iGrade: an automated short answer grading system

Tayma, A, Alqahtani*
Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia

Yasminah, H, Alali
Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia

Aljowharah, M, Alajmi
Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia

Maryam, R, Alnassar
Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia

Dina, H, Alhamed
Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia

Dina, A, Alabbad
Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia

ABSTRACT

During the COVID-19 pandemic, most countries rely on E-Learning to apply social distance policy which affects the exams evaluation process. This project aimed to assist instructors in grading the short answer questions for CCSIT courses. By implanting a website application that the instructors could use to upload the students' answers and the 'iGrade" software model will grade it. Moreover, the system will reduce the workload on the facilities members by saving time and effort as well as guarantee an objective grading for students. The model used in this project is a state-of-the-art BERT Neural Network model along with layers of BiLSTM that was trained using a dataset that has been collected from previous midterm and final exams of the CIS 211 course. The dataset consists of three categories which are (0, 0.5, 1) with around 1,128 instances. The "iGrade" test obtained an accuracy score of 85,4%, demonstrating BERT's superiority and independence from features during short answer grading as a default method in NLP.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Natural language processing**;

KEYWORDS

BiLSTM, BERT, Deep learning, Automated short answer questions grading

ACM Reference Format:
Tayma, A, Alqahtani*, Yasminah, H, Alali, Aljowharah, M, Alajmi, Maryam, R, Alnassar, Dina, H, Alhamed, and Dina, A, Alabbad. 2022. iGrade: an automated short answer grading system. In *2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2022), December 16–18, 2022, Bangkok, Thailand.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3582768.3582790

1 INTRODUCTION

Nowadays the rhythm of study changes and the trends for online quizzes and exams increase. The instructors spend lots of time marking the exams even beyond the working hours and put a lot of pressure on them. Due to the COVID19 pandemic, the universities were forced to have online lectures and conduct online quizzes as it was much safer. It was difficult to mark many essays answer questions so online quizzes usually use multiple-choice questions (MCQ). MCQs cannot measure the exact knowledge for the student which may affect the students' academic outcomes. Also, the quality of the examination states that the percentage of the MCQ questions should not exceed 60% of the exam. Accordingly, auto-scoring has attracted much attention during past years where it saves time, speeds up the process, guarantees fairness, solves the inconsistency issue, and evaluates without any bias [1]. Some of the real challenges to researchers are Automatic Essay Scoring (AES) and Short Answer Grading (SAG). AES scores essay questions that are identified by long answers where there is no reference answer is provided to compare the student answer with it for the grading process. The main factors in answer grading for AES are the relatedness to the main topic, grammar, spelling analysis, and sentence coherency. On the other hand, as mentioned by Gomaa et.al SAG handles short answer questions that usually range from 1 to 3 sentences. The main factor for SAG is measuring the similarity between the student answer and the reference answer [2]. Students usually use their understanding and rewrite the answer in different ways to convey the meaning. Even though the pace of technology is increasing every day, there are no actual auto short answer grading systems in the College of Computer Science and Information Technology as they depend on the OCR system for multiple-choice questions and manually grading which lead us to work on this project. Moreover, most of the previous works in this field were only researched as well as they do not support CCSIT courses. The "iGrade" project will aid the instructors by saving the time and effort needed for grading, therefore, increasing productivity and free time to focus more on research and teaching while raising the quality

NLPIR 2022, December 16–18, 2022, Bangkok, Thailand

ACM ISBN 978-1-4503-9762-9/22/12…$15.00
https://doi.org/10.1145/3582768.3582790

of education. “iGrade” will reduce instructors’ tension and fatigue when scoring exams and assignments which ensures the equality for all students. In this project, we propose a scheme that utilizes Natural language processing (NLP) techniques to mark a short answer question depending on the similarities of the answers given by the student and the instructor where the comparisons depend on the semantic meaning, not exact words. The main contributions of this project are Providing a scheme for the automatic grading of short answer questions based on the semantic meanings of the answers and providing a system that is tailored to the college of computer science and information technology (CCSIT) by focusing on courses that are offered to a large number of students, mainly level 3-6 courses, which require time and effort for grading.

2 RELATED WORK

2.1 Automated short answer grading system using BERT or BiLSTM model

The researcher’s [3] goal is to propose a novel ASAG model based on the Sentence BERT model. Using short answer scoring V2.0 dataset. the proposed model shows improvement in the accuracy, of M-F1, and W-F1 compared to the BERT model. Furthermore, the researcher found using the regression task function achieve a better result. As well as they found using shorter answer have better result the longer answers. A shorter answer set gave them an accuracy of 81% while a longer answer set gave them an accuracy of 67%. In the future, they proposed to expand the domain of their train dataset to accept more than the Computer Science domain.

The researcher’s [4] goal was to build a system that uses the finetuning on Roberta’s large pre-trained model using the STS benchmark dataset while evaluating the model using the SciEntBank dataset as well Mohler extended dataset. The devolved model achieves RMSE with the values of 0.56, 0.7, and Person’s Correlation of 0.79, and 0.82 respectively on both datasets. the observation the researcher found is using a batch of size 1 instead of the usual which is 16 or 32 achieve a better result.

In [5] the researchers aimed to explore the Automatic Short Answer Grading model to produce vectors representation of various methods such as Sentence-BERT, Word2Vec, and Bag-of-words. The data set used is Critical Reasoning for College Readiness (CR4CR), which consisted of 179 responses per question. To evaluate the results, they considered the accuracy and the F1 score to capture the precision and the recall. The result showed that SBERT achieves higher accuracy which was 0.62 compared to 0.575 for W2V and 0.605 for BOW.

In [6] the researchers aimed to use a neural network with BERT sentence embedded to build a model that can automatically score students’ short answers. The dataset was collected from Gadjah Mada University in Indonesia called the Ukara dataset, it contained student answers with binary assessed true or false to present if the answer is acceptable or not. The dataset consisted of two questions 1340 rows for the first question and 1500 for the second question, both divided into three categories train set which presents, 20% from the dataset, validation set presents 20%, and the rest for the test set. The result was evaluated by considering the F1-score, which achieved around 0.829. The researchers will focus more in the future on dealing with unbalanced data so the model performance can be better in students’ answers.

In [7] researchers’ goal was to build a system that can evaluate students’ answers automatically with little or no help from the instructor. There were two datasets used to conduct the paper experiments, the first dataset is the English short answer grading dataset, and the second dataset is the Indonesian short answer grading dataset. The researchers used the English dataset to find the best configuration of the model in grading the short answer then used it on the Indonesian dataset to get the best performance on grading the Indonesian dataset. For answer grading both the input answer from the student and reference answer was in form of a representation vector that then researchers used these forms to conduct two methods for similarity calculations. The first method was cosine similarity calculations, and the second method is to use logistic regression model, where calculation from student answer vector and reference answer vector became input for the logistic regression model. The best model accuracy was achieved by BERT algorithm as the pre-trained language model with MAE 1.443 and 1.893 RMSE.

In [8] researchers build GPT, GPT-2, BERT, and ELMo models to compare their efficacy on auto-grading short answers, using the Mohler dataset dividing it into 70% training and 30% testing. The evaluation was based on comparing the correlation measurements and RMSE scores of the same models with the Mohler dataset of previous works. The results showed that ELMo was the best model to use. In future work, different methods of assigning sentence embedding will be used, and eliminating stop words will be considered.

In [9] the researcher’s main objective was to create a tool to assist educators with short answer grading (ASAG) by using the base model of the BERT algorithm which has the advantage of being an open-source model as well as a pre-trained language model. The dataset they used is called DT-Grade which they have obtained from the tutorial dialogue that was constructed between students and Deep Tutor for short constructed answers. The Deep Tutor is an Intelligent Tutoring System. This dataset was from the University of Memphis Institute for Intelligent after they have preprocessed the data it results in having 994 records. The accuracy they have gained is 76% and the Cohen’s Kappa statistic is %68.4. The researchers have represented a work that they believe to be in progress to persistently investigate how they can influence AI to be an assist human decision-making.

In [10] the researchers aimed to build an ASAG model that is free from manually engineered features that used state of art models such as BERT and Extra Long Network (XLNET). The SciEntBank dataset was used which consisted of 135 questions. The advantage of using these models was that grading is performed without manually extracting features. The researchers labelled the students’ responses: 2-way task (correct and incorrect), 3-way task (correct, contradictory, or incorrect), and 5-way task (correct, partially correct, contradictory, irrelevant, or not in the domain). The highest result achieved was with the 2-way task which was 79.8%. There were some limitations such as the experiment being carried out on only one dataset. Another limitation is that the largest BERT model was not used due to limited computing power. Also, the researchers could not go beyond 10 epochs. For future work, the researchers

Table 1: Size and domain of pre-trained text corpora [11]

Corpus	# Of words	Domain
English Wiki	2.5B	General
BooksCorpus	0.8B	General
Textbooks	1.1M	Phy + Gov
QAs	0.6M	Phy + Gov
	1.3M	Phy + Gov + Psy-I,II

planned to address these limitations. Also, the researchers intended to ensemble BERT with other classifiers.

In [11] they have explored ways of improving the pre-trained BERT Model in the task of automatic short answer grading by using data augmenting which could be from the specific domain such as the textbooks or QA. Then they have presented a new way which is labeled short answer grading data in the fine-tuning step the process The dataset they have used is a large-scale industry dataset which can be described in detail in Table 1 below.

The highest accuracy they have achieved in the Phy with accuracy 83.64% while using the textbook and the QA. They have concluded that by adding a step for updating the BERT model using domain-related resources they achieved better results than relying only on fine-tuning. as well as by updating the model it becomes more specialized toward the domain.

Ye and Manoharan [12], proposed a model that marks essay-type assessments and helps provide personalized feedback promptly. The researchers used a dataset created manually about computer programming knowledge and consisted of 2129 samples. Two n-dimensional vectors stored the meaning of the answers that were given by the student and the optimal answer. To check the similarity, the vectors were compared then the grade was calculated. For extracting the semantic meaning, bidirectional LSTM was used. Finally, for providing the personalized feedback k-mean algorithm was used for partitioning the students' answers into groups depending on their meanings so the teachers can provide feedback to a few answers in each group. Then the feedback was transmitted to the other answers in the same group. The proposed model showed reasonable marking accuracy which was 88.5%. Their plan for future work was to improve accuracy further to match human markers. Also, they intended to use BERT and XLNet models in future investigations.

2.2 Automated short answer grading system using other models

In [13] the researcher's goal was to find an assessment of automatic short answers by comparing the result between three similarity methods Jaccard Similarity, Cosine Similarity, and Dice Coefficient. The data collected from a quiz in the E-business course at Amikom Purwokerto University contained four definition questions and thirty-one answers. They used the Mean Absolute Error to present the average of the forecasted result. The final results were 0.73 for the Dice Coefficient, 0.69 for the Jaccard method, the lowest value was the and Cosine Similarity which got 0.57 that proved the ability of Cosine in finding the similarity.

In [14] researchers proposed an automatic short answer grading system, using SemSpace and MaLSTM methods with Mohler ASAG dataset consisting of 87 questions and answers. As a result, RMSE and Pearson's (r) get a value of 0.04 and 0.949 respectively. Moreover, the CU-NLP dataset containing 171 answers was used to test the model generating 0.02 MAE and 0.989 Pearson values. As for future work, a new model will be generated for automatic OOV handling using external lexical semantic networks.

In [15] researchers presented an automated system for scoring short answers, using prompt, Word2Vec, content overlap, and lexical diversity features with ASAP dataset containing more than 16000 responses for 10 different questions, that divided into 70% for training, 10% for validation and 20% for testing. The results evaluation was based on the Quadratic Weighted Kappa metric which got a mean score of 0.79. As for future work, hybrid methods like the random forest model and deep neural network architecture will be used.

Gomaa and Fahmy [2], proposed an efficient short answer grading model named Ans2vec. Their goal was to introduce a simple Sort Answer Grading model (SAG). Their approach was a Skip-thought vector which was used to convert both the student's answers and the optimal answer into meaningful vectors for measuring the similarity between them. Then to predict a score, a logistic linear classifier was trained. The researchers used three datasets which are Texas, Cairo University, and SCIENTSBANK. The researchers compared their proposed system to five previous works. The model achieved promising results were with the Texas dataset, the model achieved the best correlation value (0.63) compared to all five other systems. For the Cairo University dataset, the model achieved promising results even if it wasn't the highest if we consider the other complex work performed. Also, the model achieved the best F1-score (0.58) in the Unseen Answers test set. For future work, they planned to test other approaches, merge their approach with classical machine learning and text similarity techniques, and test with other natural languages such as Arabic.

In [16] the researcher's goal was to assist the outcome of an exam if the student will pass or fail by analyzing the student's answer in natural language processing (NLP) through using machine learning technique which is SVM as supervised classification framework. This was done by using real data set from statistics exams which can be downloaded at https://zenodo.org/record/3257363#.XRsrn5P7TLY. The resulting accuracy is 89% and an F1 score of 91%. The researcher plans to integrate it into an online evaluation tool. So, it will be used for the student to submit their assignment and automatically it will score the assignment as pass or fail. As well as they have suggested using the Deep-learning model instead of the

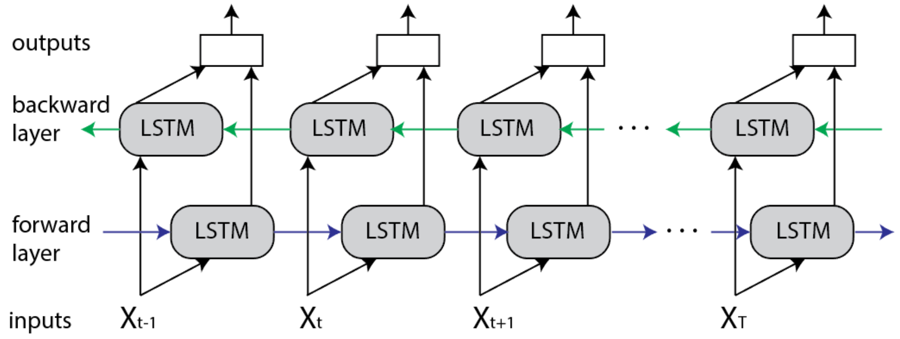

Figure 1: BiLSTM architecture *[21]*

SVM model. Furthermore, they have suggested instead of a scoring pass, or failing to make to make it to a numerical score.

3 METHODOLOGY

3.1 Bidirectional Encoder Representations from Transformers

BERT algorithm (Vaswani et al., 2017; Devlin et al., 2018) is a state-of-the-art model on natural language processing introduced by Google. BERT has the power to do more than 10 NLP tasks, including classifying the words with the disambiguation or multiple meanings [10]. Unlike recent NLP models which can read the input text sequentially from left to right or from right to left, BERT is built bidirectionality it can consider both left and right directions in all layers.

In the BERT framework there are two steps, first is the pre-training model using two unlabeled corpora which are Wikipedia that contain more than 2,500M words, also 800M words from BooksCorpus dataset. In this level, BERT used two related NLP tasks: Masked Language Model (MLM) that can predict a hidden word in a sentence by creating a mask based on the context of that hidden word. The other task is the Next Sentence Prediction (NSP) to determine whether two given sentences have a specific relationship sequential, logical or if they simply have a random relationship [17].

The fine-tuning is the second step in the BERT framework, here the parameters are initialized, and each parameter is tuned based on labeled pairs of data containing the reference answer and the student answer. Then BERT uses the embedding process to indicate whether the tokens belong to the first or second sentence. By summing the token, position, and segment embeddings associated with a token, its input representation is constructed. The result after embedding the input pair will be fed into a dense layer to optimize its efficiency by modifying the learning rate [18].

3.2 Bidirectional LSTM

Before introducing BiLSTM it is important to understand its building block which is LSTM. LSTM is an advanced type of recurrent neural network (RNN) that is presented in 1997 to overcome the vanishing gradient problem which RNN has been suffering [19]. Moreover, it consists of multiple gates that control the information flow [20].

On the other hand, Bidirectional LSTM as the name suggests is a model that takes input in from both directions which are very useful in natural language processing because of the deep understanding of the words so when the order of the words in a sentence change it will understand and produce different results for each order. It contains two layers of LSTM (forward layer and backward layer) the input flows forward and backward in the layers and can be combined in multiple ways as shown in Figure 1 [21].

3.3 Dataset

Our dataset is derived from the midterm and final exams that have been conducted in the year 2020-2021 for the course Fundamentals of Information Systems, with the course code CIS 211, which is offered by the College of Computer Science and Information Technology at Imam Abdulrahman bin Faisal University. Approximately 80 students' answers are provided for each of the 13 short answer questions. Each answer was graded 0, 0.5, or 1, and we tried to balance the dataset as much as possible.

3.4 Performance Evaluation of the Proposed Models

3.4.1 Accuracy. The accuracy is the proportion of correct graded answers of the system compared to the human grading. It is calculated using the following equation:

$$accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction} \tag{1}$$

3.4.2 Precision. The precision is defined as the ratio of the number of Positive samples correctly classified to the total number of Positive samples (either correctly or incorrectly classified). As defined in the equation:

$$Precision : \frac{TP}{TP + FP} \tag{2}$$

3.4.3 Recall. The recall is calculated based on the number of Positive samples correctly categorized as Positive compared to the total number of Positive samples. It evaluates how well the model detects positive samples. As defined in the equation:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Table 2: Optimal parameters for the proposed model

Parameters	Optimal Value chosen
Epochs	14
Batch size	32
Activation function	Softmax
Optimizer	Adam with learning rate =1e-5, epsilon=10e-8

3.4.4 F1 score. Considering the F1-score in our model is to pay attention to the ability in dealing with imbalanced data that might face in the real answers. It is calculated using the following equation:

$$F1 = \frac{2 \times Percision \times Recall}{Percision + Recall} \quad (4)$$

Also, because most previous published models consider F1-score as a performance criterion, we can use it to evaluate the resulting model and compare it with others.

3.5 Probability

Probability is the measure of how likely it is for a random variable to take on different values in the sample space, which are (0, 0.5, 1). In the output, the values are converted into probabilities using the SoftMax distribution, which totals all values in the list to one. The student will get the mark according to the highest probability on the sample space. In situations where the highest probability was below 80%, the instructor might recheck the answer.

4 EXPERIMENTAL SETUP

This model was developed in the Google Colab environment which is a Google Research product. It enables the python code to be written and executed in the browser which facilitates the work within a group and makes it a great tool for this project.

4.1 Data Prepressing

In the first pre-processing step, we lower-cased all words, and then called a function that checked for contractions, which are words that have been shortened or combined with other words such as "can't", and then replaced them with the original words that cannot. The next step is removing any special characters like "@! # . . .etc.". Then we tokenized each word in the sentence which is a technique that Breaks up each linguistic unit in the text. And lastly is removing the stop words such as (articles, prepositions, pronouns, conjunctions, etc.) however we excluded words that seem important which are ('and', 'or', 'not','until', 'while', 'during', 'both', 'same', 'against', 'between', 'but', 'dont', 'can', 'few', 'more', 'most', 'as'). Given below is used an example of a question with its answer that has been pre-processed.

4.2 Spell Checking

Because students could make some spelling mistakes due to the pressure of the limited time in the quizzes and exams. We decide to use a pre-build library used by Google called "Autocorrect Library", which is an external library in Python that can be used to develop a spell checker.

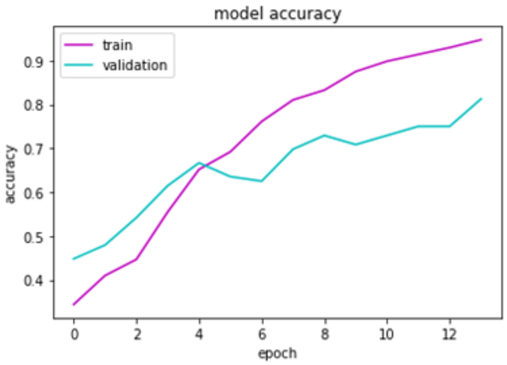

Figure 2: Model Accuracy

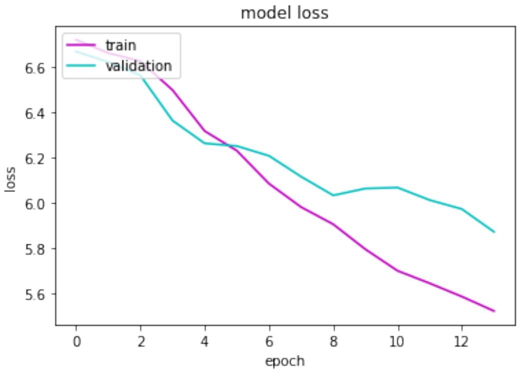

Figure 3: Model Loss

5 RESULTS AND DISCUSSION

The model is tested based on the short answers to questions that have been collected from the midterm and final exams of the CIS 211 course in CCSIT IAU. The dataset contains 1,128 answers that are divided into 80%, 10%, and 10% for training, validation, and testing respectively. Furthermore, Table 2 presents the optimal parameters chosen after several experiments and changes in the number of neurons, optimizer, epochs, and batch size.

The model was trained with 14 epochs with respect to the values of the optimal parameter that have been specified in the above table. The resulted Accuracy and the Loss of training and validation are shown in the plot figures below respectively.

As for the testing phase, the model got an accuracy of 85% as well as the macro average of precision and recall is 82% which is

Table 3: classification metrics of the proposed model

	precision	recall	F1-score	support
0	0.951	0.907	0.929	43
1	0.720	0.692	0.706	26
2	0.806	0.879	0.841	33
Macro avg	0.826	0.826	0.825	102
Weighted avg	0.845	0.843	0.843	102

Table 4: Comparison of the Proposed model with the Benchmark Studies

Study	Dataset	Model	Accuracy	Macro Average F1
[9]	DT-Grade	base-model BERT algorithm.	76%	-
[11]	Question and answers of four subjects (Human Development (Psy-I), Abnormal Psychology (Psy-II), American Government (Gov), and Physiology of Behavior (Phy))	pre-trained BERT Model	-	Psy-I 78.73 Psy-II 81.41 Gov 76.25 Phy 83.36
Proposed model	midterm and final exams of the CIS 211 course	BERT model	85%	82%

presented by the classification metrics in Table 3 below. the reason that we are concerned about the Macro average more than the weighted although the weighted gave us higher values is that using the macro average, each prediction is given the same weight when calculating loss, as well as all classes, are treated equally, regardless of support value.

For further evaluation of the iGrade performance, results were compared with previous studies. Benchmarking was done by looking at the studies that used the same model with different datasets. As Table 4 presents the accuracy and Macro Average F1(MF1) of studies [16] and [18], where the iGrade is outperformed these studies except the MF1 for the Phy subject of [18] study.

6 CONCLUSION & FUTURE WORK

The proposed project aims to facilitate the grading process of the short answer question type which is usually a time-consuming task therefore it will increase the productivity by saving the time and effort of instructors and free them from routine and repeated work to focus more on research and teaching while raising the quality of education. Moreover, "iGrade" aims to reduce instructors' tension and fatigue when scoring exams and assignments which ensures equality for all students. The project used BERT along with the BiLSTM algorithm. Additionally, a pre-build library called "Autocorrect Library" was used to handle spelling mistakes in students' answers. All functionalities of the iGrade system are developed in Python, however, the interfaces are created using HTML, CSS, and JavaScript. From the literature review, it appeared that many of the existing systems used the Mohler dataset, which is an old dataset available online. However, one of our main contributions is that we collected our dataset manually from previous exams of CCSIT college which requires excessive effort and time making it more useful for our community and for future studies to be adapted from it. This research can be expanded on in the future by increasing the number of student answers along with expanding the dataset to include more questions. Furthermore, the model can be extended to be able to analyze data from paper-based exams rather than just computer-based exams. Also, expanding the model to include more CCSIT courses as well as further enhancement of the accuracy will be taken into consideration.

ACKNOWLEDGMENTS

All these works in this project wouldn't have been possible without Allah's help and grace. Also, we thank the College of Computer Science at Imam Abdulrahman bin Faisal University for their continued support. Special thanks to our supervisor Dr. Dina Alabbad for always supporting and motivating us under her wise guidance. Without her, we could never have reached this current level of success.

REFERENCES

[1] S. Bonthu, S. Rama Sree, and M. H. M. Krishna Prasad, "Automated Short Answer Grading Using Deep Learning: A Survey," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12844 LNCS, pp. 61–78, Aug. 2021, doi: 10.1007/978-3-030-84060-0_5.

[2] W. H. Gomaa and A. A. Fahmy, "Ans2vec: A Scoring System for Short Answers," Advances in Intelligent Systems and Computing, vol. 921, pp. 586–595, 2020, doi: 10.1007/978-3-030-14118-9_59.

[3] J. Luo, "Automatic Short Answer Grading Using Deep Learning", Accessed: Apr. 22, 2022. [Online]. Available: https://ir.library.illinoisstate.edu/etd/1495

[4] M. Thakkar, A. Joorabchi, and A. Ahmed, "FINETUNING TRANSFORMER MODELS TO BUILD ASAG SYSTEM," 2021, Accessed: Apr. 22, 2022. [Online]. Available: https://github.com/mithunthakkar26/NLP-Projects

[5] A. Condor, M. Litster, and Z. Pardos, "Automatic short answer grading with SBERT on out-of-sample questions," 2021.

[6] R. A. Rajagede, "Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, vol. 6, no. 1, pp. 11–18, Feb. 2021, doi: 10.22219/KINETIK.V6I1.1196.

[7] M. H. Haidir and A. Purwarianti, “Short Answer Grading Using Contextual Word Embedding and Linear Regression,” Jurnal Linguistik Komputasional, vol. 3, no. 2, pp. 54–61, Sep. 2020, doi: 10.26418/JLK.V3I2.38.

[8] S. K. Gaddipati, D. Nair, and P. G. Plöger, “Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading,” Sep. 2020, Accessed: Oct. 19, 2021. [Online]. Available: https://arxiv.org/abs/2009.01303v1

[9] A. Condor, “Exploring Automatic Short Answer Grading as a Tool to Assist in Human Rating,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12164 LNAI, pp. 74–79, Jul. 2020, doi: 10.1007/978-3-030-52240-7_14.

[10] H. A. Ghavidel, A. Zouaq, and M. C. Desmarais, “Using BERT and XLNET for the automatic short answer grading task,” CSEDU 2020 - Proceedings of the 12th International Conference on Computer Supported Education, vol. 1, pp. 58–67, 2020, doi: 10.5220/0009422400580067.

[11] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, and R. Arora, “Pre-Training BERT on Domain Resources for Short Answer Grading,” EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 6071–6075, 2019, doi: 10.18653/V1/D19-1628.

[12] X. Ye and S. Manoharan, “Providing automated grading and personalized feedback,” ACM International Conference Proceeding Series, Dec. 2019, doi: 10.1145/3371425.3371453.

[13] T. wahyuningsih, H. Henderi, and W. Winarno, “Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice’s Coefficient,” Journal of Applied Data Sciences, vol. 2, no. 2, pp. 45–54, May 2021, Accessed: Oct. 01, 2021. [Online]. Available: http://bright-journal.org/Journal/index.php/JADS/article/view/31

[14] C. N. Tulu, O. Ozkaya, and U. Orhan, “Automatic Short Answer Grading with SemSpace Sense Vectors and MaLSTM,” IEEE Access, vol. 9, pp. 19270–19280, 2021, doi: 10.1109/ACCESS.2021.3054346.

[15] Y. Kumar, S. Aggarwal, D. Mahata, R. R. Shah, P. Kumaraguru, and R. Zimmermann, “Get It Scored Using AutoSAS – An Automated System for Scoring Short Answers,” Dec. 2020, Accessed: Oct. 01, 2021. [Online]. Available: https://arxiv.org/abs/2012.11243v1

[16] S. Menini, S. Tonelli, G. de Gasperis, P. Vittorini, †Fondazione, and B. Kessler, “Automated Short Answer Grading: A Simple Solution for a Difficult Task”, Accessed: Oct. 01, 2021. [Online]. Available: https://zenodo.org/record/

[17] M. Joshi *et al.*, “SpanBERT: Improving Pre-training by Representing and Predicting Spans”, doi: 10.1162/tacl.

[18] S. González-Carvajal and E. C. Garrido-Merchán, “Comparing BERT against traditional machine learning text classification,” May 2020, Accessed: Nov. 07, 2021. [Online]. Available: https://arxiv.org/abs/2005.13012v2

[19] “A Battle Against Amnesia: A Brief History and Introduction of Recurrent Neural Networks | by Chen Yanhui | Towards Data Science.” https://towardsdatascience.com/a-battle-against-amnesia-a-brief-history-and-introduction-of-recurrent-neural-networks-50496aae6740 (accessed Apr. 24, 2022).

[20] “Introduction to LSTM Units in RNN | Pluralsight.” https://www.pluralsight.com/guides/introduction-to-lstm-units-in-rnn (accessed Apr. 24, 2022).

[21] “Differences Between Bidirectional and Unidirectional LSTM | Baeldung on Computer Science.” https://www.baeldung.com/cs/bidirectional-vs-unidirectional-lstm (accessed Apr. 24, 2022).