IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 May 2024, accepted 20 June 2024, date of publication 1 July 2024, date of current version 22 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3420890

APPLIED RESEARCH

A Hybrid Approach for Automated Short Answer Grading

MUSTAFA KAYA AND ILYAS CICEKLI
Department of Computer Engineering, Hacettepe University, Beytepe Campus, 06800 Ankara, Türkiye
Corresponding author: Mustafa Kaya (mustafakaya05@gmail.com)

ABSTRACT With the widespread use of distance learning, technological developments have also been applied in the field of education. The need for accurate and efficient assessment methods for online exams has become even more apparent, especially with remote learning taking place during the pandemic. For a more efficient evaluation process, we propose a hybrid model of the Automatic Short Answer Grading (ASAG) system based on Bidirectional Encoder Representation of Transformers (BERT). The usage of novel state-of-the-art natural language processing (NLP) techniques in our model enhances the comprehension of text. Specifically, we employ a customized multi-head attention mechanism adapted with BERT, which enables reliable identification of semantic dependencies among words within a sentence and therefore contributes to the effectiveness and trustworthiness of the scoring system. We use a parallel connection of CNN layers in our proposed BERT based ASAG system instead of their serial connection and this usage improves the performance of the system. The proposed model is assessed using common datasets frequently used for ASAG related research projects. In this evaluation process, our model produces much better results compared to other systems available in the literature.

INDEX TERMS Automated short answer grading, BERT, CNN, LSTM.

I. INTRODUCTION

Technological developments affect the world of learning as well as every field. Distance education has become an alternative solution for students to improve their learning abilities comfortably and easily. However, global disasters such as epidemics have made distance education mandatory. It is important to evaluate student responses objectively and quickly in distance education [1]. Automatic Short Answer Grading (ASAG) systems play a crucial role in fulfilling this need and they provide a consistent and adaptive way of dealing with these challenges, given that distance education and learning organizations have a limited supply [2]. This challenge necessitates the integration of advanced natural language processing (NLP) techniques to effectively interpret and evaluate student responses.

Since BERT is released as a new platform for NLP in 2018 by Devlin et al. [3], BERT-based research has had a successful outcome in smart education systems [4]. Wang et al. [5] demonstrated the BERT model to assess the effectiveness of teachers in terms of the topics of their courses in online education. Khodeir [6] created a model of how to combine BERT with a multi-layered, reciprocal gated iterative process that helps teachers to focus on the answers to student questions quickly. Also, in the context of short-answer evaluation, Sung et al. [7] and Camus and Filighera [8] studied and compared the capabilities of BERT to various traditional neural network models and their derivatives, such as RoBERTa [9] and ALBERT [10]. In our research, we propose a new hybrid ASAG system which employs BERT. BERT is successful in determining the association between words in written passages. By combining BERT with our new customized multi-headed attention approach for this project, we attempt to give our model the ability to expand on and evaluate student responses in greater detail. In order improve our ASAG system further, we also use a parallel connection of CNN layers in our proposed BERT based ASAG system instead of their serial connection.

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

We employ two important datasets that are commonly associated with ASAG systems: SemEval-2013 Task 7 [11] and Mohler [11]. The SemEval-2013 Task 7 dataset provides a comprehensive platform for benchmarking ASAG systems with its diverse set of short-answer questions and expert-graded responses [11]. The dataset curated by Mohler et al. [12] offers a unique perspective with its focus on introductory computer science courses, presenting different linguistic challenges and grading criteria [12]. Selecting these two datasets provides a robust and varied testing ground for our proposed model, ensuring it is evaluated across different academic disciplines and question types. The diversity of studies in the literature regarding these data sets is significant because it allows us to assess the effectiveness of our model.

It is difficult to achieve a high degree of accuracy in studies in the field of ASAG. The most important reasons for this are that student answers may consist of a few words, or the student may answer the question in a different way than the instructor. Therefore, the models created for ASAG need to understand student answers in depth. In this study, we tried to contribute to this field by presenting a new hybrid approach for ASAG. The main contribution of our approach is the usage of a customized multi-head attention layer and a parallel connection of CNN layers in our proposed BERT-based model. We aim to demonstrate that the parallel connection of CNN layers, which is commonly used in the field of image processing, can also be applied in the field of NLP. In evaluations, our proposed model outperforms most of the state-of-art systems for ASAG.

The article is structured as follows. Section II discusses research on BERT-based systems in the literature. Section III describes the components that make up our proposed model. Section IV describes the evaluation methodology including the used datasets In Section V, the obtained evaluation results are discussed in detail. Section VI discusses the effects of our novel contributions such as the customized multi-head attention layer and the parallel usage of CNN layers. In the last section, we conclude the paper with possible future suggestions.

II. RELATED WORKS

BERT which is introduced by Google in 2018 [3] is a notable natural language processing model for its two-way text analysis capability. This means the model can better interpret the meaning of a word by considering the context before and after it. Also, BERT's MLM (Masked Language Modeling) pre-training enables it to learn the structure of language and relationships between words in depth. In this way, it can better interpret words and expressions that they have not encountered before. These features of BERT contribute positively to the semantic understanding of the answers given in situations where expression is provided with short answers. BERT-based approaches are developed and applied in various studies in the field of ASAG. In our study, we use a customized multi-head attention approach in accordance with our study to provide better performance. We have also increased the effectiveness of our model by using the Convolutional Neural Networks (CNN) layers in parallel.

Sung et al. [13], aimed to develop contextual representations based on BERT for ASAG. They aimed to increase efficient pre-trained BERT model by using domain-specific resources. In that study, Physiology of Behavior, American Government, with Psychology – Human Development and Abnormal Psychology textbooks were used for increasing BERT performance. The empirical study demonstrated that task-specific fine-tuning improved the performance of the pre-trained language model and thus better results were achieved for ASAG.

Condor et al. [14] conducted a study to compare the effectiveness of a modern method, Sentence-BERT (SBERT), against traditional techniques like Word2Vec and Bag-of-words. Their research found that models developed using SBERT outperformed those built with the older methods.

Zhu et al. [4] developed a four-stage framework for ASAG leveraging a pre-trained BERT model. Initially, they encoded the answer and reference texts through BERT. Following this, a Bi-directional Long Short-Term Memory (Bi-LSTM) network was employed on BERT's outputs to enhance semantic understanding. Subsequently, these outputs were merged with fine-grained token representations in a Semantic Fusion Layer. Finally, in the prediction stage, a max-pooling technique was applied to the integrated semantic representations. Their research utilized the Mohler and SemiEval datasets. The model demonstrated an accuracy of 76.5% for grading unseen answers, 69.2% for unseen domains, and 66.0% for unseen questions. Additionally, on the Mohler dataset, the model achieved a Root Mean Square Error (RMSE) of 0.248 and a Pearson Correlation Coefficient (R) of 0.89.

Cao et al. [15] aimed to determine the similarity ratio of two sentences in Vietnamese by using SBERT approach. In that study, first, output vectors were obtained from pre-trained model (SBERT) and then merged with linguistic knowledge. For paraphrase detection, Vietnamese data sets (vnPara and VNPC) were used. It was calculated distances by using output vectors from VietNet and SBERT. After evaluating the model, for VNPC, the F1 score was 95.31% and for vnPara, the F1 score was 97.62%.

Lei and Meng [16] proposed a Bi-GRU of Siamese structure based on pre-trained ALBERT. The input expression turned into word vectors by using a pre-trained ALBERT model. The obtained values were sent to the Gated recurrent unit (GRU) Network. An attention layer was added to better understand text semantic information after the Bi-GRU network. Finally, with the softmax normalization function, the result was normalized to a probability distribution to obtain better results. At the end of that study, the accuracy value of the proposed model was higher than traditional models.

Sayeed and Gupta [17] targeted an approach that leverages student and reference responses for evaluating descriptive answers, utilizing a Siamese architecture. This method is developed for ASAG using Roberta bi-encoder-based transformer models. The architecture was conceived with the

constraints of feasible computing resources in mind, specifically tailored for ASAG tasks. The model trained using the SemEval-2013 2way dataset has shown either superior or equivalent performance compared to the models it was benchmarked against.

III. PROPOSED MODEL ARCHITECTURE

In this study, we present a novel neural network model leveraging Transformer-based architectures to enhance the effectiveness of NLP tasks, specifically focusing on automated text analysis and classification. This approach builds upon the growing body of research in deep learning and NLP, including works on automated short answer grading and semantic analysis. In this section, the proposed novel model uses a BERT based model for short answer grading and its architecture includes the novel usage of a customized multi-head attention layer. Another novelty of the proposed model is the parallel connection of CNN layers instead of their serial connection of them.

We present a new hybrid approach to addressing the short answer task of grading by combining the benefits of various deep learning-based configurations. The first novelty implemented in the proposed model is the utilization of a customized multi-head attention mechanism instead of the standard multi-head attention mechanism and the multi-head mechanism is a central component of the transformer model of Vaswani et al. [18]. This multiple-headed mechanism facilitates the focus of the model on different parts of the input sequence, and it is important for comprehending the context and subtlety of language. The second new novel concept in the proposed model is the usage of the CNN layers in parallel instead of a serial connection and it also increases the performance of the system in addition to the increased performance achieved with the customized multiple head attention. The proposed ASAG system uses a combined approach that utilizes BERT, LSTM, and CNN layers, as illustrated in Figure 1. The following subsections discuss the components of the system in detail.

A. BERTBASE MODEL

The BERT model is a pre-trained language representation that was created by Google's artificial intelligence department, it combines the attributes of ELMo [19] and GPT [20] Similar to GPT, BERT has a Transformer-based design that is composed of multiple layers [18]. This architecture enhances the long-term capacity for storage at ELMo's LSTM networks, this is a significant difference. BERT's unique approach is bidirectional. Unlike the unidirectional model of GPT, BERT employs a directional dual-language model, similar to ELMo. This facilitates the capture of both direct and indirect context effectively [4].

BERT is pretrained on two unsupervised tasks: the masked language model and the next sentence predictor. It utilizes a large corpus composed of an 800 million-word book corpus and a 2500 million-word English corpus from the Internet. The pre-trained BERT model is capable of adapting to different downstream tasks by simply altering the classification layer and all pre-trained parameters together.

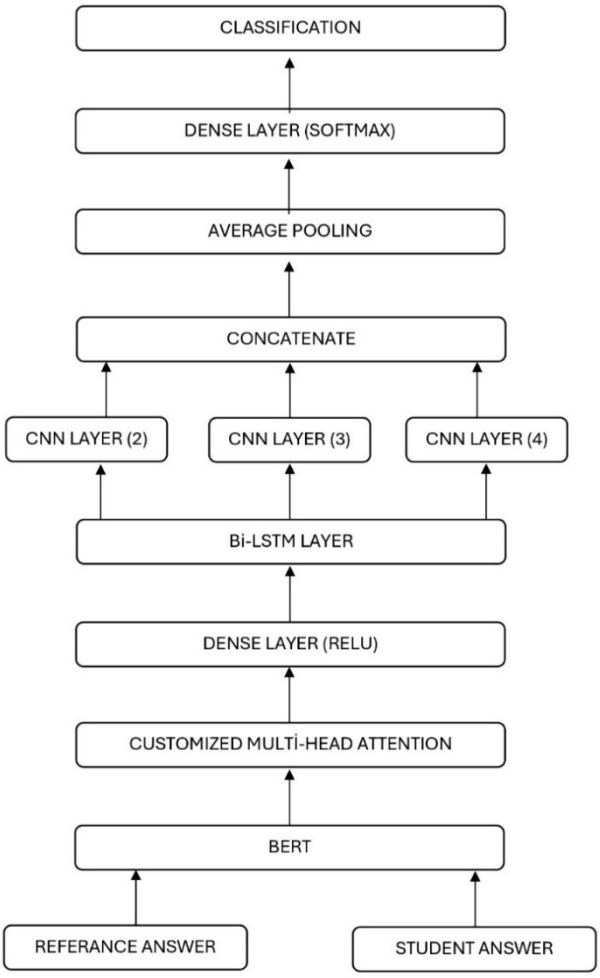

FIGURE 1. The schematic diagram of the proposed model for answer sheets evaluation.

In the study, the student answer and the reference answer are combined using special tokens and uploaded to BERT: [CLS] reference answer [SEP] student answer [SEP]. We employed the term ''bert-base-uncased'' for the pre-trained BERT. Extra layers (specifically designed to address multiple heads, dense, LSTM, and Conv1D) are incorporated into the output of the BERT model in order to make it appropriate for the task.

B. CUSTOMIZED MULTI-HEAD ATTENTION LAYER

The BERT model provides rich contextual embeddings, which are further processed by multi-head attention layers to extract complex patterns and relationships. The combination of BERT's powerful embeddings and advanced attention mechanisms improves the model's ability to perform complex NLP tasks.

Instead of using the standard multi-head attention layer introduced by the work of Vaswani et al. [18], a customized multi-head attention layer is utilized in order to increase the

TABLE 1. Differences between classical multi-head attention and proposed multi-head attention.

Standard Multi-Head Attention	Customized Multi-Head Attention
Linear Projection	
$Q=XW^Q$, $K=XW^K$, $V=XW^V$ Here, X is the input matrix and W^Q, W^K, W^V are the learnable weight matrices.	$Q=XW^Q$, $K=XW^K$, $V=XW^V$ However, here the dimensions of the W^Q, W^K, W^V matrices are fixed as (768, model_dimensions).
Head Splitting	
Q, K and V are divided for each head: $Q_i=Q[:,d_i]$, $K_i=K[:,d_i]$, $V_i=V[:,d_i]$ Here d_i is the size index for each title.	The matrix is reshaped according to the number of headings and depth dimension and the transpose process is applied. Q_i=ReshapeAndTranspose(Q), K_i =ReshapeAndTranspose(K), V_i=ReshapeAndTranspose(V)
Scaled Dot-Product Attention	
Attention (Q_i, K_i, V_i)=softmax $(\frac{Q_iK_i^T}{\sqrt{d_k}})V_i$	Attention (Q_i, K_i, V_i)=softmax $(\frac{Q_iK_i^T}{\sqrt{d_k}})V_i$
Concatenation of Heads and Final Linear Projection	
Output=Concat($head_1$, …, $head_h$)W^O	Here, "ReshapeAndTranspose" is again a transposition and reshaping operation after merging the titles and the W^O weight matrix is used. Output=ReshapeAndTranspose(Concat($head_1$,…,$head_h$))W^O

efficiency of the proposed model for the short answer grading task. The customized multi-head attention layer implements a new self-attention mechanism that multiplies the input matrix with the query, key and value matrices. The attention mechanism is then applied and the result is multiplied by the output weight matrix to obtain the final output of the multi-head attention layer. This allows for a more nuanced understanding of the input, especially in natural language processing (NLP) tasks where contextual understanding is crucial. Key aspects include improved context understanding, scalable attention mechanisms, and focusing on different parts of the input sequence simultaneously. Its similarities and differences with the classic multi-head attention layer are shown in Table 1.

In the "Head Splitting" section, the final size of the matrix is divided into two according to the number of headings and the depth of each heading. We reshape the matrix with "Reshape". Reshaping is necessary so that each header can process relevant information separately. For example, if the model has layers of size 512 and 8 attention headings, there will be a depth of size 64 for each heading. The "Reshape" process logic is as follows:

Reshape(X)

→ [batch_size, sequence_length, num_heads, depth]

After reshaping, the dimensions of the matrix are usually; [batch_size, sequence_length, num_heads, depth]. However, during the attention process, the order of the headers (num_heads) and the order of the sequence elements (sequence_length) are changed to process each head separately. This is achieved by transposing the matrix. The transpose operation facilitates parallel processing of titles on each array element. Matrix after transposing:

Transpose

→ [batch_size, num_heads, sequence_length, depth]

C. INTEGRATION WITH LSTM AND CONVOLUTIONAL LAYERS

The typical feature of Recurrent Neural Networks (RNN) is their cyclic connection, allowing them to update the current state based on past states and the current input. However, RNNs struggle with "long-term dependencies" when there's a large gap between relevant input data [21]. To address this, Hochreiter and Schmidhuber [22] introduced the Long Short-Term Memory (LSTM) cell. This innovation improved the memory capabilities of the standard recurrent cell by incorporating a "gate" mechanism. Since its inception, LSTMs have undergone numerous modifications and gained popularity through the work of various researchers, including Gers [23] and Gers and Schmidhuber [24].

CNN-based research has been employed in the classification of text and has demonstrated a superior performance than sequence-based approaches [25]. They have demonstrated an advantage over standard CNNs and LSTMs when

combined with LSTMs [26], [27]. In our implementation, three different 1D CNN layers are employed in succession to the output of the LSTM layer instead of the traditional method of employing multiple different CNN layers. This paralleled architecture enables the model to simultaneously capture multiple local text properties, unlike the serial architecture that processes information in order. Additionally, each layer in the CNN has a different-sized kernel, which enables it to capture different aspects of the input data. As a result, a more extensive extraction of information from the text is achieved. The computation for CNN layers is as follows:

$$\text{Conv1D}_k(x) = f(W_k * x + b_k) \tag{1}$$

where W_k and b_k are the weight and bias parameters of the k-th Conv1D layer, f is the activation function (relu in this case), $*$ represents the convolution operation, and x is the input vector that is the output of the Bi-LSTM layer in this case.

The outputs of the three CNN layers are combined with the ''Concatenate'' function of the keras library, this is then sent to the pooling layer. This layer calculates the average value of each feature of the aggregated output.

$$\text{GlobalAveragePooling1}(y) = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

where y is the concatenated output, and n is the number of features.

D. OUTPUT LAYER

Ultimately, the vector derived from the ''GlobalAveragePooling'' layer is inputted into a Dense layer and a softmax function is employed to classify the vector in this dense layer. This design choice enables the classification task by reducing the number of features that are extracted by previous layers to a fixed-sized output vector, which is appropriate for multi-class classification.

$$\text{Dense softmax}(z_i) = \text{RMSE} = \frac{e^{z_i}}{\sum_{j=1} e^{z_j}} \tag{3}$$

where z is the output from the ''GlobalAveragePooling'' layer.

IV. METHODOLOGY

In this section, we provide a brief overview of the methods involved in our experiments. We explain the working principle of our model and the characteristics of the data sets used for the study. We also describe in detail our analysis, comparison methods, and experimental methods.

A. DATASETS

In this study, we used 5 different datasets to evaluate the performance of the proposed model: SemEval-2013 Beetle (2-way and 3-way) datasets and SciEntsBank (2-way and 3-way) datasets [11] and Mohler dataset [12]. In 2-way datasets, answers are grouped as ''Correct'' and ''Incorrect''. In 3-way data sets, the answers are grouped as ''Correct'', ''Incorrect'' and ''Contradictory''. Test data sets may contain questions from three different groups known as unseen answers, unseen questions and unseen domains.

Unseen Answers (UA): The questions in this test data set are the same as the questions in the training data set. However, the answers belong to different students. It is used to evaluate system performance based on the answers to questions in the training data set.

Unseen Questions (UQ): There are questions that are not included in the training data set but they are related to the subject areas in the training data set. It is used to measure the performance of the system against different questions in the same subject areas.

Unseen Domain (UD): It is a test data set consisting of questions in subject areas that are not included in the training data set. It is used to measure the performance of the system against different issues in unseen domains.

The train data set of the Bettle 2-way and 3-way datasets has 47 questions and 3941 answers. The test data set consists of 2 parts: ''Unseen Answers'' and ''Unseen Questions''. There are 47 questions and 439 answers in the ''Unseen Answers'' section and 9 questions and 819 answers in the ''Unseen Questions'' section.

There are 135 questions and 4969 answers in the train data set of SciEntsBank 2-way and 3-way datasets. The test data set consists of 3 parts: ''Unseen Answers'', ''Unseen Domains'' and ''Unseen Questions''. There are 135 questions and 540 answers in the ''Unseen Answers'' section; 46 questions and 4562 answers in the ''Unseen Domains'' section; and 15 questions and 733 answers in the ''Unseen Questions'' section.

Mohler dataset contains 79 questions and 2273 student answers. 2 educators scored the answers between 0-5. In the data set, the averages of the 2 educators are available along with their grades.

B. EXPERIMENTAL SETUP

The system architecture is implemented in Python, the architecture is trained using a GPU machine, on the Cuda Tensorflow 2.10.0. BERT-base is used as the basis for the study of the model. The BERT model is comprised of a hidden dimension of 768 dimensions and has 12 attention heads, it is combined with three primary input layers: ''input_ids'', ''attention_masks'', and ''token_type_ids'', each of which collects different information important to the interpretation of the text. The student answer and the reference answer are combined using special tokens and uploaded to BERT:

[CLS] reference answer **[SEP]** student answer **[SEP]**

Post BERT processing, the sequence and pooled outputs are further refined through a dense layer followed by a bidirectional LSTM layer, enhancing the model's ability to grasp deeper textual contexts.

A critical aspect of the proposed model is the use of parallel CNN layers, as opposed to a serial configuration. The number of filters used for CNN layers is 128, and the activation function is ReLU. The kernel size is 2 for the first CNN layer,

TABLE 2. Fixed parameters for bert.

Parameter name	Values
Attention probs dropout prob	0.1
Bos token id	0
Hidden act	gelu
Hidden dropout prob	0.1
Hidden size	768
Initializer range	0.02
Intermediate size	3072
Layer norm eps	1e-12
Max position embeddings	512
Num attention heads	12
Num hidden layers	12
Pad token id	0
Type vocab size	2
Vocab size	30522

3 for the second, and 4 for the third. The unit for Bi-LSTM is set to 128. Adam optimizer is used as the optimizer to train the data and set the learning rate to 1e-5. We chose the number of epochs as 5 throughout the entire experiment. Since SemEval-2013 Beetle and SciEntsBank datasets have separate train and test sets, their test sets are used for evaluation. The 5-fold cross-validation is used for the Mohler dataset.

This model configuration, especially the use of parallel CNNs, offers a robust approach to text classification. It leverages the extensive contextual understanding from BERT and LSTM layers and combines it with the diverse local pattern recognition capabilities of parallel CNNs. This approach contrasts with traditional serial CNN models, where the sequential processing of layers might limit the scope of feature extraction to only what is passed from one layer to the next. The proposed parallel CNN strategy ensures a broader and more nuanced capture of textual features, promising enhanced accuracy and efficiency in classification tasks.

Several parameters which mainly depend on the transformer architecture (such as Vocabulary (vocab) size, Hidden dropout probability (prob), etc) are fixed with default configurations as the ''BERT- base'' transformer model to avoid heavy compute consumption. Table 2 gives these parameter values to be used for the configuration of the BERT. However, we select specific hyperparameter values suitable for the datasets used. We utilize standard parameters that yielded the best results on different datasets, as shown in Table 3. The reason for selecting standard parameters for each dataset is to test the performance of our model in different environments.

Even though the datasets had different classifications, we kept our model constant for each. Our purpose in keeping it constant is to accurately demonstrate the efficiency of performance in different data sets.

C. EVALUATION METRICS

Different approaches were applied to evaluate the proposed model. We determined three principal summary metrics,

TABLE 3. Hyper-parametres for model.

Parameter name	Values
Num_train epochs	5
Seed	42
Max seq length	128
Multiprocessing chunksize	32
Gradient accumulation steps	1
Optimizer	Adam(1e-5)
BiLSTM units	128
BiLSTM return sequences	True
CNN-1 output size	128
CNN-1 kernel size	2
CNN-1 activation	relu
CNN-2 output size	128
CNN-2 kernel size	3
CNN-2 activation	relu
CNN-3 output size	128
CNN-3 kernel size	4
CNN-3 activation	relu
Dense layers units	128
Dense layers activation	relu
Final Dense layer activation	softmax

Macro F1, Weighted F1, and Accuracy for Beetle and SciEntsBank. We calculated the Pearson correlation score and Root Mean Square Error (RMSE) for the Mohler data set.

Accuracy (Acc): It is the most intuitive performance measure. It is simply the ratio of correctly predicted observations to the total observations. Formula:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{4}$$

Macro F1 Score (M-F1): It is a method used to measure the accuracy of the model in classification tasks, especially in a dataset with multiple classes or labels.

$$Macro\ F1\ Score = \frac{1}{N}\sum_{i=1}^{N} F1_i \tag{5}$$

where $F1_i$ is the F1 score of the i-th class [28].

Weighted F1 Score (W-F1): It is a variant of the F1 Score used in classification tasks, especially when dealing with imbalanced datasets across multiple classes. Like the Macro F1 Score, it aims to provide a single metric that balances precision and recall.

$$\text{Weighted F1 Score} = \sum_{i=1}^{N} w_i x F1_i \tag{6}$$

where w_i is the weight of the i-th class, which is typically the number of true instances of the class divided by the total number of instances. $F1_i$ is the F1 score of the i-th class [28].

Pearson Correlation: It is a statistical method that measures the strength and direction of the linear relationship between two variables. It is usually symbolized by ''r'' and takes a value between -1 and +1. The value of the Pearson correlation coefficient indicates how strong the relationship between two variables is.

TABLE 4. Results of SciEntsBank 2-way.

	UA			UQ			UD		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
CoMeT	0.774	0.768	0.773	0.603	0.579	0.597	0.676	0.670	0.677
ETS	0.776	0.762	0.077	0.633	0.602	0.622	0.627	0.543	0.574
SOFTCAR	0.724	0.715	0.722	0.745	0.737	0.745	0.711	0.705	0.712
Feature ensemble	0.708	0.676	0.690	0.705	0.678	0.695	0.712	0.703	0.712
TF+SF without question	0.779	0.771	0.777	0.749	0.738	0.747	0.708	0.690	0.702
TF+SF with question	0.792	0.785	0.791	0.700	0.685	0.698	0.719	0.708	0.717
XLNET model	0.792	0.781	0.788	0.736	0.724	0.734	0.702	0.679	0.693
Roberta-lrg-v1	0.807	0.804	0.806	0.727	0.724	0.731	0.690	0.690	0.700
Proposed Model	**0.806**	**0.805**	**0.805**	**0.754**	**0.754**	**0.754**	**0.710**	**0.710**	**0.709**

Root Mean Square Error: It is the square root of the mean square of the squares of the differences between the true values and those predicted by the model or predictor. Lower RMSE values indicate that the predictions are closer to the true values, indicating better model performance.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2} \tag{7}$$

V. EVALUATION RESULTS

A. TWO-WAY TASK

The proposed model exhibited high accuracy in classifying answers into 'correct' or 'incorrect' categories. Over 80% success is achieved in both two-way datasets (SciEntsBank and Beetle). Higher or very compatible results are obtained compared to previous studies in the literature. The proposed model effectively identifies key phrases and terms that are indicative of correct answers, demonstrating its strong pattern recognition capabilities. The high accuracy rate indicates the model's proficiency in binary classification tasks. The success of the proposed model can be attributed to its advanced natural language processing capabilities, which allow it to understand and evaluate the content of the answers accurately.

Performance results (Accuracy, M-F1 and W-F1 scores) for SciEntsBank 2-way task are given in Table 4. Table 4 gives the results for the proposed model and other ASAG systems including CoMeT [29], ETS [30], SOFTCAR [31], Feature ensemble approach [32], TF+SF [33], XLNET model [34] and Roberta-large-v1 [17]. When the results of the SciEntsBank 2-way dataset were analyzed, it can be seen that the proposed model give consistent results in all test data. While other models succeeded in different test types, our model obtained either the best result or close to the best scores in all test data.

The results shown in the tables (Table 5 and Table 7) for the Beetle dataset are the performances of the teams participating in the ''Second Joint Conference on Lexical and Computational Semantic'' workshop held in 2013. Dzikovska et al. [11] published the results of this workshop and those results together with the results of our proposed system are given in those tables. When Table 5 is analyzed for the Beetle 2-way dataset, it is seen that the same score is obtained with the best result in the UA test type. However, in the UQ test data, it was seen that our model give much better results than the other models. Since the UA test type consists of only the questions in the training data set, the scores in this area are higher. However, since the UQ test data contains different questions, other models could not adapt to this test data. However, our model also adopts very well to this test data.

B. THREE-WAY TASK

In the 3-way classification task, the model successfully categorizes answers into 'correct', 'partially correct', and 'incorrect' with a notable accuracy. Table 6 and 7 give performance results of the proposed system and other ASAG systems. For SciEntsBank 3-way, better results are obtained than previous studies, with an accuracy rate of 76%. The 0.738 Macro-Average F1 score for the Beetle 3-way is a better result compared to similar studies. These results indicate the strength of our model in identifying partially correct answers, which is a challenging aspect of such tasks. The ability to identify partially correct answers highlights the model's nuanced understanding of the subject matter and the context of the questions.

The results for SciEntsBank 3-way better reflects the performance of our model. The fact that the answers had 3 options poses a significant challenge for the models. Achieving the highest score in all data types shows that our model adapts to different conditions. Since the UQ and UD test data are completely different from the training data set, the high scores in these areas show the effectiveness of our model.

C. MOHLER TASK

For the Mohler data set, we divide the scoring part into groups. We make 11 different classifications between 0, 0.5, 1, 1.5,.., 4.5, 5. Applying the model to the Mohler dataset, a standard benchmark in automated grading, resulted in a Pearson Correlation Coefficient of 0.747 with human graders. The proposed model demonstrates a strong ability to align with human judgment, particularly in grading complex, subject-specific answers. The correlation with human graders indicates the model's effectiveness in mimicking human grading patterns, a crucial aspect of real-world educational

TABLE 5. Macro-average F1 results of beetle 2-way.

	UA	UQ
CELI	0.640	0.656
CNGL	0.800	0.666
CoMeT	0.833	0.695
CU	0.778	0.689
ETS	0.833	0.702
LIMSIILES	0.723	0.641
SoftCardinality	0.774	0.635
UKP-BIU	0.608	0.481
Proposed Model	**0.830**	**0.740**

TABLE 6. Results of SciEntsBank 3-way.

	UA			UQ			UD		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
CoMeT	0.713	0.64	0.707	0.546	0.380	0.522	0.579	0.404	0.550
ETS	0.720	0.647	0.708	0.583	0.333	0.537	0.543	0.333	0.461
SOFTCAR	0.659	0.555	0.647	0.652	0.469	0.634	0.637	0.486	0.620
Feature ensemble	0.604	0.443	0.569	0.642	0.455	0.615	0.626	0.451	0.603
TF+SF without question	0.648	0.553	0.638	0.615	0.423	0.584	0.632	0.449	0.608
TF+SF with question	0.696	0.640	0.690	0.548	0.450	0.559	0.560	0.421	0.532
XLNET model	0.718	0.666	0.714	0.613	0.491	0.628	0.632	0.479	0.611
Proposed Model	**0.763**	**0.762**	**0.760**	**0.655**	**0.654**	**0.652**	**0.663**	**0.645**	**0.651**

TABLE 7. Macro-average and weighted-average F1 results of beetle 3-way.

	UA		UQ	
	M-F1	W-F1	M-F1	W-F1
CELI	0.494	0.519	0.441	0.463
CNGL	0.567	0.592	0.450	0.471
CoMeT	0.715	0.728	0.466	0.488
ETS	0.710	0.723	0.585	0.597
LIMSIILES	0.563	0.587	0.431	0.454
SoftCardinality	0.596	0.616	0.439	0.451
UKP-BIU	0.468	0.472	0.333	0.313
Proposed Model	**0.738**	**0.749**	**0.638**	**0.647**

applications. The model's performance suggests its adaptability to different academic disciplines and grading criteria. Table 8 gives the results for the proposed model and other ASAG systems including Tf-Idf [32], Lesk [35], Mohler et al. [12], Ramachandran et al. [36], Sultan et al. [32], TF+SF [without question] [33], TF+SF [with question] [33], BERT Regressor + Similarity Score [37].

TABLE 8. Results of Mohler.

	Pearson's r	RMSE
Tf-Idf	0.327	1.022
Lesk	0.450	1.050
Mohler	0.518	0.978
Ramachandran	0.610	0.860
Sultan	0.592	0.887
TF+SF [without question]	0.542	0.921
TF+SF [with question]	0.570	0.902
BERT Regressor + Similarity Score	0.777	0.732
Proposed Model	**0.747**	**0.856**

VI. EFFECTS OF CUSTOMIZED MULTI-HEAD ATTENTION AND PARALLEL CONNECTION OF CNN LAYERS

This study proposes a novel ASAG hybrid model based on BERT. Evaluation results show significant improvements in scoring accuracy and the model's ability to effectively understand and score short answers. Compared with existing ASAG models, the proposed BERT-based method is either superior or comparable in performance. Although some of the advantages of the proposed model rely on BERT's ability to interpret language, adding two novel methods to the proposed system further improves the performance of the proposed model in the ASAG task. These additions enable more nuanced and accurate evaluation of student responses, especially in understanding context and semantic meanings.

TABLE 9. Comparison of the proposed multi-head attention layer.

		Standard Multi-Head Attention	Proposed Multi-Head Attention
UA	Acc	0.741	**0.763**
	M-F1	0.740	**0.762**
	W-F1	0.737	**0.760**
UQ	Acc	0.630	**0.655**
	M-F1	0.630	**0.654**
	W-F1	0.626	**0.652**
UD	Acc	0.644	**0.663**
	M-F1	0.643	**0.645**
	W-F1	0.631	**0.651**

TABLE 10. Comparing the use of CNN layers.

		Serial CNN	Proposed Parallel CNN
UA	Acc	0.741	**0.763**
	M-F1	0.740	**0.762**
	W-F1	0.741	**0.760**
UQ	Acc	0.641	**0.655**
	M-F1	0.641	**0.654**
	W-F1	0.638	**0.652**
UD	Acc	0.643	**0.663**
	M-F1	0.642	**0.645**
	W-F1	0.637	**0.651**

The first aspect that contributes positively to the performance of the model is the customized Multi-Head Attention layer. In order to measure the contribution of the proposed Customized Multi-Head Attention layer to the performance, the evaluation results are obtained by replacing Customized Multi-Head Attention layer with the standard Multi-Head Attention layer in the proposed model. In the provided Customized Multi-Head Attention class, the weights for the query (Q), key (K), value (V) and output (O) matrices are initialized with a specific strategy (random_normal) and have a special shape defined directly in the class. This explicit and direct control overweight initialization and dimensions has a positive effect on the performance of the model. Table 9 gives the performance results of both variations for the SciEntsBank 3-way dataset. As shown in Table 9, the proposed Customized Multi-Head Attention layer outperformed the standard Multi-Head Attention layer.

The second novel approach in the proposed model is the parallel connection of CNN layers instead of their serial connection. We aim to capture a wider range of linguistic patterns using parallel CNNs. Some filters might be sensitive to short-range syntax (like phrases), while others excel at capturing longer-range dependencies (like relationships between sentences). Each CNN filter in the parallel configuration operates independently on the LSTM's output, using a different kernel size (2, 3, or 4 words). This allows our model to capture diversity of linguistic features, ranging from local syntactic patterns to broader semantic relationships between sentences. Also, parallel CNNs are less susceptible to overfitting on specific features. We put serial CNN layers instead of parallel CNN layers into the model in order to see the effect of parallel connection on performance. The performance results for both variations for the SciEntsBank 3-way dataset are given in Table 10. It can be seen that parallel linking contributes more positively to the model compared to serial linking.

VII. CONCLUSION

This research aims to contribute to the ASAG field by presenting a new BERT-based hybrid model. In this study, we use 5 different datasets to evaluate the performance of the proposed model. The language of the datasets is English. Accuracy rates for SciEntsBank and Beetle data sets are given in the analyses in Tables 4, 5, 6, and 7. Pearson and RMSE values for the Mohler data set are shown in Table 8. Our proposed approach, together with two innovative contributions, achieves results close to human evaluation in terms of grading and offers a different perspective in making sense of student responses. Our proposed model is a scalable and efficient solution for automatic annotation, but we aim to further improve its performance. Because the evaluation process in education requires effort and time, it is important to develop and improve such ASAG programs.

By studying different datasets used in the literature, we can see both the compatibility and performance of the proposed model. The performance of our model encourages us to create more sophisticated models in the future. We can see that our model can capture extra relationships between some words, which cannot be captured by the pre-trained BERT model. BERT provides the biggest impact on the performance of the model. For this reason, improving the vocabulary of BERT will increase the efficiency of the model. In future studies, for this purpose, we will try to increase the performance of the model by adding domain-specific words to BERT.

ACKNOWLEDGMENT

The OpenAI's ChatGPT system was employed solely for the purpose of improving the grammatical quality of this work.

REFERENCES

[1] A. Bozkurt, I. Jung, J. Xiao, V. Vladimirschi, R. Schuwer, G. Egorov, S. Lambert, M. Al-Freih, J. Pete, D. Olcott, and V. Rodes, "A global outlook to the interruption of education due to COVID-19 pandemic: Navigating in a time of uncertainty and crisis," *Asian J. Distance Educ.*, vol. 15, no. 1, pp. 1–126, 2020.

[2] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *Int. J. Artif. Intell. Educ.*, vol. 25, no. 1, pp. 60–117, Mar. 2015.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2019, pp. 4171–4186.

[4] X. Zhu, H. Wu, and L. Zhang, "Automatic short-answer grading via BERT-based deep neural networks," *IEEE Trans. Learn. Technol.*, vol. 15, no. 3, pp. 364–375, Jun. 2022.

[5] W. Wang, H. Zhuang, M. Zhou, H. Liu, and B. Li, "What makes a star teacher? A hierarchical BERT model for evaluating teacher's performance in online education," 2020, *arXiv:2012.01633*.

[6] N. A. Khodeir, "Bi-GRU urgent classification for MOOC discussion forums based on BERT," *IEEE Access*, vol. 9, pp. 58243–58255, 2021.

[7] C. Sung, T. I. Dhamecha, and N. Mukhi, "Improving short answer grading using transformer-based pre-training," in *Proc. Artif. Intell. Educ.*, vol. 11625, Chicago, IL, USA, Jun. 2019, pp. 469–481.

[8] L. Camus and A. Filighera, "Investigating transformers for automatic short answer grading," in *Artificial Intelligence in Education*, vol. 12164. Springer, Jun. 2020, pp. 43–48.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Represent.*, 2019, pp. 1–17.

[11] M. O. Dzikovska, R. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. T. Dang, "Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics*, vol. 2, 2013, pp. 263–274.

[12] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 752–762.

[13] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, and R. Arora, "Pre-training BERT on domain resources for short answer grading," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, 2019, pp. 6071–6075.

[14] A. Condor, M. Litster, and Z. Pardos, "Automatic short answer grading with SBERT on out-of-sample questions," in *Proc. 14th Int. Conf. Educ. Data Mining*, 2021, pp. 1–8.

[15] S. Cao, H. Vo, H. T.-T. Le, and D. Dinh, "Hybrid approach for text similarity detection in Vietnamese based on sentence-BERT and WordNet," in *Proc. 4th Int. Conf. Inf. Technol. Comput. Commun. (ITCC)*, Jun. 2022, pp. 59–63.

[16] W. Lei and Z. Meng, "Text similarity calculation method of Siamese network based on Albert," in *Proc. Int. Conf. Mach. Learn. Knowl. Eng. (MLKE)*, Feb. 2022, pp. 251–255.

[17] M. A. Sayeed and D. Gupta, "Automate descriptive answer grading using reference based models," in *Proc. OITS Int. Conf. Inf. Technol. (OCIT)*, Dec. 2022, pp. 262–267.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1–11.

[19] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2018, pp. 2227–2237.

[20] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. Accessed: Mar. 20, 2024. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[21] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[23] F. Gers, "Long short-term memory in recurrent neural networks," Doctoral thesis, Federal Inst. Technol. Lausanne, Univ. Hanover, Germany, 2001.

[24] F. A. Gers and E. Schmidhuber, "LSTM recurrent networks learn simple context-free and context-sensitive languages," *IEEE Trans. Neural Netw.*, vol. 12, no. 6, pp. 1333–1340, Dec. 2001.

[25] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 959–962.

[26] Y. Luan and S. Lin, "Research on text classification based on CNN and LSTM," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Mar. 2019, pp. 352–355.

[27] B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, "Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism," *Appl. Sci.*, vol. 10, no. 17, p. 5841, Aug. 2020.

[28] M. Laricheva, C. Zhang, Y. Liu, G. Chen, T. Tracey, R. Young, and G. Carenini, "Utterance labeling of conversations using natural language processing," in *Social, Cultural, and Behavioral Modeling* (Lecture Notes in Computer Science), vol. 13558, R. Thomson, C. Dancy, and A. Pyke, Eds., Springer, 2022, pp. 241–251.

[29] N. Ott, R. Ziai, M. Hahn, and D. Meurers, "CoMeT: Integrating different levels of linguistic modeling for meaning assessment," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics*, vol. 2, 2013, pp. 608–616.

[30] M. Heilman and N. Madnani, "ETS: Domain adaptation and stacking for short answer scoring," in *Proc. 2nd Joint Conf. Lexical Comput. Semantics*, vol. 2, 2013, pp. 275–279.

[31] S. Jimenez, C. Becerra, and A. Gelbukh, "SOFTCARDINALITY: Hierarchical text overlap for student response analysis," in *Proc. Joint Conf. Lexical Comput. Semantics*, vol. 2, 2013, pp. 280–284.

[32] M. A. Sultan, C. Salazar, and T. Sumner, "Fast and easy short answer grading with high accuracy," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, San Diego, CA, USA, 2016, pp. 1070–1075.

[33] S. Marvaniya, S. Saha, T. I. Dhamecha, P. Foltz, R. Sindhgatta, and B. Sengupta, "Creating scoring rubric from representative student answers for improved short answer grading," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2018, pp. 993–1002.

[34] H. Ghavidel, A. Zouaq, and M. Desmarais, "Using BERT and XLNET for the automatic short answer grading task," in *Proc. 12th Int. Conf. Comput. Supported Educ.*, 2020, pp. 58–67.

[35] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine code from an ice cream cone," in *Proc. 5th Annu. Int. Conf. Syst. Documentation*, 1986, pp. 24–26.

[36] L. Ramachandran, J. Cheng, and P. Foltz, "Identifying patterns for short answer scoring using graph-based lexico-semantic text matching," in *Proc. 10th Workshop Innov. Use NLP Building Educ. Appl.*, 2015, pp. 97–106.

[37] J. Garg, J. Papreja, K. Apurva, and G. Jain, "Domain-specific hybrid BERT based system for automatic short answer grading," in *Proc. 2nd Int. Conf. Intell. Technol. (CONIT)*, Jun. 2022, pp. 1–6.

MUSTAFA KAYA was born in 1986. He received the B.S. degree in computer engineering from Selcuk University, Türkiye, in 2015. He is currently pursuing the Ph.D. degree with the Department of Computer Engineering, Hacettepe University, Ankara, Türkiye. His research interests include natural language processing and deep neural networks.

ILYAS CICEKLI received the Ph.D. degree in computer science from Syracuse University, USA. He worked as a Visiting Faculty Member at Syracuse University, from 2017 to 2019, and the University of Central Florida, from 2001 to 2003. He is currently a Full Professor with the Department of Computer Engineering, Hacettepe University, Ankara, Türkiye. He has authored more than 80 journal articles in the field of natural language processing, including more than 30 journal articles. His research interests include natural language processing, text summarization, example-based machine translation, question-answering, and information extraction.