


# Indonesian automated short-answer grading using transformers-based semantic similarity

Samuel Situmeang, Sarah Rosdiana Tambunan, Lidia Ginting,  
Wahyu Krisdangolyanti Simamora, Winda Sari ButarButar

Information System Study Program, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Toba, Indonesia

Article Info	ABSTRACT
<p><b>Article history:</b></p> <p>Received Oct 7, 2024 Revised Feb 6, 2025 Accepted Jun 9, 2025</p> <hr/> <p><b>Keywords:</b></p> <p>Dimensionality reduction Hand-engineered features Indonesian Short answer grading Transformer</p>	<p>Automatic short answer grading (ASAG) systems offer a promising solution for improving the efficiency of reading literacy assessments. While promising, current Indonesian artificial intelligence (AI) grading systems still have room for improvement, especially when dealing with different domains. This study explores the effectiveness of large language models, specifically bidirectional encoder representations from transformers (BERT) variants, in conjunction with traditional hand-engineered features, to improve ASAG accuracy. We conducted experiments using various BERT models, hand-engineered features, text pre-processing techniques, and dimensionality reduction. Our findings show that BERT models consistently outperform traditional methods like term frequency-inverse document frequency (TF-IDF). IndoBERT<sub>Lite-Base-P2</sub> achieved the highest quadratic weighted kappa (QWK) score among the BERT variants. Integrating hand-engineered features with BERT resulted in a substantial enhancement of the QWK score. Utilizing comprehensive text pre-processing is a critical factor in achieving optimal performance. In addition, dimensionality reduction should be carefully used because it potentially removes semantic information.</p> <p><i>This is an open access article under the <a href="#">CC BY-SA</a> license.</i></p> <div></div>

<p><b>Corresponding Author:</b></p> <p>Samuel Situmeang Information System Study Program, Faculty of Informatics and Electrical Engineering Institut Teknologi Del Jl. Sisingamangaraja, Toba, North Sumatra, Indonesia Email: samuel.situmeang@del.ac.id</p>
---

## 1. INTRODUCTION

Literacy skills, like writing, reading, and problem-solving, are vital for everyone. Short answer questions (SAQs) are a standard way to test these skills, but grading them manually can be time-consuming and subjective. Automatic short answer grading (ASAG) systems use natural language processing (NLP) and machine learning to automate this process. However, a major challenge in ASAG is accurately understanding words and phrases that can have multiple meanings depending on the context [1]–[3]. For instance, the Indonesian word “*tahu*” can refer to either the concept of understanding or a food item. This necessitates ASAG systems capable of capturing semantic similarities between varied wordings and expressions.

Existing ASAG approaches include classical machine learning (CML) and deep learning (DL) [3]. Lexical, syntactic, and semantic features were extracted from the raw text and used as input for various CML models, including logistic regression (LR), support vector machines (SVMs), random forests (RF), extreme gradient boosting (XGBoost), and Naïve Bayes (NB) classifiers. In contrast, DL models include methods based on word embedding models, methods based on sequential models, and methods based on attention mechanisms. The models in the methods based on word embedding models (e.g., Word2Vec, GloVe,

FastText, ELMo, and Ans2Vec) generate representations that convert similar words into nearby vectors within an embedded latent space. These methods produce sentence embeddings through aggregation techniques such as summation or averaging of the individual word embeddings [4]–[7]. Compared to manually engineered features, word and sentence embeddings more effectively capture the semantic information present in textual data. To enhance feature quality and robustness of word and sentence representations, researchers have applied sequential models like recurrent neural networks (RNNs) and long short-term memory (LSTM) to the task of ASAG. By considering sentences of varying lengths and modeling longer-range word relationships within sentences, these methods are able to better capture the semantic properties of text. This enables the prediction models to make more robust and accurate inferences on the provided answers [8]–[10]. Attention mechanisms offer the ability to model long-range word relationships within a sentence and calculate the relative importance and dependencies between each word. Unlike sequential models, attention-based approaches do not explicitly consider word sequentiality. Transformer architectures, introduced by Vaswani *et al.* [11], consist of an encoder-decoder structure capable of characterizing long-range dependencies and features in sequential data. Transformers employ multiple parallel attention-head components, each learning different dependencies. Currently, the most advanced methods for ASAG leverage bidirectional encoder representations from transformers (BERT) [12]–[19]. BERT-based models have consistently demonstrated superior performance in this domain, outperforming alternative approaches on various ASAG tasks and datasets. The success of BERT can be attributed to its ability to effectively capture contextual information and long-range dependencies within the input text, leading to more accurate and nuanced predictions.

While previous research on ASAG has primarily focused on non-Indonesian languages, the study of grading using BERT in Indonesian is crucial due to the cultural nuances, the language's unique linguistic complexities, and limited high-quality datasets. These factors highlight the need for Indonesian language-specific study in ASAG. Previous research indicates that Indonesian ASAG methods have struggled to achieve consistent and robust performance across diverse domains [16], [17]. To address these challenges, we investigate the use of large language models, particularly BERT variants, combined with traditional hand-engineered features to enhance ASAG systems' accuracy. The contributions of this study are as follows:

- Assess the performance of various BERT models for text representation in the ASAG system, focusing on semantic similarity and identifying key factors contributing to their effectiveness.
- Evaluate the impact of different text pre-processing techniques on the effectiveness of a semantic similarity-based ASAG system using the optimal BERT variant.
- Evaluate the effectiveness of combining BERT embeddings with hand-engineered features in a semantic similarity-based ASAG system.
- Compare the effectiveness of dense representations (BERT embeddings) versus sparse representations such as term frequency-inverse document frequency (TF-IDF) for text representation in a semantic similarity-based ASAG.
- Evaluate the impact of dimensionality reduction techniques such as principal component analysis (PCA) on BERT embeddings for the effectiveness of a semantic similarity-based ASAG system.

## 2. RESEARCH METHOD

This section explores the proposed methodology and dataset while detailing the specific experiments to be conducted. Additionally, it outlines the evaluation metrics that will be employed to assess the outcomes. Together, these elements provide a comprehensive framework for the research process.

## 3. SEMANTIC SIMILARITY-BASED ASAG SYSTEM

We propose an approach that leverages the power of BERT [14], a state-of-the-art DL model, in conjunction with hand-engineered features [17] to enhance the accuracy and robustness of the grading process. BERT excels at capturing the semantic and syntactic nuances of text by processing words bidirectionally, enabling it to understand the context of a word based on both preceding and succeeding words. This synergistic approach harnesses the best of both worlds: the flexibility and power of DL and the precision of domain-specific knowledge. The flowchart of the proposed ASAG system is shown in Figure 1. The following are the main steps in the proposed method:

### 3.1. Text pre-processing

In this step, the input text is processed to remove noise and improve the accuracy of the assessment. The text pre-processing process consists of several steps, namely:

- a) Tokenization: we use the Treebank tokenizer to break down each question, key answer, and student answer into smaller units (tokens).
- b) Case-folding: we transform all capital letters to lowercase.
- c) Remove stop words: we remove words with no significant meaning (stop words) based on the NLTK stopwords lists [20].
- d) Remove punctuation: we remove all punctuation, including commas, periods, and question marks.
- e) Lemmatization: we use the WordNet Lemmatizer to change each word to its basic form.
- f) Stemming: we use the Porter Stemmer to change each word to its basic form.

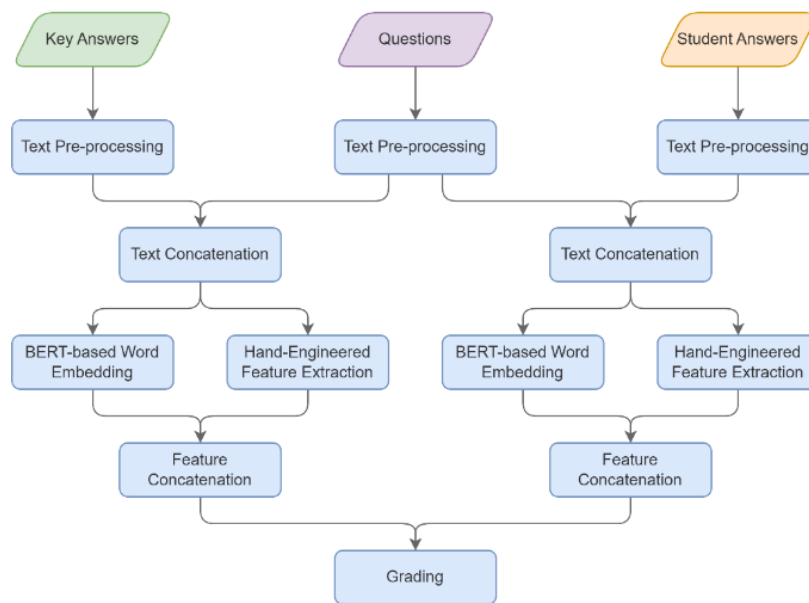


Figure 1. Flowchart of proposed ASAG system

### 3.2. Text concatenation

To provide contextual information, we combine the pre-processed question text with the pre-processed key answers and student answers. This combined text is used as input for the feature extraction process.

### 3.3. Feature extraction

#### 3.3.1. BERT-based word embedding

The combined text is processed using a BERT-based word embedding model. This model converts the text into numerical representations, or embeddings, where words with similar meanings are placed closer together in a high-dimensional space. These embeddings capture semantic and syntactic information from the text, which is crucial for downstream tasks like similarity comparison.

#### 3.3.2. Hand-engineered feature extraction

In parallel with the BERT-based word embedding, the combined text also undergoes a process of hand-engineered feature extraction. These features are presented in Table 1.

Table 1. List of hand-engineered features

No.	Feature	Source
1	Lexical overlap	[4]
2	Prompt overlap	[4]
3	Grammar error ratio	[21]
4	Average sentence length	[21]
5	Average words length	[21]
6	Answer length ratio	[22]

### 3.4. Features concatenation

The numerical representations generated by the BERT-based word embedding model are combined with the hand-engineered features. This concatenation process involves merging these two types of features into a single, larger feature vector. By combining these two sources of information, we aim to leverage the strengths of both approaches. BERT-based embeddings capture semantic and syntactic information from the text, while hand-engineered features provide more specific information that may be relevant to the task at hand. This combined feature vector will be the input to the grading step.

### 3.5. Grading

Cosine similarity measures the semantic similarity between the key and student answers. A higher cosine similarity score indicates a greater degree of similarity. A final grade can be assigned by comparing the student's answer to multiple key answers.

## 4. DATASET

We utilize the Indonesian Query Answering Dataset for Online Essay Test System [23]. This dataset encompasses diverse domains, including lifestyle, technology, politics, and sports. Each question sheet within the dataset contains questions, corresponding key answer, and a sequential numbering from 1 to 10. For each student response (answer), the dataset includes:

- Manual scores: three manual ratings from human evaluators.
- Automated scores: scores generated using various similarity metrics (Cosine, Euclidean, Jaccard) with and without stemming.
- Error analysis: detailed error analysis for each automated scoring method.
- Overall average score: a calculated average score considering both manual and automated scores.

The statistics of the dataset is shown in Table 2.

Table 2. Statistic dataset

Statistic	Topic			
	Lifestyle	Politic	Sports	Technology
Total question	10	10	10	10
Total answer	568	535	560	515
Total key answer	10	10	10	10
Min. length question	6.00	4.00	9.00	5.00
Max. length question	11.00	21.00	18.00	14.00
Avg. length question	8.30	11.50	14.0	28.30
Min. length answer	1.00	1.00	1.00	1.00
Max. length answer	385.00	556.00	297.00	457.00
Avg. length answer	25.34	29.44	26.41	21.82
Min. length key answer	24.00	12.00	14.00	12.00
Max. length key answer	134.00	49.00	75.00	39.00
Avg. length key answer	73.00	31.70	38.90	28.30

## 5. EXPERIMENT SCENARIO

This study conducts five experiments:

### a) Relative effectiveness of BERT variants

We evaluate the performance of various BERT variants, including multilingual BERT [24], IndoBERT<sub>Base-P1</sub>, IndoBERT<sub>Base-P2</sub>, IndoBERT<sub>Large-P1</sub>, IndoBERT<sub>Large-P2</sub>, IndoBERT<sub>Lite-Base-P1</sub>, IndoBERT<sub>Lite-Base-P2</sub>, IndoBERT<sub>Lite-Large-P1</sub>, and IndoBERT<sub>Lite-Large-P2</sub> [25]. The best-performing variant is selected as the foundation for subsequent experiments.

### b) Impact of text pre-processing

The optimal BERT model, identified in Experiment 1, undergoes further analysis by applying various text pre-processing techniques to examine their influence on the effectiveness of Semantic Similarity-based ASAG.

### c) Impact of hand-engineered features

We investigate the effectiveness of combining BERT embeddings with hand-engineered features. This experiment aims to determine whether the integration of domain-specific knowledge can enhance the performance of the ASAG system.

### d) Effectiveness comparison of TF-IDF and BERT

A comparative analysis is conducted between TF-IDF [26] and BERT-based representations to evaluate their relative strengths and weaknesses in the context of the ASAG system.

e) Impact of dimensionality reduction

We explore the impact of dimensionality reduction techniques, such as PCA [27], on the performance of BERT embeddings within the ASAG system.

## 6. EVALUATION

ASAG typically employs two primary metrics to assess system performance: quadratic weighted kappa (QWK) and mean absolute error (MAE) [2]. These metrics serve distinct purposes in evaluating the alignment between automated and human ratings. QWK is a measure of agreement between two raters, which, in this context, typically refers to the automated grading system and human evaluators [28]. The metric accounts for the degree of disagreement by assigning a quadratic penalty for larger differences in scores. It ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating random agreement. QWK is especially useful in ASAG as it provides a nuanced view of how closely the system replicates human judgment, considering not only exact matches but also the proximity of scores. QWK can be calculated using (1).

$$QWK = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \quad (1)$$

where:

- $W$  is a weight matrix where each element represents the penalty for each type of disagreement as in (2). The weights would increase quadratically with the distance.
- $O$  is a confusion matrix that shows the observed agreement.
- $E$  is an expected matrix that represents the agreement that would occur by chance.

$$w_{i,j} = \frac{(i-j)^2}{(n-1)^2} \quad (2)$$

where:

- $i$  is the human-assigned score.
- $j$  is the system-predicted score.
- $n$  is the number of score categories.

MAE calculates the average absolute difference between the scores assigned by the automated system and human raters [29]. It is expressed as in (3).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

where:

- $y_i$  is the human-assigned score.
- $\hat{y}_i$  is the system-predicted score.
- $n$  is the total number of answers.

MAE provides a straightforward assessment of the system's accuracy, emphasizing the magnitude of discrepancies without penalizing larger differences more heavily, as QWK does.

## 7. RESULTS AND DISCUSSION

This section presents the findings and discussions related to the implemented methods. It highlights the results obtained from the implementation process and examines their significance. Together, these discussions provide a comprehensive analysis of the outcomes.

## 8. RELATIVE EFFECTIVENESS OF BERT VARIANTS

This subsection presents the findings of our experiment aimed at identifying the optimal BERT model for our task. We evaluated nine different BERT model configurations, assessing their performance using MAE and QWK metrics. Table 3 provides a detailed comparison of the performance achieved by each model.

Based on Table 3, IndoBERT<sub>Lite-Base-P2</sub> is the best model with the highest QWK score of 0.0942, demonstrating better accuracy in understanding and assessing answers than other models. The embedding size and

number of layers in BERT models affect ASAG performance. Models with a large embedding size, such as BERT<sub>Base-Multilingual</sub> and IndoBERT<sub>Base-P1</sub> with an embedding size of 768, can capture complex text representations but have low QWK values (0.0835 and 0.0912), indicating non-optimality in understanding language context. In contrast, IndoBERT<sub>Lite-Large</sub>, with an embedding size of 128, is more efficient in capturing patterns relevant to ASAG, although the QWK improvement remains low (0.0842). The number of layers also matters; models with 24 layers, such as IndoBERT<sub>Lite-Large</sub> and IndoBERT<sub>Large-P1</sub>, can capture more context but only sometimes improve performance if the embedding size is small. Models with fewer layers, such as IndoBERT<sub>Lite-Base</sub> (12 layers), are more efficient in short text processing. The QWK improvement of IndoBERT<sub>Lite-Base-P2</sub> over IndoBERT<sub>Lite-Base-P1</sub> (0.0942 and 0.0638) suggests that additional training helps context understanding. Although more layers and large embedding sizes can increase the model's capacity, optimal computational performance and context understanding remain the main challenges in improving ASAG performance. Table 4 shows the effectiveness of the IndoBERT<sub>Lite-Base-P2</sub> model across various domains.

Table 4 indicates that text topic complexity significantly influences model performance. Models generally excel at evaluating more straightforward topics like lifestyle but struggle with more complex ones like politics. We must refine text processing techniques and enhance model training to optimize performance across varying text complexities. Future research will delve deeper into the factors contributing to these performance disparities between topics, enabling us to develop more robust and versatile ASAG systems.

Table 3. Performance comparison across various BERT models

Model	MAE	QWK
BERT <sub>Base-Multilingual</sub>	0.4117	0.0835
IndoBERT <sub>Base-P1</sub>	0.4121	0.0912
IndoBERT <sub>Base-P2</sub>	0.4121	0.0918
IndoBERT <sub>Large-P1</sub>	0.4116	0.0717
IndoBERT <sub>Large-P2</sub>	0.4116	0.0721
IndoBERT <sub>Lite-Base-P1</sub>	0.4114	0.0638
IndoBERT <sub>Lite-Base-P2</sub>	0.4115	0.0942
IndoBERT <sub>Lite-Large-P1</sub>	0.4111	0.0793
IndoBERT <sub>Lite-Large-P2</sub>	0.4112	0.0842

Table 4. The IndoBERT<sub>Lite-base-P2</sub> model effectiveness across various domains

Evaluation	Topics			
	Lifestyle	Politics	Sport	Technology
MAE	0.2533	0.5300	0.4351	0.4278
QWK	0.1327	0.0536	0.0977	0.0929

## 9. IMPACT OF TEXT PRE-PROCESSING

This section presents the results of our experiment, comparing the performance of the IndoBERT<sub>Lite-Base-P2</sub> model when applied to different text pre-processing combinations. The performance of each combination is evaluated and presented in Table 5. Based on Table 5, the IndoBERT<sub>Lite-Base-P2</sub> model without any text pre-processing exhibits the most optimal evaluation results. It achieves the lowest MAE value of 0.3540, significantly outperforming models with various pre-processing combinations (MAE values ranging from 0.4113 to 0.4116). This finding suggests that aggressive pre-processing techniques, namely stop word removal, stemming, and lemmatization, can inadvertently strip away crucial contextual information. While pre-processing can improve data quality by reducing noise and word variability, it is essential to strike a balance to preserve semantic meaning. In this case, the model benefits from the raw, unprocessed data, allowing it to capture nuanced linguistic cues.

Table 5. Comparison of performance for text pre-processing combinations

Combination of text pre-processing	MAE	QWK
Without pre-processing	0.3540	-0.1208
Tokenization, remove stop words, stemming	0.4115	0.0819
Tokenization, case folding, remove punctuation, stemming	0.4115	0.0819
Tokenization, remove stop words, remove punctuation, lemmatization	0.4115	0.0824
Tokenization, remove punctuation, lemmatization, stemming	0.4113	0.0332
Tokenization, case folding, remove stop words, lemmatization	0.4116	0.0821
Tokenization, case folding, remove stop words, remove punctuation, stemming, lemmatization	0.4115	0.0942

The table shows that the QWK score for the scenario involving tokenization, punctuation removal, lemmatization, and stemming is relatively low at 0.0332. This lower performance can be attributed to the absence of essential pre-processing steps like case folding and stop word removal. With case folding, the model can process text with consistent capitalization, positively impacting its ability to recognize word variations and their semantic meanings. Additionally, the stop word removal removes numerous uninformative words (e.g., “*dan*”, “*yang*”, “*di*”) that can boost the model’s focus on the core semantic content. The combined effect of these factors limits the model’s effectiveness in accurately evaluating the text, ultimately leading to a lower QWK score.

## 10. IMPACT OF HAND-ENGINEERED FEATURES

This section presents the results of our experiment, comparing the performance of the IndoBERT<sub>Lite-Base-P2</sub> model with and without the incorporation of hand-engineered features. The performance of both models is evaluated and presented in Table 6. Table 6 compares the IndoBERT<sub>Lite-Base-P2</sub> model’s performance with and without hand-engineered features. While both models exhibit similar MAE values (0.3534 with hand-engineered features and 0.4115 without hand-engineered features), a significant difference emerges in the QWK metric. The model incorporating hand-engineered features achieves a substantially higher QWK score of 0.2993 than without features, which is 0.0942. The hand-engineered features employed in this study include prompt overlap, lexical overlap, grammar error ratio, average sentence length, average word length, and answer length ratio. These features serve as additional linguistic cues, providing the BERT model with valuable information to enhance its decision-making process.

Table 6. Performance impact of hand-engineered features on IndoBERT<sub>Lite-base-P2</sub>

Model	MAE	QWK
IndoBERT <sub>Lite-Base-P2</sub> with hand engineered features	0.3534	0.2993
IndoBERT <sub>Lite-Base-P2</sub> without hand-engineered features	0.4115	0.0942

## 11. EFFECTIVENESS COMPARISON OF TF-IDF AND BERT

This section presents the results of our experiment, which compared the performance of our proposed model to a traditional TF-IDF model. The performance of both models is evaluated and presented in Table 7.

Table 7. Performance comparison: TF-IDF vs. IndoBERT<sub>Lite-base-P2</sub>

Model	MAE	QWK
IndoBERT <sub>Lite-Base-P2</sub>	0.4115	0.0942
TF-IDF	0.3564	0.0249

Table 7 reveals that while the TF-IDF model has a lower MAE value (0.3564) compared to IndoBERT<sub>Lite-Base-P2</sub> (0.4115), indicating better grading accuracy, the latter significantly outperforms TF-IDF in terms of QWK (0.0942 vs. 0.0249). Dense representations, like those employed by IndoBERT<sub>Lite-Base-P2</sub>, leverage transformer architecture to capture intricate word relationships and contextual nuances. This mechanism enables the model to better understand the semantic meaning of text compared to sparse representations like TF-IDF, which rely solely on word frequencies. While TF-IDF excels at minimizing grading errors, it struggles to grasp complex linguistic patterns and semantic relationships. In contrast, despite a slightly higher MAE, IndoBERT<sub>Lite-Base-P2</sub> demonstrates a superior ability to align system grading scores with actual scores, as evidenced by the higher QWK score. The discrepancy between MAE and QWK might be attributed to factors such as uneven data distribution or the presence of outliers. Dense representations like BERT embeddings prove more effective for ASAG tasks based on semantic similarity due to their superior ability to capture complex contextual relationships and meanings.

## 12. IMPACT OF DIMENSIONALITY REDUCTION

This section presents the results of our experiment, comparing the performance of the IndoBERT<sub>Lite-Base-P2</sub> model when using PCA and when not using PCA. The performance of both models is evaluated and presented in Table 8. Applying PCA to the IndoBERT<sub>Lite-Base-P2</sub> model reduced MAE from 0.4115 to 0.3584, indicating improved grading accuracy. However, this came at the cost of a significant decrease in QWK from 0.0942 to 0.0118. This finding suggests that dimensionality reduction using PCA while improving accuracy, may compromise the model’s ability to maintain consistent judgments relative to actual labels.

Table 8. Impact of PCA on IndoBERTLite-base-P2 model performance

Model	MAE	QWK
IndoBERT <sub>Lite-Base-P2</sub>	0.4115	0.0942
IndoBERT <sub>Lite-Base-P2</sub> with PCA	0.3584	0.0118

### 13. CONCLUSION

This study examines the influence of various components-namely BERT model variants, text pre-processing strategies, hand-engineered features, sparse representations, and dimensionality reduction-on the effectiveness of an ASAG system grounded in semantic similarity. While differences in BERT model selection do lead to some performance variation, the results indicate that meticulous text pre-processing and the inclusion of carefully crafted linguistic features have a more substantial effect on improving grading accuracy. The implementation of comprehensive pre-processing stages, including tokenization, normalization, and stop word removal, consistently enhanced the system's performance. Moreover, the addition of hand-engineered features, such as lexical diversity, semantic depth, and grammatical characteristics, further improved the system's ability to replicate human-like evaluations by enriching its linguistic comprehension. However, although dimensionality reduction techniques such as PCA reduce data complexity and aid numerical prediction, they occasionally introduce inconsistencies in grading outcomes, suggesting a need for balance when applying such methods.

Looking ahead, there are several promising directions to enhance ASAG systems. One such direction involves developing adaptive BERT models tailored to diverse topics and linguistic complexities. This approach involves either fine-tuning existing pre-trained models with domain-specific data or applying transfer learning strategies to adapt their general language understanding for specialized tasks. Additionally, future systems should move beyond semantic similarity alone by incorporating syntactic and pragmatic features, allowing for a more comprehensive and nuanced assessment of student responses. Another prospective path involves multi-modal approaches that integrate semantic, syntactic, and practical analyses with sophisticated representation techniques to create more robust and balanced grading frameworks. Advancements in these areas will help build ASAG systems that are not only more accurate and consistent but also better aligned with human evaluative standards, ultimately enhancing the quality and reliability of automated assessment in educational contexts.

### ACKNOWLEDGEMENTS

The authors would like to thank the Research and Community Service Unit (LPPM) of Institut Teknologi Del for their sponsorship and financial support. Furthermore, special acknowledgment is given to Rosni Lumbantoruan and Susi Eva Maria Purba for their valuable feedback on this work.

### FUNDING INFORMATION

This research was funded by the Research and Community Service Unit of Institut Teknologi Del. The funding agency had no role in the study design, data collection and analysis, decision to publish, or manuscript preparation.

### AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Samuel Situmeang	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓
Sarah Rosdiana	✓	✓		✓	✓	✓				✓		✓		
Tambunan														
Lidia Ginting			✓		✓			✓	✓		✓			
Wahyu Krisdangolyanti			✓		✓	✓		✓	✓		✓			
Simamora														
Winda Sari ButarButar			✓		✓			✓	✓		✓			

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : Writing - **O**riginal Draft

E : Writing - Review & **E**ditting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition



## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## DATA AVAILABILITY

The data generated and analyzed during this study are available from the corresponding author upon reasonable request.




## REFERENCES

- [1] L. Yan *et al.*, "Practical and ethical challenges of large language models in education: a systematic scoping review," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90–112, Jan. 2024, doi: 10.1111/bjet.13370.
- [2] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, Mar. 2022, doi: 10.1007/s10462-021-10068-2.
- [3] S. Haller, A. Aldea, C. Seifert, and N. Strisciuglio, "Survey on automated short answer grading with deep learning: from word embeddings to transformers," vol. 1, no. 1, Mar. 2022, [Online]. Available: <http://arxiv.org/abs/2204.03503>.
- [4] Y. Kumar, S. Aggarwal, D. Mahata, R. R. Shah, P. Kumaraguru, and R. Zimmermann, "Get it scored using autosas - An automated system for scoring short answers," *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, vol. 33, no. 01, pp. 9662–9669, Jul. 2019, doi: 10.1609/aaai.v33i01.33019662.
- [5] S. Hassan, A. A. Fahmy, and M. El-Ramly, "Automatic short answer scoring based on paragraph embeddings," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 397–402, 2018, doi: 10.14569/IJACSA.2018.091048.
- [6] W. H. Goma and A. A. Fahmy, *Ans2vec: a scoring system for short answers*, vol. 921. Springer International Publishing, 2020.
- [7] P. Shweta and K. Adhiya, "Comparative study of feature engineering for automated short answer grading," in *Proceedings - 2022 IEEE World Conference on Applied Intelligence and Computing, AIC 2022*, Jun. 2022, pp. 594–597, doi: 10.1109/AIC55036.2022.9848851.
- [8] C. Cai, "Automatic essay scoring with recurrent neural network," in *ACM International Conference Proceeding Series*, Mar. 2019, pp. 1–7, doi: 10.1145/3318265.3318296.
- [9] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang, "An automatic short-answer grading model for semi-open-ended questions," *Interactive Learning Environments*, vol. 30, no. 1, pp. 177–190, Jan. 2022, doi: 10.1080/10494820.2019.1648300.
- [10] U. K. Chakraborty and A. Mishra, "Automatic short answer grading using a LSTM based approach," in *Proceedings - 2023 IEEE World Conference on Applied Intelligence and Computing, AIC 2023*, Jul. 2023, pp. 332–337, doi: 10.1109/AIC57670.2023.10263899.
- [11] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 5999–6009, 2017.
- [12] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12164 LNAI, 2020, pp. 43–48.
- [13] M. Thakkar, "Finetuning transformer models to build ASAG system," *Arxiv*, Sep. 2021, [Online]. Available: <http://arxiv.org/abs/2109.12300>.
- [14] X. Zhu, H. Wu, and L. Zhang, "Automatic short-answer grading via BERT-based deep neural networks," *IEEE Transactions on Learning Technologies*, vol. 15, no. 3, pp. 364–375, 2022, doi: 10.1109/TLT.2022.3175537.
- [15] D. H. Alhamed, A. M. Alajmi, T. A. Alqahtani, Y. H. Alali, M. R. Alnassar, and D. A. Alabbad, "iGrade: an automated short answer grading system," in *ACM International Conference Proceeding Series*, Dec. 2022, pp. 110–116, doi: 10.1145/3582768.3582790.
- [16] H. R. Salim, C. De, N. D. Pratamaputra, and D. Suhartono, "Indonesian automatic short answer grading system," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 11, no. 3, pp. 1586–1603, Jun. 2022, doi: 10.11591/eei.v11i3.3531.
- [17] S. I. G. Situmeang, R. M. G. T. Sihite, H. Simanjuntak, and J. Amalia, "A deep learning-based regression approach to Indonesian short answer grading system," in *ACM International Conference Proceeding Series*, Oct. 2023, pp. 201–209, doi: 10.1145/3626641.3626929.
- [18] I. D. Mardini G, C. G. Quintero M, C. A. Vilorio N, W. S. Percybrooks B, H. S. Robles N, and K. Villalba R, "A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions," *Education and Information Technologies*, vol. 29, no. 4, pp. 4565–4590, Mar. 2024, doi: 10.1007/s10639-023-11890-7.
- [19] C. Park *et al.*, "ThisIsCompetition at SemEval-2019 task 9: BERT is unstable for out-of-domain samples," *NAACL HLT 2019 - International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop*, vol. 1, no. 1, pp. 1254–1261, 2019, doi: 10.18653/v1/s19-2220.
- [20] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. ntlk.org. O'Reilly Media, Inc., 2023.
- [21] H. Chimingyang, "An automatic system for essay questions scoring based on LSTM and Word Embedding," in *Proceedings - 2020 5th International Conference on Information Science, Computer Technology and Transportation, ISCTT 2020*, Nov. 2020, pp. 355–364, doi: 10.1109/ISCTT51595.2020.00068.
- [22] A. Filighera, T. Steuer, and C. Rensing, "Fooling automatic short answer grading systems," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12163 LNAI, 2020, pp. 177–190.
- [23] F. Rahutomo *et al.*, "Open problems in Indonesian automatic essay scoring system," *International Journal of Engineering & Technology*, vol. 7, no. 4.44, p. 156, 2018, doi: 10.14419/ijet.v7i4.44.26974.
- [24] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, vol. 1, pp. 4171–4186.




- [25] B. Wilie *et al.*, “IndoNLU: benchmark and resources for evaluating Indonesian natural language understanding,” *arXiv*, Sep. 2020.
- [26] A. Aizawa, “An information-theoretic perspective of TF-IDF measures,” *Information Processing and Management*, vol. 39, no. 1, pp. 45–65, Jan. 2003, doi: 10.1016/S0306-4573(02)00021-3.
- [27] G. Marentakis and J. Hözl, “Compression efficiency and signal distortion of common PCA bases for HRTF modelling,” *Proceedings of the Sound and Music Computing Conferences*, vol. 2021-June, pp. 60–67, 2021.
- [28] N. A. Kurdhi and A. Saxena, “Evaluating quadratic weighted Kappa as the standard performance metric for automated essay scoring,” *International Educational Data Mining Society*, no. July, pp. 103–113, 2023.
- [29] A. Prabhudesai and T. N. B. Duong, “Automatic short answer grading using siamese Bidirectional LSTM based regression,” in *TALE 2019 - 2019 IEEE International Conference on Engineering, Technology and Education*, Dec. 2019, pp. 1–6, doi: 10.1109/TALE48000.2019.9226026.

## BIOGRAPHIES OF AUTHORS






**Samuel Situmeang**    received his bachelor's degree in the Information Technology Study Program, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Indonesia, in 2014. He received his master's degree from the Department of Computer Science and Information Engineering, College of Engineering, National Chung Cheng University, Taiwan, in 2017. He currently lectures in the Information System Study Program, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del. His main research interest is AI-driven solutions for education and cultural heritage, which intersect areas of data science, natural language processing, and computer vision. He can be contacted at email: samuel.situmeang@del.ac.id.






**Sarah Rosdiana Tambunan**    received her bachelor's degree in the Information System Study Program, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Indonesia, in 2019. She received her master's degree from the Computer Sciences Master Program, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia, in 2023. She currently lectures in the Information System Study Program, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del. Her research focuses on software engineering and data security. She can be contacted at email: sarah.tambunan@del.ac.id.






**Lidia Ginting**    her bachelor's degree in the Information System Study Program, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Indonesia, in 2024. Her research interests are natural language processing, data mining, and education technology. She can be contacted at email: lidiaginting243@gmail.com.



**Wahyu Krisdangolyanti Simamora**    her bachelor's degree in the Information System Study Program, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Indonesia, in 2024. Her research interests are natural language processing, data mining, and education technology. She can be contacted at email: wahyukrisdangolyantisimamora@gmail.com.



**Winda Sari ButarButar**    her bachelor's degree in the Information System Study Program, Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, Indonesia, in 2024. Her research interests are natural language processing, data mining, and education technology. She can be contacted at email: windadrini26@gmail.com.