Review of Computer Engineering Research
2026 Vol. 13, No. 3, pp. 200-211
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v13i3.5148

A hybrid embedding approach for short-answer grading using BERT and universal sentence encoder

Chandralika Chakraborty[1+]
Atowar ul Islam[2]
Bhairab Sarma[3]

[1]*Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, India.*
[1]*Email: chandralika.c@gmail.com*
[2]*University of Science and Technology Meghalaya, Meghalaya, India.*
[2]*Email: atowar91626@gmail.com*
[3]*Assam Royal Global University, Guwahati, Assam, India.*
[3]*Email: sarmabhairab@gmail.com*


(+ Corresponding author)

Article History
Received: 16 January 2026
Revised: 20 April 2026
Accepted: 2 June 2026
Published: 4 September 2026

Keywords
Automatic short-answer grading
Bidirectional encoder representations from transformers
Cosine similarity
Educational data mining
Text embeddings
Universal sentence encoder.

ABSTRACT

Evaluating learner responses is vital in assessment, especially for free-text answers, which are most challenging due to their demand for reasoning and expression. These responses reflect the learner's understanding of the subject matter. However, manual evaluation is time-consuming and also prone to human error, highlighting the need for automation. Automatic Short-Answer Grading (ASAG) is a rapidly advancing area within the domain of educational technology, focusing on automating the evaluation process of student responses. Although numerous approaches have been proposed in recent years, the challenge of identifying the most accurate and reliable method for short-answer grading still remains an open research question. This work investigates the use of advanced natural language processing techniques, specifically text embeddings generated by Bidirectional Encoder Representations from Transformers (BERT) and the Universal Sentence Encoder (USE), for automatic grading of short answers. Furthermore, the work explores the possibilities for identifying the dependency of the method on data and proposes a novel hybrid model that appropriately evaluates student responses. A sample of 500 student responses was extracted from the 1,672 set-1 records of the publicly available HP: Short Answer Scoring (SAS) dataset. By integrating BERT and USE models to measure semantic and contextual similarity against reference answers, the study automated the scoring process. This hybrid approach achieves an accuracy of 0.554 and a quadratic weighted Kappa score of 0.999. The high coefficient reflects categorical agreement success through threshold-based binning, rather than granular decimal precision. The results are also compared with existing benchmarks established using the HP: SAS dataset.

Contribution/Originality: This study contributes to the existing literature by exploring the use of BERT and USE text embeddings in grading short answers. It offers new insights into the generation of automated scores for the student answers and enhances understanding of the scoring process through an empirical study, making responses more interpretable.

1. INTRODUCTION

Educational assessment in the digital age has observed a rise in the combination of Artificial Intelligence (AI) and Machine Learning (ML) for improving efficiency and scalability. A promising application in this domain is Automatic Short-Answer Grading (ASAG), which makes use of a machine (i.e., a computer) to evaluate a student's short text response with consistency and accuracy [1, 2]. Automating the grading process significantly lessens the workload of educators, thus letting them focus on many other critical teaching activities [2]. ASAG provides

consistent and objective grading, maintaining uniformity, unlike human graders who are at times influenced by fatigue or bias – thereby improving reliability [3]. Such systems enable students to receive immediate feedback on their responses, leading to improvement in learners' understanding in real time. Large-scale assessments are possible using ASAG, where manual grading seems impractical.

Traditional ASAG systems employed rule-based algorithms and statistical models to compare student responses with reference answers [2, 4]. Some reported works compared student answers with reference answers using not only keywords but also the preceding and following words [5]. These approaches also included stop words in the comparison. However, these methods are unable to capture the deeper meaning of the responses, thus limiting their effectiveness in grading varied student responses [6]. ASAG has evolved over time, transitioning from rule-based approaches to machine learning techniques, and more recently to deep learning architecture. The early rule-based methods relied on manually defined linguistic patterns for evaluating student responses [2]. These methods were interpretable, but they lacked handling the diverse language or information expression.

Subsequently, statistical and ML-based approaches like Latent Semantic Analysis (LSA), Support Vector Machines (SVMs) were used for ASAG, resulting in improved response similarities, but these approaches had difficulty in generalizing across different question types and subject areas. The systematic review of Galhardi and Brancher highlights the growing interest and advancements in applying Machine Learning techniques to automatic short-answer grading [7]. The authors suggest that future research in ASAG should focus on new deep learning architectures and create datasets to increase the robustness and applicability of ASAG systems.

Recent years have seen significant progress in the promotion of Natural Language Processing (NLP), which has significantly enhanced ASAG performance. ASAG utilizes computational techniques to evaluate open-ended free text responses. Linguistic variations, synonyms, and differences in sentence structure are some of the challenges ASAG has to take care of Roy, et al. [8]. The fundamental challenge is to obtain a correct score for a student's response even when it is phrased differently from a model answer. One of the most significant advancements in enhancing the effectiveness of ASAG is the inception of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) [9], USE (Universal Sentence Encoder) [10], and GPT (Generative Pre-Trained Transformer) [11]. These models make use of contextualized text embeddings to capture the meaning of the responses, enabling precise query understanding and relevance ranking of text [12]. This makes them highly effective for grading short answers. Over the past few years, transformer-based models have revolutionized ASAG. Pre-trained models like BERT and USE have shown superior performance in capturing contextual information and providing acceptable grading of responses [9, 10]. These models encode text (words and sentences) into high-dimensional vectors, due to which the minute details of the text can be represented as contextual vectors for comparison between student responses and model answers for correct grading.

Unlike traditional bag-of-words models, transformers generate embeddings that preserve word relationships within sentences, improving semantic accuracy [13]. Pre-trained transformers can be fine-tuned [9] on diverse educational datasets, allowing for adaptability to different subjects and question types. In spite of these improvements, there are challenges faced by transformer-based models. Explainability being an area of concern, the use of an appropriate method for a given task would reduce computational overhead and result in efficient system design.

The exploration of automated approaches for grading short answers is also found to be addressed by educational data mining researchers [14]. Educational Data Mining (EDM) is an application of data mining in the educational field. It is used to classify, analyse, and predict students' academic performances for improving the teaching-learning process [15]. As the field of ASAG and EDM evolves, the need for scalable, high-precision tools has grown.

Prior research has explored ensemble ASAG systems, often pairing BERT with computationally intensive architectures like CNNs or LSTMs, while our work identifies and exploits a specific functional interaction between BERT and USE. Our analysis reveals a variation in performance: BERT demonstrates higher correlation with high-scoring human-rated responses, whereas USE exhibits higher correlation with low-scoring human-rated responses.

By integrating these two specific embeddings, the proposed model achieves balanced calibration across the entire score spectrum. Furthermore, the combination of BERT and USE remains underexplored in ASAG literature, offering a more computationally efficient alternative to traditional deep-learning ensembles.

The contribution of this work is:

- This work contributes to ASAG in particular and EDM in general by proposing an approach to automate the evaluation of short text responses.
- This research evaluates student responses through comparative analysis of embeddings of transformer-based BERT and Deep Averaging Network (DAN)-based model USE. After assessing the initial machine-generated scores, the work investigates the dependency of the methods on data for understanding how the model performance scales with varying input quality.
- Consequently, a robust grading framework is proposed that integrates a Transformer-based and Deep Averaging Network-based model to optimize short-answer scoring accuracy.

To guide this investigation, the following research questions are addressed:

Research Question 1 (RQ1): To what extent can standalone Transformer-based (BERT) and Deep Averaging Network (USE) models effectively grade short-answer responses without the combination of supplementary neural layers (e.g., CNN, LSTM, or RNN)?

Research Question 2 (RQ2): To what extent is a sample of 500 student responses representative and sufficient for evaluating the performance of a BERT-USE hybrid model using HP: SAS dataset (Question Set 1)?

The remainder of this paper is organized as follows. Section 2 discusses the literature survey. Section 3 provides the methodology, along with describing the data set and model building. Section 4 introduces the proposed model for automatic short-answer grading. Results and discussions are presented in Section 5. Finally, Section 6 concludes the paper and discusses limitations and future work.

2. LITERATURE SURVEY

The work on computer-assisted grading started in 1966 with the work of Page [16] where the author introduced Project Essay Grade (PEG), a pioneering system designed to assess essays using a statistical approach [16]. Experiments showed that computers could grade essays as reliably as human evaluators by analyzing linguistic and statistical features. The high correlation between machine and human scores indicated that automated grading is a valid alternative. This research became a foundation for advances in Natural Language Processing and machine learning. This paved the way for developing models that can understand and assess short text responses effectively.

A comprehensive survey by Burrows, et al. [2] reported the development in ASAG during the years that followed [2]. The authors identified and categorized the developments of ASAG into five different themes and six dimensions till 2015. The work identified five key themes: concept mapping, information extraction, corpus-based approaches, machine learning, and evaluation. It also highlighted six dimensions: Datasets, NLP, model development, grading mechanisms, model assessment, and overall effectiveness. The study also identified key trends such as increasing reliance on shared datasets for evaluation, the role of competitions like the Automated Student Assessment Prize (ASAP) in advancing and building new methodologies, and growing integration of linguistic and statistical features for improving system performance. The authors also highlighted that advancements in ASAG research are highly influenced by the availability of a standardized dataset.

Most traditional methods for short-answer grading relied on normally crafted regular expressions to recognize specific terms and notions in student responses, which can be time-consuming and limited in scope. The approach by Ramachandran, et al. [17] introduced a contrasting innovative approach by employing graph-driven lexico-semantic approaches for text matching to automatically extract meaningful patterns derived from rubric texts and top-performing student answers. Evaluation was performed on the Automated Student Assessment Prize Short Answer Scoring dataset, achieving a Quadratic Weighted Kappa (QWK) of 0.78. It also performed on the Mohler dataset,

with a Pearson correlation of 0.52 and Root Mean Square Error (RMSE) of 0.98, indicating effectiveness and generalizability across datasets.

Neural model-based approaches subsequently started gaining popularity with an increased number of applications. Riordan, et al. [18] explored the effectiveness of neural models like LSTMs, CNNs, and attention mechanisms for automated short-answer scoring on three datasets: the Automated Student Assessment Prize Short Answer Scoring dataset, Powergrading, and SRA. Key findings highlight the effectiveness of pretrained word embeddings, bidirectional LSTMs, and attention mechanisms in enhancing short answer scoring performance. Results may vary across datasets, which is an area the proposed method aims to address. Kumar, et al. [19] presented AutoSAS, a scalable automated short-answer scoring system that integrated various features such as Word2Vec, Doc2Vec, weighted keywords, and word frequency to improve grading accuracy. The authors assessed the system's performance using the Automated Student Assessment Prize Short Answer Scoring dataset and achieved an average Quadratic Weighted Kappa (QWK) score of 0.79. The system not only assigned a score to the responses but also provided valuable feedback to the users about the characteristics of a response. Süzen, et al. [20] explore Data Mining Techniques for automatic grading of answers to students of an introductory computer science class at the University of North Texas. The authors applied standard data mining methods to assess the similarity between student responses and model answers and provide useful feedback on answers. Hamming distance similarity measure was performed between the student answers and the model answer. Answers classified as correct and incorrect with high correlation between machine-generated and human-rated, using the Pearson correlation coefficient evaluation metric.

The study by Ndukwe, et al. [21] investigates the application of Sentence-BERT (SBERT), a pre-trained neural network language model for automatic grading of short-answer questions in an Introduction to Networking subject. The grading of answers in this study focused on three types of questions: Comparison, description, and listing. The short responses were a few words (20-30 words). Only one model answer was considered. On a scale of 0 to 5 (discrete values), the course instructor graded the answers manually, and a subset of these graded answers was used to fine-tune the SBERT model. Quadratic Weighted Kappa (QWK) was employed to assess the agreement between human and model-generated scores. The model predicts comparison questions with the highest accuracy, followed by descriptive questions, and listing questions with the lowest accuracy. Zhang, et al. [22] introduce a novel framework for automatic short-answer grading in mathematics. The authors highlight the limitations of existing ASAG models, which either lack adaptation to math questions or require separate models for each question, leading to inefficiencies. To address these issues, the research proposes a framework that fine-tunes MathBERT for mathematical content adaptation and incorporates in-context learning for developing an ASAG approach using multi-task and meta-learning tools. The work demonstrates that meta in-context learning results in highly generalizable automated scoring models by explicitly using example responses and scores as input, facilitating easier learning for the model. Evaluations on student response datasets show that this approach significantly outperforms existing models but does not surpass standard BERT, indicating the need for improved mathematical language models.

A foundational pilot work of the first author [23] proposes the use of a language model, Universal Sentence Encoder (USE), and establishes the feasibility of using the model for ASAG. Their experiments demonstrated that the generated scores aligned with human-graded scores, and this approach is effective for sizeable datasets, suggesting its reliability for ASAG tasks. Bilal and Almazroi [24] investigate the efficacy of fine-tuning the BERT model to classify online customer reviews as helpful and unhelpful. The work addresses the problem of information overload on review platforms and proposes a generalized approach that takes advantage of BERT's deep learning capabilities. Experiments conducted on Yelp shopping reviews demonstrated that fine-tuned BERT classifiers notably surpassed bag-of-words methods in differentiating between helpful and unhelpful reviews. The study reveals that the sequence length used in the BERT classifier impacts performance significantly, highlighting the importance of this parameter in NLP tasks. Jiang and Bosch [25] investigate the use of the GPT model, gpt-4-1106-preview, on the Automated Student Assessment Prize Short Answer Scoring dataset for short-answer scoring. The work evaluates different

prompt configurations, such as including correct answers, scoring examples, and rationale generation before assigning a score. A quadratic weighted kappa (QWK) of 0.677 was achieved for the best-performing setup, indicating substantial agreement with human scoring. This research is part of a larger project dedicated to automating short-answer evaluation. While our previous study [26] focused on the individual performance of BERT and USE, the current work introduces an integrated hybrid model to address precision in automated assessment. Furthermore, the empirical analysis extends beyond raw similarity scores by implementing normalization and discretization techniques to ensure machine-generated outputs better align with human-graded rubrics. These pedagogical insights and structural refinements formally contextualize the current work within the domain of Educational Data Mining.

While different approaches report successful implementations, the core question of the choice of the most effective method still remains unanswered. The users' choice is still largely limited to a black-box approach, and a data-centric decision is found by counting.

3. METHODOLOGY

To implement the research work focused on grading short answers, the following steps were undertaken.

i) Selection of student responses and corresponding model answers using a publicly available dataset.
ii) Conduction of initial experiments to develop a suitable model.
iii) Identifying evaluation metrics to assess the effectiveness of the proposed model.
iv) Proposing a method to compute the degree of similarity and generate machine-assigned scores, which can then be compared with the scores assigned by human evaluators.

The following sub-sections provide a detailed description of the steps outlined above.

3.1. Data Set and Sampling Strategy

The dataset used in this work is a publicly available one – HP: Short Answer Scoring dataset [27]. It consists of 17,198 records of student responses, and is stored in a tab-separated value (TSV) format containing the following fields.

i) Student ID: A unique number.
ii) Question set number: Sets are numbered from 1 to 10.
iii) Two human-evaluated scores: Scores range from 0 to 3 (discrete values).
iv) Student's textual response: Responses are in ASCII text.

The work reported in this paper used student responses from question set '1'. All responses are written by tenth-grade students. For experimentation, a sample of 500 student responses was selected from question set '1', along with five full-scoring answers (score of 3), designated as model answers. These model answers can be replaced by answers drafted by human evaluators. The student responses are evaluated by two human raters with scores 0, 1, 2, and 3, with an inter-evaluator correlation value of 0.943. The sample responses have an average length of 60 words per response. A subset of 500 records was extracted from a total of 1,672 available records using a stratified sampling approach. As explored in RQ2, this sample is a reliable representation of the full Question Set-1 records. The frequency of human-assigned scores (0, 1, 2, and 3) within the sample closely mirrored the original distribution, with percentage differences ranging from -3.7% to +2.9%, as shown in Table 1.

Table 1. Comparative Score Distribution of HP: SAS Question Set 1 vs. Experimental Sample.

Score	HP: SAS Question Set 1 (N=1,672)	Experimental Sample (N=500)	Difference in percentage
0	380 (22.7%)	121 (24.2%)	+1.5%
1	429 (25.7%)	143 (28.6%)	+2.9%
2	524 (31.3%)	138 (27.6%)	-3.7%
3	339 (20.3%)	98 (19.6%)	-0.7%

3.2. Experimentation and Model Building

Initial experiments were conducted using transformer-based Bidirectional Encoder Representations from Transformers (BERT) and Deep Averaging Network (DAN) model Universal Sentence Encoder (USE) separately to predict the scores for student responses. The DAN model Universal Sentence Encoder is applied in this work as it is less computation-intensive with marginal loss of accuracy [10]. The initial model structure using BERT and USE separately is illustrated in Figure 1. After identifying five full scoring model answers and five hundred student answers from question set 1, each is vectorized using BERT and USE models separately. Vector comparison of each student's answer vector with the model answer vectors is performed using cosine similarity. After comparing the vectors, the predicted score is normalized according to the algorithm below. The normalized score is then classified as low, medium, or high.

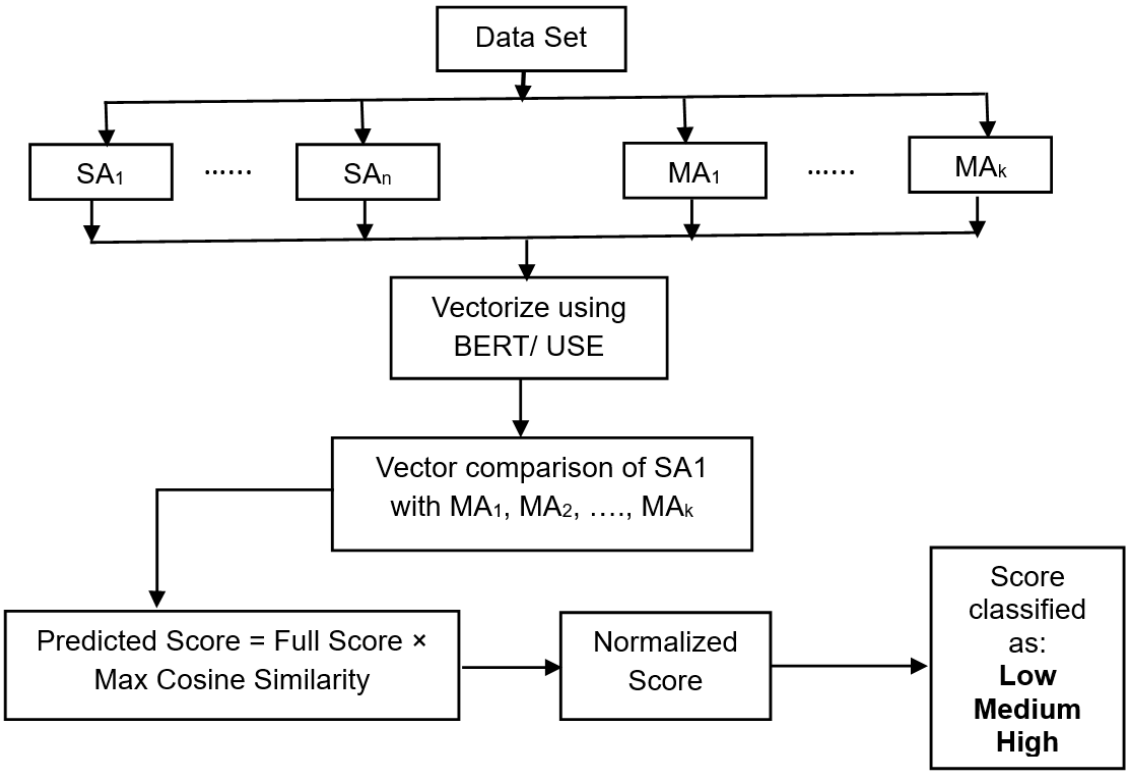

Figure 1. Initial model using BERT and the USE model separately.

Note: SA: Student Answer; MA: Model Answer.
n: Number of student answers; k: Number of model answers.

A comparison of the predicted scores, as shown in Table 2, reveals that BERT performs better for high-scoring responses, while the USE model is more effective for low-scoring responses. Based on this analysis, the BERT model can be used for predicting high scores, while the USE model is better suited for low scores. This evaluation was possible because the dataset contained scores assigned by human graders. However, for unlabeled datasets, distinguishing between high- and low-scoring responses beforehand becomes challenging, making it difficult to determine which model to use. To address this limitation, the authors propose a hybrid approach that integrates both BERT and USE models for score prediction of responses, as detailed in the next section. The comparison of student answer vectors with model answer vectors generates the predicted score according to Equation 1.

$$\text{Predicted Score} = \text{Max}\left[f_{CS}(SA_j, MA_1), f_{CS}(SA_j, MA_2) \ldots\ldots f_{CS}(SA_j, MA_k)\right] \times \text{Full Score} \quad (1)$$

Where f_{CS} denotes the function computing Cosine Similarity.

The predicted score generated above, however, is precise up to three decimal places, which does not match standards practiced by human evaluators. Under normal circumstances, human-generated scores are in multiples of 0.5.

Hence, the predicted score generated by the ASAG model has been normalized using a process f_{norm}(Predicted Score), which is as shown below.

Algorithm for Normalized Score: f_{norm}(Predicted Score)

Input: Predicted Score

Output: Normalized Score

1. Begin
2. Initialize normalized valid scores as V = [0.0, 0.5, 1, 1.5, 2, 2.5, 3.0]
3. Index =0; Diff = Max Score
4. While (Index <= length(V))
   - 4.1 If (|Predicted Score – V[Index]|) < Diff
     - 4.1.1 Diff = |Predicted Score – V[Index]|
     - 4.1.2 NormIndex = Index
   - 4.2 EndIf
   - 4.3 Index ++
5. EndWhile
6. Normalized Score = V[NormIndex]
7. End

Table 2. Comparison of BERT and USE models predicted scores.

Sl. No.	Student ID	Human Evaluator Score 1	Human Evaluator Score 2	Max of Human Evaluator	BERT predicted Score	USE predicted Score
1	16	0	0	0	1.13	0.17
2	43	3	3	3	2.8	2.1
3	63	1	1	1	1.65	0.95
4	77	3	3	3	2.77	2.29
5	120	2	3	3	2.75	1.88
6	124	1	1	1	1.62	0.58
7	212	0	0	0	0.68	0.37
8	225	0	0	0	0.12	0.12
9	233	0	0	0	1.24	0.46
10	385	0	0	0	1.29	0.34
11	387	0	0	0	0.76	0.26
12	390	3	3	3	2.75	2.26
13	449	3	3	3	2.75	2.02

3.3. Evaluation Metrics

Various evaluation metrics are utilized to measure the effectiveness of the proposed ASAG model. The metrics include correlation coefficients, Root Mean Square Error (RMSE), accuracy, precision, recall, F1-score, and Quadratic Weighted Kappa (QWK). The Spearman correlation coefficient is used to measure the agreement between machine-generated scores and human scores. RMSE quantifies the error in score generation. Accuracy, precision, recall, and F1-score evaluate the classification performance across different score categories. The score categories used in this work are Low, Medium, and High. These categories result from discretizing continuous outputs into ordinal categories. QWK is a metric to assess the model's reliability, as it accounts for the degree of agreement between generated and human-graded scores. The integration of BERT and USE models demonstrates improvements in the performance of low-scoring and high-scoring responses, as shown in Table 3.

4. PROPOSED ASAG MODEL

The proposed model for automatic short-answer grading, as shown in Figure 2, integrates both BERT and USE models to address the limitation of model selection in unlabeled datasets, as mentioned earlier. The approach involves parallel prediction of responses using both models, followed by averaging the normalized scores to determine the final score. This method enables the system to leverage BERT's strength in predicting high scores and USE's effectiveness for low scores without prior knowledge of the score distribution. The proposed model design ensures robust performance, as demonstrated in Table 3.

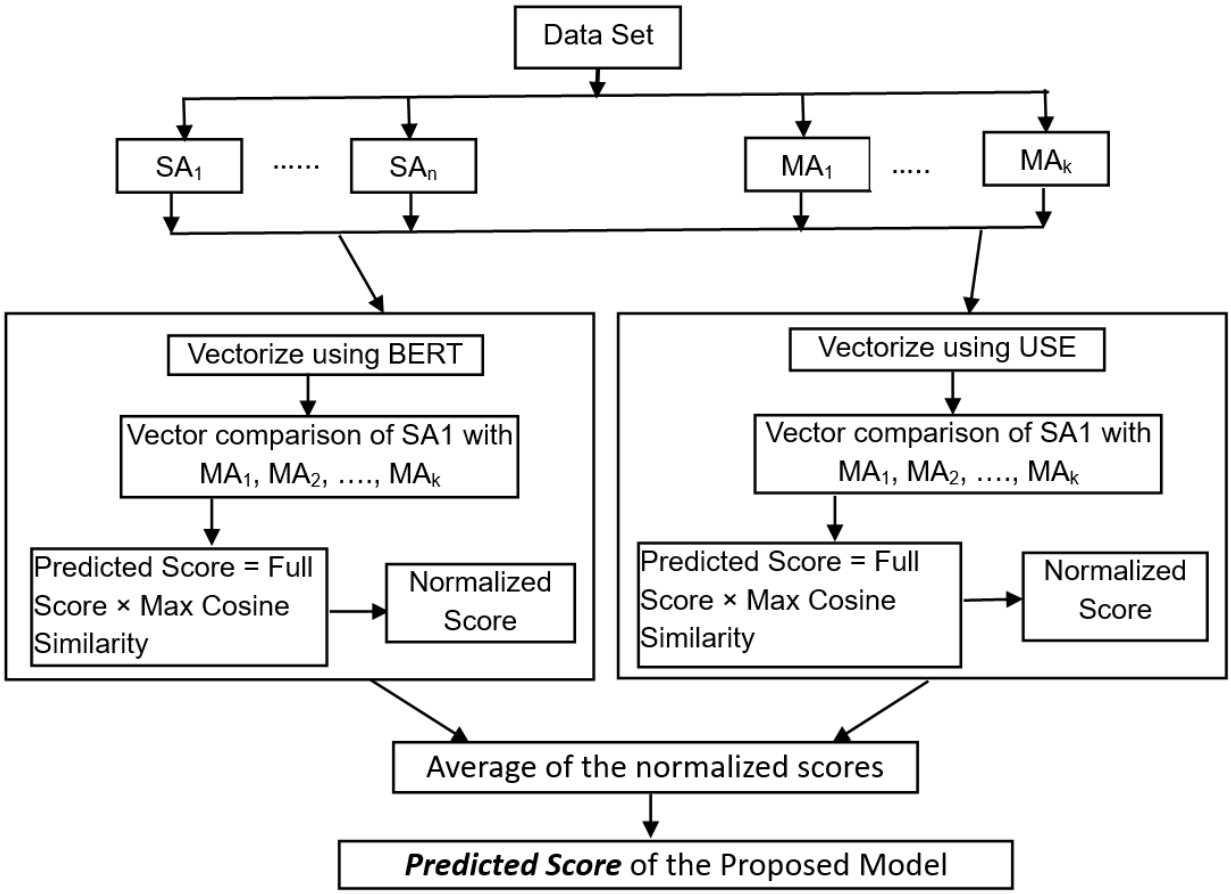

Figure 2. Proposed Model for predicting scores of student responses.

5. RESULTS AND DISCUSSIONS

The machine-generated scores using BERT and USE models separately align with human-rated scores, with Spearman Correlation coefficients of 0.515 and 0.583, respectively, thus addressing RQ1. The analysis of the individual models, BERT and USE, revealed that BERT-generated scores had a higher correlation with human evaluators for high-scoring answers. Conversely, USE showed a higher correlation for low-scoring answers. The proposed method, integrating BERT and USE models, is more balanced as it can handle the entire spectrum of scores more reliably, as shown in Table 3. It was also observed that BERT-based similarity scoring emphasizes the presence of important keywords from the model answers. The model tends to assign higher scores to such student answers, even if the sentences are poorly structured or contain grammatical errors. In contrast, USE focuses more on the overall semantic structure of responses, resulting in higher scores than BERT-based scores for such responses.

Table 3. Evaluation metrics for BERT and USE text embeddings and the proposed method.

ASAG Model using	Correlation Spearman Coeff.	RMSE	Accuracy	Precision	Recall				F1	QWK
					Low	Medium	High	Overall		
BERT embedding	0.515	0.952	0.514	0.598	0.034	0.115	0.967	0.372	0.459	0.998
USE embedding	0.583	0.713	0.576	0.626	0.163	0.468	0.833	0.488	0.549	0.999
Integration of BERT & USE model (Proposed)	0.599	0.768	0.554	0.610	0.060	0.317	0.922	0.433	0.506	0.999

Table 4. Comparison of the proposed method with one of the recent works based on the accuracy evaluation metric (Utilizing the HP: SAS dataset).

Automatic Short-Answer Scoring Model reference	Accuracy
Proposed Method (Integration of BERT & USE model)	0.554
Jiang and Bosch [25]	0.530

Table 5. Comparison of the proposed method with prior works utilizing the HP: SAS dataset.

Automatic Short-Answer Scoring Model reference	Model used		Accuracy	QWK	Remarks
	BERT	USE			
Ramachandran, et al. [17]	No	No	NA	0.78	
Riordan, et al. [18]	No	No	NA	0.72	
Kumar, et al. [19]	No	No	NA	0.79	
Jiang and Bosch [25]	No	No	0.53	0.660	
Proposed Method (Integration of BERT & USE model)	Yes	Yes	0.554	0.999	Integration of BERT and USE

Source: Barbara, et al. [27].

The accuracy of the proposed model is higher than that of previous work using the GPT-4 model on the same dataset, as shown in Table 4. Table 5 presents the utilization of BERT and USE models in prior studies involving the HP: Short Answer Scoring dataset. It indicates that the works listed in the table did not incorporate BERT or USE in their experiments. After reviewing the comparison in Table 5, the proposed model displayed a higher QWK metric of 0.999, indicating a greater agreement between machine-predicted scores and human raters than previous works. This elevated QWK score reflects strong categorical agreement between human and machine scores, as the metric rewards proximity within ordinal scales rather than exact point-to-point identity. The three-category binning used for calculating the score is shown in Table 6.

Table 6. Three-Category binning.

Threshold range for machine-generated scores	Category (Bin)	Human Score Equivalent
[0.00, 0.99]	Low	0
[1.00, 1.99]	Medium	1
[2.00, 3.00]	High	2 & 3

Table 7 shows the predicted scores of the proposed method compared to the predicted scores of the system using BERT and USE models separately. Values in Column (1) are the maximum scores of the two human-evaluated scores from the dataset [27]. The category of this score is mentioned in Column (2). Columns (3) and (4) are the normalized scores determined using BERT and USE models, respectively. (5) is the average of the BERT and USE model scores. The normalized score of (5) is in Column (6), which is the predicted score of the proposed method. Column (7) represents the category of the predicted score.

Table 7. Predicted Scores (Classified as Low, Medium, and High) against Maximum Human scores. (Few samples shown below).

(1) Max of Human Evaluator	(2) Human Rater_Category	(3) BERT Normalized Score	(4) USE Normalized Score	(5) Avg Mc Score	(6) (Predicted Score) Normalized Avg Score	(7) Avg Norm Sc_Category
1	Medium	2.5	2.5	2.48	2.5	High
1	Medium	2.5	1.5	2.125	2	High
1	Medium	2	1.5	1.915	1.75	Medium
2	High	2.5	2	2.205	2.25	High
1	Medium	2.5	2	2.08	2.25	High
1	Medium	2.5	2	2.24	2.25	High
3	High	1.5	2	1.905	1.75	Medium
3	High	2.5	2	2.085	2.25	High
2	High	2.5	2	2.215	2.25	High
3	High	2.5	2	2.36	2.25	High
2	High	2.5	2	2.28	2.25	High
2	High	1.5	1.5	1.52	1.5	Medium
0	Low	1	0.5	0.65	0.75	Low
3	High	2.5	2	2.405	2.25	High
3	High	2	2	2.15	2	High
2	High	2.5	2.5	2.31	2.5	High

6. CONCLUSION

In this work, the authors investigated the use of BERT and USE text embeddings for automatic short-answer grading, taking advantage of transformer-based and DAN-based text representations to improve grading accuracy. The experimental results show that the integration of BERT and USE models provides more reliable scoring by capturing both contextual and semantic similarities in student responses. The proposed approach obtains a higher correlation with human scores and displays better generalization across student responses. Hence, it can be stated that the proposed method is a robust model that reduces the model's dependency on data.

The empirical analysis also shows that the BERT model tends to assign higher scores to student responses that contain important keywords present in the reference answers, even with poorly structured sentences. This suggests that BERT places significant emphasis on contextual word representations and keyword overlap when computing similarity. In contrast, USE focuses more on the overall semantic coherence of the sentence. As a result, USE is more sensitive to grammatical correctness, sentence framing, and the clarity of the expressed meaning.

The findings highlight the potential of advanced text-embedding models to enhance automated assessment systems in educational environments. This approach contributes to the development of intelligent educational tools that support efficient evaluation and enable faster feedback in large-scale learning settings such as digital learning platforms. The pedagogical insights provided by the analysis of this work contribute to the field of Educational Data Mining (EDM) for developing robust, scalable tools for automated assessment systems. These tools can function effectively in environments where labeled data may be scarce. From an EDM standpoint, utilizing both BERT and USE allows for a multi-dimensional analysis of student responses, providing more granular data for assessment of student responses.

Transformer-based approaches enhance ASAG performance, but some challenges remain. The proposed model was evaluated on a limited dataset, which may affect its generalizability across different subjects, question types, and educational levels. Additionally, transformer models require substantial computational resources, limiting real-time deployment in resource-constrained environments. Another challenge is the limited interpretability of deep learning models, making it difficult to explain how scores were generated.

Future research can address these limitations by evaluating the model on larger and more diverse datasets. Lightweight architectures can be explored to reduce computational costs and improve model transparency. Furthermore, integrating automated feedback mechanisms for student responses could make ASAG systems more

pedagogically useful. Such advancements would improve the accuracy and efficiency of ASAG systems. As future work, these models can be integrated into human-in-the-loop systems. Such systems can support a collaborative grading environment where AI-driven insights augment, rather than replace, the instructor's expertise. This advancement will transition ASAG from an evaluation tool into a comprehensive engine for automated learning analytics within the EDM framework.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] P. Selvi and A. K. Bnerjee, "Automatic short-answer grading system (ASAGS)," *arXiv preprint arXiv:1011.1742*, 2010. https://doi.org/10.48550/arXiv.1011.1742

[2] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60-117, 2015. https://doi.org/10.1007/s40593-014-0026-8

[3] M. A. Sultan, S. Bethard, and T. Sumner, "DLS@CU: Sentence similarity from word alignment and semantic vector composition," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 148-153.

[4] C. Leacock and M. Chodorow, "C-rater: Automated scoring of short-answer questions," *Computers and the Humanities*, vol. 37, no. 4, pp. 389-405, 2003. https://doi.org/10.1023/A:1025779619903

[5] U. K. Chakraborty, S. Roy, and S. Choudhury, "A novel semantic similarity based technique for computer assisted automatic evaluation of textual answers," in *Advanced Computing, Networking and Informatics-Volume 1: Advanced Computing and Informatics Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014), Springer*, 2014, pp. 393-401.

[6] M. O. Dzikovska *et al.*, "Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 263-274.

[7] L. B. Galhardi and J. D. Brancher, "Machine learning approach for automatic short answer grading: A systematic review," presented at the Ibero-American Conference on Artificial Intelligence, Springer, 2018.

[8] S. Roy, S. Dandapat, and Y. Narahari, "A fluctuation smoothing approach for unsupervised automatic short answer grading," in *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, 2016, pp. 82-91.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) Association for Computational Linguistics*, 2019, pp. 4171-4186.

[10] D. Cer *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018. https://doi.org/10.48550/arXiv.1803.11175

[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving language understanding by generative pre-training*. United States: OpenAI, 2018.

[12] T. Chakraborty, *Introduction to large language models: Generative AI for text*, 1st ed. New Delhi, India: Wiley India, 2025.

[13] A. Vaswani *et al.*, *Attention is all you need. In Advances in Neural Information Processing Systems*. United States: Curran Associates, Inc, 2017.

[14] C. Ye, T. H. Borman, and G. D. Borman, "Automating self-affirmation essay coding: Fine-tuned BERT performance comparable to human coders and comparison with GPT-4," *Journal of Educational Data Mining*, vol. 18, no. 1, pp. 66-88, 2026.

[15] S. M. Dol and P. M. Jawandhiya, "Classification technique and its combination with clustering and association rule mining in educational data mining—A survey," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106071, 2023. https://doi.org/10.1016/j.engappai.2023.106071

[16] E. B. Page, "The imminence of... grading essays by computer," *The Phi Delta Kappan*, vol. 47, no. 5, pp. 238-243, 1966.

[17] L. Ramachandran, J. Cheng, and P. Foltz, "Identifying patterns for short answer scoring using graph-based lexico-semantic text matching," in *Proceedings of the 10th Workshop on Innovative use of NLP for Building Educational Applications*, 2015, pp. 97-106.

[18] B. Riordan, A. Horbach, A. Cahill, T. Zesch, and C. M. Lee, "Investigating neural architectures for short answer scoring," in *Proceedings of the 12th Workshop on Innovative use of NLP for Building Educational Applications*, 2017, pp. 159-168.

[19] 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," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 9662-9669, 2019. https://doi.org/10.1609/aaai.v33i01.33019662

[20] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, "Automatic short answer grading and feedback using text mining methods," *Procedia Computer Science*, vol. 169, pp. 726-743, 2020. https://doi.org/10.1016/j.procs.2020.02.171

[21] I. G. Ndukwe, C. E. Amadi, L. M. Nkomo, and B. K. Daniel, "Automatic grading system using sentence-BERT network," presented at the International Conference on Artificial Intelligence in Education, Springer, 2020.

[22] M. Zhang, S. Baral, N. Heffernan, and A. Lan, "Automatic short math answer grading via in-context meta-learning," *arXiv preprint arXiv:2205.15219*, 2022. https://doi.org/10.48550/arXiv.2205.15219

[23] C. Chakraborty, R. Sethi, V. Chauhan, B. Sarma, and U. K. Chakraborty, "Automatic short answer grading using universal sentence encoder," presented at the International Conference on Interactive Collaborative Learning, Springer, 2022.

[24] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned BERT model in classification of helpful and unhelpful online customer reviews," *Electronic Commerce Research*, vol. 23, no. 4, pp. 2737-2757, 2023. https://doi.org/10.1007/s10660-022-09560-w

[25] L. Jiang and N. Bosch, "Short answer scoring with GPT-4," in *Proceedings of the 11th ACM Conference on Learning@ scale*, 2024, pp. 438-442.

[26] C. Chakraborty, A. U. Islam, and B. Sarma, "Automatic short answer grading for enhancing educational assessment using BERT and USE models," *Indian Journal of Science and Technology*, vol. 19, no. 6, pp. 384–393, 2026. https://doi.org/10.17485/IJST/v19i6.1167

[27] P. Barbara, H. Ben, M. Jaison, V. Lynn, and D. S. Mark, *The Hewlett Foundation: Short answer scoring*. United States: Kaggle, 2012.