Deep Learning in Automated Short Answer Grading: A Comprehensive Review

Rupal Chaudhari[1*], Manish Patel [2]

Assistant. Professor, Department of Computer Engineering, Sankalchand Patel College of Engineering, Visnagar/SPU, Gujarat, India[1]

Associate Professor, Department of Information Technology, Sankalchand Patel College of Engineering, Visnagar/SPU, Gujarat, India[2]

rrchaudharice_spce@spu.ac.in[*1], mmpatelit_spce@spu.ac.in[2]

Abstract: Automated Short Answer Grading (ASAG), more generally referred to as ASAG, is a method that evaluates the written short answers provided by students through the use of certain computer algorithms. This particular component of ASAG has been the subject of study for a considerable amount of time [4]. A significant obstacle in ASAG is the low availability of relevant training data inside the domain. This is one of the most significant obstacles. There are a number of different approaches that may be taken to address this problem. These approaches can be broadly classified into two categories: traditional methods that rely on handcrafted characteristics and Deep Learning-based approaches [22]. Over the course of the past five years, there has been a significant increase in the number of researchers in this field that have adopted Deep Learning techniques in order to address the ASAG challenge [6]. The purpose of this research is to determine whether or whether strategies based on Deep Learning are superior to traditional methods across 38 different publications. Additionally, the study intends to provide a full review of the many deep learning methodology that have been investigated by academics in order to address this issue [19]. In addition to this, the study provides an analysis of a number of state-of-the-art datasets that are ideal for ASAG tasks and makes recommendations for evaluation metrics that are suitable for regression and classification situations.

Keywords: Automated Short Answer Grading (ASAG), Computer Algorithms, Deep Learning-Based Approaches, Training data, Research Problems

I. INTRODUCTION

"Automatic Short Answer Grading" (ASAG) is the process of assessing the short responses that students have submitted, and it is performed via the use of "Natural Language Processing" (NLP) techniques [1]. An acronym for "Automatic Short Answer Grading" ASAG has attracted a lot of attention from a broad variety of research organisations owing to the fact that its major objective is to automatically score the free-text responses that students make in accordance with the reference answers that connect to those responses. This has resulted in a lot of interest being generated by ASAG. The scary spread of the new Corona virus, which had a domino effect, caused a number of different enterprises to be placed in a situation of crisis. This was the outcome of these firms. In addition to being impacted by the pandemic, the education sector has also been impacted by it [13]. During the lockdown time, many educational institutions have transitioned to the kind of online teaching, which has established itself as a popular way of instruction. This is done in order to facilitate learning. Evaluation, on the other hand, has become a substantial challenge throughout this time period due to the strategy that has been taken. The objective of assessment may be accomplished by the use of a wide range of techniques, such as essay responses, short answers, one-word answers, and multiple-choice questions (MCQ). As an alternative to multiple-choice question (MCQs), it has been noted that the use of short and essay responses may be a more comprehensive method of evaluating the knowledge of students. As a consequence of the advanced technology that is now available to the general public, massive open online courses, often known as MOOCs, are gaining prominence [7]. As a mechanism for correctly scoring the answers provided by students, the majority of online courses make use of peer grading. Despite this, there is a large level of variability in the results that are generated due to the fact that peer graders have different mentalities. Human graders are needed to review a soft copy, which makes the process of appropriately assessing short answers and essay responses more arduous. This is because the procedure is becoming more difficult. The field of education is seeing a growing need for the use of automated systems for the evaluation of both short and essay responses. Making use of this approach allows for review to be carried out in a way that is not only easy but also unbiased, expeditious, and devoid of any form

of prejudice. There is also a need for instructors to provide students constructive feedback on their solutions, which is yet another responsibility that many individuals find challenging to do.

The assessment of students' knowledge along the course of the learning process is one of the most essential components of effective teaching. A significant amount of time is often required for the manual scoring procedure; however, the process of providing feedback that is pertinent to the situation demands even more time. The manual scoring of responses may be prone to discrepancies since the human grader is needed to infer meaning from the candidate's answer, which is a free text consisting of the candidate's own words [1]. This means that the human grader is forced to make an inference. Additionally, the human grader could incur stress after reading a limited number of answers, and the method in which he corrects the remaining responses might also change as a result of this shift. On the other hand, a widely acknowledged method of assessment that need to be employed throughout the whole of the learning process is the acceptance of free text responses from students. This is due to the fact that these kinds of questions are helpful in enhancing the cognitive capacities of students and also in demonstrating information in brief texts via the usage of short texts. Because of this, there is a high need for the development of technologies that are capable of addressing these concerns in the evaluation process.

In spite of the fact that ASAG is not a fresh technique, it is very necessary for it to include the most recent technology breakthroughs; this is because the current conditions cannot be ignored. Over the course of a number of years, a number of different academics have placed a significant amount of attention on the problem of assessing quick responses. An extensive variety of approaches were provided, starting with the more traditional hand-crafted features and advancing to the more modern deep learning models and their combination. The methods that were provided comprised a broad range of methodologies. At one point in time, rule-based and statistical methods were thought to be the most suitable ways for resolving the ASAG problem, as can be shown in Figure 1. 2009 was the year when machine learning was first put into practice, while 2015 was the year that deep learning was first put into practice. Deep Learning is becoming more popular in the disciplines of computer vision, speech, and text, and as a consequence, researchers have started to use it in this aspect of the field in order to produce automatic grading systems that are accurate. As a result of the fact that there is no survey that is exclusively committed to deep learning methods, we made the decision to conduct research on the works that used deep learning techniques in order to address the ASAG problem.

Thinking of the ASAG problem as a supervised learning problem is one way to approach modelling the problem. It is possible to read it as either a regression task or a classification task, depending on whether the student answer is granted a mark or grade or whether it is categorised into categories such as "correct," "partially correct," or "incorrect." Both of these interpretations are plausible but not guaranteed. In order to address this issue, you will need to provide two concise solutions, which are referred to as a Reference Answer (Qr) and a Student Answer (Qs) to a question Q. Input is provided in the form of these replies. As a result of the degree of connection that exists between Qr and Qs, the output is either a label (classification) or a grade (regression), and it is decided by the extent of the link.

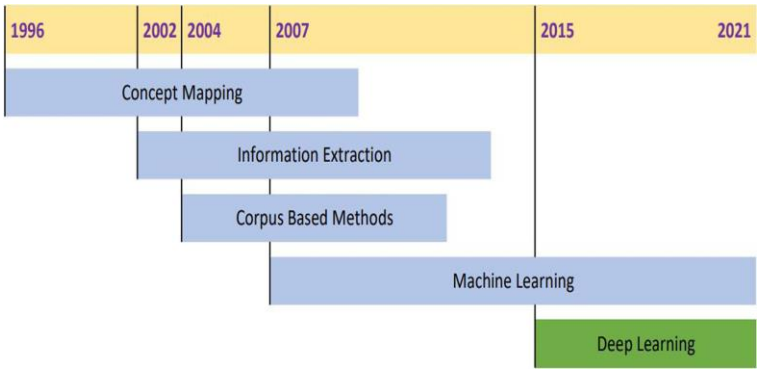

Fig.1 The ASAG systems

For regression models, the objective is to learn

$$Y = f(\vec{X}, \vec{w}) \quad (1)$$

In this context, X represents the input feature vector that is produced by determining the degree of similarity between, Q_r and Q_S. $\vec{X} = X_1, X_2, \ldots, X_n$ which is an n-dimensional similarity vector of a pair, and $(Q_r, Q_S).\vec{w}$ represent model parameter are to be estimated.

For classification issues, the goal is to express the score with a category k for each and every occurrence of the data, such as

$$\text{score}\ (X_i, k) = \beta_k . X_i \quad (2)$$

In the given context, the vector of weights belonging to category k is denoted as β_k, whereas the feature vector X_i is determined by the similarity between the data instance i and the pair (Q_r, Q_s). The final category is determined by taking into account the category that received the greatest score."

$$k^* = \underset{i}{\text{argmax}}\ \text{score}(X_i, k) \quad (3)$$

The purpose of this research is to investigate the Deep Neural Network techniques that were used for this job and the influence that these approaches had on ASAG in comparison to more conventional machine learning approaches [2]. These Deep Neural Network approaches are able to automatically infer syntactic and semantic elements from the text. Our best information leads us to believe that there are three literature studies on the subject of Short Answer Grading that do not place any limitations on the methodology. In order to answer the ASAG challenge, there is no research that is specifically devoted to Deep Learning techniques.

The organisational structure of the remaining article is as described below. This section explains the methodology of the study, the section that addresses the current corpora to carry out this work, the section that provides an overview of the evaluation metrics, and the section that examines the different state-of-the-art Deep Learning techniques is the section that is offered in the second section [5]. The observations that the authors have made in regard to the defined research questions are presented in Section 6. As a last point of interest, the findings of the study are presented in Section 7.

II. METHODOLOGY

With a particular emphasis on ASAG works that made use of deep learning strategies, the purpose of this study is to investigate, analyse, and get an understanding of the present state of the art in ASAG. For the purpose of addressing the purpose of the survey, the following research questions have been developed:

RQ1: "What are the various datasets available to perform ASAG?"

RQ2: "What are the various Evaluation Metrics used to measure the performance of ASAG tasks?"

RQ3: "Which Deep Learning approaches are used?"

RQ4: "What are the results obtained?"

The identification of search terms is accomplished by using the preliminary research as a basis. The following are some of the keywords that have been identified: "Automatic, Short Answer Grading, Scoring, Assessment, Natural Language Processing, Deep Learning, Question, Answer, Response, etc." When searching for research contributions, a search query is constructed by grouping terms from the keywords that are similar to one another and then using Boolean operators to build the query. Search engines such as Google Scholar, IEEE Xplore, ACM Digital Library, Elsevier-ssrn, EBSCO, and ACL Anthology are the key online databases that were taken into consideration as sources for this study [1]. Because it contains 64000 articles, the ACL Anthology is an excellent resource for this review because it focuses specifically on the study of natural language processing (NLP) [15]. It has been determined that a total of two thousand documents have been retrieved from the databases. We were left with 676 documents after we eliminated the duplicates from the collection. The next step was the elimination of a few articles on the basis of their titles. This was done since the survey primarily focuses on Deep Learning implementations. We quickly went over the titles, abstracts, introductions, contributions, model architecture, uniqueness, published venue, and other aspects of the 87 papers that were produced as a consequence. After going through the process of screening, there are a total of 38 articles that are being considered for this survey. On the basis of these 38 publications, more research is conducted on the datasets, metrics, and deep learning algorithms in relation to the Research Questions that were provided.

III. CORPORA

When it comes to finding a solution to the ASAG problem, one of the most major challenges that must be conquered is the absence of datasets that include natural responses. The publications that were analysed make use of a broad variety of datasets, and the studies that were analysed display a substantial amount of variation in terms of the language that was used, the subject matter of the questions, the grading system, the number of questions, and the reference answers. The next section discusses six of the datasets of the English language that are used the most often, which were selected from the 38 papers that were reviewed [1]. In addition, a summary of the distinct characteristics of each dataset, as well as the benefits and drawbacks associated with each dataset, is presented. The majority of the datasets that are used by ASAG come from competitions. Some examples of these datasets are ASAP, SemEval-2013, and Joint SRA (Beetle & ScientsBank). The significance of this

discovery lies in the fact that it need to be taken into account. By glancing at Table1, you will be able to get a glimpse of some sample records from the Texas dataset as well as the SemEval-2013 dataset. A summary of the ways in which the Deep Learning community makes use of these datasets is shown in Table 2, which can be accessed at this location.

A. ASAP

As part of the Hewlett Foundation's Automated Assessment Prize competition, Kaggle offers a dataset for performing ASAG that goes by the moniker ASAP-AES.

TABLE I

AS AN EXAMPLE, BELOW IS A MODEL QUESTION, ALONG WITH REFERENCE ANSWERS AND STUDENT RESPONSES TAKEN FROM THE SEMEVAL-2013 AND TEXAS DATASETS [1]

Texas	
Question	What is the role of a prototype program in problem solving?
Model Answer	To simulate the behaviour of portions of the desired software product
Model Vocabulary	Simulate, behaviour, portion, desire, software, product
Student Answer 1	High risk problems are address in the prototype program to make sure that the program is feasible. A prototype may also be used to show a company that the software can be possibly programmed
Student Answer 2	It simulates the behaviour of portions of the desired software product
SemEval-2013	
Question	Lee has an object he wants to test to see if it is an insulator or a conductor. He is going to use the circuit you see in the picture Explain how he can use the circuit to test the object
Reference Answer	If the motor runs, the object is a conductor
Student Answer	He could know if it works

(ASAP) on the Kaggle app." You are able to get it by downloading it from the official website of Kaggle. 10686 examples are included in this dataset, and they are all derived from a single challenge. These samples come from eight distinct sets of essays. These essays have an average length of between 150 and 550 words per answer, with the length ranging from low to high. Each and every one of the essays that were included in the dataset was manually graded by either two or three different teachers. The fact that each set has its own unique grading scale is the most significant difficulty associated with this dataset.

B. Beetle and ScientsBank

Students Response Analysis (SRA) was a process that included annotating student-authored replies with categories. "These categories, in turn, would assist dialogue systems in generating appropriate and helpful feedback on faults. (1) BEETLE data, which is entirely based on transcripts of students interacting with the BEETLE II tutorial dialogue system, and (2) SCIENTSBANK data, which is based on the corpus of student answers to assessment questions collected by Nielsen et al. The SRA corpus is primarily composed of two distinct corpora: (1) BEETLE data and (2) SCIENTSBANK data. It is estimated that there are over three thousand student responses to the 56 questions that make up the BEETLE corpus. These questions are mostly related to the field of basic electrical and electronics and need replies that are consisting of one or two sentences. There are roughly 10,000 responses to 197 evaluation questions spread over 15 different scientific fields that are included in the SCIENTSBANK corpus. A trained human being uses a technique that directly translates to SRA annotations in order to manually annotate student responses that are included inside the BEETLE corpus. A fine-grained method that automatically labels using a series of question-specific heuristics and also manually revises them depending on the definition of the class is used in order to convert the labels of the SCIENTSBANK corpus into SRA labels. This scheme helps to ensure that the labels are accurate and accurate." Researchers that are interested in working with these datasets are need to do further filtering and transformations on the corpus in order to generate training and test data sets [1].

C. Texas

This dataset is comprised of eighty questions that were gathered from a course entitled Data Structures that was taken by undergraduate students. A total of 2,273 student responses were collected, and the questions were dispersed throughout 10 distinct assignments and two examinations, each of which covered a connected pair of topics. The students were two human graders who evaluated the responses, and they took into consideration the example solution that was provided for each question. As the final gold score for each student's answer, the average of the two scores that were assigned by human graders is applied [1]. A copy of the dataset is available for download from the archive that is located on the website.

D. Cairo

An overall number of 610 questions are included in the dataset that is made available by Cairo University. Ten replies are supplied for each of the 61 questions that are included in the dataset. “This material is derived from a single chapter of the official Egyptian curriculum for the Environmental Science course. It is published here for your convenience. Answers that are submitted by students typically consist of 2.2 sentences, which is comparable to 20 words or 103 characters. This is the usual length of a response. A collection of replies from students is included in the dataset, along with their grades, which range from 0 to 5 based on the assessments generated by two human evaluators. The dataset also includes the total number of students. In addition, an English version of the data set that was compiled by Cairo University is accessible for use in the process of doing research in this particular area.” Access to the dataset in question may be obtained via the website's download section.

E. Power grading

The Powergrading dataset includes ten separate prompts taken from the United States immigration examinations, each of which had around seven hundred replies. There is at least one reference answer that is included with each question. Because the responses in this dataset are so brief and the proportion of right answers is so high, the responses are repetitious. The model's capacity to provide accurate results from very brief replies may be evaluated with the help of this dataset. This dataset does not have any state-of-the-art score results available since it was first utilised for the job of clustering, which was done without supervision.

F. Statistics

Statistical analysis is the name of the dataset that Stefano Menini and his colleagues made available in order to carry out the process of grading short answer questions. This dataset is available on the internet for download by members of the general public. This information was acquired in part by utilising data from real statistics tests that were taken over a period of time spanning several years. Additionally, the authors of this study have contributed to the extension of this data throughout the course of their research. The dataset contains a collection of phrases that were created by students and will be included in the collection. Every single sentence is assigned a one-of-a-kind sentence ID, the kind of statistical analysis to which it is related, its degree on a scale ranging from 0 to 1, and the conclusion of whether or not it was successful. Additionally, the dataset has a gold standard that was determined manually, which is the solution that was selected.

IV. EVALUTION METRICS

Throughout “the whole of the assessment process, evaluation metrics are used in order to statistically quantify the performance of the ASAG models or to compare the performance with the baselines. In accordance with the method in which the ASAG system is designed, namely as a classifier or as a regressor, a number of metrics may be used, as is seen in Figure 2.” In the next section, you will get knowledge of the metrics that may be used with the ASAG regressor and classifier system [1].

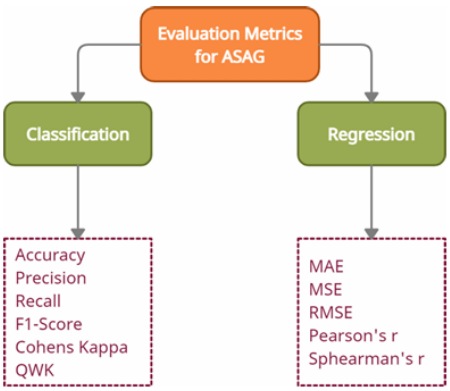

Fig.2 The ASAG task's evaluation metrics

A. Regression Metrics

Root Mean Square Error (RMSE): A popular metric of how well a regression model is doing is the root mean square error (RMSE) [11]. In accordance with the formal definition, it is defined as the root of the residual sum of squares that is derived by comparing the predictions y with the ground truth y. This is the definition from the formal definition. In order to calculate the root mean square error (RMSE), one must use equation 4.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1} n \frac{(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

When it comes to the ASAG job, the equation $\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_n}$ represents the projected grades, whereas $y_1, y_2, \ldots, y_n$ represents the grades that were awarded by human graders. "The number n represents the total number of observations. Calculating the root mean square error (RMSE) is a straightforward process;" nevertheless, the RMSE number is contingent on the size of the observed data, which is a downside.

Correlation Co-efficient: The correlation coefficient is a statistical tool that analyses the degree of relationship between two groups of data [9]. Indicating the degree of relationship, the coefficient's value may range from +1 to -1, and it can even be negative. A positive value is necessary for the ASAG position since the job demands a strong correlation between the reference's replies and the students' own. Researchers use either the Pearson correlation coefficient or Spearman's correlation coefficient when assessing activities with short replies. Using the equation given in Eq.5, one can get the Pearson's r statistic.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (5)$$

This is where r stands for the Pearson's correlation coefficient, xi for the grade given by the human grader, x for the average of the x-variable values, yi for the grader predicted by the model, and ¨y for the average of the y variable.

Equation 6 is used for the purpose of calculating Spearman's ρ.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (6)$$

Let us assume that the symbol ρ represents the Spearman's correlation coefficient, the symbol di represents the difference between the grade that was given by humans and the grade that was predicted, and the symbol n indicates the total number of observations.

In contrast to Spearman's ρ, which is derived from the ranking values of each variable rather than the raw data, Pearson's r is used for the purpose of analysing the linear connection that exists between two variables of interest."

B. Analytics for Classification

F1-Score: The F1-score is often used as a measure for classifier performance. This number is calculated by taking the harmonic mean of Precision and Recall, and it is done by using the seventh equation.

$$\text{F1 Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (7)$$

Where,

$$\text{Precision} = \frac{\text{True Positive}}{\text{Tue Positive}+\text{False Positive}}, \text{Recall} = \frac{\text{True Positive}}{\text{Tue Positive}+\text{False Positive}}$$

Whenever there is an imbalance in the class distribution, the F1-score metric is the one that is preferred. MacroF1 and weighted-F1 scores are the two variations of the F1-score that are used for the tasks associated with the ASAG.

To assess the quality of problems with many binary labels or classes, one may use the macro-averaged-F1 score, abbreviated as macro-F1, which is the geometric mean of the per-class F1 scores; nevertheless, it gives equal weight to each label or class. The most desired value is 1 , however it may take on values between 0 and 1. We utilise equation (Eq.8) to find it.

$$\text{Macro} - \text{F1} = 2 * \frac{\text{Macro Average Precision}*\text{Macro Average Recall}}{Macro\ Average\ Precision^{-1}+Macro\ Average\ Recall^{-1}} \quad (8)$$

Where,

$\text{Macro Average Precision} = \frac{\sum_{k=1}^{K} \text{Precision}_k}{K}$, $\text{Macro Average Recall} = \frac{\sum_{k=1}^{K} \text{Recall}_k}{K}$, $\text{Precision}_k = \frac{TP_k}{TP_k+FP_k}$, $\text{Recall}_k = \frac{TP_k}{TP_k+FN_k}$

The weighted-average-F1 score, also known as weighted-F1, is calculated by dividing the weighted F1 score of each class by the total number of samples that come from that class [1].

Cohen's Kappa: The kappa statistic devised by Cohen is able to handle difficulties involving several classes as well as unbalanced classes well. It was designed to take into consideration the potential that response graders may make an educated estimate on at least one variable owing to uncertainty. It is determined by applying the equation 9 to the data.

$$k = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \quad (9)$$

where p_o and p_e are the degrees of agreement that are observed and predicted, respectively. A value of -1 to +1 is possible for the kappa value. A trained response grader is considered to be ineffective if the values are less than or equal to zero. Despite the fact that it is the statistic that is used the most, the acceptable level of kappa value is being called into question in a few disciplines, such as health research.

Quadratic Weighted Kappa: In order to determine the degree of concordance that exists between two ratings, the QWK measure is used [10]. When it comes to the ASAG job, it may be used to determine the degree of concordance that exists between the grades that were predicted and the actual results. Additionally, it takes into account the possibility that two raters will award the same grade to a sample due to random differences. In most cases, it falls somewhere between 0 and 1, although it may also result in a negative value if there is less agreement. For the purpose of calculating QWK, the weight matrix W must first be built in accordance with Equation 10.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \quad (10)$$

Where i represent the rating that was given by the human evaluator, j represents the rating that was predicted, and N represents the total number of ratings. After that, the QWK is computed using the equation by following the steps.

$$k = 1 = \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \quad (11)$$

The scores that were observed are included in the matrix O, where the human grader is responsible for assigning rating i, and the model is responsible for assigning rating j. The adoption records that have a rating of i and projected a rating of j are the ones that match to the Oij designation. The histogram matrix of anticipated ratings is denoted by the letter E. This matrix is accomplished by multiplying the histogram vectors of both the human grader score and the model projected score.

V. DEEP LEARNING APPROACHES

Over the course of the past few years, the researchers have utilised a variety of different processes, such as "Transfer Learning, Siamese LSTM, clustering, Latent Semantic Analysis, Bidirectional Transformers, Paragraph Embedding, Deep Auto encoders, Attention Networks, and Transformer-based pretraining [17]. These processes were utilised in order to accomplish their goals. As a consequence of recent advancements in the area of deep learning for natural language processing (NLP), it is promising to apply deep learning structures for more difficult natural language processing tasks. Some examples of these designs include the Attention mechanism and Transformer. Figure 3 illustrates a few of the possible contributions that were produced in 2016 via the use of experiments that were based on Deep Learning. Table 2 offers a description of the degree of performance that these experiments achieved." The Long Short-Term Memory (LSTM) and its variants have been the focus of investigation by a considerable number of researchers [21].

TABLE 2

RESULT FOR INDIVIDUALS ON ASAG

Year	Approach	Corpus	Evaluation							
			Accuracy	Macro-F1	Weighted-F1	QWK κ	Cohen's κ	Pearson's r	Spearman's ρ	RMSE
2016	LSTM	SICK	–	–	–	–	–	0.88	0.83	–
2016	LSTM	ASAP	–	–	–	–	0.96	0.96	0.91	2.4
2017	Bi-LSTM	Sem-Eval	–	–	–	–	–	0.55	–	0.75
		Texas	–	–	–	–	–	0.64	–	0.83
2017	LSTM	ASAP	–	–	–	0.74	–	–	–	–
		Powergrading	–	–	–	90.36	–	–	–	–
		SRA	–	74.54	–	–	–	–	–	–
2017	Bi-LSTM	Sem-Eval	0.76	–	–	-	–	–	–	–
2018	LSTM	Sem-Eval	79.26	78.58	79.1	–	–	–	–	–
		Texas	–	–	–	–	–	0.57	–	0.9
		LSI	66.36	63.09	65.58	–	–	–	–	–
2019	LSTM	K12	88.9	81.5	–	–	–	–	–	–
2019	LSTM	Texas	–	–	–	–	–	0.63	–	0.89
		Cairo	–	–	–	–	–	0.79	–	0.92
2019	Transformers	Custom	80.17	81.5	–	–	–	–	–	–
2019	Transformers	ScientsBank	75.9	72.0	75.8	–	–	–	–	–
2020	Transformers	SemEval	79.7	79.1	79.7	–	–	–	–	–
		ScientsBank	76.0	75.0	76.0	–	–	0.65	–	0.88
2021	LSTM	Ukara	–	–	–	–	–	–	–	–

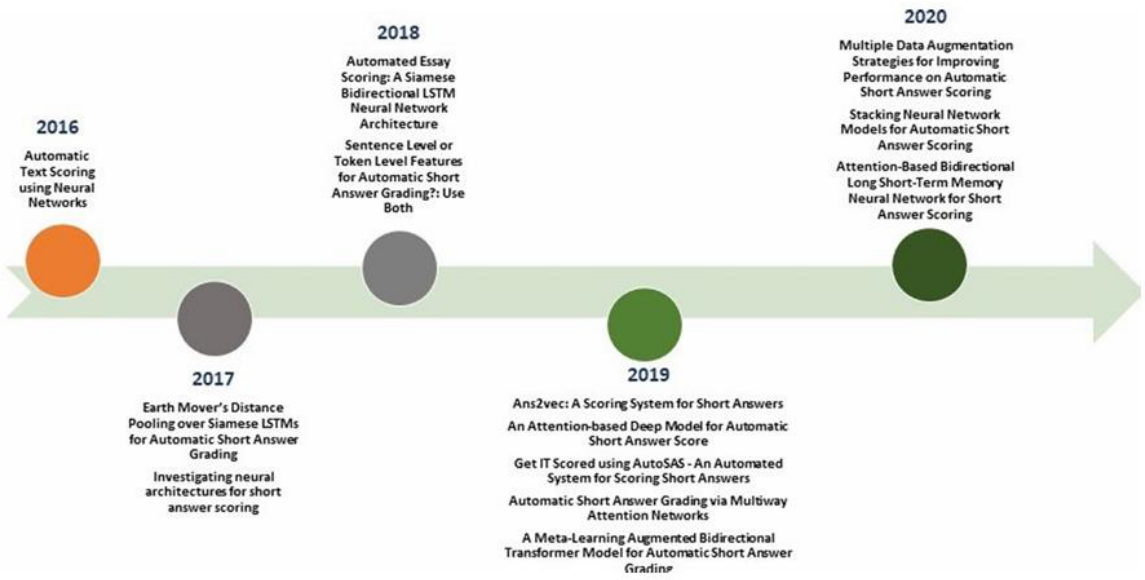

Fig.3 Prominent achievements using Deep Learning [1]

A. LSTM

Large-scale long-term memory (LSTM) units are complex activation units that have the ability to selectively remember or forget information and facts. Kumar et al. Siamese bidirectional LSTM, a layer based on Earth-Mover-Distance (EMD), is used to cover all hidden states of the LSTM of the Siamese network, and the final transformation process is used to provide points to solve the ASAG light problem. This is done to generate points. To improve training, the authors use a method called task-specific profile augmentation. In addition, they were tested on publicly available data (SemEval) and achieved a RMSE score of 0.830, which is higher than the LSTM baseline. Using our publicly available data, Riordan et al. The performance of different topologies was investigated. These documents are ASAP-SAS, Power grading and SRA. They provide answers to questions such as whether convolutional processes produce features of interest, whether we can use fewer layers, the power of bidirectional LSTMs, and maintainability. They concluded that the neural model based on learning previous embedding using LSTM is an excellent architecture for short response scoring. That's what

they did. Connaught et al. A new joint multineuron model (JMD-ASAG) has been proposed for ASAG. The model developed by scientists. It achieves this goal by using various registry-specific corpora (such as SemEval-2013) and does not require a large target audience. So it achieves its purpose well. Prabhudesai and his colleagues proposed the idea that such a regressor would be useful as it would provide the benefits of deep learning and good performance. The authors used a new method to improve training [23]. This approach involves supplementing data with highly effective responses.

B. Attention

In the fields of natural language processing (NLP) and deep learning, it is the concept that has received the greatest attention among researchers. For the purpose of generating the context vectors that are required for the decoders, each and every state of the intermediate encoders is used [12]. The idea of attention may be broken down into one essential aspect. In order to automatically collect linguistic information from the responses of both the student and the reference, as well as precisely model the semantic linkages that exist between the student and the reference, it was suggested by Liu and colleagues that a generalised end-to-end ASAG framework be developed. This framework would be used to accomplish both of these goals. The evaluation of this model is carried out with the use of a dataset that is comprised of K-12 pupils from the real world. “The model that they developed made use of the multi-way attention and transformer layers in order to improve the matching between words that were included inside a phrase. This model outperformed several state-of-the-art baselines, particularly Logistic regression, Gradient boosted decision tree, Multichannel convolutional neural networks, Sentence embedding by Bidirectional Transformer block (Bi-Transformer), Multiway Attention Network (MAN), and Manhattan LSTM with max-pooling (MaLSTM), with an accuracy score of 0.8899 and an AUC score of 0.9444. The authors made the decision to use area under the curve (AUC) and accuracy as metrics, and they chose to use these metrics. In addition, Gong et al. proposed a method that is based on deep learning and adds an attention mechanism in order to solve this issue. This method was proposed in order to finish the challenge. An embedding word vector that has been pre-trained and a Recurrent Neural Network (RNN) model that has attention to learn answer vector are both components of the approach that has been proposed. The learnt response answer vector and the reference answer vector are then entered into the logistic regression model in order to make a prediction about the response answer score.” As measured by Quadratic Weighted Kappa (QWK), the authors assert that they have achieved a relative 10% improvement in performance in contrast to the results of the baseline model by more than 8% in some of the question prompts. This demonstrates that the authors have achieved performance that is comparable to that of humans.

C. Transformer

In the article titled "Attention Is All You Need," the innovative architecture known as Transformer was revealed to the reader. Making use of the attention mechanism in the manner that the term says it should be used. “BERT, which is an acronym that stands for "Bidirectional Representation of the Transformer," is widely regarded as the most cutting-edge model for the purpose of acquiring knowledge about linguistic representations. Wang et al. devised the ml-BERT strategy for grading short answer questions in order to advance the existing models in ASAG in circumstances when the training data is inadequate. This was done in order to make progress regarding the models. A training framework that takes use of supplemental data and education assignments is called meta-learning. The authors merged BERT with meta-learning in order to improve the performance of the model in circumstances when there is a limited quantity of labelled data. As a consequence of the incorporation of meta-learning, the model was able to achieve an impressive accuracy of 80.17% and an F1 score of 0.815 than it had previously achieved. When compared to the baseline BERT model, which has an accuracy of 77.8%, this result is much higher. In addition, they proved that models that were trained via the process of knowledge distillation are capable of being used in the grading of brief answers. Camus and colleagues conducted research on Transformers for ASAG by fine-tuning a number of Transformer-based designs that had been pretrained. They also made this finding, which is interesting. For the objective of upgrading the pre-trained BERT language model for the short response grading, Sung et al. proposed two distinct ways. These approaches were intended to be used, respectively. Both unstructured data from textbooks and data from labelled question-answer sessions were used by the authors in order to provide an explanation for the evolution of the model. When compared to the results of the state-of-the-art, the benchmarking dataset of SemEval2013 reveals an absolute improvement of up to 10% in macro-average-F1. This is according to the findings that they obtained. In addition to this, the authors examined the length of time that is necessary for training in order to create the perfect model.” In addition to this, they said that task-specific transfer is only seen during the first few epochs of their training.

A few writers conducted experiments in which they employed deep learning models to evaluate skip-thought vectors with a range of embeddings. During the course of their investigation, Saha and his colleagues developed novel token-level features that are specifically designed to interpret words that are only partially correct. Ans2vec is the name that Gomaa and his colleagues gave to the grading system that they developed for

short comments. This technique makes use of skip-thought vectors in order to convert the replies of the model and the students into meaningful vectors. These vectors can then be used in order to ascertain the degree of similarity that exists between the two sets of responses [8]. This model was given a Pearson correlation value of 0.63 when it was applied to the dataset that was collected from Texas. This model was examined using three different benchmarking datasets to determine its performance. For the first time, Yaman and his colleagues presented a supervised regression model that they named AutoSAS. It is possible to utilise this technique in a classroom setting in order to swiftly score students' short replies. Within the scope of this investigation, the features that were used the most often were Weighted Keywords and Word2Vec/Doc2Vec embeddings. This category includes characteristics such as prompt information, weighted keywords, lemmatized response, and lexical overlap, to name only five instances of such characteristics. According to the authors' results, additional variables such as the frequency of the word, the difficulty of the word, the statistics of the word, and the length of the sentence do not have a big effect in either the ranking or the accuracy values being greatly affected. The authors assert that AutoSAS has attained higher levels of performance and has acquired the ability to operate at the cutting edge of technology.

VI. RESULT

The scholars that are interested in applying deep learning to find a solution to the ASAG issue are the ones who are doing this investigation. In order to accomplish this objective, we formulated a group of four research questions, which are provided in Section 2. We have examined a collection of 38 recent publications that focus on Deep Learning strategies for resolving the issue of short answer grading. Our results and observations are outlined in the following paragraphs.

RQ1: “What are the various datasets available to perform ASAG?” - In order to provide an answer to this issue, we have conducted an analysis of the advantages and disadvantages of the datasets that are used by the bulk of the research community, and the specifics of this analysis are described in Section 3. When it comes to training the model, Deep Learning demands a greater amount of data than regular machine learning does. A great percentage of the datasets that are intended for the grading of brief answers have a small number of entries. Only a handful of the datasets have student replies that are either well-formed or nonsentential respectively. In order to implement deep learning, suitable corpora that include a large number of training records are required. One other thing that can be seen is that the questions that are included in the datasets are gathered from courses such as environmental science, data structures, and other related topics. There is not a single dataset that contains questions from programming areas that can be used to accurately evaluate the capabilities of Deep Learning techniques.

RQ2: “What are the various Evaluation Metrics used to measure the performance of ASAG tasks?” - The evaluation metrics for the ASAG task change depending on the approach that is used to address the challenge. The metrics for the ASAG task's Regression and Classification settings are discussed in Section 2, which serves to offer an insight into these metrics. The majority of ASAG challenges articulate the metric that will be used to evaluate performance. It is not difficult to compute all of the ASAG metrics, and the majority of them are included in the open-source libraries that are accessible.

RQ3: “Which Deep Learning approaches are used?” - The methodologies that were used in each of the 38 studies were investigated. the advantages and disadvantages that were noticed in models that were offered in publications that performed well. In order to effectively tackle the ASAG challenge, researchers have used a wide range of deep neural networks, including Fully Connected Networks, Convolutional Neural Networks, Recurrent Neural Networks such as LSTM, and transformers. Although we have seen that a number of writers have achieved high performance via the use of LSTM with careful attention, it is important to note that LSTM are sequential in nature and cannot profit from the use of GPUs.

RQ4: “What are the results obtained?” - The Deep Learning model will not be the only factor that determines whether or not the ASAG system pipeline is successful in production.

A number of factors contribute to the model's performance, including its repeatability on a comparable but smaller dataset [16], the simplicity of fine-tuning, and the latency of the inference time. When pitted against several state-of-the-art models that relied on manually produced features and minimised the time needed for feature extraction, many of the ASAG systems suggested utilising Deep Learning methodologies fared very well. No deep learning model has been trained on the advantage of automatically sending students feedback on their replies, despite researchers discussing it. There is need for new approaches and strategies to get better performance, even if numerous Deep Learning models have suggested and attained accuracy. Models that made use of an attention mechanism outperformed the others. Learning a language revolves on encoding text. The encoding models, such as Deep Averaging Network (DAN), are made to be as versatile as they can be by using multi-task learning. What this means is that researchers have a lot of leeway to experiment with various combinations of self-attention, co-attention, hierarchical-attention, and multiple-attention in order to optimise

performance. New tokenizations such as Byte-Pair encoding, Word-Piece encoding, SentencePiece encoding, etc., have been released in the last year, following the Transformer paradigm with Reformer, Long former, GTrXL, etc., and BERT with XLNet, RoBERTa, T5, etc. Electric motors are quadratic. These may all be evaluated using the current ASAG system's deep learning models to see how they perform in terms of accuracy." Due to its size, GPU inferences may be expensive, and GPT3 is very large with 175GB parameters. The inference time may be reduced by a factor of 10 to 40 with the use of quantization, pruning approaches, and onnx. Given the inference latency requirements, it's possible that we don't always need a model of BERT-base size, and that smaller models will be necessary. Finding the optimal balance between transferring knowledge and training the model from scratch should be the goal of the research community.

VII. CONCLUSION AND FUTURE CHALLENGES

In today's school system, there is a pressing demand for an ASAG system that is not only more accurate but also requires less time to train and infer. Over the course of the last five years, academics have begun to use Deep Learning strategies in order to address this research challenge [5], which has been around for many decades. The current models that are based on deep neural networks, the corpora, and the evaluation metrics that are utilised for this job are all things that are investigated in this research. Among the models that were investigated, the ones that used attention mechanisms did very well. Additionally, it was discovered that there are a restricted number of corpora suitable for use in this field. It is necessary to develop fresh, high-quality datasets that also include contemporary topics such as programming languages. Furthermore, there are several opportunities to continue researching in this field by using enhanced encoding methods, generalised principal component analysis (GPT), quantization, pruning, and several other techniques.

The authors of the current Deep Learning techniques for ASAG were the ones who made the methodological choices that were addressed in Section 5. In Section 6, we also expressed our opinions on the articles that were previously examined. Using the findings of this study, we have developed a research agenda that includes possible next actions that may be taken to improve the outcomes. The following are the primary difficulties that have been found and that have to be solved in further research: As a result of the fact that many of the university campuses are preparing their students for placements, it has been determined that there is a need for a corpus that includes a good number of training instances and covers technical concerns. Secondly, an ASAG model that is able to learn from its surroundings and requires less time for training and inference. (3) A methodology that offers quick feedback in addition to the grade provided.

REFERENCES

[1] Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, pp. 715–725. Association for Computational Linguistics, August 2016. https://doi.org/10. 18653/v1/P16-1068. https://www.aclweb.org/anthology/P16-1068
[2] Angelov, P., Sperduti, A.: Challenges in deep learning. In: ESANN (2016)
[3] Basu, S., Jacobs, C., Vanderwende, L.: Powergrading: a clustering approach toamplify human effort for short answer grading. Trans. Assoc. Comput. Linguist. **1**, 391–402 (2013)
[4] Beltagy, I., Peters, M.E., Cohan, A.: Longformer: the long-document transformer.arXiv preprint arXiv:2004.05150 (2020)
[5] Benesty, J., Chen, J., Huang, Y., Cohen, I.: Pearson correlation coefficient. In: Benesty, J., Chen, J., Huang, Y., Cohen, I. (eds.) Noise Reduction in Speech Processing. STSP, vol. 2, pp. 1–4. Springer, Heidelberg (2009). https://doi.org/10. 1007/978-3-642-00296-0 5
[6] Brenner, H., Kliebsch, U.: Dependence of weighted kappa coefficients on the number of categories. Epidemiology 199–202 (1996)
[7] Brown, T.B., et al.: Language models are few-shot learners. arXiv preprintarXiv:2005.14165 (2020)
[8] Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answergrading. Int. J. Artif. Intell. Educ. **25**(1), 60–117 (2015)
[9] Camus, L., Filighera, A.: Investigating transformers for automatic short answergrading. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Mill´an, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 43–48. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7 8
[10] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learningof universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364 (2017)
[11] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-XL: attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019)
[12] Dwivedi, C.: A study of selected-response type assessment (MCQ) and essay typeassessment methods for engineering students. J. Eng. Educ. Transform. **32**(3), 91– 95 (2019)
[13] Dzikovska, M.O., et al.: Semeval-2013 task 7: the joint student response analysisand 8th recognizing textual embodiment challenge. In: Second Joint Conference on Lexical and Computational Semantics (* SEM): Seventh International Workshop on Semantic Evaluation (SemEval 2013), vol. 2. Association for Computational Linguistics (2013)
[14] Gomaa, W.H., Fahmy, A.A.: Ans2vec: a scoring system for short answers. In: Hassanien, A.E., Azar, A.T., Gaber, T., Bhatnagar, R., F. Tolba, M. (eds.) AMLTA 2019. AISC, vol. 921, pp. 586–595. Springer, Cham (2020). https://doi.org/10. 1007/978-3-030-14118-9 59
[15] Gong, T., Yao, X.: An attention-based deep model for automatic short answerscore. Int. J. Comput. Sci. Softw. Eng. **8**(6), 127–132 (2019)
[16] Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: anoverview. arXiv preprint arXiv:2008.05756 (2020)
[17] Guerra, L., Zhuang, B., Reid, I., Drummond, T.: Automatic pruning for quantizedneural networks. arXiv preprint arXiv:2002.00523 (2020)

[18] Hasanah, U., Permanasari, A.E., Kusumawardani, S.S., Pribadi, F.S.: A review ofan information extraction technique approach for automatic short answer grading. In: 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 192–196. IEEE (2016) 19. Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. Int. J. Forecast. **22**(4), 679–688 (2006)
[19] J. Luo, "Automatic Short Answer Grading Using Deep Learning", Accessed: Apr. 22, 2022. [Online]. Available: https://ir.library.illinoisstate.edu/etd/1495
[20] Kaggle: The Hewlett Foundation: Automated Essay Scoring—Kaggle. https:// www.kaggle.com/c/asap-aes/. Accessed 04 Oct 2021
[21] Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: the efficient transformer. arXivpreprint arXiv:2001.04451 (2020)
[22] Kumar, S., Chakrabarti, S., Roy, S.: Earth mover's distance pooling over SiameseLSTMs for automatic short answer grading. In: IJCAI, pp. 2046–2052 (2017)
[23] Kumar, Y., Aggarwal, S., Mahata, D., Shah, R.R., Kumaraguru, P., Zimmermann,R.: Get it scored using autosas-an automated system for scoring short answers. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9662– 9669 (2019)
[24] Li, Z.; Zhang, C.; Jin, Y.; Cang, X.; Puntambekar, S.; and Passonneau, R. J. 2023. Learning When to Defer to Humans for Short Answer Grading. In International Conference on Artificial Intelligence in Education, 414–425. Springer.
[25] Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G.Y., Liu, Z.: Automatic shortanswer grading via multiway attention networks. In: Isotani, S., Mill´an, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), vol. 11626, pp. 169–173. Springer, Cham (2019). https://doi.org/10.1007/978-3-03023207-8 32
[26] Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. arXivpreprint arXiv:1907.11692 (2019)
[27] Lopez, M.M., Kalita, J.: Deep learning applied to NLP. arXiv preprintarXiv:1703.03091 (2017)
[28] Lun, J., Zhu, J., Tang, Y., Yang, M.: Multiple data augmentation strategies forimproving performance on automatic short answer scoring. In: AAAI, pp. 13389– 13396 (2020)
[29] 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
[30] Putnikovic, M.; and Jovanovic, J. 2023. Embeddings for Automatic Short Answer Grading: A Scoping Review. IEEE Transactions on Learning Technologies.
[31] 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.