INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

OPEN ACCESS

Received: 25/06/2025

Accepted: 06/02/2026

Published: 21/02/2026

Citation: Chakraborty C, Islam A, Sarma B (2026) Automatic Short Answer Grading for Enhancing Educational Assessment Using BERT and USE Models. Indian Journal of Science and Technology 19(6): 384-393. https://doi.org/10.17485/IJST/v19i6.1167

* **Corresponding author**.

chandralika.c@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Automatic Short Answer Grading for Enhancing Educational Assessment Using BERT and USE Models

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

1 Department of Information Technology, Sikkim Manipal Institute of Technology, Sikkim Manipal University, Sikkim, 737136, India
2 Department of Computer Science, University of Science and Technology Meghalaya, Baridua, Meghalaya, 793101, India
3 Department of Computer Science & Engineering, The Assam Royal Global University (RGU), Guwahati, Assam 781035, India

Abstract

Objectives: This work explores the application of two advanced state-of-the-art models, BERT (Bidirectional Encoder Representations from Transformers) and USE (Universal Sentence Encoder), to automate the grading of short answers. **Methods:** This work investigates the use of BERT and USE models for automatically grading short answers. The research utilizes HP:SAS dataset containing manually graded responses by two human evaluators. The student responses as well as model answers responses of question 1 and question set 6 are then processed using the BERT and USE models, with scores generated based on cosine similarity measures between student answers and predefined model answers. **Findings:** The work demonstrates that BERT and USE embeddings can effectively capture contextual and semantic similarity, their performance is heavily dependent on the function which generates the score. Our finding reveal that a non-linear mapping function mimics the human grading more than a linear mapping function. Such a function enhances accuracy (0.67) and reduces the error (0.617) by computing Pearson correlation coefficient and RMSE respectively. Notably, longer responses achieved higher Pearson correlations (0.67) than shorted answers (0.59). The results bring out usability and choice aspects of BERT and USE in relation to ASAG, contributing to the understanding of their application across various answers. We conclude with a weighted ensemble method combining BERT and USE with subject-specific strictness parameter (k) provides a robust framework for automated assessment. **Novelty:** Evaluates and compares two deep learning models for automatic short answer grading, a scarcely explored area. A novel contribution is the granular analysis across different scoring ranges across two question sets of the dataset. The novelty of this work lies in the transition from linear scoring to non-linear mapping framework. This approach introduces a tunable sigmoid- based ensemble that would replicate human assessment. Finally, a comparison with existing studies demonstrates very limited research.

Keywords: Bidirectional Encoder Representations from Transformers (BERT); Universal Sentence Encoder (USE); Transformer; word embedding; Non-linear mapping; deep learning; short answer grading

1 Introduction

Objective questions form an important aspect of evaluation of learning outcomes. These require recognition and recall and usually cover the lowest level of Bloom's Taxonomy. Different forms of questions can be employed for the purpose of testing recognition and recall, including multiple choice types, sentence completion, match the columns, true-false, fill-in-the blanks etc. However, these tests have the drawback of allowing the learner to score even through sheer guesswork.

The shortcomings mentioned above can be overcome through short answers, which can be defined as short natural language responses to objective questions. Essentially covering recall and recognition, short answers can also move up the Bloom's Taxonomy ladder and be employed to test understanding.

Considering the recent advancements in the field of machine learning and the penetration of internet based technologies, online learning is gaining a foothold in academics. Students prefer to opt for anytime anywhere learning over the classical time bound closed learning environment. However, the computational complexity involved in the evaluation of answers in natural language has made the academic world take to the easy to assess methods like multiple choice questions. To add value to the evaluation process and ease the efforts of the evaluators, automation of grading short answers is highly desired [1]. The growing acceptance of Massive Open Online Courses (MOOC) make the goal even worthier. There are other benefits of automatic grading as well. This provides fast, cost-efficient, effective and unbiased evaluation for both summative and formative assessments. Human errors, creeping in while evaluating large class-sizes, owing to fatigue or immediate contrasting answer scripts penalty can also be avoided.

Over the last couple of decades, much research has concentrated on evaluation of text based answers and however, owing to the differences in purpose and style of presentations, all subjective answers cannot be evaluated in the same manner [2].

Automatic grading has its origins in the mid 1960s, when Page [3] in his work introduced computational methods for providing grades to student essays and also predicted the large-scale use of computers for the same. Over the years much research has gone into the field and answer grading is now treated differently for short answers and essays.

Several studies have explored the use of deep learning techniques for Automatic Short Answer Grading (ASAG). This highlights the growing significance of attention-based models and pretrained language representations. Al-Rouf *et al.* [4] demonstrated that transformer architectures can replace RNNs in short-answer grading tasks, with their deep (64-layer) transformer model achieving state-of -the art results on benchmarks such as text8 and enwik8 due to its ability to transmit information efficiently across sequences. A broader survey on ASAG with deep learning techniques emphasizes word embeddings and the introduction of transformer-based attention mechanisms as a pivotal advancement, for better semantic representation than traditional methods [5]. However, the application of such models specifically to ASAG remains relatively unexplored. Bonthu *et al.* [6] addressed this gap by proposing a system that integrates Pretrained Language Models (PLMs) and data augmentation techniques to enhance the grading process. Xinghua *etal.* [7] proposed a BERT-based deep neural network framework incorporating bidirectional LSTM and Capsule network which achieves superior results on SemEval2013 and Mohler datasets, despite limited training data. Another practical approach using Universal Sentence Encoder (USE) was explored by Wijanto and Yong [8], which led to a notable improvement in grading accuracy, and also reveals the model's ability to handle diverse linguistic expressions effectively. Further, Chakraborty, *et al.* showed that automatic evaluation of text answers using the USE model performs reliably when applied to sizeable datasets, reinforcing its suitability and can be considered as a reliable approach [9].

Most work carried out in the earlier part of the current millennium used hand-crafted patterns or large datasets to train evaluation models [2]. Recent trends have however shifted towards Deep Neural Models being employed for the task. The use of Neural models have brought a revolution in the area of automatic short answer grading (ASAG) by leveraging various deep learning techniques for evaluating student responses with greater precision. To understand or evaluate the semantic content of the answers, these neural models implement Recurrent Neural Networks (RNNs), transformer based architectures such as USE, BERT. The current paper is based on a study conducted on a question-answer response of five hundred students which were blindly evaluated by two human evaluators on a scale of three marks. Automated evaluation of these responses are performed using BERT and USE models, and a granular comparison is carried out on the machine generated scores obtained.

Few works have shown the use of a Deep Averaging Network USE model for ASAG. But, there is limited work found which utilizes BERT and USE in automated evaluation of question-answer responses. So, this study tries to employ BERT and USE for ASAG in generating scores for short answers and then comparing the predictions from both models – an approach not much explored in existing literature. To achieve this, questions assessed by two human evaluators were subsequently evaluated using an automated method that employed cosine similarity measures between the vectors generated from the answers utilizing BERT and USE. This work bridges the gap between raw semantic and contextual embeddings and the nuanced human grading by providing empirical evidence for superiority of non-linear mapping functions. The work demonstrates that the threshold based

function better model the S-curve of academic assessment while preserving the student ranking fairness – this fine grained insight is rarely explored in existing literature.

2 Methodology

To actualize the research work for grading short answers, data collection is the first step, where human evaluated student answers are available. Next, model answers are to be identified for grading the students answers. This is followed by finding a suitable technique for evaluating the students answers with respect to model answers. And then determining a method for obtaining the amount of degree of similarity and generating a machine generated score – which can be compared with the human evaluated scores.

2.1 Data Set

The records for student answers are considered from dataset[10]. The dataset is a tab-separated value file which contains a total of 17,198 records of student responses. The data set includes the following:

1. student id,
2. question set number,
3. two human evaluated scores, and
4. the answer given by student.

There are 10 distinct question sets in the dataset, each representing the answer to a single question. In this work, we have considered answers of question set 1(related to chemistry) and 6 (related to biology). The scores of the students' answers are in the range 0 to 3. The full score of the answers is '3'. The average length of the answers in question set 1 and 6 are approximately 60 and 90 words respectively. Our aim is to determine the machine generated score of 'student answers' using deep learning models – Bidirectional Encoder Representations from Transformers and Universal Sentence Encoder, with the help of cosine similarity measure, and analyse how the models performs grading of answers with this method.

2.2 Algorithm for the proposed model

Input : Model answers and Student answers

Output : Predicted scores using BERT model and USE model

1. Begin

   // Dataset selection
2. Load the dataset

   // Initialization of models
3. Initialize embedding models : BERT and USE

   // Embedding Generation
4. For each model m ε {BERT, USE}

   (a) Generate embeddings for each student answer
   (b) Select five full scoring answers as reference (model) answers

   // Similarity computation
5. For each student answer

   (a) Compute cosine similarity between student answer and each of the reference answer for both BERT and USE
   (b) Identify the maximum cosine-similarity score across all reference answers for each model.

   // Proposed Ensemble Method
6. Predicted Score $= \sigma(w_1 \cdot \cos\_sim_{BERT} + w_2 \cdot \cos\_sim_{USE})$

   Where σ is the non-linear mapping using Sigmoid function, w_1 and w_2 are weights, which are determined based on empirical performance observed while analysing each model).

   //Combined weighted cosine similarity

7. Combine the individual similarities into a unified score using a weighted average:
   $cos\_sim_{combined} = w_1.cos\_sim_{BERT} + w_2.cos\_sim_{USE}$, where, $w_1 + w_2 = 1$
   // Non-Linear Mapping using Sigmoid function
8. Compute the predicted machine score using the Sigmoid function:

$$Score_{combined} = \frac{Full\ Score}{1 + e^{-k(cos\_sim_{combined} - \theta)}}$$

   where k = steepness, which determines the strictness of the evaluation; and
   Θ = the minimum cosine similarity threshold for partial correctness.
   // Output Generation
9. Generate output by producing the final predicted score for each student answers
10. End

Cosine similarity is a widely used metric in automatic short answer grading (ASAG) and has shown to perform well in assessing textual similarity, which is crucial in ASAG tasks[2].

Pearson's correlation coefficient is used to compute the similarity between the human raters and the machine predicted score. And, RMSE quantifies the average deviation between the machine generated answers and Human graders.

A set of 500 answers are considered from the dataset[10] for Question Set 1 and Question Set 6. The answers in the dataset are already evaluated by two human evaluators, where the scores vary between 0 to 3. Any five full scoring answers from the answer set have been considered as model answers for each question set. The merit of the model answers have not been checked manually.

The student answers and all five model answers are vectorized using BERT and USE models separately, as shown in Figure 1. i^{th} student answer vector is compared with each of the j^{th} model answer vectors, using Cosine similarity measure. BERT model fine tuning was not done to keep the model simple and easy to use for the non-technical human evaluator.

3 Results and Discussion

In this paper, we present the BERT and USE model results for computing the machine generated scores, as the preliminary experiments that provide the empirical justification for our proposed ensemble method. The models BERT and USE have need separately tested across different question sets of the HP:SAS dataset, and for linear and non-linear mapping functions. It was observed that both models generate scores for the students answers which align with human rater scores.

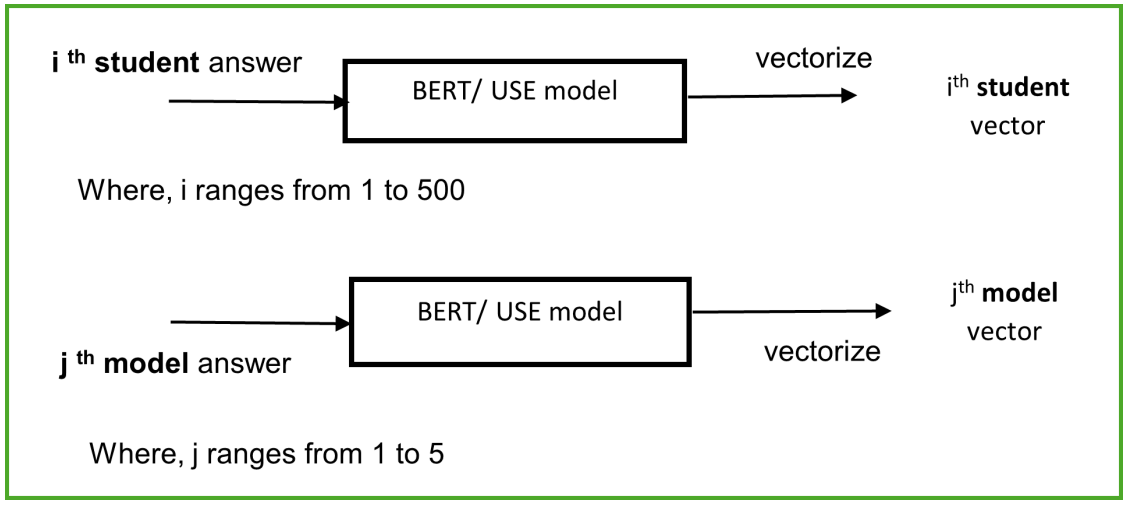

Fig 1. Representation of a student answer and a model answer

Table 1. Analysis of score mismatches between human evaluation and BERT/USE models (only few samples shown below)

Student ID	Avg Human Evaluation	NonLinear Score_ BERT	Machine Generated Score using BERT	Analysis of score mismatches between human evaluation and BERT/USE models	NonLinear Score_USE	Machine Generated Score using USE
16	0	**0.28**	**1.13**	No mention of keywords. Answer is containing only few words, which does not display proper meaning. Hence human evaluator has provided zero marks. Other than the keywords, remaining words belong to the model answers. Hence, may be due to the presence of this match of words in the model answers, BERT provides a higher score and due to improper meaning of the sentence USE generates a very low score. This may be because BERT uses word piece tokenization where each word in the input sentence are broken down into sub-word tokens.	**0.03**	**0.17**
167	0	**2.25**	**2.57**	Keyword Vinegar is not present. Referred to as just chemical. And wrong spelling of words used– replicate is mentioned as deplicate; and difficient. Due to wrong spelling of words, meaning of words not clear.Hence Human Evaluator have given a score of 0. BERT being a token level embedding model scores the answer higher may be due to the presence of certain words.	**0.94**	**1.76**
374	0	**2.26**	**2.58**	Vinegar word is present. Keywords available in model answer is present but meaning of sentence is not correct. Wrong spelling –pour is mentioned as poor. Due to the presence of keywords in student answer BERT provides are greater score than USE score.	**1.34**	**2.01**

Continued on next page

Table 1 continued

422	0	**2.12**	**2.47**	Word vinegar is not present. Short form of different used — diff. Certain keywords present, but meaning of sentence expressed in general terms. Hence, in compared to st id 374 the BERT score is less as keyword vinegar not present in the student answer. But, USE score is higher, may be because the sentence's proper meaning is conveyed through the presence of certain words in the sentence. Overall, the machine generated score(BERT) is higher because of the presence of the keywords.	**1.65**	**2.19**
453	0	**2.16**	**2.5**	Word vinager is present. Certain keywords present , but meaning of sentence not clear. Hence, the presence of keywords generates a higher score using BERT in compared to answers without the mention of keywords(as in st id 422). But USE does not generate a higher score simply due to the presence of certain keywords in the answer of std id 422 - it also checks the meaning of the sentences. Thus a low score here.	**0.65**	**1.55**
6	0.5	**1.95**	**2.36**	Word vinegar not present. Very few keywords are present, which are less than in the student answer 453 hence a BERT score towards full score but less than of student id453. Meaning of answer is not clear. Sentences does not provide proper meaning, so we can find a lower USE score.	**0.99**	**1.79**
7	0.5	**1.96**	**2.37**	Word vinegar is not present. Wrong spellings of words. Certain keywords are present but does not give any clear meaning. Presence of certain keywords in the answer generates a higher score using BERT, and a lower USE score due to lack of sentence meaning.	**1.53**	**2.11**

The predicted machine generated score was initially computed by a linear mapping function of maximum cosine similarity multiplied by the maximum score of the answers. Then it was observed that a small increase in the cosine similarity would correspond to a fixed increase in the score. This often fails to make a distinction between a partially correct answer and perfectly correct answer. Hence, the similarity score is not strictly proportional to the human assigned scores. By applying a non-linear function would better capture the distribution of human scoring which might reduce error and increase the correlation between human raters and the machined generated scores. The same is shown in Table 1 and Table 2. In the search for a non-linear

mapping function, a power based function was used, as it would effectively separate low and high scores. With power function, low cosine similarity scores would be disproportionately reduced while higher scores are preserved, which would ensure answers with truly high similarity are mapped to high scores, medium similarity to medium scores and low similarity to low scores. To mimic human grading, another non-linear mapping function used was the sigmoid function. The sigmoid mapping function is used as a function of the weighted sum of the maximum cosine similarity score of each model, to determine the predicted machine generated score. A comparison of the mapping function in this work is shown in Table 3.

Table 2. Appropriate evaluation by machine against human evaluation.

Student ID	Average of Human evaluators	Non-Linear Score_ BERT	Machine generated score using BERT	Non-Linear Score_ USE	Machine generated score using USE	Remarks
44	3	1.43	2.06	1.32	2.00	Answer not so detailed. Important key information mentioned –hence full marks must have been awarded by human evaluator. But few points missing. Machine evaluation using BERT and USE is appropriate.
99	3	1.07	1.84	0.21	0.99	Incomplete information with no meaning. No mention of vinegar. Sentence ended abruptly. Human evaluator must have overlooked the missing information and have awarded full marks. Machine evaluation is correct.
204	3	1.54	2.12	1.14	1.89	Not same as the model answers, but contains the keywords, vinegar, container, mass. Human evaluator overlooks the sentence framing, and thinks that the student has knowledge about the answer and hence awarded marks. As BERT model performs token-level embedding, so due to the presence of the keywords in the student answer, the score generated is also higher.
497	0	0.86	1.71	0.55	1.46	Does not contain the word vinegar. Contains other keywords. Sentences are meaningful and relevant, but does not provide all required information. Human evaluator could have given 1 mark out of 3. So, machine seems to give a better judgement.

Table 3. Comparison of mapping functions used in this work.

Sl. No.	Mapping function type	Impact on scores
1	Linear	Scores were proportional to cosine similarity ; high error (RMSE)
2	Power function	Distinct separation between low and high scores; improved correlation
3	Sigmoid function	Better correction (Pearson) and low errors(RMSE)

Performance of BERT and USE on question set 1 and 6 are as shown Table 4 and Table 5.

Table 4. Evaluation metrics for HP:SAS dataset, **Question Set 1**

	Linear Mapping			Non-Linear Mapping (Power=2.0)			Non-Linear Mapping (Sigmoid function, k=7, Θ=0.8(for BERT) and k=7,Θ=0.6(for USE))		
Model/ Mapping Method	Pearson Coeff	Spearman Coeff	RMSE	Pearson Coeff	Spearman Coeff	RMSE	Pearson Coeff	Spearman Coeff	RMSE
BERT	0.54	0.55	1.22	0.55	0.55	0.95	**0.55**	0.55	**0.907**
USE	0.57	0.58	0.95	0.58	0.58	0.96	**0.59**	0.58	**0.911**

As both the models showed strong Pearson correlations (upto 0.6722 for USE in question set 6) and a low RMSE with Sigmoid mapping (0.907 by BERT in questions set 1), the models were combined into the proposed ensemble method for maximizing scoring accuracy.

Table 5. Evaluation metrics for HP:SAS dataset, **Question Set 6**

	Linear Mapping			**Non-Linear Mapping (Power=2)**			**Non-LinearMapping (Sigmoid, k=18, Θ=0.9)for BERT and (k=18, Θ=0.8) for USE**		
Model/ Mapping Method	Pearson Coeff	Spearman Coeff	RMSE	Pearson Coeff For Power=2	Spearman-Coeff	RMSE For Power=2	PearsonCoeff For k=18, Θ=0.9) for BERT and (k=18, Θ=0.8) for USE	Spearman-Coeff	RMSE For k=18, Θ=0.9)for BERT and (k=18, Θ=0.8) for USE
BERT	0.53	0.63	1.84	0.57	0.62	1.33	**0.65**	0.63	**0.636**
USE	0.54	0.58	1.43	0.61	0.58	0.90	**0.67**	0.58	**0.617**

For both models, the sigmoid mapping function produced better results, achieving a low RMSE of 0.907 in question set 1 using BERT model, and 0.617 in question set 6 using USE model. A high Pearson correlation of 0.59 and 0.67 by USE model was obtained in question set 1 and question set 6 respectively.

The text in question set 1 is related to chemistry subject and for question set 6 the subject is biology. The average size of the answers is 60 words and 90 words for question set 1 and question set 6 respectively.

It was observed that Pearson correlation is higher for longer answers than shorter ones as shown in Table 6. Longer answers may provide more semantic and more contextual information, allowing the models to create more distinct embeddings.

Table 6. Comparison of the evaluation metrics among the question sets of the HP:SAS dataset.

Evaluation metric	Question set 1 (Chemistry contents)	Question set 6 (Biology contents)	Remarks
Average length of answers	60 words	90 words	Longer answers provide better semantic and contextual information
Maximum Pearson Correlation coefficient	0.59 (obtained using USE)	0.67 (obtained using USE)	Biology (descriptive text) ranked better than technical (keywork specific text (Chemistry)
Minimum RMSE	0.907 (obtained using BERT	0.617 (obtained using USE)	Error is lower in longer / Biology set
Value of k in the sigmoid function	K=7	K=18	Descriptive text has sharper correctness threshold.

Spearman coefficients was found to be stable across mappings. This shows that the sigmoid function improves score accuracy without disrupting the relative ranking of the students.

In the sigmoid mapping function, k and Θ serve as the control mechanism for aligning machine generated similarity scores with human assigned grades. Θ is the midpoint of the grading curve, which defines the cosine similarity value where the student is expected to receive exactly 50% of the full score. Θ is set higher for BERT (0.8 or 0.9) than for USE(0.6 or 0.8), suggesting that BERT's similarity scores are naturally higher and require a threshold to prevent over-scoring. k determines the steepness of the curve or in other words the strictness of the grading, by controlling how rapidly the score increases once the threshold Θ is met. A higher k value creates a steeper curve, that is the model makes a sharp distinction between correct and incorrect answers, as indicated in question set 6. This is better for descriptive answers where once the core concepts are identified, the score achieves full credit. A low k value as in question set 1, creates a lenient or gradual scoring slope, where partial credit is awarded more incrementally. This is more suited for text where partial credit is awarded for the presence of specific keywords or phrases. Keeping k and Θ as parameters, allows for flexibility and gives additional control to human evaluators.

To optimize the ensemble methods predictive power, the weights w_1 and w_2 will be assigned based on the individual Pearson correlation coefficients observed during experimentation. The weights will be calculated as given below:

$$w_1 = \frac{Pearson\ Coeff\ for\ BERT}{Pearson\ Coeff\ for\ BERT + Pearson\ Coeff\ for\ USE};$$

$$w_2 = \frac{Pearson\ Coeff\ for\ USE}{Pearson\ Coeff\ for\ BERT + Pearson\ Coeff\ for\ USE}$$

This would ensure higher empirical accuracy.

Finally, Table 7 compares existing studies with our approach, showing that limited research works have examined the combined use of BERT and USE models for automatic short answer grading.

Table 7. A comparison of this study with previous works.

Automatic Short Answer Scoring Model reference	Year	Model used		Evaluation metrics				Remarks
				Pearson Correlation Coefficient	RMSE	Accuracy	QWK	
		BERT	USE					
Gamit, et. al.[11]	2025	Yes	Yes	0.56 (USE) 0.31(BERT)	0.997(USE) 1.062(BERT)	–	–	
Wang D, Chen G[12]	2025	Yes	No	–	–	0.86	–	Used for analyzing classroom dialogic moves
Li, H.[13]	2025	Yes	No	–	–	0.97	–	Used for evaluating reading strategies
Jiang & Bosch[14]	2024	No	No	–	–	0.53	0.660	
Meena & Tsunenori[15]	2023	Yes	No	–	–	0.357	0.798	Used versions of BERT: SBERT, AraBERT
Wilianto & Girsang[16]	2023	No	No	–	0.219	–	–	
Kumar,Y. et al.[17]	2019	No	No	–	–	NA	0.79	
Proposed Method		Yes	Yes	0.67 (USE) 0.65 (BERT)	0.617 (USE) 0.636 (BERT)	–	0.99	For question set 6 of HP:SAS dataset

4 Conclusion

Traditional forms of automatic grading such as multiple-choice questions are commonly used due to ease of automated grading, however, they often fail to examine the understanding of a student on the knowledge of the subject, and can also lead to guesswork. Short answer questions, on the other hand, provide a better understanding of a student's grasp of the subject matter, but the evaluation of short answers is complex and labor-intensive.

The use of deep learning models has opened new avenues for more accurately automating the grading of short answers, thus enhancing the evaluation process. This work explored the application of two cutting-edge models – BERT and USE, to automatically grade short answers by leveraging their capabilities to understand context and semantic meaning through cosine similarity measures.

In this work both BERT and USE models were used to evaluate student answers, and then these automated scores were compared with human evaluator scores. The results demonstrated that while both models can effectively generate grading scores, BERT consistently produced higher scores than USE. The comparison revealed that BERT's ability to capture fine-grained contextual information often resulted in scores that aligned closely with human evaluations, particularly in cases where key domain-specific terms were present in the answers.

Despite some discrepancies between the two models and between machine-generated and human scores, the findings suggest that automated grading using these deep learning models is a viable approach.

This work investigates the efficacy of transformer based BERT embeddings, Deep Averaging Network version of USE embeddings, and non-linear functions for automated short answer responses in chemistry and biology. The experimental results show that BERT and USE provide strong contextual and semantic representations, but their performance is influenced by the mapping function used to translate the cosine similarity score to the machine predicted score. A non-linear mapping function proved essential for reducing the error, and effectively align machine generated score with human grading. This was one of the key conclusions of this work. Another is, the models performed better, resulting in higher Pearson coefficients for longer answers which contained more contextual and semantic data. Also, the parameters k and theta are important controls, which allows the system to adapt to specific requirements of different academic domains. Based of these findings, this work proposes an Ensemble method that integrates BERT and USE using weighted average. Further, this work introduces a domain-aware ensemble methodology that optimizes BERT and USE embeddings through subject- specific parameters.

This work suggests that a combination of these can be used together for high stake examination. This also leaves scope for future study and development in this direction.

References

1) Ghavidel H, Zouaq A, Desmarais M. Using BERT and XLNET for the Automatic Short Answer Grading Task. 2020. Available from: https://pdfs.semanticscholar.org/8e4c/bc85a19c7d32c1426ae9df558b21461c7dc9.pdf.
2) Mohler M, Mihalcea R. Text-to-text Semantic Similarity for Automatic Short Answer Grading. Athens, Greece. 2009. Available from: https://aclanthology.org/E09-1065/.
3) Page EB. The imminence of grading essays by computer. *The Phi Delta Kappan*. 1966;47(5):238–243. Available from: https://www.sciencedirect.com/science/article/abs/pii/S8755461505800581.
4) Al-Rfou R, Choe D, Constant N, Guo M, Jones L. Character-Level Language Modeling with Deeper Self-Attention;vol. 33. 2019. 10.1609/aaai.v33i01.33013159.
5) Sung C, Dhamecha TI, Mukhi N. Improving Short Answer Grading Using Transformer Based Pre-training. Cham. Springer International Publishing. 2019.
6) Bonthu S, Sree SR, Prasad MHMK. Improving the performance of automatic short answer grading using transfer learning and augmentation. *Engineering Applications of Artificial Intelligence*. 2023;123. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0952197623004761.
7) Zhu X, Wu H, Zhang L. Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies*. 2022;15(3):364–375. 10.1109/tlt.2022.3175537.
8) Wijanto MC, Yong HS. Combining Balancing Dataset and Sentence Transformers to Improve Short Answer Grading Performance. *Applied Sciences*. 2024;14(11):4532. 10.3390/app14114532.
9) Chakraborty C, Sethi R, Chauhan V, Sarma B, Chakraborty UK. Automatic Short Answer Grading Using Universal Sentence Encoder;vol. 633. Cham. Springer. 2023. 10.1007/978-3-031-26876-2_49.
10) Barbara, Hamner B, Morgan J, lynnvandev, Shermis M. The Hewlett Foundation: Short Answer Scoring. 2012. Available from: https://www.kaggle.com/competitions/asap-sas/data?select=train_rel_2.tsv.
11) Gamit NC, Upadhyaya AN. Evaluating Pretrained Embeddings for Automatic Short Answer Grading. *Journal of Information Systems Engineering and Management*. 2025;10:150–155. 10.52783/jisem.v10i18s.2897.
12) Wang D, Chen G. Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy. *British Journal of Educational Technology*. 2025 May. 10.1111/bjet.13604.
13) Li H. Automatic evaluation and enhancement of reading strategies in English reading comprehension based on the BERT model. *Journal of Computational Methods in Sciences and Engineering*. 2025;25(1):794–807. 10.1177/14727978251322023.
14) Jiang L, Bosch N. Short answer scoring with GPT-4. 2024. 10.1145/3657604.3664685.
15) Fateen M, Mine T. In-Context Meta-Learning vs. Semantic Score-Based Similarity: A Comparative Study in Arabic Short Answer Grading. Singapore (Hybrid). Association for Computational Linguistics. 2023. 10.18653/v1/2023.arabicnlp-1.28.
16) Wilianto D, Girsang AS. Automatic short answer grading on high school's e-learning using semantic similarity methods. *TEM Journal*. 2023;12(1):297–302. 10.18421/TEM121-37.
17) Kumar Y, Aggarwal S, Mahata D, Shah RR, Kumaraguru P, Zimmermann R. Get IT Scored Using AutoSAS — An Automated System for Scoring Short Answers;vol. 33. 2019. 10.1609/aaai.v33i01.33019662.