Automatic Short Answer Grading using pretrained BERT model

Esraa Hassan Ahmed
Electronics and Communications department
Faculty of Engineering Cairo University
Cairo, Egypt
esraaarafat2020@gmail.com

Eman Mahmoud Abdel- Halim
Electronics and Communications department
Faculty of Engineering Cairo University
Cairo, Egypt
emannegm2021@gmail.com

Noran Waleed Mohamed
Electronics and Communications department
Faculty of Engineering Cairo University
Cairo, Egypt
nouranwaled28@gmail.com

Eslam Khaled Mohamed
Electronics and Communications department
Faculty of Engineering Cairo University
Cairo, Egypt
eslam5khaled16@gmail.com

Mohamed Magdy Abd-El Ghaffar
Electronics and Communications department
Faculty of Engineering Cairo University
Cairo, Egypt
Mmagdy191@gmail.com

Norhan Ahmed Hamed
Electronics and Communications department
Faculty of Engineering Cairo University
Cairo, Egypt
norhan.ahmedeece228@gmail.com

Mohsen A. Rashwan
Electronics and Communications department
Faculty of Engineering Cairo University
Cairo, Egypt
Mrashwan@rdi-eg.ai

***Abstract*—in this paper we tried to get benefit from pre-trained transformer-based model, specifically BERT model, in ASAG task, Automatic Short Answer Grading is challenging fields in NLP, the purpose of ASAG is to build a model can give a numeric grade to students for their answers on a certain question with presence of the model answer, this method of grading depends on the semantic similarity between the model answer and the student answer, in this paper many experiments are done to provide acceptable solutions for ASAG task, using different datasets to train the model, we reached excellent results compared to previous work done using different techniques, the main problem is leak of data, we tried to use transfer learning to solve part of this problem.**

Keywords—Transformer, BERT, ASAG, Transfer Learning

I. INTRODUCTION

A. *Motivaton*

Assessing the efficiency and quality of the educational process and the extent to which students understand the course is very important. The most common used assessment method is exams. The questions included in these exams may exist in two forms, either closed (e.g., essays or short answers) or open questions (e.g., multiple-choice). Open-ended questions require more details to fully answer the question, so they force the student to provide a brief description of their knowledge.

But the manual assessment process is a tedious process, especially in this type of open-ended questions, for several reasons: the assessment process is a very time-consuming process, and the correction process is unfair, as it varies from one teacher to another, and may differ for the same teacher. This leads us to think about how to make this assessment process automatic, but we will focus on the Automatic Short Answer Grading (ASAG). The benefits that accrue to us from using automatic grading are many. Making the assessment process automatic leads to formalization of assessment criteria and speeding up the assessment process.

The automatic grading varies in difficulty according to the nature of the question. Short answer question grading is more difficult than other the other types of open questions because it requires creating a model capable of understanding natural language. Therefore, we took the risk of this challenge and tried to build a model that can grade the short answer.

B. *Scope of the project*

The scope of our project, or in other words, the determinants that determine the framework in which the project operates is as follows: First, all the input data for the model is English data only, and therefore the model grades only the English question. Second, the answer length should be roughly between one phrase and one paragraph. Thirdly, the assigned degree will be in the range from 0 to 5. Fourth, the reference answer must be sufficient because the grading process in our project is based on the semantic similarity between the student answer and reference answer, and therefore the question is not considered part of the data.

C. Problem statement

Automatic short answer grading (ASAG) is the task of assessing short natural language responses to objective questions using computational methods. According to a short answer question is defined as the question that can be considered as meeting at least five specific criteria. First, the question must require a response that recalls external knowledge instead of requiring the answer to be recognized from within the question. Second, the question must require a response given in natural language. Third, the answer length should be roughly between one phrase and one paragraph. Fourth, the assessment of the responses should focus on the content instead of writing style. Fifth, the level of openness in open-ended versus closed-ended responses should be restricted with an objective question design.

The general form of ASAG process is described in 11 steps as shown in **Figure 1** First Test or exam sitting for students (1) Then all the required data for the assessment model such as questions, reference answers and student answers are gathered (2). All data sets are collected in XML file or other similar format (3). Natural Language Processing (NLP) techniques are applied on datasets (4) to produce a well- preprocessed data (5). Some of these datasets will be the material to be used in model building and, in the training, process using the supervised machine learning methods (6). The reminder of data will be used for the test process in which the automatically graded are done (7/8) providing us with a predicted scores or labels (9) which will be used in model evaluation (10) finally we are able to calculate the efficiency of the model (10). [1]

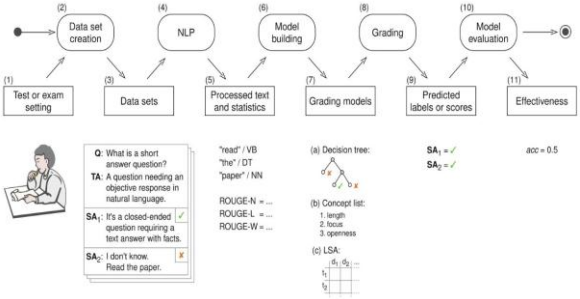

Figure 1: An ASAG system development pipeline

Thus, given the block diagram shown in **Figure 1**, we can state the problem clearly as a set of data and requirements. The data for our problem are labeled datasets that include students answer, reference answer and grade. The requirements are preprocessing the data well as it affects the performance of the model directly, designing a model with an effective, high-performance training approach, create a model that deals efficiently with data sparsity and make the model decisions comprehensible and explainable.

D. General view

Research and contributions have been presented in the field of ASAG. In [2] they first apply standard data mining techniques to the corpus of student answers for the purpose of measuring similarity between the student answers and the model answer. This is based on the number of common words. In [6] they study multi-task training methods and domain adaptation on Automatic Short Answer Grading (ASAG) using the text-to-text transfer transformer model (T5). They fine-tuned a multi-task model that is trained on a profound selection of related tasks and an extensively pre-trained model. In our project we use un-based Bert transformer to design the ASAG model.

E. Paper organization

The remainder of this paper is organized as follows. In Chapter II we analyze the existing related work in the field of ASAG and we mention data sets required for the model. Chapter III we deeply illustrate the concept of machine learning approaches, transfer learning, fine tuning and transformers, more details have been presented in a specific type of transformers , Bert transformer . Then, in Chapter IV the data preprocessing steps are illustrated clearly then we describe the underlying methodology used to build a model capable of implementing ASAG. This is followed by a detailed discussion of the Experiments and results associated with the work presented in Chapter V. Finally, in Chapter VI, conclusions and recommendations for future work are given.

II. LITERATURE REVIEWS

A. Related works

[1] previews the task of ASAG and here in this chapter we will discuss the main systems which is used in our task. The early automatic short answer grading system was based on the mapping concept like Burstein and C-Rater: which was treating the student answer as a group of concepts and system's role to check these concepts and compare it to reference answer concepts to detect the grade of the student.

After that some papers became based on information extraction concept which was built on finding the main ide of the student answers by doing some matching operations and here one of examples of systems that uses the information extraction concept is WebAS system [14] which its role to get the main segments of the model answer and student answers and use regular expressions to match these segments, Also for another system of information extraction is CoSeC-DE (Comparing Semantics in Context) [13] which uses the lexical semantic method to understand the meaning and compared this meaning to reference answer meaning.
Later the concept of Corpus-Based Method started to be used as it depending on using statistical properties of large document corpra to facilitate understanding the meaning of the answers and to predict the grades and some papers that used this concept are Mohler [15] which developed several systems to get unsupervised grading methods by using eight knowledge based and also it modified the system by mixing the best student answers with the reference one to expand the model answer vocabulary, another system also used the same concept is BELU with LSA which is a matric used to evaluate the student answer to model answer.

There is another concept is represented which is Machine learning which uses some of measurements extracted from natural language processing techniques and combines them to get the grade using classification or regression model and one of the systems that uses this concept is Content assessment module [16] which uses a K-nearest neighbor classifier and features that measure the overlap of content on linguistic levels between the student

and reference answers. Also, the e-Examiner [17] uses ROUGE which are a set of metrics and software package which are combined as linear regression.

After the previous overview from Burrow's paper there are also some systems which also related to our task like [12] which uses text mining tools, clusters approach and bag of words model to transform the text into sets of data and convert them to numeric values and then using the cosine similarity to calculate the similarity between the reference answer and the student answer. Also in [7] it explains the concept of transfer learning and transformer and compare between the main models like BERT,ELMO and GPT and how they pretrained on huge corpus to extract the semantic context and how to use these model to required data to be used in our task

B. Datasets

There are main 6 data sets which are used in our task and in following they will be discussed in details.

1) Mohler Dataset: An English data set in computer science field which consists of 2442 student answers which were collected from 2 examinations and 10 assignments from university of North Texas.Two teachers graded the students answer and gave them a grade from 0 to 5.

2) PT-ASAG Dataset: A Portguese data set in biology field which consists of 9864 student answers for 15 questions.The answers are graded by 2 teachers and the grades are from 0 to 3.

3) AR-ASAG: An Arabic data set consistof 1967 student answers of 48 questions which depend on explanation and comparison concepts in security and hacking field and the grades from 0 to 5.

4) STS Dataset: It consists of English sentences pairs and their scores expresses how similar they are and these scores are from 0 to 5.This dataset consist of 8623 sentaence pairs in genral field.

5) SICK Dataset: It consists of about 10.000 sentences pairs and its similarity evaluation is done by score from 0 to 5.

6) SNLI Dataset: it consist of 570 K sentence pairs with classification with entailment , neutral and contradiction label to express the similarity of these sentences.

Data Problems AND Proposed Solution

In this section we mentioned the problems in datasets and the proposed solutions if they are available.

The first main problem we faced in the project is the leakage of datasets which has an obvious effect on the accuracy. As a result we proposed some solutions to overcome this problem. First, we translate the foreign datasets which are not in English language like AR-ASAG and PT-ASAG using specific tools to guarantee the efficiency of the translation and keep the semantic meaning of sentences. Second, we convert the classified data which is labeled in entailment and contradiction and neutral labels to numeric scores to be able to use them. The second problem in the data section was the imbalance between grades. That is because the number of samples with grades equal to 5 is more than the rest of samples and this problem affects the accuracy of the model. To solve this problem we use the "Augmentation process" by generating samples with the same meaning of the origin sample using an efficient tool to rephrase the sentence to balance the distribution of the data. The third problem is that there are number of reference answers for each question which is confusing as we use one reference answer and to solve this problem we reviewed these reference answers and students' answers and decide to choose the reference answer which is almost includes what is common of concepts of the rest of the reference answers The last problem is that there is more than one teacher grading the answers so we had to take the average of these grades to use it and pass it to the model.

III. REVIEW SCIENTIFIC ALGORITHMS

A. Machine learning approaches

In the machine learning field, there are two approaches that are mainly used to pre-train (or learn) the model. Supervised and unsupervised learning approaches. The main difference between those two approaches is that one of them uses labeled data to predict the output while the other doesn't.

Supervised learning is defined as a machine learning approach which uses a labeled data set either to classify data or to predict a numerical output in an accurate way. Therefore, supervised learning methods can be used in two different types of problems according to what output is desired to be predicted to the input labeled data. Those two types of problems are classification and regression problems. **Figure 2** could simply visualize different outputs resulting from using either Classification algorithm or Regression algorithm.

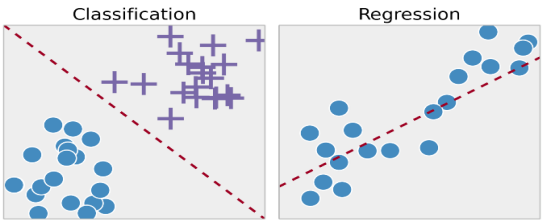

Figure 2:Supervised learning algorithms, Classification and Regrision

Unsupervised learning approach is more complicated to implement than the supervised one as it analyses the unlabeled datasets then try to discover the hidden pattern of data without the presence of human intervention, in other words machine depends on itself to totally to learn, that is why we call it "unsupervised learning approach [5]. Unsupervised learning algorithms serve three main tasks: Clustering, Association and Dimensionality reduction.

So, from the previous comparison and the problem statement described clearly in introduction chapter our problem is considered a regression problem and the model will use the supervised learning approach.

B. *Transfer learning*

As the traditional machine learning approach is an isolated learning approach where the model is trained to solve a single task. With this approach, no knowledge is retained or accumulated. Consequently, the learning approach relies only on the single task. On the other hand, transfer learning approach lets the model to make use of knowledge gained from previous tasks into a new and unknown task. The main idea behind is that an extensively trained base model can be used for new tasks. This makes typical training from scratch unnecessary, and knowledge can be retained. [4]

C. *Attention-based transformer architecture*

Attention technique is commonly used in NLP as it was designed mainly for the context of Neural Machine Translation (NMT) to enhance the performance of the Encoder-Decoder architecture by solving the disadvantage of Seq2Seq Models (long dependencies issue) which is represented by the fixed-length context vector than can't be able to memorize the long sentences as this technique forgets the earlier elements of the input sequence once it has processed the complete sequence.[24][25]

The Transformer is a model that uses attention to boost the speed with which these models can be trained. The biggest benefit, however, comes from how The Transformer lends itself to parallelization. A transformer essentially has two major components Encoder and Decoder. Each encoder consists of Multi Head-Attention layer and Feed Forward Neural Network and each decoder can be similarly decomposed with the addition of an encoder-decoder attention layer between the Multi Head-Attention layer and the feed-forward as shown in **Figure 3**. [27]

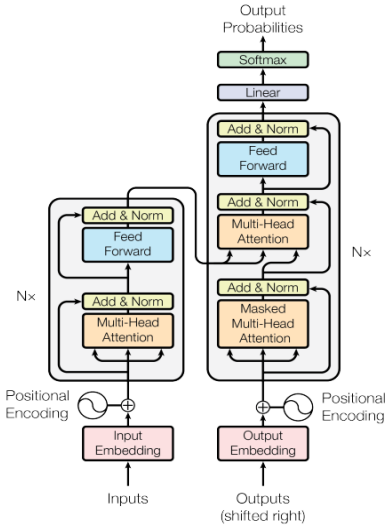

Figure 3:Transformer architecture

Input first enters the Multi Head-Attention layer of encoder which is the method the Transformer uses to bake the "understanding" of other relevant words into the one we're currently processing to facilitate the prediction process of the next word. Then the output is fed to the feed forward network. In The Multi Head-Attention of the decoder future tokens are "masked" when computing attention for that token.

Such attention-based transformer architecture with capability to pay attention to a specific subset of the sequential input data helped improve the performance of several NLP tasks. These models became state-of-the-art for most NLP tasks due to their better and more efficient performance in terms of computational resources. As described the most popular type of attention-based networks are the transformers which handle sequential data simultaneously rather than just sequentially like RNNs

D. *Bert (Bidirectional Encoder Representations from Transformers) model*

BERT refers to Bidirectional Encoder Representations from Transformers; This Model is one of the most effective types of the transformers. Due to its bidirectionality, this model can understand the overall context, so it can define the meaning of each word from the context both to the right and to the left of the word.

The Model that we used in our ASAG project is BERT – Base, Uncased Model. This model was pre-trained on English language using a masked language modeling (MLM) objective. Uncased model as it treats with "english" as "EnGlish". Pre-training Process depends on two effective bases Masked Language Modeling (MLM) where model masks some words randomly in the input sentence then runs the masked sentence to predict the masked words and Next sentence prediction (NSP) where model predicts the relation between each two sentences and decides that they are following each other or not.

The training processes that the model goes through are as follows: BERT-Base, Uncased /Cased – our chosen mode – specially pretrained on Book Corpus, dataset consisting of 11,038 unpublished books and English Wikipedia (excluding lists, tables, and headers). The model characteristics are 12 layers, 768 hidden layers, 12 heads and 110M parameters. Preprocessing steps are applied on the dataset mentioned before such as Lower Casing and Tokenization using WordPiece and the vocabulary size that outs from preprocessing step are 30,000 vocabularies. The preprocessing step results in formatting the data in two sentences as following: [CLS] Sentence A [SEP] Sentence B [SEP], the two combined sentences length must be less than 512 tokens.

The Masking procedure is the next step after preprocessing. This procedure is applied for each sentence as following: 15% of the tokens are masked, in 80% of the cases, the masked tokens are replaced by [MASK] and in 10% of the cases the masked tokens are replaced by a random token (different) from the one they replace. In the 10% remaining cases, the masked tokens are left as is.

The final procedure is the Pre-training procedure in which 4 cloud TPUs (16 TPU chips total) are used, one million steps with a batch size of 256, The sequence length was limited to 128 tokens for 90% of the steps and 512 for the remaining 10%, The optimizer used is Adam and the Hyper Parameters are: learning rate of 1e-4, $\beta 1=0.9$,

β2=0.999, weight decay of 0.01 and learning rate warm-up for 10,000 steps

After fine-tuning, this model recorded the following results presented in **Table 1**based on GLUE test.

Task	MNLI	QQP	QNLI	SST-2	CoLA	STS-B
	84.6/83.4	71.2	90.5	93.5	52.1	85.8

Table 1:Evaluation Results on GLUE test

Most of the performance improvements - including BERT itself - are either due to increased data, computation power, or training procedure. There is a tradeoff between computation (training time) and prediction metrics (performance). Fundamental improvements that can increase performance while using fewer data and compute resources are needed.

IV. METHODOLOGY

The main used approach for the ASAG problem in our research is to get benefit from the pre-trained transformer-based models, BERT is chosen in our research to use its base model (Bert-base-uncased). BERT base model is used for many tasks; it provides an excellence in most of the proposed tasks, so in our research we focused on using the BERT Tokenizer and BERT model.

We use Keras API in our research for all configurations, building, training, and testing the model. Keras is a deep learning API written in python running on top of the machine learning platform TensorFlow. The main advantages of Keras are simplicity which makes it user friendly to understand and use and only focuses on the main problem, the flexibility where simple workflows should be quick and easy, while arbitrarily advanced workflow should be possible via a clear path that builds upon what you have already learned, and it provides a powerful performance and scalability.

The workflow for building our system is done in the following steps: collecting the appropriate datasets related to our tasks, applying the cleaning and preprocessing algorithm on datasets, passing the cleaned and well preprocessed data to the model, building the custom model to solve ASAG problem, training the last layers of the model, fine-tuning the entire model, and finally testing the efficiency of the designed model. All these steps will be explained in detail below

A. Step1: Collecting datasets

Table 2 shows the number of samples for all the collected datasets.

Data set	Number of samples
AR-ASAG	1966
Mohler	2198
PT-ASAG-2018	7396
Sick2014	9428
STS	8579

AR-ASAG	1966

Table 2:the dats sets used in ASAG model

B. Step 2: Text Data Cleaning and Preprocessing

Text preprocessing in natural language processing (NLP) is an important step because it makes the text more understandable by the algorithms of machine learning.so we must prepare our text data by cleaning and preprocessing it.

Before Text preprocessing, focusing on finding suitable and good data sets was among our goals. We focused to prepare the data in this way: 3 columns "reference answers, student answers and scores". Once we prepare it and check that it is in the way as we want, then we can begin to clean and preprocessing it by these steps:

1) REMOVING DUPLICATE ROWS

We get datasets from different sources and some data we create by ourselves, so we must check if we have duplicate rows. If they are found, we remove it to make our data more efficient.

2) TEXT NORMALIZATION

In our project, Text normalization is the main step to preprocessing the data because it allows us to put sentences on equal footing where it transforms text into its canonical form.in this case, we choose some steps from text normalization which are beneficial in our project.

- **Removing Punctuations:** Punctuations will not add value to our data; also, they may create a problem. They will make differentiating between words difficult. The punctuations which we removed are (!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~).
- **Converting all answers to lowercase** helps us to make all text in the same format. It treats ‘good’, ‘GOOD’ and ‘Good’ in the same way.
- **Removing stop words** helps us to make our model focus on key features. In addition, they are meaningless and considered a load on our data because they don’t carry any information.

3) Lemmatization

At first, we choose between stemming and lemmatization. Both are processes of reducing inflection. Stemming removes the suffix, for example “works” converts to “work” and “information” converts to “inform”. Sometimes stemming makes words not suitable in English language like “airliner” converts to “airlin”. For this reason, we decided to use lemmatization. Lemmatization converts a word to its canonical form. It is provided by the word and its pos “Part of Speech” tag which defines its type “verb, noun, adjective etc.” and gives results dependent on this. [26]

- In stemming:

The original word	After applying stemming to the word
Airliner	Airlin
change, changes, changing	Chang

- In lemmatization:

Word + POS tag	Result
better + adjective	Good
change + verb	Change

It is obvious that lemmatization gives better results in our project. The lemmatization process is slower than the stemming process because lemmatization does more difficult processes in text, but it gives the desired result.

4) **Removing multiple white spaces**: removing all multiple white spaces and replaces them by single space to make our data more normalized.

5) **Removing Noise**: removing all noise which may be harmful to our model. Some of these noises are URLS, HTML tags, emojis and emoticons.

C. *Step 3: Passing the data to the model*

The data is passed to the model using the class (Bertsemanticdatagenerator) this class is implemented by keras team, this class contains functions that convert the input data (data frame) into tuple contains the input of the input layers to the BERT layers.
These inputs are input ids, attention mask and token type ids. This work is done by Bert tokenizer, which converts every word to its numeric vectors to be fed to the Bert layers.

D. *Step 4: building the model*

The input layers of the model are input ids, attention mask and token type ids. Then the (bert-base-uncased) is fed by these input layers. Then Bert model is frozen in the training step, the training is done to the last layers are added. The following step is to create a bidirectional LSTM layer with size of 64, followed by layer contains (max pooling and average pooling concatenated), these layers are the last layers before the Dense layer added, one node is used in this layer and the activation for this is (ReLU). The loss function used is the model is the mean square error (MSE), the optimizer Is Adam optimizer and the metrics used are Pearson correlation and RMSE (root mean square error).

E. *Step 5: Training the last layers of the model*

In this step the training is done, the Bert model is frozen, and the training done only on the last trainable layers.
The Hyper parameters used in model are the Batch size equals 32, the Dropout rate equals 0.3, the Number of epochs is 5 and the learning rate equals 0.001

F. *Step 6: Fine-tuning the entire model*

This is the last step where BERT model is unfreezed and retrained with a very low learning rate. This can deliver meaningful improvement by incrementally adapting the pretrained features to the new data. All the parameters are trainable now, not just the last layers. Using the hyper parameters: Batch size is 32, Dropout rate 0.3, Number of epochs is 5, Learning rate 0.00001, The loss function used is mean square error (MSE), The optimizer Is Adam optimizer, The metrics used are Pearson correlation and RMSE (root mean square error).

G. *Step 7: testing the model*

The model is evaluated by the evaluation method from keras, this is done by the test set, and the test set passed the class (Bertsemanticdatagenerator) to be in the proposed format.

V. EXPERIMENTS AND RESULTS

Firstly, our dataset is collected from different sources, as shown in Table 2, the total number of dataset samples is 28151, the used dataset in the first experiment is all the mentioned samples for training our model, the second experiment use the same data in experiment 1, but 1000 sample is added, these samples are created from Mohler dataset by augmentation process is done on it.

Table 3 shows the number of samples for each experiment.

Experiment	Number of data samples
Experiment 1	28151
Experiment 2	29151
Experiment 3	25949

Table 3 : number of samples for each experiment.

The fourth experiment we did in this paper is to use transfer learning to get better results, we trained BERT on SNLI data set with analogous task with classification end layer, then then the last layer is replaced by another output layer fits our labels, the result of this experiment provides the best Pearson Correlation compared to the other experiments we did.

The dataset used in experiment 4 is the same dataset used in experiment 3. The grades distribution of the data used in experiment 3, 4 is shown in **Figure 4**

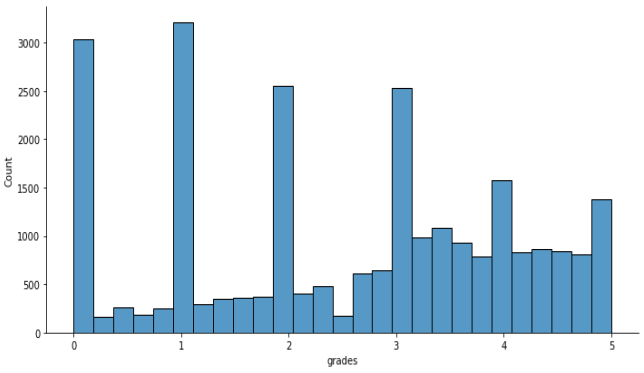

Figure 4:Grades distribution of experiment 3,4 dataset

The results of our experiments are shown in **Table 4** that experiment 4 is the best, with RMSE 0.139 and Pearson Correlation 0.896, the worst experiment is the second one, and these results encourage us to remove Mohler dataset from the data.

Experiment	RMSE	Pearson Correlation
Experiment 1	0.153	0.871
Experiment 2	0.153	0.812
Experiment 3	0.142	0.889

Experiment 4	**0.139**	**0.896**

Table 4 : Results of our experiments

Comparing our results with the previous work[7], [8], [9] **Table 5** shows that our Pearson Correlation is the best comparing to them and RMSE in our model is the best too.

Model	RMSE	Pearson Correlation
Mohler [9]	1.018	0.411
Gaddipati [7]	0.978	0.485
D.B.Leila [8]	1.03	0.722
Our model	**0.139**	**0.896**

Table 5 : Comparing the results in ASAG

VI. Conclusion and Futer work

Automatic short answer grading aims to help the educational processes to be fast, fair, and automated, so in this research we tried to introduce an acceptable solution for this task. The pretrained transformer-based models added a huge impact on natural language processing tasks, so we used BERT model to be fine-tuned using ASAG dataset.

In this paper, we propose simple and effective model for grading answers in presence of the model answer, the results are excellent compared to the result of other research worked in the same problem with different methods. Many experiments are done using our model, most of them give good results, The most effective experiment we did is fine-tuning BERT on SNLI dataset, which works on classification output, then replacing the output layer by another output layer fits our data, then retrain the entire model, results on this experiment is 0.139 RMSE and 0.896 Pearson Correlation.

The most effective element in this research is BERT model, but in future we hope to fine-tune another model to see if there are better results can be provided, for example, Roberta and XLNet may help introducing better results than BERT.

One of the big problems faced our research is data, the available data in ASAG (graded answers) is not sufficient, so in the next step we would work on datasets from other tasks to make it suitable for our task, like converting the format of labels of SNLI dataset to make it numeric values for regression task, or any dataset can be converted to the ideal format. This would help increasing the number of samples used in ASAG task, consequently the training process would be better.

Building a big database for ASAG task is mandatory to improve the performance of ASAG models, this database should contain answers from many relevant fields to make the learning focus on certain area. Determining one criterion in grading is important for building the database.

References

[1] 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, 2015.

[2] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, “Automatic short answer grading and feedback using text mining methods,” Procedia Comput. Sci., vol. 169, pp. 726–743, 2020.

[3] S. Haller, Automatic Short Answer Grading using Text-to-Text Transfer Transformer Model. Faculty of Electrical Engineering, Mathematics & Computer Science, 2020.

[4] S. Ruder, Neural Transfer Learning for Natural Language Processing. Dublin 2, Ireland: National University of Ireland, Galway, 2019.

[5] “Supervised vs. Unsupervised learning: What’s the difference?,”Ibm.com.[Online].Available: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning. [Accessed: 21-Jul-2021].

[6] K.Rungta, “Supervised vs Unsupervised Learning: Key differences,” Guru99.com, 01-Jan-2020. [Online]. Available: https://www.guru99.com/supervised-vs-unsupervised-learning.html. [Accessed: 22-Jul-2021].

[7] S. K. Gaddipati, D. Nair, and P. G. Plöger, “Comparative evaluation of pretrained transfer learning models on automatic Short Answer Grading,” arXiv [cs.CL], 2020.

[8] D. B. Leila OUAHRANI, AR-ASAG An ARabic Dataset for Automatic Short Answer Grading Evaluation. Algeria: Bouira University, 11–16 May 2020

[9] M. Mohler, R. Bunescu and R. Mihalcea, "Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments", vol. 11, 2011. [Accessed 22 July 2021].

[10] L. Galhardi, R. C. T. De Souza, and J. Brancher, “Automatic grading of Portuguese short answers using a machine learning approach,” in Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação (Anais Estendidos do SBSI 2020), 2020.

[11] Semantic_similarity_with_bert.Py at master keras-team/keras-io. .

[12] Suzen, Neslihan, et al. “Automatic Short Answer Grading and Feedback Using Text Mining Methods.” Procedia Computer Science,2020

[13] Hahn, M., & Meurers, D. (2012). Evaluating the meaning of answers to reading comprehension questions: a semantics-based approach. In J. Tetreault, J. Burstein, C. Leacock (Eds.), Proceedings of the 7th workshop on building educational applications using NLP (pp. 326–336). Montreal: Association for Computational Linguistics.

[14] Bachman, L.F., Carr, N., Kamei, G., Kim, M., Pan, M.J., Salvador, C., Sawaki, Y. (2002). A reliable approach to automatic assessment of short answer free responses. In S.C. Tseng, T.E. Chen, Y.F. Liu (Eds.), Proceedings of the 19th international conference on computational linguistics, volume 2 of COLING ’02 (pp. 1–4). Taipei: Association for Computational Linguistics.

[15] Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In A. Lascarides, C. Gardent, J. Nivre (Eds.), Proceedings of the 12th conference of the european chapter of the association for computational linguistics(pp. 567–575). Athens: Association for Computational Linguistics.

[16] Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein, R. De Felice (Eds.), Proceedings of the 3rd ACL workshop on innovative use of NLP for building educational applications (pp. 107–115). Columbus: Association or Computational Linguistics.

[17] G¨utl, C. (2007). e-Examiner: towards a fully-automatic knowledge assessment tool applicable in adap tive e-learning systems. In P.H. Ghassib (Ed.), Proceedings of the 2nd international conference on interactive mobile and computer aided learning (pp. 1–10). Amman

[18] "BERT, RoBERTa, DistilBERT, XLNet: Which one to use? - KDnuggets," Kdnuggets.com. [Online]. Available: https://www.kdnuggets.com/2019/09/bert-roberta-distilbert-xlnet-one-use.html. [Accessed: 22-Jul-2021].

[19] "3 Reasons Why BERT NLP Will Be Revolutionary" Revuze.it.[Online].Available: https://www.revuze.it/blog/bert-nlp. [Accessed: 22-Jul-2021].

[20] "Bert-base-uncased hugging face," Huggingface.co. [Online]. Available: https://huggingface.co/bert-base-uncased. [Accessed: 22-Jul-2021].

[21] S. Ruder, "The state of Transfer Learning in NLP," Ruder.io, 18-Aug-2019. [Online]. Available: https://ruder.io/state-of-transfer-learning-in-nlp/. [Accessed: 22-Jul-2021].

[22] "Transfer learning in natural language processing," in Transfer Learning, Cambridge University Press, 2020, pp. 234–256.

[23] J. Brownlee, "A gentle introduction to transfer learning for deep learning," Machinelearningmastery.com, 19-Dec-2017. [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/. [Accessed: 22-Jul-2021].

[24] N. S. Chauhan, "Attention mechanism in Deep Learning, Explained - KDnuggets," Kdnuggets.com. [Online]. Available: https://www.kdnuggets.com/2021/01/attention-mechanism-deep-learning-explained.html. [Accessed: 22-Jul-2021].

[25] sarthakvajpayee, "㊟ attention mechanism! Japanese → English," Kaggle.com, 20-Mar-2021. [Online]. Available: https://www.kaggle.com/sarthakvajpayee/attention-mechanism-japanese-english. [Accessed: 22-Jul-2021].

[26] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," arXiv [cs.CL], 2017.

[27] Manning Publications, "Deep Transfer Learning for NLP with transformers - manning," Manning.com, 27-Mar-2021. [Online]. Available: https://freecontent.manning.com/deep-transfer-learning-for-nlp-with-transformers/. [Accessed: 23-Jul-2021].