Automated Subjective Answer Evaluation Using Machine Learning

1st Ishika Aggarwal
Department of Computer Science & Engineering
KIET Group of Institutions
Ghaziabad, India
ishagrawal24@gmail.com

2nd Pallav Gautam
Department of Computer Science & Engineering
KIET Group of Institutions
Ghaziabad, India
pallav.technorex46@gmail.com

3rd Gaurav Parashar
Department of Computer Science & Engineering
KIET Group of Institutions
Ghaziabad, India
Orcid Id: 0000-0003-4869-1819

***Abstract*—A written exam or a test is a technique to check or assess the knowledge of a student or skills and abilities. With every passing year the methods to take examinations change, but what does not change is how those exams are evaluated. It's the same physical and exhausting mode of evaluation. We used previous research to come up with a solution where the evaluations can be done automatically and with precision and with the least error. We assessed the problems of evaluators and reduced them to almost 90% saving them a huge amount of time. We converted handwritten notes to text and generated an engine that would take those answer texts and question papers as input and then evaluate and give marks. The result we get is on a scale of 0 to 9 then convert them intelligently to percentages.**

***Index Terms*—Automatic Subjective Evaluation, Intelligent scoring, Automatic feedback**

I. Introduction

Educational assessments are very important in the learning process for students. The evaluation and scoring is a tedious process that also takes most of the valuable time of professors. Because of this, the scoring is not optimum, lacks fairness, and is the cause of mental and physical fatigue of the evaluator. Also in physical evaluation, the chances of getting feedback are very difficult. The primary goal of a teacher is, simply, to teach, but this evaluation system is taking away the most important resource a teacher has time.

In this work, we present a system where the answer sheets are evaluated automatically, Automatic Subjective Evaluator. How is an answer sheet actually evaluated? From a teacher's or evaluator's point of view, it's 3 step process. First, you read an answer, then match it with the existing answer we have or judge how close it is to the actual answer, which words are closest to the correct answer, how much is the deviation from "to the point answer" etc., then score according to that. Is a computer program capable of doing all this with precision? Maybe. This task can be achieved using an ability known as Artificial Intelligence (AI). AI is the capability of a computer machine to perform certain tasks without human intervention. AI is the science that learns about human thinking. It is a part of computer science technology that emphasises on creating intelligent and quick solutions that can execute those tasks that normally require intelligence like humans, for example, learning, solving problems, smart decision making, and language understanding. AI has the power to revolutionize many industries/markets and change the way we live our life and work. It is already helpful in various fields including self-driving cars, virtual assistants, and image processing and speech recognition.

We use natural language processing (NLP) for our system. NLP is a sub-part of artificial intelligence that improves the relationship between computers and humans with the utilization of natural language. It involves the development of algorithms and models which can analyze, judge, and lead to human language, allowing computers to interpret and respond to human requests and commands. NLP has a vast range of applications, which includes language translation, summarization of text, analysis of sentiments, and development of chatbots. It is a flexible field that joins computer science, grammatical, and cognitive psychology to permit computer machines to exercise and understand human language in a familiar way to how we humans do. We convert those answer sheets into texts then summarise and use them as the input to the engine. After that, it is matched with the answer found on various websites by our engine then uses the similarity index technique to find similarity and score a particular answer.

II. Literature Review

An automatic subjective answer sheet evaluator, also known as an automatic essay grader or automated scoring system, is a tool that uses NLP(Natural Language Processing) and algorithms of machine learning to evaluate and grade written responses to open-ended questions. These tools are primarily used in educational settings to grade student essays, but they

have also been used in other areas such as language proficiency testing [1].

A literature review of automatic subjective answer sheet evaluators reveals that these tools have been widely researched and developed over the years. Many studies have been conducted to evaluate their effectiveness in grading student essays.

Research has shown that automatic subjective answer sheet evaluators are able to grade written responses with a high level of accuracy. Studies have found that the scores generated by these tools are very close to those produced by human graders.

Past research shows that because of the performance of teachers' measures, which is the test scores, harms their job fulfilment, together with increased distress and health issues.

If we agree with these kinds of assumptions, there is an automatic trust that numerical measures, like test scores, show the actual reality of the condition and that the result of education is the outcome of the individual working and is not affected by greater societal contexts or family occurrences.

The constant pressure to enhance test scores leads to changes in the probabilities by which the educator profession, and educators, can be valued, and the ways that they can finally ultimately be and relate themselves related to their work.

Many studies have found how the social expectations of educators, and the usual culture of testing, are correlated with a hike in the workload of the teacher and constant, work-related pressure, mental fatigue and stress, and reduced job satisfaction.

A study has linked risky accountability to extravagant fatigue and teacher yield, proclaiming that their participants showed disappointment, the reality of teaching is poorer than we know, and the nature (rather than the quantity) of the workload, associated with performance and accountability, being an important factor for why teachers were leaving the profession.

Bringing together the issues associated with statistics by using the test scores of the student in teacher accountability and the pressures often attached to such systems, we say that the testing culture is bringing an environment where teacher satisfaction is adjusted. [1].

Generally, many of the AWE technologies use (a) Natural Language Processing tools to extract linguistic, syntax-related, semantics-related, or different attributes of text related to writing quality and (b) statistics-related or machine-learning algorithms to create scores and feedback based on patterns discovered among those features.

Natural language processing(NLP) is a subpart of Artificial Intelligence that works with the interaction between humans through natural language. NLP enables computers to understand, explain, and generate human language, whether it is written or spoken. NLP involves various tasks for example classification of text, analysis of sentiments, called entity recognition, language translation, and speech recognition. It uses machine learning algorithms and statistical models to analyze and derive meaning from human language.

Initially, NLP was not for text retrieval, because it included high-level techniques to search large data sets. As technological advancement grew, that information retrieval and NLP were merged. That is the reason NLP today is so vast. [2]

The goal of NLP of to do human-like processing. Originally the field of NLP was called Natural Language Understanding (NLU). Some of the jobs of NLU were - paraphrasing an input text and translating the text into another language. The concepts that might be used in automatic subjective answer sheet evaluation may be extracting text from an image or a pdf and paraphrasing the questions and answers, drawing inferences from those texts like a human being (the evaluator) does. [3]

Machine learning is the area of study where computers have the ability to learn things without being hard coded. Arthur Samuel was famous for his checkers playing program. ML teaches computers how to handle data correctly. The purpose of ML is to learn from the given data. For that it used algorithms. ML can be supervised or unsupervised. For fewer data, supervised learning is preferred. [4]

An automated answer sheet evaluator helps students get automated scores and feedback, improving their work and writing both in terms of looks and content. It is equally beneficial for the teachers as it gives them the power to monitor and access students more efficiently. Teachers who used AWE gave more effective feedback on answer writing skills than teachers who didn't use it. Due to lack of time and energy, they provide insufficient feedback on marks, answers etc. Still, AWEs are not yet adopted by schools, and universities because of less accurate scoring algorithms. Well, some people argue that if students use automated feedback then their answers will be very much formal and not natural. Some say that NLP-based evaluation will restrict the checking because it will incur only those results that it really can from the methods/algorithms, nothing beyond that. Currently, there is very less amount of studies conducted on AWE.

Questions have also been raised on the conclusions driven by the AWE because it is tested on a certain amount of answers and only in the English language. But it increases the quality of answers produced by the students. It was seen that students who received regular automated feedback produced fewer errors in grammar, style and answer quality. [5]

When tied with tutorials or educational games, learning management features or peer assessment, AWEs potentially offer flexible, robust, and time-saving additions to the writing curriculum.

One of the solutions is to generate each answer again and again for a specific question. It is done using a randomisation array and a key space array. The implementation process is in 3 parts. Generating answers, distributing answers and starting evaluation. Generation starts at the server side by selecting a question. Then the answer is extracted and given to the engine for evaluation. There can be an explanation for this. Giving feedback is different from using feedback. It was seen that 50% of the students didn't use feedback but those who used improved in a lot of areas.

People should not just consider the cost of buying the AWE software but also what will happen if they will not buy this

software. Also with automated feedback student is motivated to write because he/she know that this feedback is unbiased so it increases the effort and outcome of the student. [6]

Now how will this system work? There will be many modules on which the system will work. There may be a management module which has the user information, and login details of both the evaluator and student. The maintainer can put the details of the users and grant permissions. Teachers can maintain their profiles like change passwords, change their subjects, upload study materials, schedule surprise tests etc. The student management portal allows the student to check their result, see feedback both automated and teacher generated, submit solutions, view scores etc.

The answer is scored using the cosine similarity. For every question, there is an answer that the system gets itself from search engines like Google or provided by the teacher. Then both the answers, given by students and by the teacher or the search engine are sent forward to the engine for summarization and other evaluations. Then the similarity between the summarized text is calculated using some algorithms. [7]

AWE is more objective and more accurate while grading an answer. AWS can also score the answer according to the level of the student's mind and his/her needs. It can be programmed for that too. AWE can check a very large number of copies in very less. However, AWE can give unclear judgement and feedback if the student is not familiar with the software's way of evaluating. The way AWE evaluates can be different from what a human evaluates. Also, there will be a fixed number of feedbacks given by the AWE because that feedback will be given from the inbuilt feedback options. The team in this [8] study, worked on the data of 28 undergraduate students. This study showed that there was an increase in the results of the students. The AWE increased their awareness of spelling and grammar. [8]

Most of the research is done by the programmers to make their software work. AWE's feedback may not be useful for everyone or all students. One of the main purposes of AWE is to teach students to go through their answers before submitting their answer sheets which students seriously lack. [9] AWE feedbacks also focus on the revision techniques of the students. It can change the way students revise their copies. [10]

III. Experiment

The system can be used in educational institutions such as schools, colleges, and coaching centres for evaluating the answer sheets of students. In this proposed model, the data set includes typed answers and paragraphs taken from students themselves and from various websites and blogs for comparison. The data set needs to be massive amount so it comes from students through google forms, we also targeted various websites that contain subjective questions answer to train our model.

In our proposed system, we compare the marks given by a machine and a human. Evaluation is done by the algorithm proposed and assign the marks. Then a human will evaluate the same answer and assign the marks. Then a comparison is made between both the marks to find out the accuracy and correctness of the machine evaluation process.

The overall system consists of five sub-processes. All these five sub-process have their own working and methodology based on specific techniques. The idea behind the machine evaluation is that firstly we summarize the subjective question answers written by students and find the first ground truth. Then we search for the actual answer to the subjective question through various websites and platforms in order to find out the second ground truth. A comparison is made and the similarity Index is calculated between both the ground truths obtained using the Cosine Similarity Index. Now, the classes are made based on some defined ranges. Using the value obtained from Cosine similarity, we decide the class and marks are given to the student's answer.

Let us understand each process in detail-

A. Process 1

In this sub-process, the answers written by the students are summarized using the Text summarizer tool QuillBot AI. It is an online AI-based tool that provides the text you need using Natural Language Processing (NLP) by maintaining the context of it. It acts as a Summarizer, Paraphraser, Grammar checker, Plagiarism checker and Citation generator. It is a powerful tool used for research and writing purposes. After summarization, the first ground truth is obtained.

B. Process 2

In this sub-process, the actual answer to the question is searched over the websites and other platforms and the URL is obtained. The URL is parsed to extract the text from it. The Process is called Web Scraping which is done using the beautifulSoup4 package in Python. Web Scraping is an automated process used to extract the massive amount of unstructured data from websites and store it in a structured format. To do web scraping in our process, we use Python, its packages and libraries.

C. Process 3

In this sub-process, the parsed text obtained after web scraping is summarized using the summarizing tool QuillBot AI. This will give more precise and brief text. This text is called the second ground truth, which is used for comparison with other ground truths.

D. Process 4

This sub-process is related to the calculation of the similarity index between two Ground truths(GT) obtained using the Cosine similarity Index. It is implemented using Python. The cosine similarity between two numerical sequences is a metric of similarity. It is defined as the cosine of the angle between the sequences, which is the dot product of the vectors divided by the product of their lengths. The sequences are viewed as vectors in an inner product space. As a result, only the angle of the vectors matters for the cosine similarity, not the magnitudes. Cosine similarity is always found in the range [-1,1].

The term frequency vectors of the documents ae treated as the attribute vectors X and Y for text matching. When comparing documents, cosine similarity may be thought of as a way to normalise document length. It is used to calculate the similarity between two text documents or tokenized texts. The raw data has to be tokenized and a similarity matrix has to be generated which can be passed to the cosine similarity metric to check the similarity between the texts.

E. Process 5

This is the final sub-process which includes classifying similarity output in ranges to allocate marks to individual answers. When similarity multiplies by 100, the percentage similarity between two texts is obtained. When the similarity is multiplied by 10, we get a number ranging from 0 to 9. Classes are made and marks are allotted based on the classes automatically by machine.

After the machine evaluation, the same answers are evaluated by a teacher, and marks are given. Both the marks are compared in order to find the correctness of the proposed system.

IV. Results & Discussion

a) Division of classes in ranges and allocating Marks : To allocate the marks after calculating the cosine similarity, Classes are made based on the ranges on a scale of 0 to 9. These classes tell us the category of answer that whether it is Good, Very Good or Excellent. This will be helpful in finding out the quality of answer based on the category or class, which makes our process of allocation of Marks more smooth and organised. This is shown in table I.

TABLE I
Class with Ranges

Class	Range
Good	0-3
Very Good	4-6
Excellent	7-9

b) Calculation of Cosine similarity: Using the proposed system, cosine similarity is obtained using the data set from various students. When comparing documents, cosine similarity may be used as a way to normalise the length of the document. The Cosine similarity is calculated between two tokenized texts. Approximate marks or score is given based on the comparison made between the student's answer and the actual answer. Cosine similarity is always found in the range [-1,1]. This number when multiplied by 10 generates the marks scored by the student. This is shown in table II.

c) Comparison between Human Evaluation and Proposed System: In this proposed system, we first show the result obtained by the algorithm developed using the tools and technology. After that, we also evaluated the student's responses manually with the help of teachers in order to check the consistency of the result. This step plays an important role in checking the accuracy of the proposed system. This is shown in table III.

TABLE II
Cosine similarity Calculation

Question	Student	Cosine Similarity
Q1	1	0.7035
	2	0.5687
	3	0.4157
	4	0.4747
	5	0.3527
	6	0.5119
	7	0.6080
	8	0.7424
	9	0.7689
Q2	1	0.6424
	2	0.5801
	3	0.6859
	4	0.8891
	5	0.6083
	6	0.6246
	7	0.6080
	8	0.5452
	9	0.9049
Q3	1	0.8314
	2	0.5787
	3	0.7197
	4	0.6624
	5	0.3905
	6	0.7841
	7	0.5746
	8	0.6000
	9	0.9055
Q4	1	0.7428
	2	0.5411
	3	0.6890
	4	0.7105
	5	0.5568
	6	0.6194
	7	0.8528
	8	0.7659
	9	0.5567
Q5	1	0.9028
	2	0.5678
	3	0.6722
	4	0.8567
	5	0.6780
	6	0.5257
	7	0.4896
	8	0.9040
	9	0.6789

Conclusion

This paper provides a clear approach to evaluate the answer sheets of the students based on the algorithm proposed using machine learning and natural language processing. Our system provides a convenient way to do evaluations and generate almost accurate results. Various tools and techniques are used in this paper which are better in providing the results.

In this, we can evaluate the answer of any length and there are no word limits. Our system processes and generates the most deserving scores without any partiality. Furthermore, we also make a comparison between the machine score and the human score in order to check the certainty and accuracy of the system.

With sufficient training, the model can work on its own and generate scores without the need for any manual semantics checking.

TABLE III
COMAPRISON BETWEEN HUMAN AND MACHINE SCORE

Student	Machine Score	Human Score
1	7.0	8.0
2	5.6	8.0
3	4.1	6.0
4	4.7	6.0
5	3.5	6.0
6	5.1	8.0
7	6.0	8.0
8	7.4	8.0
9	7.6	9.0
1	6.4	9.0
2	5.8	8.0
3	6.8	9.0
4	8.8	9.0
5	6.0	9.0
6	6.2	9.0
7	6.0	6.0
8	5.4	6.0
9	9.0	9.0
1	8.3	9.0
2	5.7	6.0
3	7.1	8.0
4	6.6	8.0
5	3.9	4.0
6	7.8	8.0
7	5.7	6.0
8	6.0	8.0
9	9.0	9.0
1	7.4	8.0
2	5.4	6.0
3	6.8	8.0
4	7.1	8.0
5	5.5	6.0
6	6.1	8.0
7	8.5	9.0
8	7.6	8.0
9	5.5	6.0
1	9.0	9.0
2	5.6	6.0
3	6.7	8.0
4	8.5	9.0
5	6.7	9.0
6	5.2	6.0
7	4.8	6.0
8	9.0	9.0
9	6.7	9.0

REFERENCES

[1] W. C. Smith and J. Holloway, "School testing culture and teacher satisfaction," *Educational Assessment, Evaluation and Accountability*, vol. 32, no. 4, pp. 461–479, 2020.

[2] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[3] E. D. Liddy, "Natural language processing," 2001.

[4] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[5] J. Wilson and R. D. Roscoe, "Automated writing evaluation and feedback: Multiple metrics of efficacy," *Journal of Educational Computing Research*, vol. 58, no. 1, pp. 87–125, 2020.

[6] H. Albazar, "A new automated forms generation algorithm for online assessment," *Journal of Information & Knowledge Management*, vol. 19, no. 01, p. 2040008, 2020.

[7] T. M. Tashu, J. P. Esclamado, and T. Horvath, "Intelligent on-line exam management and evaluation system," in *Intelligent Tutoring Systems: 15th International Conference, ITS 2019, Kingston, Jamaica, June 3–7, 2019, Proceedings 15*. Springer, 2019, pp. 105–111.

[8] L. Parra G and X. Calero S, "Automated writing evaluation tools in the improvement of the writing skill." *International Journal of Instruction*, vol. 12, no. 2, pp. 209–226, 2019.

[9] M. Warschauer and P. Ware, "Automated writing evaluation: Defining the classroom research agenda," *Language teaching research*, vol. 10, no. 2, pp. 157–180, 2006.

[10] R. D. Roscoe, M. E. Jacovina, L. K. Allen, A. C. Johnson, and D. S. McNamara, "Toward revision-sensitive feedback in automated writing evaluation." in *EDM*, 2016, pp. 628–629.