An Automated Approach to Subjective Answer Evaluation Using ML and NLP

Jyoti Metan
Associate Professor
Dept of ISE
Atria Institute of Technology
Bengaluru, India
Jyoti.m@atria.edu

Dhanya A N
Information Science and Engineering
Atria Institute of Technology
Bengaluru, India
Dhanyaan02@gmail.com

Dinesh Kumar
Information Science and Engineering
Atria Institute of Technology
Bengaluru, India
Dkjan536@gmail.com

Hemant Kumar
Information Science and Engineering
Atria Institute of Technology
Bengaluru, India
Hemantkrsinha01@gmail.com

Abstract – There are many approaches to subjective answer evaluation and in this project, one of these effective approaches is used and put into best use. In simple words machine evaluates and grades the answer that is provided. In this proposed system cosine similarity algorithm is used to get the amount of similarity between the answers and grade them accordingly. When the error percentage is observed after comparing the model with human valuation nearly there is 87% accuracy. Which means it has pretty low error percentage. There are many algorithms that is used for subjective answer evaluation but proposed algorithm has least time complexity and also proposed system has more accuracy than the existing systems. The main aim is to bring the subjective answer evaluator into use as much as possible and work on further improvement.

Keywords- Subjective answer evaluation, machine learning, natural language processing.

I. INTRODUCTION

Subjective exams are considered as scary and complex by both students and teachers because of the context. In case of these answers, it is necessary to check every word in answer for scoring and also factors like metal health of the evaluator plays an important role in their score. The answers, naturally, are not bound to any restriction and students are free to write them. Subjective answer evaluation often involves human graders or evaluators who evaluate the response's quality according to some specific standards like accuracy, completeness, organisation, coherence, and use of evidence or examples. These are considered to have some similarity in grading and reduce partiality, they may keep particular things as base while evaluating.

Traditional essay grading, while crucial, can be subjective and time-consuming. Subjective Answer Evaluation offers a solution by utilizing machine learning and natural language processing to evaluate more efficiently.

One challenge is the beautiful messiness of human language. Students might express the same idea in many other ways, which can trip up the system's understanding. Another hurdle is the idea of grading itself tricky. Teachers use their own ideas, not just right or wrong answers when they are grading. Different teachers might have slightly different This paper contributes by enabling insights on how NLP can be used to create subjective answer evaluator. Through this paper it is just try to bring an overview to reader about what all research is already done and what all they can improvise.

The organization of the paper is as follows: Section II presents the background and existing system and the literature review. Section III provides the proposed methodology. Section IV presents the implementation and results. And section V concludes the paper.

II. BACKGROUND AND EXISTING SYSTEMS

When other systems are observed, there are few features and drawbacks that were necessary to be mentioned for further improvisation. Imagine a world where grading essays doesn't take forever and isn't biased by a teacher's mood. That's the promise of Automated Essay Scoring (AES) systems, the cool new tech changing how writing is evaluated. This system uses artificial intelligence, like the kind in those brainy robots, to understand student writing. Think of ASE (Automated Scoring Engine) as a super-powered grammar checker on steroids. This helps ensure all essays are graded fairly, no matter who reads them first. Another one is the Pearson Education's IntelliMetric takes a different approach that is think there is a list of things that are considered to conclude that an essay is best, IntelliMetric uses this list like checklist with scores to evaluate writing.

B. Literature Review

In the research work through the drawbacks or limitations, looking at those systems author, Li et al. [1] present a system where a computer model generates questions for human evaluation. Three experts reviewed this system concluded they were satisfied with the questions generated but they didn't like the grading quality. Which made them believe that the human monitoring was necessary for evaluation purpose as generating questions are done easily but they didn't like the gradings given.

In this paper using machine learning in exams proposed by Lu et al. [2] in this system they did their best to achieve the performance of system. While they tried to get the best performance possible, they had to face issues like quality of the results generated were not that similar to expected. This system also was lacking the quality of question generated. The inability in training of BERT layer all together was affecting the system. In this system they used Euclidean distance measure and scoring. It may seem easy but there will be some or other minor problems in quality or evaluation while creating these models.

The next survey examines advancements in Open-domain Question Answering (Open QA) powered by recent breakthroughs in neural language models. Lee et al. [3] propose two approaches: YONO, a single powerful model that retrieves information from external sources, and a method using shared representations for finding the best answer. These approaches performs better than prior models, but limitations like overfitting and knowledge expansion challenges remain.

In the future, the author Hwang et al. [4] hope to implement their model in different industrial fields, including interactive chatbots, robotic process automation (RPA), and Internet of Things (IoT) services. In this paper author Kocoń et al. [5] shows the analytical skills of ChatGPT, an AI language model. This survey gives us ideas on how well ChatGPT tackles Natural Language Processing (NLP) tasks, which involves understanding language. These tasks fall into two categories that is meaning and influence. "Meaning" tasks test if ChatGPT understands the literal meaning of words in a sentence. "Influence" tasks are trickier, asking if ChatGPT can predict how someone might interpret a message and how it might affect them.

Imagine driving down the highway, feeling perfectly at ease in your car. That's the dream for automakers with their fancy "intelligent cockpits"! This new study helps us understand how comfy these high-tech cockpits truly are. The research Yang et at. [6] brainstorms a cool system using a "cloud model" to consider all the things that make a ride pleasant think noise, light, and how toasty you are. It's like a comfort meter that takes everything into account! While this system shows promise, it's still learning.

As the result of the brain's understanding of the spatiotemporal state that is condition of something at specific location and time of objective things, and information is the part of a message that the receiver does not know in advance and can decode and understand its meaning. The authors Shi et al. [7] gives a model with more concerning on semantic decomposition and composition. That is semantic decomposition meaning breaking down the whole sentence into small parts and understanding their meaning, and semantic composition means combining all those smaller parts.

Grading subjective papers manually is challenging task because of the difficulty in understanding and accepting the data This paper Bashir et al. [8] introduces an approach using machine learning and natural language processing techniques, along with tools like Wordnet, Word2vec, word mover's distance, multinomial naive bayes, and TF-IDF. Through which they achieved nearly 88% accuracy without the multinomial naive bayes model. The accuracy is further improved by 1.3% with the addition of multinomial naive bayes.

Careful evaluations are needed to assess whether this expectation has been fulfilled. In this work of Andreas et al. [9] evaluates whether explanations can improve human decision-making in practical scenarios of machine learning model development. In a mixed-methods user study involving image data to evaluate saliency maps generated by SmoothGrad, GradCAM, and an oracle explanation on two tasks: model selection and counterfactual simulation. There was no evidence of significant improvement on these tasks when users were provided with any of the saliency maps, even the synthetic oracle explanation designed to be simple to understand and highly indicative of the answer.

On this research by Bech et al. [10] author talks about the people from different background may disagree with some tasks like for things like understanding language and judging things for example hate speech etc. On this paper, Mengting et al, [11] interestingly the thought process of one evaluator matched to that of the system that is he would have graded similarly meanwhile others disagreed with the evaluators gradings. So, it cannot be concluded that the system was completely wrong.

The limitations of ChatGPT, the language model technology used by this chatbot, Oral et al, [12] include the fact that prompts may not always be strict and precise enough, post-processing may be required due to less prompt precision, the system sometimes evaluates press reports and quotes without considering the broader context, there are some disapproved words, and the sequence of prompts provide limited control over them Khan et al, [13] the model's ability to generate appropriate questions will depend on its ability to understand the context of the conversation.

Likewise, while it is trained with large amount of dataset over a period of time changing it and making it better improves the system performance. Afzal et al, [14] Here researchers want to test it out in more situations, like bumpy roads, and include things like the car's smell and air quality. Any system can perform better after a certain period of bringing them to use and improvising it periodically. Sirts et al, [15] The survey on this also says some people using this technique find a lot of confusions but overall, this can be improvised and increase peoples understanding.

III. PROPOSED METHODOLOGY

1) System Architecture

In system design here is the system's overall structure and organisation. In Fig. 1 the system is using SQL as database. Next, there is evaluator. The contents of the evaluator can be divided into three parts they are Question Specific Evaluation, Grammar Checking and Keyword Evaluation. All these three are linked to Database that is question has to be accessed and the model answers are in database, then there is grammar checking for this the data is accessed and made evaluation and the results are sent back. Then the keyword evaluation here the model answers keywords are considered that is stored in database and the whole evaluation process takes place, also it checks weather the answer has any keywords or not. The predictor and the result updating part that is basically running side by all those processes. Here predictor analyse all the data that is collected after the entire evaluation part. These stored data is predicted to give a particular score and the result is updated.

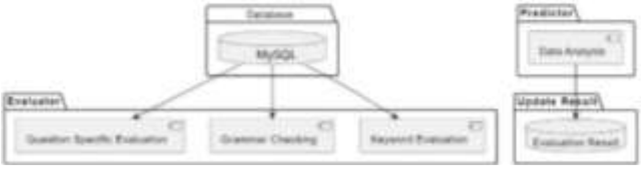

Fig. 1. Architectural Diagram

2) Data Flow

Understanding the flow of data and logic within the system is paramount for ensuring its functionality and performance. Data flow diagrams provide a visual representation of the flow of data within the system, illustrating how information moves between processes, data stores, and external entities. Here in Fig. 2 there is data flow diagram in which it can be observed where the data is flown.

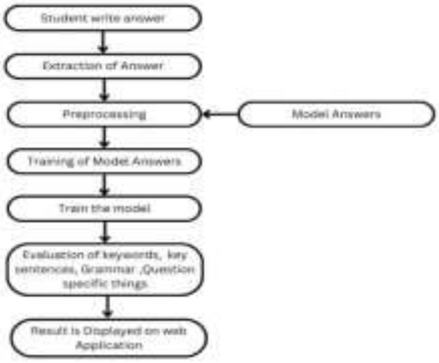
Student write answer
Extraction of Answer
Preprocessing
Model Answers
Training of Model Answers
Train the model
Evaluation of keywords, key sentences, Grammar ,Question specific things
Result is Displayed on web Application

Fig. 2 Data Flow Diagram

3) Interactions Between System Components

Sequence diagrams depict the interactions between system components over time, illustrating the sequence of messages exchanged during the execution of a particular scenario.

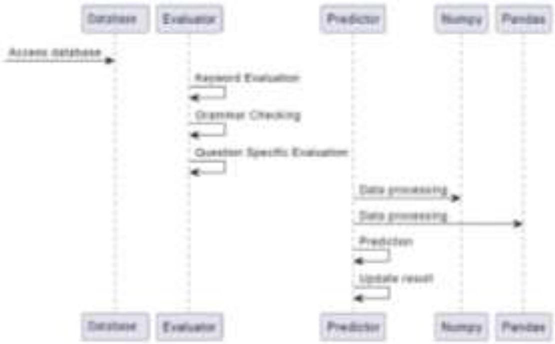

Fig. 3 Sequence Diagram

Figure 3 has a sequence diagram which has the whole process. That is accessing the database and then there is evaluator that evaluates in three parts that is keyword evaluation, grammar checking, question specific evaluation. After this process of evaluation there must be prediction of marks and uploading. This data is visible to us as result.

IV. IMPLEMENTATION AND RESULTS

A large and varied dataset of subjective responses must be gathered, feature engineering techniques must be developed and evaluated, various supervised learning algorithms must be developed and evaluated, a user-friendly interface must be created, the system's performance must be optimised, it must be deployed on a server or cloud-based platform, its functionality must be tested and validated, and it must be regularly updated.

The proposed system uses cosine similarity algorithm. Let A = ($a_{1,}$ $a_{2\ldots}$.a_n) and B = ($b_{1,}$ b_2,…b_n) be two vectors in an n dimensional space the cosine similarity between a and be can given as.,

$$\text{Cosine\_similarity (A, B)}= \frac{\sum_{i=1}^{n} a_i, b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \qquad (1)$$

Through this equation (1) how much two things match in a particular direction using dot products can be obtained. Then the normalization is done after this process if the result is “1” it means they are very similar, if the result is “0” it means they are different and if it is “-1” it means the two things are completely opposite to each other.

Algorithm:

```
function cosine_similarity(R, S):
  # Preprocessing (assume stopwords already removed)
  vocab = unique_words(R, S)

  # Initialize vectors
  VR = [0] * len(vocab)
  VS = [0] * len(vocab)

  for word in vocab:
    VR[vocab.index(word)] = count(word, R)
    VS[vocab.index(word)] = count(word, S)

  # Similarity Calculation
  dot_product = 0
  R_mag = 0
  S_mag = 0
  for i in range(len(vocab)):
    dot_product += VR[i] * VS[i]
    R_mag += VR[i] ** 2
    S_mag += VS[i] ** 2

  if R_mag == 0 or S_mag == 0:
    return 0

  similarity = dot_product / (math.sqrt(R_mag) *
math.sqrt(S_mag))

 return similarity
```

When the time and space complexity of different similarity algorithm is measured, following conclusion can be drawn that is plotted in Fig. 4 and Fig. 5 that is cosine similarity algorithm has constant time and space complexity. Which means it runs slightly fast and does not need lot of memory. There are several algorithms, it cannot be concluded that one is better than other because it depends on where and how these algorithms are used. Like for speed cosine similarity is best and for memory there is DTW algorithm.

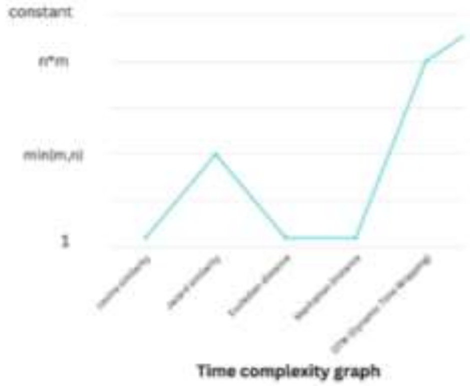

Fig. 4 Time complexity graph

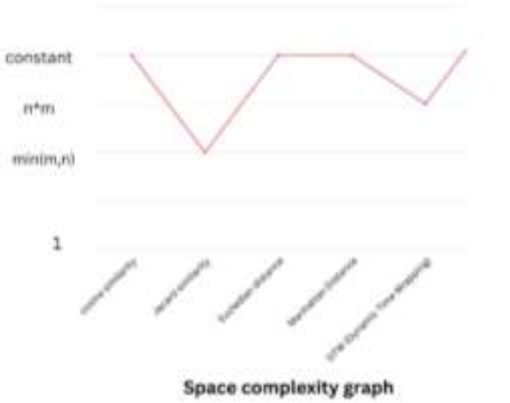

Fig. 5. Space complexity graph

From the system the following results can be obtained. Here when the length of the answer is considered, subjectivity level and score obtained. Here subjectivity level means to what extent the human interpretation and personal opinions influence evaluator.

TABLE. I. Results obtained by the system

Student ID	Question Number	Answer Length (words)	Subjectivity level (1 to 10)	Score (out of 10)
001	1	150	3	7.5
001	2	200	5	8.2
002	1	180	4	7.8
002	2	220	6	8.5
003	1	160	3	7.2
003	2	190	5	8.0

So, in this Table. I with increase in length of answer there is increase in subjectivity level that is because shorter the answer it is straighter forward. Here after using same questions but different answers and the difference is seen in the score for each answer. Depending on the capacity and volume of the training data used to build the model, the outcomes of this model may be precise and reliable. To predict the grade or score of a new subjective answer, machine learning models can be trained to look at factors such linguistic complexity, coherence, and relevance to the topic.

In the Fig. 6 that shows the errors in scores evaluated using the cosine similarity approach without any model suggestion. The results show an accuracy of 87%, these are for some of the values that is obtained but the error depends on amount and capacity of training data, with much more better training data the model can perform better than most other models.

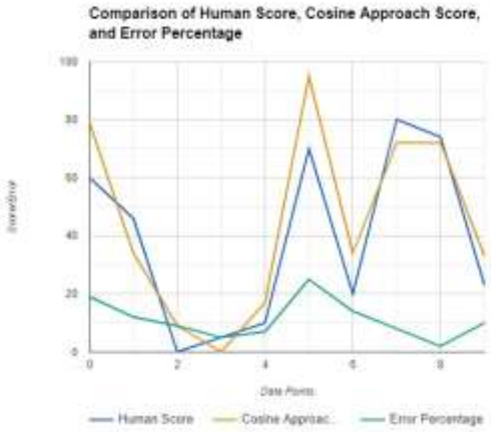

Fig. 6. Comparison of cosine and human approach

This data is represented below in Table II. Shows us that the error percentage is not that much varied compared to human valuation from the data.

TABLE. II. Score prediction using cosine similarity model

Human Score	Cosine approach score	Error percentage
60	79	19
46	34	12
0	9	9
5	0	5
10	17	7
70	95	25
20	34	14
80	72	8
74	72	2
23	33	10

When the accuracy of the system is compared through the results, there are some interesting figures to look at. As shown in the Fig. 7 proposed system achieves an accuracy of 87% outperforming IntelliMetric and AES and Qmark which have accuracy 80% and 75% and 71 respectively.

The time saved compared to traditional method and using evaluator even of low efficiency is much more beneficial than traditional method. Either it is low efficiency one or high it is better to use any model instead of following traditional method to decrease the burden.

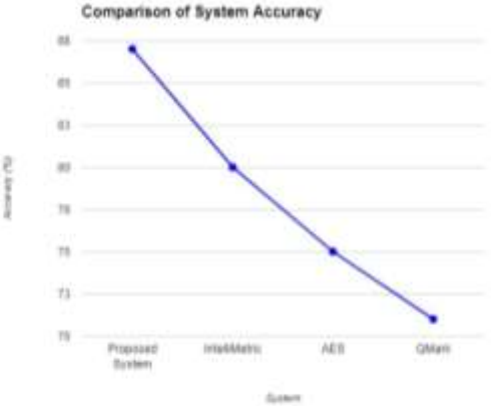

Fig. 7. Accuracy Graph of SAE

V. CONCLUSION

Traditionally, this domain is dependent more on human judgment, that is any answer is being evaluated based on the persons view towards other persons perspective. This is a small try to make this process simple for people and possibly save some time. This is just a step towards making system more useable. There can be a lot of improvement that can be made so that it is perfect for use.

With current knowledge it is an attempt to improvise as much as possible. Using python libraries and tools like scikit-learn, the project constructs complicated models for evaluation. The project is dependent mainly on two key factors that is the data's quality that is used for training and the careful selection of features for prediction. Machine learning effectively assists in the domain of subjective evaluation, augmenting human expertise and streamlining the process.

Scope for Future Enhancement:

The model can be further trained to domain-specific models to learn the terminology and subtleties for a given subject or industry. By utilising prior knowledge and transfer learning abilities, fine-tuning pre-trained language models may enhance the system's performance. In this project the process is automated and made easier. For example, in future, an entire test platform can be implemented which notifies student about the test and they have to take test from the system and get it evaluated. Even it can be made to document the entire things and sent to the faculties the scores of all the students. Also, there is necessity to improve the quality of the system so it works more efficiently. Finding the ideal balance between model complexity and making it simple to use is very tricky because a model that is too basic could miss important details while a model that is too complicated might overfit the data and struggle to generalise to new data. The system's intended use case must be reflected in the assessment metrics selection, which is equally crucial.

REFERENCES

[1] Minghuan Li, Guihua Wen, Jiahui Zhong, Pei Yang, "Personalized Intelligent Syndrome Differentiation Guided By TCM Consultation Philosophy", Journal of Healthcare Engineering, vol. 2022, Article ID 6553017, 11 pages, 2022. https://doi.org/10.1155/2022/6553017

[2] Lu, Yining & Qiu, Jingxi & Gupta, Gaurav. (2022). ProtSi: Prototypical Siamese Network with Data Augmentation for Few-Shot Subjective Answer Evaluation. https://doi.org/10.48550/arXiv.2211.09855

[3] Lee, Haejun, Akhil Kedia, Jongwon Lee, Ashwin Paranjape, Christopher Manning and Kyoung-Gu Woo."You Only Need One Model for Open-domain Question Answering." *Conference on Empirical Methods in Natural Language Processing* (2021) https://doi.org/10.48550/arXiv.2112.07381

[4] Hwang, Myeong-Ha, Jikang Shin, Hojin Seo, Jeong-Seon Im, Hee Cho, and Chun-Kwon Lee. 2023. "Ensemble-NQG-T5: Ensemble Neural Question Generation Model Based on Text-to-Text Transfer Transformer" *Applied Sciences* 13, no. 2: 903. https://doi.org/10.3390/app13020903.

[5] Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., & Kazienko, P. (2023). Chatgpt: jack of all trades, master of none. Information Fusion, 99, 101861. https://doi.org/10.1016/j.inffus.2023.101861

[6] Yang, Jianjun; Wan, Qilin; Han, Jiahao; Xing, Shanshan (2023). The judgment matrix data of light environment.. PLOS ONE. Journal contribution. https://doi.org/10.1371/journal.pone.0282602.s002

[7] Shi, Guangming & Gao, Dahua & Yang, Minxi & Xiao, Yong & Xie, Xuemei. (2023). Mathematical Characterization of Signal Semantics and Rethinking of the Mathematical Theory of Information. https://doi.org/10.48550/arXiv.2303.14701

[8] Bashir, Muhammad & Arshad, Hamza & Javed, Abdul Rehman & Kryvinska, N. & S. Band, Shahab. (2021). Subjective Answers Evaluation Using Machine Learning and Natural Language Processing. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3130902. https://doi.org/10.1109/access.2021.3130902

[9] Zhou, Jianlong, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. "Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics" *Electronics* 10, no. 5: 593. https://doi.org/10.3390/electronics10050593

[10] Tilman Beck; Hendrik Schuff; Anne Lauscher; Iryna Gurevych; "How (Not) to Use Sociodemographic Information for Subjective NLP Tasks", ARXIV-CS.CL, 2023. https://doi.org/10.48550/arXiv.2309.07034

[11] Han, Mengting & Zhang, Xuan & Yuan, Xin & Jiang, Jiahao & Yun, Wei & Gao, Chen. (2020). A survey on the techniques, applications, and performance of short text semantic similarity. Concurrency and Computation: Practice and Experience. 33. 10.1002/cpe.5971. **https://doi.org/10.1002/cpe.5971**

[12] B. Oral, E. Emekligil, S. Arslan, and G. Eryigit, ''Information extraction from text intensive and visually rich banking documents,'' Inf. Process. Manage., vol. 57, no. 6, Nov. 2020, Art. no. 102361.

[13] Khan, Hanif & Asghar, Muhammad & Zubair, Mohd & Srivastava, Gautam & Reddy, Praveen & Gadekallu, Thippa. (2021). Fake Review Classification Using Supervised Machine Learning. 10.1007/978-3-030-68799-1_19. https://doi.org/10.1007/978-3-030-68799-1_19

[14] Sara Afzal, Muhammad Asim, Abdul Rehman Javed, Mirza Omer Beg, and Thar Baker. 2021. URLdeepDetect: A Deep Learning Approach for Detecting Malicious URLs Using Semantic Vector Models. J. Netw. Syst. Manage. 29, 3 (Jul 2021). https://doi.org/10.1007/s10922-021-09587-8

[15] K. Sirts and K. Peekman, ''Evaluating sentence segmentation and word Tokenization systems on Estonian web texts,'' in Proc. 9th Int. Conf. Baltic (HLT) 328, U. Andrius, V. Jurgita, K. Jolantai, and K. Danguole, Eds. Kaunas, Lithuania: IOS Press, Sep. 2020, pp. 174–181. https://doi.org/10.48550/arXiv.2011.07868