Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 20, 2025

AUTOMATED STUDENT ANSWER SCORING USING GLOVE-LSTM AND HYBRID SIMILARITY METRICS

I Gede Susrama Mas Diyasa*	School of Magister Information Technology, University of Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia	igsusrama.if@upnjatim.ac.id
Mohammad Idhom	School of Informatics, University of Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia	idhom@upnjatim.ac.id
Ahmad Sofian Aris Saputra	School of Informatics, University of Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia	ahmadsofianarissaputra@gmail.com
Deshinta Arrova Dewi	INTI International University, Selangor, Selangor, Malaysia	deshinta.ad@newinti.edu.my
Tresna Maulana Fahrudin	Department of Information and Communication Systems, Okayama University, Okayama, Japan	tresnamf@s.okayama-u.ac.jp

* Corresponding author

ABSTRACT

Aim/Purpose	This study aims to develop and evaluate an automated scoring model for Indonesian student answers that enhances objectivity, accuracy, and adaptability – addressing persistent challenges in manual assessment, such as subjectivity, inconsistency, time inefficiency, and the increasing grading workload faced by teachers.
Background	The proposed model combines GloVe word embeddings with a Long Short-Term Memory (LSTM) network, supported by evaluation algorithms including

Accepting Editor Dimitar Grozdanov Christozov | Received: July 10, 2025 | Revised: September 2, September 17, 2025 | Accepted: September 19, 2025.
Cite as: Mas Diyasa, I. G. S., Idhom, M., Saputra, A. S. A., Dewi, D., A., & Fahrudin, T. M. (2025). Automated student answer scoring using GloVe-LSTM and hybrid similarity metrics. *Interdisciplinary Journal of Information, Knowledge, and Management, 20,* Article 31. https://doi.org/10.28945/5629

	ROUGE Score, TF-IDF, and cosine similarity, to form a robust hybrid scoring system.
Methodology	The methodology involves designing the model architecture, assembling a proprietary dataset of 3,420 student answers (processed into 3,152 samples) from online sources, practice books, and public repositories, applying standard NLP preprocessing techniques, and training the model using TensorFlow and Keras. A comparative baseline using a manually implemented LSTM with NumPy was also explored.
Contribution	This research contributes a tailored hybrid model for automated scoring in the Indonesian language, providing a foundational analysis that highlights both the model's potential and its key limitations, thereby informing future system improvements.
Findings	The model achieved a mean absolute error (MAE) of 0.0761 and a Pearson correlation of 0.8429, indicating a strong alignment with manual grading in terms of relative ranking. However, it tends to overestimate scores for low-quality or irrelevant responses. It struggles with the use of synonyms, variations in answer length, and minor linguistic errors.
Recommendations for Practitioners	As a proof-of-concept, the model shows promise as a supportive grading tool that can help reduce teachers' correction workload and provide fairer, faster assessments in digital learning environments.
Recommendations for Researchers	Future research should prioritize expanding the dataset's size and diversity, enhancing architectural components, and integrating more advanced linguistic features. Investigating contextual embeddings, such as BERT, may also address current semantic limitations.
Impact on Society	A reliable automated scoring system could significantly reduce teachers' grading workload, enabling them to dedicate more time to qualitative learning activities and fostering fairer, more efficient assessments.
Future Research	Further efforts should focus on enhancing the model's precision, particularly in identifying and penalizing low-quality answers, through improved hybrid architecture design, rigorous hyperparameter optimization, and the exploration of more sophisticated embedding techniques.

Keywords	automated scoring, GloVe, LSTM, ROUGE score, TF-IDF, cosine similarity, natural language processing, deep learning, automated essay scoring, Indonesian language, education quality, innovation

INTRODUCTION

Assessment is an integral part of the educational ecosystem. Assessment serves the purpose of measuring educational outcomes. There are fundamental issues, including bias, inconsistency, and time-consuming adjudication, that plague the evaluation processes conducted manually by teachers (Misgna et al., 2024). The situation is far more dire where teachers have to tackle several classes, each filled with a good number of students. Automated systems seek to provide relief in such scenarios (Buditjahjanto et al., 2022). In addition, the multilingual background of Indonesian students and their use of synonyms, relative sentence structures, and phrase variation render unbiased adjudication perplexing and decisively complicated (Wibowo et al., 2024).

The recent introduction of machine learning and deep learning technologies into natural language processing (NLP) has fostered an interest in automated essay scoring (AES) systems that evaluate

student responses automatically. Research demonstrates that AES systems based on semantic, thematic, and linguistic representations of content can significantly improve scoring accuracy and even mimic human understanding of the content (Wang, 2024). The most dominant recent design transformations have been brought about by BERT and GPT, which have greatly improved the contextual semantic relationship capturing capabilities as compared to the keyword matching utilized in older algorithms. Their bidirectional context understanding has made them extremely influential on the entire NLP and BERT and GPTGPT-powered tools evolution has transformed the NLP and AES tools landscape (Kim et al., 2022).

However, these models have their shortcomings. As shown by Wangkriangkri et al. (2020), even sophisticated models like BERT are responsive to changes in keywords and synonyms, often yielding inconsistent outputs when students provide answers in paraphrased or alternate sentence structures (Beseiso & Alzahrani, 2020). Further research has shown that transformers still have difficulties with long, unstructured, and diverse phrasings (Nasreen et al., 2024).

In recent years, scholars have tried to bridge these gaps in the Indonesian setting. For example, Sriyanto and Kusrini (2025) proposed a hybrid universal sentence encoder–cosine similarity model that performed well. Aisyah et al. (2025), with their vision-language and large language model hybrid handwriting OCR broad assessment capability, sought to enable assessment in remote rural classrooms. These modern solutions, although inventive and effective, tend to overlook the capacity to deal with the very flexible, synonymy, shallow parsing, and answer semantic depth spanning the Indonesian language.

Addressing these gaps, this study seeks to formulate a hybrid model merging Global Vectors for Word Representation (GloVe) with a Long Short-Term Memory (LSTM) network (Basha et al., 2025), given that GloVe has been shown to capture global semantics in Indonesian applications (Singgalen, 2024) and LSTM excels in modeling sequential context (Dewi et al., 2024). Prior research has shown that CNN-LSTM models capture text patterns effectively, but they tend to struggle with synonymy (Kusumaningrum et al., 2024). To improve robustness, the GloVe-LSTM model is fortified with similarity measures, such as cosine similarity, TF-IDF, and ROUGE Score – metrics designed and used for assessing sentence-level, keyword-weighted, and structural similarity.

To guide this work, the following research questions are posed:

> **RQ1:** To what extent can the hybrid model simulate human scoring of Indonesian student answers?
>
> **RQ2:** How effectively can the model handle linguistic variation, including synonyms, structural differences, and minor errors?
>
> **RQ3:** What is the impact of integrating similarity metrics with the GloVe-LSTM core on scoring accuracy and reliability?

The study leverages a proprietary dataset of 3,420 Indonesian student answers (reduced to 3,152 after preprocessing), evaluated using both quantitative metrics (e.g., MAE, correlation, QWK) and scenario-based testing. By addressing the shortcomings of existing approaches, this research contributes to NLP through the adaptation of hybrid models for a morphologically rich language and to education by offering a proof-of-concept system that can reduce teacher workload and promote fairer, more efficient assessment.

RELATED WORKS

In recent years, research in automated essay scoring (AES) based on natural language processing (NLP) has evolved rapidly, transitioning from rule-based and statistical methods toward machine learning and deep learning architectures (Misgna et al., 2024). These changes demonstrate an initiative to improve scoring precision, uniformity, and scalability across numerous student responses. From

the beginning, systems heavily depended on shallow language patterns and handcrafted features, which, though interpretable, were semantically shallow. Consequently, the field has gradually moved toward developing more advanced models that can learn complex textual representations and capture more intricate meanings beyond mere keyword matching (Dhiman et al., 2024; Misgna et al., 2024).

In the Indonesian context, this evolution has been mirrored by localized efforts to adapt AES methods to linguistic and infrastructural realities. Sriyanto and Kusrini (2025) proposed a hybrid framework combining the universal sentence encoder with cosine similarity, achieving strong performance on LMS-based grading tasks. Meanwhile, Aisyah et al. (2025) explored a multimodal pipeline that integrates handwriting OCR with vision–language models and large language models to support assessments in rural classrooms. These studies highlight the potential of hybrid and multimodal approaches, but also reveal key gaps, particularly in handling synonym variation, morphological richness, and the structural flexibility inherent to the Indonesian language. Addressing these linguistic challenges remains central to developing AES systems that are not only accurate but also contextually robust for diverse Indonesian student responses.

Machine Learning

Algorithms can glean insights from data to render predictions without explicit instructions to do so. This capability of machine learning (ML) makes it a vital subdivision of artificial intelligence (Nahin et al., 2025). In the scope of automated essay scoring (AES), ML techniques were some of the first approaches to evaluate the students' answers based on handcrafted linguistic and statistical features (Misgna et al., 2024). Though these methods were computationally inexpensive and easier to interpret, their use of shallow features stalled generalization in richly morphologically complex languages like Indonesian. ML's ability to analyze education and text (Buditjahjanto et al., 2022; Xu et al., 2025) demonstrates its practical, tempered promise while exposing the difficulty of capturing nuanced semantic meaning. This is why hybrid approaches that integrate ML with deep learning have become popular. They utilize ML components, such as TF-IDF or cosine similarity, alongside neural models (Faseeh et al., 2024).

Deep Learning

Deep learning (DL) builds upon ML by utilizing multi-layered architectures of neural networks that can learn hierarchically from raw text. In automated essay scoring (AES), capturing the linguistic, semantic, and syntactic features of student writing through the application of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), including their variants LSTMs, has become common practice (Abdullah et al., 2024; Kumar & Boulanger, 2020). These models have been applied in AES and show improvement over ML benchmarks. As advanced as these models are, their performance comes at a cost due to the inherent challenges that a model based on deep learning faces, such as the need for significant amounts of data, overfitting, and reliance on surface-level heuristics, like sentence and keyword counting. As discussed by Kusumaningrum et al. (2024), while the CNN-LSTM models do capture the text patterning common in student responses, they do not do well in addressing synonym-based equivalences. These problems are the reason why there exists a demand for hybrid models that combine the strengths of DL and other complementary frameworks to ensure semantic stability.

Natural Language Processing: NLP

NLP techniques form the basis of AES, as they enable computer systems to understand and analyze human language. In the field of education technology, the features of NLP have been utilized for automated text summarization of Indonesian scientific articles (Sihombing et al., 2024) and for distinguishing idiomatic and literal expressions of language through deep learning (Abarna et al., 2022). However, NLP struggles the most when it is applied to students' responses, specifically in Indonesian, due to the pervasive informal and colloquial grammar and the diverse vocabulary used. This variability presents problems for shallow or rule-based NLP systems, which aim to automate scoring,

leading to reliable and unbiased results. Therefore, approaches that combine vector-based embeddings, such as GloVe, with sequence models like LSTM and statistical similarity measures are increasingly viewed as solutions to these language problems.

Word Representation: GloVe

Converting human language into machine-readable formats, word embeddings enable processing by changing words into numerical vectors. Global Vectors for Word Representation (GloVe) is an empirical approach of an unsupervised algorithm to create embeddings using global co-occurrence statistics (Pennington et al., 2014). Unlike other models, which are restricted to local contexts, GloVe builds a word–word co-occurrence matrix for a larger corpus and then factorizes the matrix to create dense vectors that retain both semantic and syntactic relationships. One of the well-known examples is the operation king–man + woman ≈ queen, which demonstrates semantically meaningful operations on GloVe (Pennington et al., 2014).

GloVe's weaknesses stem from areas that require a refined understanding and critical interpretation of content. Its accuracy and efficacy have been tested on various Indonesian NLP applications, where GloVe facilitates semantic comprehension of the language despite its morphological complexity (Abdullah et al., 2024; Singgalen, 2024). This is depicted in Figure 1.

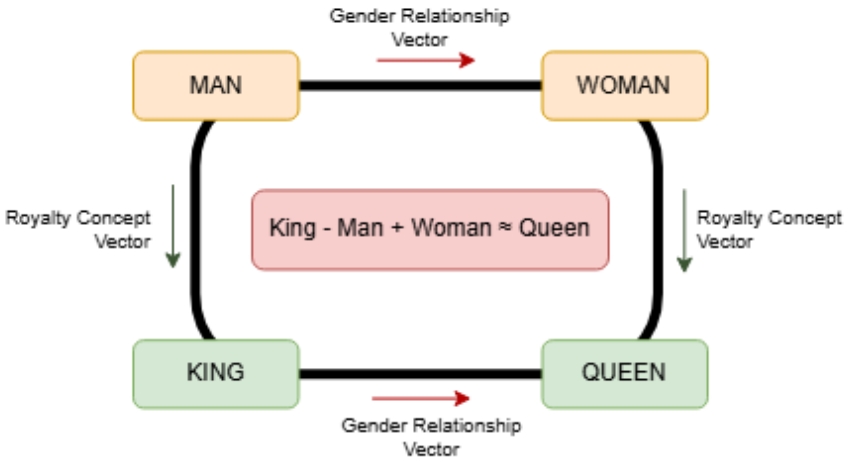

Figure 1. GloVe semantics between words

Regardless of their benefits, GloVe embeddings on their own are not adequate for context capture at the sentence level. Context capture is particularly important for Indonesian AES because the language's use of affixation and flexible word order often changes meaning. Therefore, in this research, GloVe is paired with sequential models in order to express a richer context.

Sequence Modeling: LSTM

Long short-term memory (LSTM) networks are a subtype of recurrent neural network (RNN) with the capability to learn long-term dependencies in sequential data due to the solution it provides to the vanishing gradient problem in traditional RNNs. The main idea of LSTM networks is the memory cell (cell state), which is managed by three gates: the input gate that controls the flow of information into the cell, the forget gate that decides which information to discard, and the output gate that determines the final representation. Text processing, conversation modeling, and other tasks, such as text classification, require LSTM to understand contextual meaning over long sequences, which is due to the gating mechanism.

In the area of AES, LSTM networks have shown considerable success in capturing the progression in students' responses. To illustrate, Kusumaningrum et al. (2024) proposed a CNN-LSTM architecture. Although successful in capturing text patterns, it faced challenges with synonymous expressions, resulting in a divergence from human scoring alignment. In a similar vein, Dewi et al. (2024) also highlighted the considerable strength of LSTM in processing ordered data, further underscoring the importance of LSTM as a primary sequence model. Additional design elements are in Table 1.

Table 1. Architecture of LSTM

Component	Description
Input Gate (i_t)	Controls the amount of new information to store in the cell state.
Forget gate (f_t)	Decides which information to discard from the cell state.
Output Gate (O_t)	Controls which information will be removed from the cell state.
Call State (C_t)	Stores important information over time (long-term memory).
Hidden State (h_t)	Representation of the output of the LSTM cell at time t (short-term memory).
Candidate State ($\tilde{C}_t$)	Candidate values that may be added to the cell state.
Sigmoid Activation (σ)	Activation function that produces a value between 0 and 1 (for gates).
Tanh Activation	Activation function that produces a value between -1 and 1 (for transform).

Although LSTM is powerful in modeling sequential dependencies, its limitations in handling synonym variation and subtle semantic equivalence make it insufficient as a standalone solution (Kusumaningrum et al., 2024). For Indonesian AES, these weaknesses necessitate integration with semantic embeddings and similarity metrics, forming the basis of the hybrid approach proposed in this study.

ROUGE Score

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used for the automatic evaluation of text quality by comparing a candidate text against one or more reference texts. It is commonly used to measure content overlap based on n-grams. For instance, ROUGE-N calculates precision, recall, and F1-score based on the overlap of n-grams between the candidate and reference texts. The formulas for ROUGE-N recall and precision are as follows:

$$ROUGE - N_{recall} = \frac{\sum S \in \{References\} \sum gram_n \in S \, Count_{match} \, (gram_n)}{\sum S \in \{References\} \sum gram_n \in S \, Count \, (gram_n)}$$

$$ROUGE - N_{precision} = \frac{\sum C \in \{Condidate\} \sum gram_n \in C \, Count_{match} \, (gram_n)}{\sum S \in \{Condidate\} \sum gram_n \in C \, Count \, (gram_n)}$$

Where $Count_{match}\,(gram_n)$ is the number of co-occurring n-grams in both the candidate and reference texts, and $Count\,(gram_n)$ is the total number of n-grams in the respective texts. While initially developed for text summarization (Onan & Alhumyani, 2024), its principles are adapted here to assess how well a student's answer captures key phrases from a reference answer (Kim et al., 2022).

Despite its utility in measuring structural overlap (Onan & Alhumyani, 2024), ROUGE remains limited to surface-level phrase matching. It fails to capture semantic meaning, mainly when synonyms or paraphrasing are used (Kim et al., 2022). In Indonesian AES, where linguistic variation is prevalent, ROUGE functions best as a complementary metric rather than a primary scoring mechanism.

TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical method to evaluate the importance of a word within a document relative to a collection of documents (corpus). The TF-IDF score for a term *t* in a document *d* is calculated by multiplying two metrics:

1. Term Frequency (TF) measures how frequently a term appears in a document.

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{.Total\ number\ of\ terms\ in\ document\ d}$$

2. Inverse Document Frequency (IDF) measures how important a term is by diminishing the weight of terms that appear very frequently across all documents.

$$IDF\ (t,D) = \log(\frac{Total\ number\ of\ document\ |D|}{Number\ of\ documents\ contraining\ term\ t\ |\{d \in D: t \in d\}|})$$

The final TF-IDF score is the product of these two values:

$$TF - IDF\ (t,d,D) = TF\ (t,d) x\ IDF\ (t,D)$$

This method forms the basis of information retrieval. It has been successfully implemented in various areas, such as public complaint analysis (Putra et al., 2024) and in automated question-answering systems (Ahmed et al., 2022).

As noted in Putra et al. (2024) and Ahmed et al. (2022), the TF-IDF method successfully determines the relevance of keywords within and across documents. However, it fails to incorporate contextual meaning and synonymy. This is especially challenging in Indonesian, a language where different words often mean the same thing. For this reason, in this research, TF-IDF is incorporated in the hybrid model as a supporting feature and not as the main scorer.

COSINE SIMILARITY

Cosine similarity is a metric that measures the similarity between two non-zero vectors in a multidimensional space. Rather, it measures cosine similarity, which is a measure of the angle between two vectors. In the field of NLP, this is a highly effective method for identifying similarities between two documents or sentences, regardless of the length of the sentences or documents being compared. The formula is:

$$Similarty\ (\vec{A}, \vec{B}) = \cos(0) = \frac{\vec{A} \cdot \vec{B}}{\left|\left|\vec{A}\right|\right| x\ \left|\left|\vec{B}\right|\right|} = \frac{\sum_{i=1}^{n} A_i\ B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where $\vec{A}$ and $\vec{B}$ are the vector representations of the texts (e.g., from TF-IDF or GloVe). A value of 1 means the vectors are identical in orientation, 0 means they are orthogonal (unrelated), and -1 means they are diametrically opposed. It is commonly applied to determine textual relevance in tasks such as automated question-answering (Ahmed et al., 2022). It has been integrated with deep learning models to enhance assessment consistency (Abdullah et al., 2024).

Cosine similarity offers a simple but effective method for determining the semantics between two sentences (Abdullah et al., 2024; Ahmed et al., 2022). However, it does not check for factual or contextual accuracy, which is critical for educational scoring. Thus, in this research, cosine similarity is used as one of the evaluators aimed at improving the accuracy of the scoring system.

Collectively, these approaches illustrate an inherent trade-off: each has its strengths, which, when applied singularly, become weaknesses. GloVe captures semantics but lacks contextual integration. LSTM captures the order of a sequence but does not capture synonym equivalence. ROUGE, TF-

IDF, and Cosine, while useful, emphasise shallow measures of similarity. The impact of these weaknesses is increased in Indonesian AES due to the rich morphology, affixation, and flexible phrasal order. To overcome this, the current study implemented a hybrid GloVe–LSTM model, accompanied by a multi-metric evaluation model, aiming for more accurate, equitable, and flexible automated scoring.

DESIGN AND IMPLEMENTATION

OVERVIEW

This section outlines the research methodology employed to design, develop, and evaluate the GloVe–LSTM model for automated scoring of Indonesian student answers. The study adopts a research and development (R&D) approach, which is suitable for iteratively refining artificial intelligence models in natural language processing (NLP). The methodology is structured into five main stages: dataset collection, data preprocessing, model architecture design, model training, and comprehensive evaluation through multiple testing scenarios.

To provide clarity, a schematic overview of the workflow is presented in Figure 2, which illustrates the end-to-end process from dataset preparation to model evaluation. Each subsequent subsection provides a detailed elaboration on these stages.

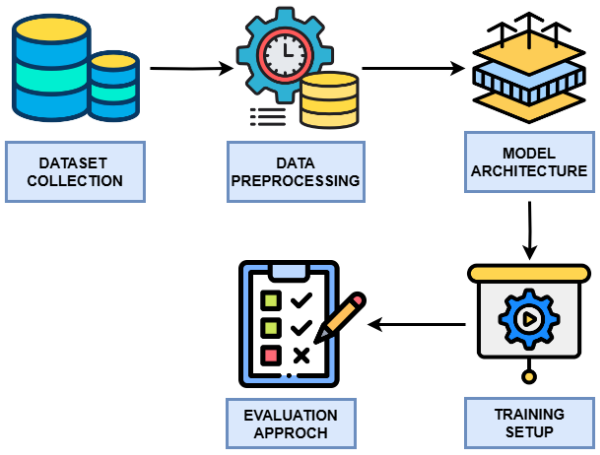

Figure 2. Workmanship flowchart

DATASET COLLECTION AND DESCRIPTION

The researcher independently compiled the dataset to evaluate the model's scoring capabilities across a variety of student responses. Data sources included online repositories, practice question books, and publicly available datasets from the Kaggle platform. The initial dataset contained 1,140 unique questions, each paired with three answer variations: (1) a complete answer, (2) a moderately complete answer, and (3) an incomplete answer. This structure produced a total of 3,420 student answer items before preprocessing.

Each data entry was stored in JSON format with the following fields: "id," "pertanyaan" (question), "kata_kunci" (keywords), "jawaban_siswa" (student answer), and "skor" (score). A representative subset of this structure is presented in Table 2, which demonstrates how different answer qualities (high, medium, and low) are mapped to corresponding manual scores.

Human raters manually applied the scoring rubric based on the completeness and relevance of student responses to the provided keywords and question context. The rubric was defined as follows:

- 90–100: Highly complete, clear, and relevant answers, where all keywords were thoroughly explained.
- 70–89: Good answers but lacking depth or requiring further elaboration.
- 50–69: Partial answers containing some keywords but with insufficient or less relevant explanations.
- 30–49: Limited answers missing many keywords or containing largely irrelevant content.
- 0–29: Completely irrelevant answers with little or no alignment to the expected keywords.

Table 2. Example of dataset structure

Data component	Example
Question	Jelaskan apa yang dimaksud dengan zodiac dalam konteks astronomi.
Keywords	Zodiac, Astronomi, Rasi bintang, Posisi, Matahari
High-Quality Answer Example	Zodiac dalam konteks astronomi merujuk pada dua belas rasi bintang yang terletak di sepanjang ecliptic, jalur yang dilalui Matahari. Setiap rasi bintang mewakili satu bulan dalam kalender astrological dan dikaitkan dengan tanda zodiak dalam astrologi. Rasi bintang ini berfungsi sebagai referensi untuk mengamati posisi Matahari, Bulan, dan planet-planet, serta untuk menentukan pergerakan mereka dalam kaitannya dengan bintang-bintang latar belakang.
Manual Score	93
Medium-Quality Answer Example	Zodiac adalah dua belas rasi bintang yang berada sepanjang ecliptic. Ini digunakan untuk memetakan posisi Matahari dan planet dalam kaitannya dengan bintang-bintang.
Manual Score	81
Lower-Quality Answer Example	Zodiac adalah dua belas rasi bintang yang terletak di ecliptic dan digunakan untuk menentukan posisi Matahari dan planet dalam astrologi.
Manual Score	68

Following preprocessing, the dataset was reduced to 3,152 valid samples, which were partitioned into three subsets: a 70% training set (2,206 samples), a 15% validation set (473 samples), and a 15% test set (473 samples). For model input, each sequence was padded to a maximum length of 150 tokens, resulting in a vocabulary size of 3,497 tokens derived from the tokenizer.

DATA PREPROCESSING

Before being used for training, the collected text data underwent a preprocessing pipeline designed to clean and standardize the input. This step is essential in preparing Indonesian text for NLP tasks, as it reduces noise and ensures that the model processes linguistically consistent input. The pipeline consisted of:

1. Tokenization, which involves splitting text into words or subwords, enables the analysis of discrete linguistic units.
2. Case Folding, converting all characters to lowercase (e.g., "Zodiac" → "zodiac") to avoid duplication from capitalization.
3. Stopword Removal, eliminating common words (e.g., "yang", "dan", "di") that contribute little semantic value, thereby reducing noise.
4. Stemming, using the Sastrawi library to reduce words to their base form (e.g., "menentukan" → "tentu"), which helps unify inflections and shrink vocabulary size.
5. Padding/Truncation, ensuring uniform sequence length of 150 tokens, which is required for batch training in the LSTM model.

This sequential process transformed raw text into a structured numerical format, enabling practical training while preserving semantic content. The pipeline is illustrated in Figure 3, where each step systematically refines the raw dataset into model-ready sequences.

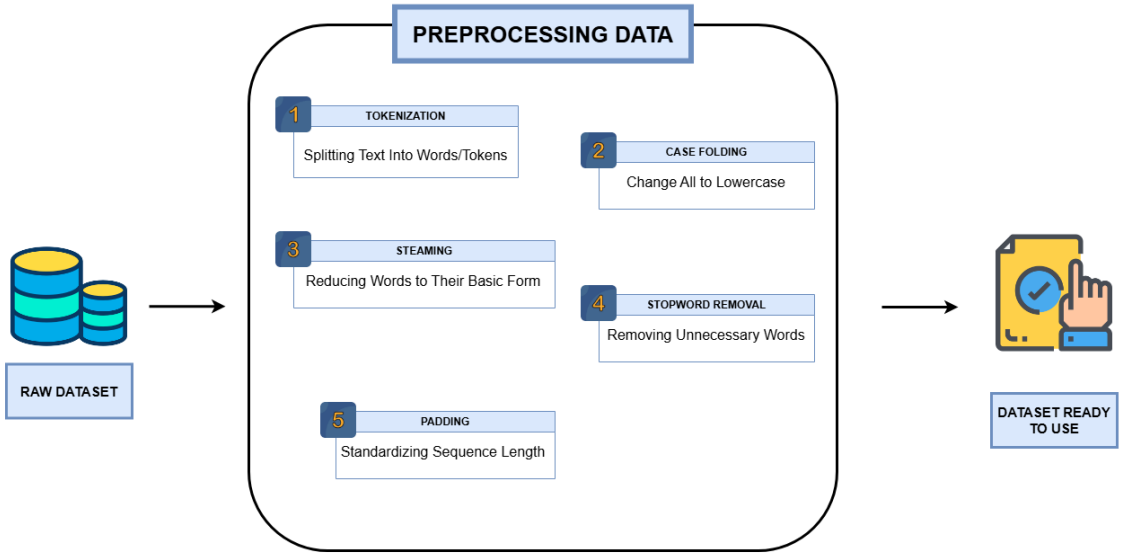

Figure 3. Data preprocessing flow

Model Architecture

The automated scoring system is built on a sequential deep learning architecture designed to capture both semantic content and contextual relationships in student answers. The proposed architecture is illustrated in Figure 4.

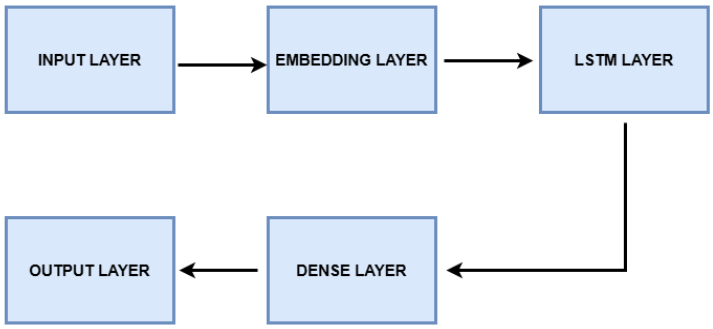

Figure 4. Proposed hybrid model architecture

The data flows through the architecture as follows:

1. **Input Layer**
   The model first receives the preprocessed and padded text sequences. Each sequence is a fixed-length numerical representation (integer tokens) of the combined question, student answer, and keywords.

2. **Embedding Layer**
   This layer is responsible for converting the integer tokens into dense vector representations. It is initialized with a pre-trained GloVe matrix of 300 dimensions, which is effective for

capturing semantic relationships in the Indonesian language (Singgalen, 2024). Crucially, the weights of this layer are frozen (trainable=False) to preserve the rich semantic information learned from the large external corpus and to prevent overfitting on the smaller proprietary dataset.

3. **Bidirectional LSTM Layer**
   The sequence of word vectors is then processed by a Bidirectional Long Short-Term Memory (LSTM) layer with 256 units. A bidirectional approach was chosen because it enables the model to learn context from both forward and backward directions within the sequence, resulting in a more robust contextual understanding of the text (Aburass et al., 2024). To combat overfitting, which is a common challenge in NLP tasks (Abarna et al., 2022), dropout and recurrent dropout rates of 0.3 are applied within this layer.
4. **Dense and Dropout Layers**
   The output from the Bi-LSTM layer is passed through two consecutive fully connected (Dense) layers with 128 and 64 units, respectively. These layers use the Rectified Linear Unit (ReLU) activation function, which is computationally efficient and effective at learning non-linear feature combinations from the LSTM's output. Following each Dense layer, a Dropout layer (with rates of 0.4 and 0.3) is applied as an additional regularization technique to further prevent the model from becoming overly specialized to the training data.
5. **Output Layer**
   Finally, the architecture concludes with a single-neuron Dense layer using a sigmoid activation function. The sigmoid function is ideal for this task as it squashes the output into a scalar value between 0 and 1, which directly corresponds to the normalized predicted score for the student's answer.

TRAINING SETUP

The model was trained using a set of specific hyperparameters, selected to optimize performance and ensure stable convergence. For the optimization algorithm, Adam was chosen due to its efficiency. He established effectiveness in training deep neural networks for text processing tasks (Onan & Alhumyani, 2024). The Mean Squared Error (MSE) was selected as the loss function, as it is a standard and robust metric for regression tasks in NLP that aim to minimize the error between predicted and actual continuous scores (Xu et al., 2025). The complete list of hyperparameters used during training, including learning rate, batch size, and callback configurations, is detailed in Table 3.

As detailed in Table 3, the model was configured to train for a maximum of 50 epochs using the Adam optimizer, with a learning rate of 0.0005 and a batch size of 64. To monitor the model's generalization capability, its performance was continuously evaluated on a validation set throughout the training process.

A dynamic EarlyStopping mechanism was implemented, configured to monitor the validation Mean Absolute Error (val_mae). This callback automatically terminated training once performance plateaued and restored the model weights from the epoch that yielded the best validation score. This approach ensures that the final model represents its peak generalization performance, rather than being overfit to the training data.

Table 3. Model configuration and training hyperparameters

Komponen	Parameter	Nilai
Embedding Layer	input_dim (Vocabulary Size)	3498
	output_dim (Vector Size)	300
	input_length (Sequence Length)	150
	weights (Source)	Pre-trained GloVe Matrix
	trainable	False

Komponen	Parameter	Nilai
LSTM Layer	LSTM Type	Bidirectional
	Units	256
	return_sequences	False
	dropout	0.3
	recurrent_dropout	0.3
Dense Layer 1	Units	128
	Activation Function	relu
	Dropout Rate	0.4
Dense Layer 2	Units	64
	Activation Function	relu
	Dropout Rate	0.3
Output Layer	Units	1
	Activation Function	sigmoid
Parameter Pelatihan	Optimizer	Adam
	Learning Rate	0.0005
	Loss Function	mean_squared_error
	Additional Metrics	mean_absolute_error, Root-MeanSquaredError, R2Score
	Batch Size	64
	Epochs	50
	Callbacks	EarlyStopping (monitor: val_mae, patience: 5), ReduceLROnPlateau (monitor: val_loss, patience: 3)

The overall training behavior is depicted in Figure 5 (Mean Squared Error) and Figure 6 (Mean Absolute Error). During the initial epochs, both training and validation errors show a steep decline, indicating effective learning of patterns from the data. As training progressed, these curves began to converge and stabilize, signaling that the model approached an optimal fit. The tight alignment between training and validation curves suggests that the model was able to generalize well to unseen data and did not suffer from significant overfitting.

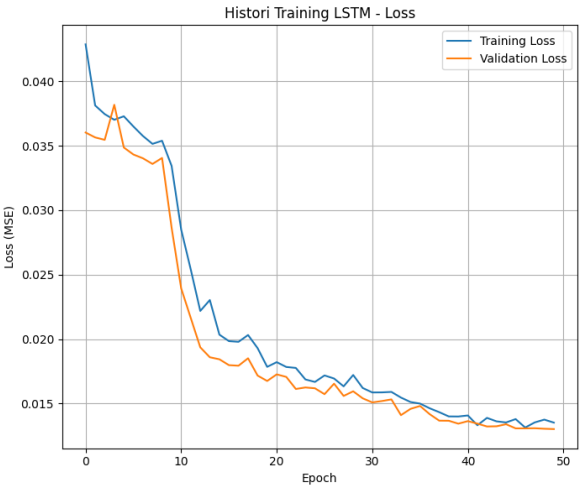

Figure 5. Mean squared error

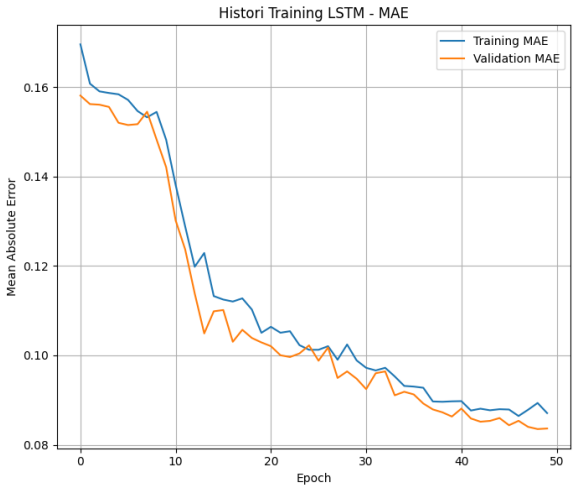

Figure 6. Mean absolute error

EVALUATION EXPERIMENT

To comprehensively evaluate the effectiveness and robustness of the proposed GloVe–LSTM hybrid model in the context of automated student answer scoring, a dual evaluation strategy was adopted. The first component involves quantitative analysis using widely accepted regression, correlation, and agreement metrics to objectively measure how closely the model's predicted scores align with human-assigned scores. This enables a numerical evaluation of the model's absolute accuracy, consistency, and reliability in scoring tasks.

The second component of the evaluation focuses on scenario-based qualitative testing, in which the model is exposed to carefully constructed cases that simulate real-world challenges commonly encountered in student responses. These include variations in answer length, use of synonyms, grammatical errors, and the presence or absence of explicit keywords. Each scenario is designed to probe a specific aspect of the model's semantic understanding and decision-making capability. Overall, these assessment techniques provide balanced and holistic insights into the model's performance, highlighting areas where it excels, such as scoring consistency, and pinpointing areas where it struggles, such as linguistic variability. In the next section, the qualitative scenarios are detailed, along with the exercises where they were conducted, and the model's actions are explored in detail under various scenarios.

EVALUATION SCENARIOS

To add to the quantitative metrics, the model's performance within certain, well-defined scenarios of heightened difficulty warranted a thorough qualitative assessment. For a complete evaluation of a well-functioning automated scoring engine, six such test scenarios were developed, each with a focus on an essential component. The purpose of each scenario is outlined below:

1. Scenario 1 (Baseline Functionality)
   This initial scenario was designed to verify the model's fundamental ability to assign scores proportionally across a broad spectrum of answer quality, from highly complete to entirely incorrect, establishing a performance baseline.
2. Scenario 2 (Answer Length Variation)
   This scenario aimed to assess the model's consistency and potential bias when presented with answers of significantly different lengths but equivalent semantic content and quality.
3. Scenario 3 (Synonym and Structural Variation)
   To test the model's semantic understanding, this scenario evaluated its ability to recognize and consistently score answers that were conceptually identical but expressed using different vocabulary (synonyms) and sentence structures.
4. Scenario 4 (Robustness to Noise)
   This test was designed to measure the model's tolerance for common linguistic imperfections, such as minor spelling mistakes and grammatical errors, which do not alter the core meaning of the answer.
5. Scenario 5 (Conceptual Understanding without Keywords)
   This scenario validated the model's ability to move beyond simple keyword matching by assessing answers that were conceptually correct but deliberately avoided the explicit keywords found in the question or reference answer.
6. Scenario 6 (Discriminatory Power against Irrelevant Answers)
   As a critical test of contextual understanding, this final scenario evaluated the model's ability to identify and assign appropriately low scores to responses that were irrelevant, factually incorrect, or consisted only of filler phrases

EVALUATION METRICS

In order to evaluate the trained model's performance, a blended evaluation approach was utilized to address all the components of the assessment comprehensively. In this particular approach, a set of

the model's performance evaluation criteria, which was set to a certain level of scoring granularity (absolute) and scoring correlation (aligned) with manual grading, is further examined on the level of agreement obtained on categorical grading systems. The selected metrics, each serving a distinct purpose as detailed in Table 4, are grouped into three primary categories.

Table 4. Evaluation metrics used in the study

Metric category	Metric	Purpose
Regression	Mean Absolute Error (MAE)	Measures the average absolute difference between predicted scores and actual scores, indicating the magnitude of error.
	Mean Squared Error (MSE)	Measures the average of the squares of the errors, giving higher weight to larger errors.
	R^2 Score	Indicates the proportion of the variance in the actual scores that is predictable from the model's predictions.
Correlation	Pearson Correlation	Measures the linear relationship between the predicted scores and the actual scores.
	Spearman Rank Correlation	Measures the monotonic relationship between the predicted scores and the actual scores, focusing on rank order.
Agreement	Quadratic Weighted Kappa (QWK)	Measures the agreement between two raters (the model and human scorer) on an ordinal scale, accounting for the degree of disagreement.

RESULTS AND DISCUSSION

QUANTITATIVE RESULT

The proposed GloVe–LSTM model achieved a Mean Absolute Error (MAE) of 7.61%, a Pearson correlation of 0.87, and a Quadratic Weighted Kappa (QWK) of 0.82. These values demonstrate a strong alignment between automated scores and human ratings, particularly in terms of correlation and consistency across scoring ranges.

Notably, an MAE of 7.61% corresponds to an average deviation of approximately 0.38 points on a 5-point grading scale, meaning the system's errors could shift a score from B+ to A- or C to C+. While this is within acceptable margins for supportive use, it highlights the necessity of human oversight in high-stakes grading. Quantitative performance is summarized in Table 5 and visualized in Figure 7.

Table 5. Overall performance metrics of the main model on the test set

Metric category	Metric	Value	p-value	Note
Regression metrics	Mean Squared Error (MSE)	0.0106	-	Lower is better
	Mean Absolute Error (MAE)	0.0761	-	Lower is better
Correlation metrics	Pearson Correlation Coefficient	0.8429	< 0.001	Strong and statistically significant
	Spearman Rank Correlation	0.7637	< 0.001	Strong and statistically significant
Agreement metric	Quadratic Weighted Kappa (QWK)	0.7332	-	Good agreement between raters

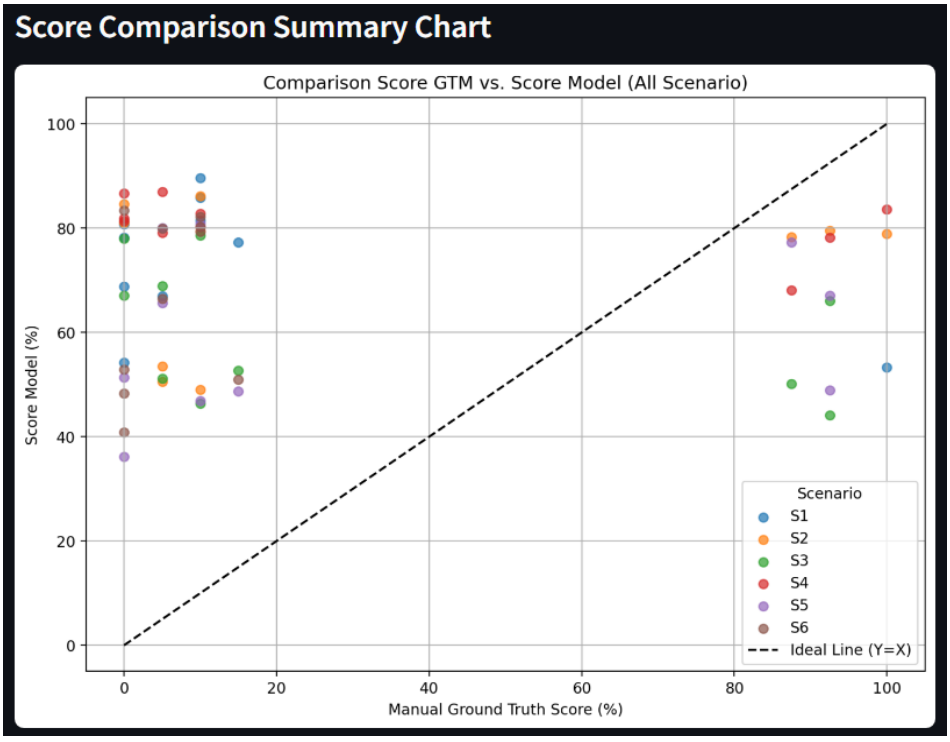

Figure 7. Scatter plot of predicted scores vs. ground truth scores

Scenario-Based Evaluation

To further assess robustness, the model was tested under six scenarios (S1–S6): synonym substitution, sentence restructuring, length variation, grammatical errors, incomplete answers, and irrelevant responses. A consolidated summary of representative test cases is provided in Table 6, which shows ground-truth versus predicted scores, deviations, and key observations.

Table 6. Scenario-based evaluation: Ground-truth vs predicted scores

Scenario	Test case example	Ground truth	Predicted	Deviation	Note
S1 – Synonym Substitution	"pupil" vs "student"	4.5	3.3	-1.2	Underestimation when synonyms differ in distribution
S2 – Sentence Restructuring	Passive to Active form	4.0	3.6	-0.4	Works for simple reordering, fails on complex
S3 – Length Variation	Redundant long answer	2.0	3.2	+1.2	Overestimation due to verbosity; std. dev = 0.88
S4 – Grammatical Errors	Minor mistakes	3.5	3.3	–0.2	Resilient to small errors; weaker on severe ones
S5 – Incomplete Answer	Missing keywords	1.0	1.2	+0.2	Closely follows the rubric, slight inflation
S6 – Irrelevant Response	Off-topic with keyword overlap	0.5	1.0	+0.5	Correctly penalized, but inflated if the keyword is present.

The results indicate that the most challenging conditions were S1 (Synonym Substitution) and S3 (Length Variation), where deviations reached up to 1.2 points. Variability was highest (standard deviation = 0.74 and 0.88, respectively). By contrast, the model was more reliable in S5 (Incomplete Answers) and S6 (Irrelevant Responses), aligning closely with human grading. In S4 (Grammatical Errors), the system demonstrated resilience to minor mistakes, while S2 (Sentence Restructuring) yielded mixed results, depending on the complexity.

Synthesis of Findings

The integration of quantitative and qualitative results paints a dual picture of the model's strengths and limitations. On the one hand, overall performance was strong, with a high correlation to human scores (Pearson's r = 0.87) and reliable grading for incomplete and irrelevant answers (S5, S6). The system also demonstrated robustness to minor grammatical errors (S4).

On the other hand, critical weaknesses remain, particularly in handling synonym diversity (S1) and length variation (S3). These errors reveal that while GloVe embeddings effectively capture global semantics, the Bi-LSTM struggles with lexical variation and verbosity. Taken together, the findings suggest that the hybrid GloVe–LSTM model provides a promising foundation for Indonesian AES, but further refinement is needed to ensure fairness in diverse linguistic contexts.

Practical Implications

In practice, the system's accuracy (MAE 7.61%) is sufficient for use as a supportive tool in educational contexts. Potential applications include:

- *Batch screening:* rapidly processing large sets of student answers.
- *Teacher assistance:* generating preliminary scores that educators can review.
- *Borderline detection:* flagging ambiguous cases for human review.

However, the model should not replace teachers in high-stakes grading (e.g., national exams), due to weaknesses in handling synonyms and sentence length. Instead, it should be positioned as a first-pass filter that reduces grading time, supports fairer evaluation, and allows teachers to focus on qualitative feedback.

By situating the system in this way, the research contributes to both NLP, through the adaptation of hybrid architectures for morphologically rich languages, and to education, by demonstrating how AI can augment teacher capacity while preserving fairness in assessment.

DISCUSSION

This study aimed to evaluate whether a hybrid GloVe–LSTM model can simulate human scoring (RQ1), handle linguistic variation (RQ2), and demonstrate the impact of integrating embeddings with similarity measures (RQ3). The findings reveal both promising strengths and critical weaknesses, which are best understood when grouped under two broad themes: discrimination & fairness and linguistic robustness.

RQ1: Simulation of Human Scoring

Quantitative evaluation demonstrated strong agreement with human raters, yielding high Pearson correlation coefficients (0.8429) and Spearman correlation coefficients (0.7637). These findings indicate the model's capability to capture the relative ranking of the set of student answers and assign scores in accordance with the evaluative hierarchy, that is, higher scores to those judged as superior and lower scores to weaker responses. The ability to capture this ranking is additional proof that the GloVe–LSTM architecture is capable of capturing signals from properly and meaningfully answered questions (Kumar & Boulanger, 2020).

Their consistency with findings of relative scoring with AES in other studies is notable. These studies highlight the application of hybrid architecture approaches, such as AES, as an approximate human evaluator in relative scoring, which has been documented (Beseiso & Alzahrani, 2020; Wangkriangkri et al., 2020). Within this Indonesian context, the outcomes demonstrate the feasibility of human-like relative scoring with lightweight models and highlight the potential of AES as an assistive tool.

RQ2: Handling of Linguistic Variation

The scenario-based evaluation exposed the model's weaknesses in handling synonyms, structural variation, and text length. Scores fluctuated by up to 15.97% for answers that were semantically equivalent but differed in synonyms or sentence structures. This inconsistency reflects known limitations of LSTM-based systems in capturing fine-grained semantic equivalence (Kusumaningrum et al., 2024).

Indonesian linguistic characteristics amplify this challenge. Affixation and free word order make semantic equivalence less straightforward, a difficulty also highlighted by Wibowo et al. (2024). The model sometimes favored longer answers, even when redundancy did not contribute to correctness, revealing a bias toward verbosity. Similarly, answers with minor spelling or grammatical errors were penalized in some cases and rewarded in others, indicating inconsistent robustness to textual noise. These results align with prior observations that robust semantic handling remains a central challenge for AES models across languages (Beseiso & Alzahrani, 2020; Wangkriangkri et al., 2020).

RQ3: Impact of the Hybrid Approach

The integration of GloVe embeddings with similarity measures (ROUGE, TF-IDF, and cosine similarity) provided clear benefits in discriminating between incomplete and irrelevant responses. For example, irrelevant answers were consistently penalized, aligning closely with human scores. This demonstrates that lightweight similarity metrics add applicable constraints to purely neural representations.

Compared with Sriyanto and Kusrini (2025), who adopted a universal sentence encoder hybrid emphasizing efficiency, the current study emphasizes semantic depth. While both approaches achieved encouraging results, the novelty of this work lies in prioritizing robustness against Indonesian linguistic diversity rather than computational minimalism. Similarly, while Aisyah et al. (2025) explored the integration of OCR for handwritten responses, their focus was on technical feasibility rather than semantic scoring. By contrast, the present model contributes a balanced approach, coupling semantic embeddings with interpretable similarity measures.

Discrimination and Fairness

Despite its strengths, the model demonstrated poor discriminatory power in absolute scoring, particularly in its tendency to overestimate answers that are poor or irrelevant. Scenario testing revealed stark failures: one irrelevant answer with a ground truth of 15% was scored at 77.28%. In contrast, a completely incorrect answer (ground truth: 0%) received 83.45%.

This pattern suggests that the model relies heavily on surface-level features, such as keyword presence, sentence length, and syntactic structure, rather than validating the contextual relevance of the content. Similar overestimation problems have been reported in other AES systems, reflecting broader NLP challenges in distinguishing literal from idiomatic or off-topic content (Abarna et al., 2022; Wangkriangkri et al., 2020). Such systematic overestimation undermines fairness and highlights the risk of deploying the model for autonomous high-stakes grading.

Linguistic Robustness

The model's inconsistent handling of synonyms, restructured sentences, and noisy inputs underscores its limited linguistic robustness. While GloVe embeddings capture global semantics (Pennington et al., 2014), the Bi-LSTM component remains sensitive to superficial variations, especially in Indonesian, where affixation and free word order are pervasive (Wibowo et al., 2024).

These findings reinforce the need for AES systems to move beyond surface-level pattern matching toward deeper contextual understanding, a priority also identified in recent surveys (Misgna et al., 2024). Without such robustness, models risk perpetuating bias against students who use alternative but valid linguistic forms.

IMPLICATIONS AND RECOMMENDATIONS

Taken together, the findings suggest that the hybrid GloVe–LSTM model represents a promising proof-of-concept, but is not yet reliable for high-stakes, absolute scoring. Its ranking strength can be leveraged for assistive purposes, such as pre-sorting student responses or flagging ambiguous cases for teacher review.

Pedagogically, the tool can reduce grading workload, provide faster feedback, and support formative assessment, but human oversight remains essential. From a policy perspective, AES deployment should be framed as assistive technology, with safeguards for transparency and accountability.

For future development, several directions are recommended:

- Incorporating Transformer-based embeddings (e.g., IndoBERT) while retaining lightweight similarity metrics.
- Addressing synonym and verbosity biases via data augmentation and semantic regularization techniques.
- Establishing standardized Indonesian AES benchmarks enables fairer cross-study comparisons and accelerates progress in this underexplored domain.

CONCLUSION

This study proposes a hybrid GloVe–LSTM model, combined with ROUGE, TF-IDF, and cosine similarity, for the automated scoring of Indonesian student answers. The model achieved strong alignment with human scoring (MAE = 7.61%, Pearson = 0.87, QWK = 0.82) and demonstrated consistent ranking ability. Scenario-based testing further revealed that it effectively penalizes incomplete or irrelevant answers. However, it remains sensitive to the use of synonyms, structural variation, and verbosity.

The novelty of this research lies in its adaptation of a hybrid AES framework to the Indonesian context. This morphologically rich language has been underexplored in prior work. To our knowledge, this is the first study to integrate GloVe–LSTM with traditional similarity measures for Indonesian AES and to validate it through scenario-based evaluations (S1–S6), offering a more pedagogically grounded perspective beyond aggregate metrics.

Despite its contributions, the model exhibits limitations, particularly in overestimating poor responses and lacking robustness to linguistic variation. These weaknesses highlight the need for caution in deploying such systems for high-stakes assessments. However, they hold promise as assistive tools to reduce teacher workload, support formative feedback, and streamline large-scale evaluation.

Future research should prioritize enhancing linguistic robustness through Transformer-based embeddings such as IndoBERT, data augmentation to handle synonym and structural diversity, and the development of standardized Indonesian AES benchmarks. Expanding toward multimodal assessment, such as integrating OCR for handwritten inputs, may also strengthen practical applicability.

In conclusion, this study delivers a proof-of-concept hybrid AES model tailored to Indonesian education, demonstrating both the potential and challenges of automated scoring. Its contributions span both the NLP domain, by adapting hybrid architectures to a low-resource language, and the educational domain, by advancing fairer and more efficient student assessment methods.

REFERENCES

Abarna, S., Sheeba, J. I., & Devaneyan, S. P. (2022). An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning. *Measurement: Sensors, 24*, 100434. https://doi.org/10.1016/j.measen.2022.100434

Abdullah, A. S., Geetha, S., Abdul Aziz, A. B., & Mishra, U. (2024). Design of automated model for inspecting and evaluating handwritten answer scripts: A pedagogical approach with NLP and deep learning. *Alexandria Engineering Journal*, *108*, 764-788. https://doi.org/10.1016/j.aej.2024.08.067

Aburass, S., Dorgham, O., & Rumman, M. A. (2024). An ensemble approach to question classification: Integrating Electra transformer, GloVe, and LSTM. *International Journal of Advanced Computer Science and Applications*, *15*(1), 507-514. https://doi.org/10.14569/IJACSA.2024.0150148

Ahmed, M., Khan, H. U., Iqbal, S., & Althebyan, Q. (2022, November). Automated question answering based on improved TF-IDF and cosine similarity. *Proceedings of the Ninth International Conference on Social Networks Analysis, Management and Security, Milan, Italy,* 1-6. https://doi.org/10.1109/SNAMS58071.2022.10062839

Aisyah, N., Kautsar, M. D. A., Hidayat, A., Chowdhury, R., & Koto, F. (2025). *Evaluating vision-language and large language models for automated student assessment in Indonesian classrooms.* PsyArXiv. https://www.arxiv.org/pdf/2506.04822

Basha, S. A. K., Durai, R. V. P. M., Suleiman, I. M., Asokan, V., Eddie, E. H. S., Qusai, S., & Alshurideh, M. T. (2025). Exploring deep learning methods for audio speech emotion detection: An ensemble MFCCs, CNNs and LSTM. *Applied Mathematics and Information Sciences Journal*, *19*(1), 75-85. https://doi.org/10.18576/amis/190107

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

Buditjahjanto, I. G. P. A., Idhom, M., Munoto, M., & Samani, M. (2022). An automated essay scoring based on neural networks to predict and classify competence of examinees in community academy. *TEM Journal*, *11*(4), 1694-1701. https://doi.org/10.18421/TEM114-34

Dewi, C., Kristiantoro, C. S. A., Christanto, H. J., & Riantama, D. (2024). Improving service quality through classifying chatbot messages based on natural language processing: A bidirectional long short-term memory network model. *International Journal of Applied Science and Engineering, 21*(2), 1-14. https://doi.org/10.6703/IJASE.202406_21(2).006

Dhiman, P., Kaur, A., Gupta, D., Juneja, S., Nauman, A., & Muhammad, G. (2024). GBERT: A hybrid deep learning model based on GPT-BERT for fake news detection. *Heliyon*, *10*(16), e35865. https://doi.org/10.1016/j.heliyon.2024.e35865

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

Kim, J., Chung, S., Moon, S., & Chi, S. (2022, December). Feasibility study of a BERT-based question answering chatbot for information retrieval from construction specifications. *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management,* 970-974. https://doi.org/10.1109/IEEM55944.2022.9989625

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, *5*, 572367. https://doi.org/10.3389/feduc.2020.572367

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

Misgna, H., On, B.-W., Lee, I., & Choi, G. S. (2024). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, *58*, Article 36. https://doi.org/10.1007/s10462-024-11017-5

Nahin, K. H., Nirob, J. H., Taki, A. A., Haque, M. A., Singh, N. S. S. S., Paul, L. C., Alkanhel, R. I., Abdallah, H. A., Ateya, A. A., & El-Latif, A. A. A. (2025). Performance prediction and optimization of a high-efficiency tessellated diamond fractal MIMO antenna for terahertz 6G communication using machine learning approaches. *Scientific Reports*, *15*, Article 4215. https://doi.org/10.1038/s41598-025-88174-2

Nasreen, G., Khan, M. M., Younus, M., Zafar, B., & Hanif, M. K. (2024). Email spam detection by deep learning models using novel feature selection technique and BERT. *Egyptian Informatics Journal*, 26, 100473. https://doi.org/10.1016/j.eij.2024.100473

Onan, A., & Alhumyani, H. A. (2024). FuzzyTP-BERT: Enhancing extractive text summarization with fuzzy topic modeling and transformer networks. *Journal of King Saud University - Computer and Information Sciences*, *36*(6), 102080. https://doi.org/10.1016/j.jksuci.2024.102080

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10.3115/v1/D14-1162

Putra, R. R., Putri, N. A., & Putra, A. D. (2024). Web-based analysis of public complaints in Pertumbuhan Village using cosine similarity and TF-IDF techniques. *Jurnal Sains Komputer dan Informatika*, *8*(1), 748-761.

Sihombing, J. J., Arnita, A., Al Idrus, S. I., & Niska, D. Y. (2024). Implementation of text summarization on Indonesian scientific articles using TextRank algorithm with TF-IDF web-based. *Journal of Soft Computing Exploration, 5*(3), 310-319.

Singgalen, Y. A. (2024). Implementation of Global Vectors for Word Representation (GloVe) model and social network analysis through wonderland Indonesia content reviews. *Jurnal Sistem Komputer dan Informatika*, *5*(3), 559-569. https://doi.org/10.30865/json.v5i3.7569

Sriyanto, A., & Kusrini. (2025). Hybrid deep learning and USE algorithm for essay scoring: Accuracy and performance analysis. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, *9*(2), 914-922. https://doi.org/10.33395/sinkron.v9i2.14784

Wang, Q. (2024). A multifaceted architecture to automate essay scoring for assessing english article writing: Integrating semantic, thematic, and linguistic representations. *Computers and Electrical Engineering*, *118*, 109308. https://doi.org/10.1016/j.compeleceng.2024.109308

Wangkriangkri, P., Viboonlarp, C., Rutherford, A. T., & Chuangsuwanich, E. (2020, November). A comparative study of pretrained language models for automated essay scoring with adversarial inputs. Proceedings of the *IEEE Region 10 Conference*, *Osaka, Japan*, 875-880. https://doi.org/10.1109/TENCON50793.2020.9293930

Wibowo, M. E., Rokhman, N., & Sihabudin, A. (2024). Combining multiple text representations for improved automatic evaluation of Indonesian essay answers. *Scientific Journal of Informatics*, *11*(3), 681-692. https://doi.org/10.15294/sji.v11i3.9440

Xu, W., Yao, Z., Ma, Y., & Li, Z. (2025). Understanding customer complaints from negative online hotel reviews: A BERT-based deep learning approach. *International Journal of Hospitality Management*, *126*, 104057. https://doi.org/10.1016/j.ijhm.2024.104057

AUTHORS

I Gede Susrama Mas Diyasa earned his doctoral degree in Electrical Engineering from the Sepuluh Nopember Institute of Technology Surabaya, where he specialized in intelligent systems. He serves as a faculty member at the Magister Program in Information Technology, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur. His primary research interests lie in intelligent systems, biomedical engineering, and smart city applications. Much of his work explores the design of algorithms and computational models to support medical diagnosis and treatment, as well as the implementation of intelligent technologies for urban management and facial detection systems.

Mohammad Idhon received his bachelor's and master's degrees in Information Technology from Universitas Pembangunan Nasional "Veteran" East Java. He is a lecturer at the School of Informatics, Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur. His research interests include software engineering, information systems, artificial intelligence, and technology-based education. He has been actively involved in academic research, publications, and community service projects that focus on the application of digital technology to solve real-world problems.

Ahmad Sofian Aris Saputra is a final-year student in Informatics at the Faculty of Computer Science, Universitas Pembangunan Nasional Veteran Jawa Timur. Currently in his final semester, he is focusing on his final project. His research interests include cloud computing, backend development, and system design, with a particular focus on the development and deployment of efficient and scalable information systems.

Deshinta Arrova Dewi is an academic and researcher specializing in artificial intelligence (AI) and big data analytics. She earned her doctorate from Universiti Kebangsaan Malaysia (UKM) with a focus on applying AI in education, economics, and healthcare. Her research includes award-winning solutions such as predictive machine learning algorithms and adaptive learning models that support technology-based education. As an educator, she actively inspires students through practical teaching and frequently serves as a keynote speaker at international AI conferences.

Tresna Maulana Fahrudin is a lecturer at the Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur. He is currently pursuing his doctoral degree in the Department of Information and Communication Systems, Okayama University, Japan. His academic interests include information systems, communication technologies, and digital transformation. In addition to his teaching activities, he has been actively engaged in research collaborations and publications, focusing on developing innovative approaches to strengthen the role of information and communication technologies in education, business, and society.