, M. I, et.al. 2023).

Teachers' competencies are dynamic and evolve in line with the times, so assessments should also adapt to the evolving competencies required or educational standards set. Changes that can be made to teacher assessments include a framework that serves as the basis for developing instruments that are in line with the times, and the use of technology in the evaluation system (Permana, et.al. 2021).

Automated Essay Scoring

Automated essay scoring has been a growing topic in the field of Natural Language Processing (NLP) with various approaches developed to predict text scores. Approaches with Recurrent Neural Network (RNN)-based models are used to capture semantic relationships between text sections and patterns of essay structure (Wang J, 2025; Elks Tim, 2021) while Convolutional Neural Networks (CNN) are utilized to extract sentence coherence in paragraphs (Muangkammuen P, et al., 2020). More recent approaches use Transformer based models (CAvva Reddy RK, et al., 2024; Ludwig S, et al., 2021) and Large Language Models (LLMs) (Dini L, et al., 2025).

A commonly used data in automated essay grading is the Automated Student Assessment Prize (ASAP). Based on this data, various studies have compared the state-of-the-art performance to evaluate the effectiveness of the algorithms used. For example, compared a hybrid approach with basic models such as LSTM and BERT. The results show that combining handcrafted features such as essay length and spelling errors with deep neural networks (DNNs) can improve model performance, with the Quadratic Weighted Kappa (QWK) value of the hybrid approach being 5% higher than LSTM and 1% higher than BERT. The approach proposed in (Wang Y, et al., 2022) represents the BERT model with a joint learning scheme to extract essay features at various levels of granularity, namely token, sentence, and paragraph levels. The model is then compared with ten other deep learning models. The evaluation results show that the modified BERT with joint learning can achieve an accuracy

*Andi Nurfadillah Ali

e-ISSN : 2541-2019
p-ISSN : 2541-044X

of 77% for essays with a length of more than 510 words and increases to 78% for essays with a length of less than 510 words. The Neural Pairwise Contrastive Regression (NPCR) approach is a model that combines regression objectives by utilizing contrastive learning to maximize the margin between essays with different scores and minimize score prediction errors (Xie J, et al., 2022). This model is able to achieve 81% accuracy and shows better performance compared to several other models such as CNN+LSTM, SKIPFLOW and R2BERT (Dini L, et al., 2025). Although previous models have widely utilized Large Language Model (LLM)-based approaches and Transformer architectures in essay assessment, they have less emphasis on specific linguistic features such as text coherence.

A sentence-level embedding approach designed to encode textual coherence and cohesion into vector representations (Ramesh et al., 2024). These vectors were then used to train LSTM and Bi-LSTM models to identify both internal and external sentence connections, including sentence structuring, relationships among sentences, and relevance to the given prompt. After training, the model was evaluated using various adversarial responses and produced satisfactory performance.

The normalization process is intended to ensure that the values produced are significantly consistent with the values assigned by human graders. For each student's answer, the system will obtain Cosine Similarity for the number of questions. While deep models such as BERT and NPCR learn implicit coherence patterns, none of the reviewed methods incorporate explicit discourse-level coherence as a predictive variable. This creates a methodological blind spot that the present study aims to fill.

Artificial Intelligence in Education

Creativity is one of the important aspects in education that can help learners to develop new ideas, explore ideas, and produce innovative solutions. Thus, it is hoped that the existence of artificial intelligence will make students as the next generation of the nation who are always creative and make learning fun. Education in the current era of digitalization, artificial intelligence (AI) is one of the innovations that has the potential to make students have 21st century skills, one of which is that students must have creative skills, with these skills it is hoped that students can develop in the learning process, because learning through an independent curriculum is tailored to the needs, interests and learning styles of students (Saputra, Rahmat, and Komalasari, 2024). The application of artificial intelligence in education also enables an adaptive learning approach. AI can identify students' level of understanding in real time and adjust learning content according to individual needs. This allows students to develop their maximum potential and overcome difficulties more effectively (Cahyadi et al., 2025).

Machine Learning in Education

The application of Machine Learning in education offers great potential to enhance learning effectiveness and help overcome various challenges faced by both students and teachers. With the support of this technology, students' learning experiences can be improved through personalized learning, individual assistance, and more efficient classroom management (Nurul et al., 2024). Machine learning can be organized that computer applications and algorithms receive math by learning from data and making predictions of future data. Machine learning, a subset of artificial intelligence, focuses on algorithms that learn from data to make predictions or decisions without being explicitly programmed.

METHOD

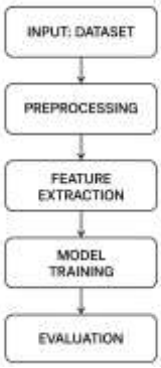

Fig 1. Research flow

The research flow in figure 1, began with the collection of an essay dataset containing texts and final scores from human raters. The data then underwent preprocessing, including text cleaning and normalization, to ensure

*Andi Nurfadillah Ali

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15483
e-ISSN : 2541-2019
p-ISSN : 2541-044X

input consistency. Next, feature extraction was performed, covering general linguistic features as well as coherence features that were specifically constructed as the main contribution of the research.

These features were used in training a regression model to predict essay scores, with SMOTE applied to address the imbalance in the score distribution. The final stage was model evaluation using the QWK metric to assess its alignment with human assessment. This workflow ensured that the resulting model was not only accurate but also capable of capturing important structural aspects of essay writing.

Research Stages

1. Pre-processing Dataset
   Dataset Pre-processing The data used is Automated Student Assessment Prize (ASAP) 2.0. The ASAP 2.0 dataset includes approximately 24,000 argumentative essays in English written by students that are tailored to the assessment standards appropriate for today's students.
   The dataset also includes samples from different economic backgrounds and geographical regions to reduce potential algorithm bias.

2. Prediction Score

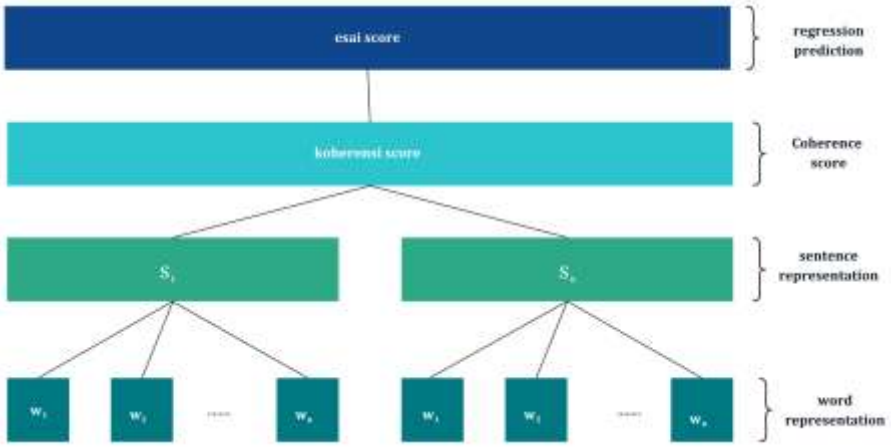

Fig. 2 Prediction score Development

This research will use coherence score as a feature that will be used in predicting the score of an essay. The approach that will be used to perform essay score prediction is multiclass logistic regression. This method was chosen because it is suitable for classification cases where the target variable (essay score) consists of six score levels (scale 1-6) according to the labels in the available training data. This approach allows the system to estimate the probability of each score level, and the level with the highest probability will be selected as the final prediction result. Mathematically, multiclass logistic regression uses a softmax function to calculate the probability of an input x belonging to class j, with the following formula:

$$P(y = j \mid x) = \frac{exp(z_j)}{\sum_{i=1}^{m} exp(z_i)} \quad (1)$$

Where:

$$x \quad : input\ vector$$
$$z_j \quad : regression\ coefficient\ vector\ for\ class\ j$$
$$m \quad : number\ of\ classes$$
$$P(y = j \mid x) : probability\ of\ essay\ falling\ into\ class\ j$$

The model will predict the class with the highest probability as the final label for an essay:

$$\hat{y} = argmax\ P(y = j \mid x) \quad (2)$$

This model was selected because it is computationally efficient, fast to train, and highly interpretable, making it suitable for an initial study focusing on coherence feature evaluation. In contrast, deep learning models require substantially higher computational resources and training time, and offer lower interpretability, making them less aligned with the objectives of this research.

*Andi Nurfadillah Ali

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15483
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

Pre Processing Dataset

The data used in this study is ASAP 2.0, which is a standard and commonly used dataset in Automated Essay Scoring (AES) research. This dataset consists of approximately 24,728 essays written by 6th to 10th grade students with non-native English speaker backgrounds, covering 7 different topics (prompts) with a score range from 1 to 6. A score of 1 indicates the lowest quality of the essay, while a score of 6 indicates an excellent quality of the essay, both in terms of the writer's understanding of the assignment, the suitability of the information with the reading source, and the use of appropriate sentences and words. The following is the number of essays from each topic in the ASAP 2.0 dataset. The distribution of the dataset can be seen in Table 1 and Fig. 3.

Prior to model training, this dataset goes through a series of pre-processing stages to ensure data quality and reduce noise in the text. The pre-processing stages include:

a. Word cleaning, such as correcting typos and removing irrelevant words.
b. Lowercasing, which is converting all letters into lowercase letters to maintain word consistency.
c. Removing non-alphabetic characters, such as numbers, unnecessary punctuation marks, or special symbols.
d. Tokenization, which breaks sentences into word tokens to simplify the analysis process.
e. Stop words removal, by removing common words that do not provide significant meaning to the essay score, such as is, the, or and.

In its implementation, this research uses several Python libraries such as NLTK, spaCy, and re (regular expression) for text cleaning. The results of this process will be used as the basis for extracting features related to writing quality, which are then used in the score prediction process.

Table 1. Essay Distribution by Topic

Topic	Total Essay
A Cowboy Who Rode the Waves	2175
Car-free cities	1959
Does the electoral college work?	2046
Driverless cars	6170
Exploring Venus	4480
Facial action coding system	4883
The Face on Mars	3015
Total	**24728**

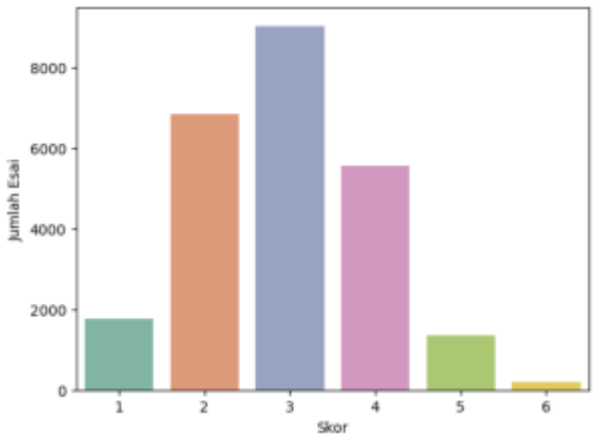

Fig. 3 Essay Distribution against Score

Feature Extraction

a. Word Representation

*Andi Nurfadillah Ali

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15483
e-ISSN : 2541-2019
p-ISSN : 2541-044X

This research takes several features from essays on the ASAP 2.0 dataset to be used as input to the score prediction model. One of the feature categories used is word representation and several linguistic features, which are adapted from the research of Uto et al. (2020).

Fitur Word Embeddings

An essay (E) to$j$ that has been processed into a sequence of tokens:

$$E_j = (w_{j1}, w_{j2}, \ldots, w_{jn}) \quad (3)$$

Where each essay is converted into a sequence of word tokens, which are then converted into numerical vectors so that they can be analyzed by machine learning models or artificial intelligence systems.

Each token $w\ jn$ has a Word2Vec embedding of dimension d :

$$w_{jn} \in \mathbb{R}^d \quad (4)$$

The goal is to build a vector with essay features $w_{jn} \in \mathbb{R}^d$ from word embedding which will later be combined with other features. The next step is to combine the vectors of each word (Word2Vec) in the essay into a single vector so that the entire essay can be represented numerically for modeling. into a single vector so that the entire essay can be represented numerically for modeling. The approach used is to average the vectors by considering the word weights. If a token appears more frequently, it will be considered more representative. Token frequency w in the essay

$$w_{jw} = tf_{jw} \quad (4)$$

then the average vector with frequency weights can be represented by :

$$e_j = \frac{\sum_{t=1}^{n_j} w_{jw} w_{jn}}{\sum_{t=1}^{n_j} w_{jw}} \quad (6)$$

Lingustik Fiture of Uto et al (2020)

- Total word count (wc_j)
- Number of sentences (sc_j)
- Average sentence $(aveS_j)$
- Number of verbs $(Noun_j)$
- Number of adjectives (Adj_j)
- Grammar Error (GE_j)

Furthermore, these linguistic features are stored in a vector$u$ whose dimension is p

$$u_j = [wc_j, sc_j, aveS_j, Noun_j, Adj_j, GE_j] \in \mathbb{R}^p \quad (7)$$

Then we have the word representation feature:

- Average of frequency weighted vectors (e_j)
- Linguistic feature $(u_j \in \mathbb{R}^p)$

b. Sentence Representation

An essay (E) consists of$n$ sentences:

$$E = (S_1, S_2, \ldots, S_n) \quad (8)$$

Each sentence Si is represented with an S-BERT embedding vector:

$$e_i = SBERT(S_i), e_i \in \mathbb{R}^c \quad (9)$$

Where c is the embedding dimension. Coherence between sentences is calculated using cosine similarity between neighboring sentences which can be represented by:

$$c_i = cos \frac{(e_i \cdot e_{i+1})}{|e_i|\ |e_{i+1}|}, i = 1, \ldots, n-1 \quad (10)$$

Then after getting the coherence score between sentences, the average coherence in the essay is calculated with :

$$C(E) = \frac{1}{n-1} \sum_{i=1}^{n-1} c_i \quad (12)$$

*Andi Nurfadillah Ali

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15483
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Then the coherence feature $C(E)$ as sentence representation will be used in training data along with word representation feature and Uto, et.al (2020) feature.

c. Feature Merging
Each essay will have features will be stored in x and the correspondence with each y_j

$$y_j\ , x_j = \left[e_j\ , u_j\ , C(E)\right]^T\ \epsilon\ \mathbb{R}^{1+p+1}\ , \forall E \quad (12)$$

Train Model

Balancing the dataset

The distribution of essays in each score class shows significant imbalance, as shown in Fig. 2. For example, the amount of data in class 1, 5, and 6 are relatively less than the other classes, with the initial distribution: class 1 is 1,751, class 2 is 6,847, class 3 is 9,021, class 4 is 5,553, class 5 is 1,356, and class 6 is only 200 data. To address this imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE) [20] which generates synthetic data on minority classes. After the application of SMOTE, the data distribution becomes more balanced, namely: class 1 is 5,782, class 2 is 6,847, class 3 is 9,021, class 4 is 5,535, class 5 is 6,623, and class 6 is 7,367. The data distribution graph can be seen in Fig. 4.

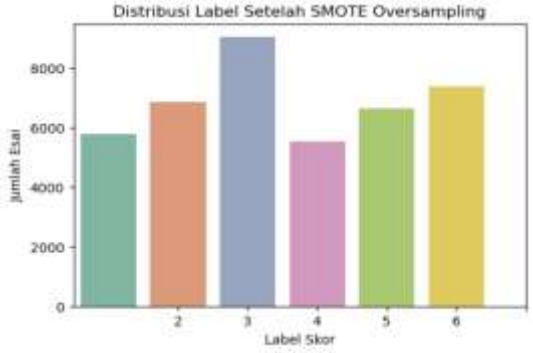

Fig. 4 Distribution of Essay Score

In this experiment, researchers conducted two test scenarios, namely using original data and data that has been processed with SMOTE. Both experiments received the same treatment, where the data was divided into 75% for training and 25% for testing. In addition, the features used remained the same, and the model applied to the training process was logistic regression.

Logistic Regression Prediction

These features are used to calculate the probability of each essay $P(y_j)$ for each class k against the given features x_j and parameter $\theta_k\ \epsilon\ \mathbb{R}^{1+p+1}$

$$P(y_j = k \mid x_j) = \frac{exp\,(\theta_k^T\ x_j)}{\sum_{i=1}^{m}\ exp\,(\theta_k^T\ x_j)}\ , k\ =\ 1, \ldots, K \quad (13)$$

These probabilities are then used to predict the class$k$ with

$$\hat{y}_j = arg_{k\epsilon\{1,\ldots,K\}}\, P(y_j = k \mid x_j) \quad (14)$$

Evaluation and Results

This study uses two types of evaluation: machine evaluation and human evaluation. Below is an explanation of each type of evaluation used.

1. **Machine Evaluation**

To assess the quality of the model, this study uses Quadratic Weighted Kappa (QWK). QWK is appropriate for tasks involving ordinal classification, where labels represent categories in a specific order. This metric considers the degree of discrepancy between predicted and actual scores by penalizing larger discrepancies. For example, a prediction error from a score of 6 to 5 would be considered less severe than a prediction error from a score of 6 to 1. QWK is formulated as:

$$QWK\ =\ 1\ -\ \frac{\Sigma_{i,j}\, W_{i,j}\ O_{i,j}}{\Sigma_{i,j}\, W_{i,j}\ E_{i,j}} \quad (15)$$

with O_{ij} is the observation matrix, $E_{ij}$is the expectation matrix, and $W_{ij} = \frac{(i-j)^2}{(M-1)^2}$ is the quadratic weight.

*Andi Nurfadillah Ali

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15483
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In addition, to ensure more reliable model evaluation, this study uses the k-fold cross validation technique. In this method, the dataset is divided into$k$ subsets (folds), i.e. at each iteration one-fold is used as test data and 1-fold is used as test data. test data and $k-1$ other fold is used as training data. This process is repeated k times so that each fold acts as test data once. The QWK value is calculated at each iteration, and the average of all QWKs becomes a more stable measure of model performance and is not biased towards a particular data division. The machine evaluation results show that the QWK value on the original test data is 0.69, while that on the test data with the application of SMOTE reaches 0.89. This improvement is due to the use of SMOTE (Synthetic Minority Over-sampling Technique) which helps balance the class distribution. in the dataset. In the original data, the distribution of essay scores tends to be unbalanced, with some score categories having far fewer samples than others. This imbalance makes it more difficult for the model to learn patterns in categories with a limited amount of data, resulting in low prediction performance in minority categories. With SMOTE, synthetic data for minority classes is generated adaptively so that the model obtains a better representation for all classes. As a result, the model is better able to classify across the entire range of scores, resulting in an improved fit of predictions to actual scores as reflected in the increase in QWK scores.

The increase in the QWK value indicates that the model is increasingly consistent with human rater assessment patterns. In practical terms, this signifies an increase in the reliability of the model, as the score predictions are not only numerically accurate but also follow the evaluation logic used in authentic assessment. Higher reliability means that the model is capable of providing stable scores across a variety of essay types, including scores that were previously underrepresented in the data.

2. Human Evaluation

The manual evaluation in this study involved three raters with relevant backgrounds as detailed in Table 2. The dataset consists of approximately 24,728 essays. Each rater holistically scored the same set of 100 student essays, providing a single overall score reflecting the general quality of the writing (integrating reading coherence, grammatical accuracy, and vocabulary variation into one comprehensive judgment). Each evaluation was conducted independently without discussion among raters.

Table 2. Human Evaluation Team

No	Nama Rater	Highest Level of Education	Occupation
1.	Wiwit Melayu	Master's Degree - English Education	Teacher at English Corner
2.	Enggar Pangesti Wibowo	Master's Degree - English Education	Teacher at DHIS Secondary Islamic Boarding School
3.	Andini Safitri	Bachelor's Degree - English Education	Teacher at SMPN 03 Satap Bulutaba

To obtain the human reference score (gold score), the researcher calculated the average of the three raters' scores for each essay, then rounded it to the nearest category on the scoring scale. If s_i^r represents the score given by the r-th rater to the i-th essay, with r = 1,2,3 the human average score $\underline{s_i}$ can be expressed as:

$$\underline{s_i} = round\left(\frac{s_i^1 + s_i^2 + s_i^3}{3}\right) \quad (16)$$

The results showed an accuracy of ±1 point tolerance of 0.82 for the model with SMOTE, while the model without SMOTE only achieved 0.67 accuracy within the same tolerance. Thus, applying SMOTE improves the model's prediction quality to better align with human scoring patterns.

Next, to assess inter-rater consistency, pairwise Quadratic Weighted Kappa (QWK) was calculated for each pair of raters. The results were as follows: Rater 1 and Rater 2 achieved a QWK of 0.82; Rater 1 and Rater 3 scored 0.79; and Rater 2 and Rater 3 scored 0.81. These values indicate a high level of agreement among raters, suggesting the stability of the assessment instrument used.

*Andi Nurfadillah Ali

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15483
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In general, the machine and human evaluation results are summarized in Table 3.

Table 3. Machine and Human Evaluation Results

Evaluation Type	Description	Result
Machine Evaluation	Summary of model performance using test label data	Without SMOTE = 0.69 With SMOTE = 0.89
Prediction Against Gold Score	Summary of model performance against gold score	With SMOTE: Accuracy ±1 point = 0.82 Without SMOTE: Accuracy ±1 point = 0.67
Pairwise QWK Between Raters	Consistency level among human raters on 100 essays	Rater 1 vs Rater 2 = 0,82 Rater 1 vs Rater 3 = 0,79 Rater 2 vs Rater 3 = 0,81

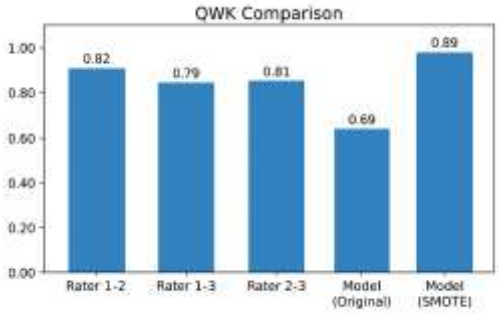

Fig 4. QWK Comparison

The figure 4 shows a comparison of Quadratic Weighted Kappa (QWK) values between human rater consistency and automatic model performance. It can be seen that the human rater pairs (Rater 1–2, Rater 1–3, and Rater 2–3) have a high level of agreement with QWK values ranging from 0.79 to 0.82. The machine learning model shows a QWK value of 0.69 on the original data, but increases sharply to 0.89 after applying SMOTE. This increase shows that equalizing the score distribution helps the model capture assessment patterns more accurately, thereby approaching—and even surpassing—the consistency between human assessors.

DISCUSSIONS

This study proposes another alternative approach by building a representation of the coherence features as predictor variables in a logistic regression model. Logistic regression was chosen because it is easy to interpret and effective in classifying classes into certain classifications. This study demonstrates that adding coherence features as predictor variables in a machine learning model can significantly improve the accuracy of essay score predictions compared to models that use only conventional features. Evaluation results show that integrating coherence features yields higher QWK values, both in machine evaluation and in agreement among human raters, indicating the validity and stability of the assessment instrument used.

The results show that adding coherence features makes the model more consistent with human assessment patterns, as indicated by an increase in QWK that approaches inter-rater consistency. This is consistent with the literature, which confirms that discursive features improve the model's ability to capture argument organization. The use of SMOTE also plays a role in improving generalization in unbalanced score classes, although the risk of overfitting to synthetic patterns needs to be considered. Methodologically and practically, these findings open opportunities for further research, including combining various types of coherence features (entity-based, discursive relations, neural coherence), developing datasets with aspect labels, and evaluating the model's robustness and fairness towards authors with different linguistic backgrounds. An interpretability approach is also

*Andi Nurfadillah Ali

important to ensure that AES systems are not only statistically accurate but also accountable in an educational context.

The improvement in QWK demonstrates that explicit coherence representation bridges the gap between surface linguistic features and discourse-level evaluation, aligning model predictions more closely with human scoring behaviour. This finding suggests that incorporating coherence signals enables the model to capture deeper argumentative structure and logical flow, which are key elements valued by human raters. Consequently, coherence-aware AES models can provide more reliable and human-like assessments across diverse essay qualities.

However, this study has some limitations, such as the use of a school-level English essay dataset that may not represent a wide variety of writing styles or difficulty levels. Therefore, future research should expand the data coverage and consider other structural features that can enrich the automated scoring model.

The benefit of this research lies in the application of a more objective and efficient Automated Essay Scoring system, especially in the context of large-scale English language learning. This system has the potential to help teachers provide consistent assessments while saving time. Based on the results obtained, it is recommended that the development of the next AES models combine various linguistic and structural features and conduct testing on more diverse datasets to improve the model's generalization and accuracy.

CONCLUSION

This study shows that essay scoring models can be improved through an approach that considers structural aspects of writing, particularly coherence. By applying coherence feature representations to regression models, automated scoring systems are able to generate more stable score predictions that closely resemble human scoring patterns. These findings confirm that the development of AES does not only depend on improving metric performance, but also on selecting features that reflect the authentic evaluation process in writing assessment.

In addition to providing an empirical basis for the application of coherence features, this research contributes to the development of language evaluation technology that is more transparent and pedagogically oriented. The results obtained open up opportunities to expand the design of AES that is not only accurate but also capable of providing structural feedback to language learners. Thus, this research is a first step towards a more mature integration between NLP and educational assessment in the Indonesian context.

The theoretical contribution of this research is the introduction of explicit coherence representation in regression-based AES. The practical contribution is increased reliability for teachers and the implementation of automatic assessment. Future research includes transformer-based coherence modeling for discourse processing and multilingual AES models specific to prompts.

ACKNOWLEDGMENT

This research is fully supported by the internal grant from the Novice Lecturer Research (PDP) program, funded by LPPM-PM Institut Teknologi Bacharuddin Jusuf Habibie, under the Decision of the Rector of Institut Teknologi Bacharuddin Jusuf Habibie, Number: 050/IT13.A.SK/SU.01.00/2025, regarding the Determination of Recipients of Internal Research Funding for the Year 2025. The authors would also like to express their deepest gratitude to all parties who have provided support and contributed to the smooth execution of this research. Their invaluable assistance and guidance have been crucial to the successful completion of this work.

REFERENCES

Arifuddin, M. R., Rafiq, I. A., Mubarok, R., & Susilo, P. H. (2023.). Sistem Cerdas Penilaian Ujian Essay Menggunakan Metode Cosine Similarity. In *Generation Journal* (Vol. 7, Issue 1).

Cahyadi, Purnomo, D., Dewi Sahara Nasution, & Fitri anggraini. (2025). PENILAIAN ESAI MATA KULIAH BAHASA INGGRIS BERBASIS MACHINE LEARNING MENGGUNAKAN ALGORITMA REGRESI LINIER. *INFOTECH Journal*, *11*(1), 68–72. https://doi.org/10.31949/infotech.v11i1.13014

CAvva Reddy RK, et.al. 2024. A *Transformer-Based Approach for Enhancing Automated Essay Scoring*. 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET). DOI: 10.1109/ACET61898.2024.10730000

Dini L, at.al. 2025. TEXT-CAKE: *Challenging Language Models on Local Text Coherence*. Proceedings of the 31st International Conference on Computational Linguistics. 4384-4398. Available at: https://aclanthology.org/2025.coling-main.296/

EYue C, Hanqi J, Xiaojun W, and Zhiwei Y. 2020. *Domain-adaptive neural automated essay scoring*. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, page 1011–1020. Available at: https://dl.acm.org/doi/10.1145/3397271.3401037

*Andi Nurfadillah Ali

lks Tim. 2021. *Using Transfer Learning to Automatically Mark L2 Writing Texts*. Proceedings of the Student Research Workshop Associated with RANLP. Available at: https://aclanthology.org/2021.ranlp-srw.8/

Ludwig S, et.al. 2021. *Automated Essay Scoring Using Transformer Models*. Cornell University. Available at: https://doi.org/10.3390/psych3040056

Muangkammuen P, et.al. 2020. *A Neural Local Coherence Analysis Model for Text Clarity Scoring*. Proceedings of the 28th International Conference on Computational Linguistics. pp 2138-2143. Available at: https://aclanthology.org/2020.coling-main.194/

Mubarok, M. I, et.al. 2023. PENERAPAN ALGORITMA K-NEAREST NEIGHBOR (KNN) DALAM KLASIFIKASI PENILAIAN JAWABAN UJISAN ESAI. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Issue 5).

Nurul Latifatul Inayati, Anisha Nurul Fatimah, Salma Emilia Azzahra, & Imaniar Risty Alamsyah. (2024). Implementasi Tes Essay Dalam Evaluasi Pembelajaran Pendidikan Agama Islam. *Khatulistiwa: Jurnal Pendidikan Dan Sosial Humaniora*, *4*(1), 114–120. https://doi.org/10.55606/khatulistiwa.v4i1.2724

Permana, et.al. 2021. Penggunaan Penskor Jawaban Esai Otomatis dalam Pengukuran Pengetahuan Guru. *Jurnal IPA & Pembelajaran IPA*, *5*(4), 279–292. https://doi.org/10.24815/jipi.v5i4.22724

Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artif. Intell. Rev. **55**(3), 2495–2527 (2021). https://doi.org/10.1007/s10462-021-10068-2

Shen A, et.al. 2021. *Evaluating Document Coherence Modeling.* Transactions of the Association for Computational Linguistics, Volume 9. pp 621-640. Available at: https://aclanthology.org/2021.tacl-1.38/

Wang J, Liu J. 2025. *T-MES: Trait-Aware Mix-of-Experts Representation Learning for Multi-trait Essay Scoring.* Proceedings of the 31st International Conference on Computational Linguistics, pages 1224-1236. Available at: https://aclanthology.org/2025.coling-main.81/

Wang Y, et.al. 2022. *On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation*. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp 3416-3425. Available at: https://aclanthology.org/2022.naacl-main.249/

Xie J, et.al. 2022. *Automated Essay Scoring via Pairwise Contrastive Regression*. In Proceedings of the 29th International Conference on Computational Linguistics. pp 2724-2733. Available at: https://aclanthology.org/2022.coling-1.240/

Yancey PK, et.al. 2023. *Rating Short L2 Essays on the CEFR Scale with GPT-4*. Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). pp 576-584. Available at: https://aclanthology.org/2023.bea-1.49/

*Andi Nurfadillah Ali