STM JOURNALS

International Journal of
Electrical and Communication
Engineering Technology

Volume 1, Issue 1, 2023
January–June
DOI (Journal): 10.37591/IJECET

https://journals.stmjournals.com/ijecet

Review IJECET

Answer Checking System Using Artificial Intelligence and Machine Learning: A Review

A.K. Sreeja[1], Bhavana V.[2], Bhoomika Manjunath Vaidya[2],
H. Prakhya[2,*], Nettem Lasya[2]

Abstract

Automatic answer checker software that evaluates and scores written responses in a manner comparable to that of a human. This software programme is designed to evaluate subjective responses provided in an online test and to award points to the user after the answer has been validated. For the sake of the system, you must save the initial response. The administrator is given access to this resource. The system's admin can add questions and the corresponding subjective responses. In notepad files, these solutions are kept. A user is given questions and a place to post his or her responses when taking the test. The system checks the user's uploaded answers to the original answers stored in the database and then assigns points accordingly. You do not have to use the same term in both answers exactly.

Keywords: Subjective answer verifier, answer verifier, similarity check, artificial intelligence, K-nearest neighbours

INTRODUCTION

Educational institutions hold a variety of exams each year, including competitive, institutional and non-institutional exams. Online tests and exams are becoming common these days to ease the workload associated with exam evaluation. The inquiries on the internet-based assessments may be multiple-choice or objective. However, the tests only have multiple-choice or objective questions. Due to the intricacy and effectiveness of the review procedure, however, subjectively based questions and answers are not included. In the current modern period, it is vital to have an automatic answer checker application that verifies written responses and assigns weights in a manner comparable to that of a human. Therefore, it may be more beneficial to assign marks to the user after validating the answers for online exams using software tools designed to assess subjective responses. Although automatic grading is not a new method, it is now crucial to adapt it to the most recent technology. Partially or fully automated grading systems utilising computational approaches have been developed and have grown into a significant area of research as technology has quickly gotten more powerful on scoring exams and essays, notably from the 1990s onward. Tools that may be used to automatically assess these responses are especially needed, given the demand for scoring natural language responses.

***Author for Correspondence**
H. Prakhya
E-mail: prakhyahubballi@gmail.com

[1]Assistant Professor, Department of Computer Science, BNM Institute of Technology, Bengaluru, Karnataka, India
[2]Student, Department of Computer Science, BNM Institute of Technology, Bengaluru, Karnataka, India

Received Date: July 29, 2023
Accepted Date: October 03, 2023
Published Date: October 13, 2023

Citation: A.K. Sreeja, Bhavana V., Bhoomika Manjunath Vaidya, H. Prakhya, Nettem Lasya. Answer Checking System Using Artificial Intelligence and Machine Learning: A Review. International Journal of Electrical and Communication Engineering Technology. 2023;1(1): 8–13p.

MOTIVATION

Fast processing, reduced labour requirements, independence from shifts in the psychology of

human raters, and simplicity of extraction and record-keeping are some of the factors driving automation of descriptive response rating. Additionally, it guarantees consistency in rating regardless of human assessors' mood swings or shifts in perspective.

The demand for virtual settings as the automated evaluation studies of essays that have previously shown promising outcomes are emerging together with the rise of remote education. The transfer of essay grading systems to short replies has not shown results with the same level of accuracy, therefore simulating the judgements of a human grader in short answers is still difficult. The need for virtual settings increases as distant education grows, as seen by the automated essay evaluation projects that have shown encouraging outcomes thus far. Though essay grading approaches can be used to short response questions, the accuracy of the results has not increased, making it difficult to replicate a human grader's conclusions when dealing with short replies.

LITERATURE SURVEY

The most recent technologies must be adapted for automatic grading. It has become crucial, especially since most classes have moved online (MOOCs). On the Quora Question Pairs and UNT Computer Science Short Answer datasets, eight experiments were run. Without fine-tuning the pre-trained models by the USE, the best accuracy in the first four experiments was 77.95%. Methodology Equations Similarity Checker algorithm is utilised here. The paper's ability to evaluate equations and multiple-choice questions was an advantage. The written-answer scripts in the study by Hossam and Saafa were not assessed, which was a drawback [1].

In the article presented by Suzen *et al.*, the authors emphasised the idea of automatically assessing short answer questions, which are prevalent in the UK's GCSE system, and giving students insightful feedback on their responses [2]. They provide experimental findings using data from the University of North Texas' basic computer science course. To determine how closely the student answers resemble the model answer, Suzen *et al*. used common data mining techniques on the corpus of student responses [2]. Based on the quantity of widely used words. The relationship between these parallels and the scores given by scorers is then assessed. They considered a method that clusters student responses. Each cluster would receive the same grade, and each cluster's answers would receive the same feedback. In this way, they show that clusters represent groups of students who receive the same or comparable grades. To demonstrate that groups were created based on how frequently and which words from the model answer have been used, the words in each cluster were compared. The primary innovation in the study is the model they developed to forecast marks based on the similarity between student responses and the model response. We contend that rather than replacing human scoring, computational technologies should be employed to increase its accuracy. The system must be calibrated by humans, and they must also handle difficult scenarios. Computational approaches can reveal which student responses will be considered difficult and thus be an area where human judgement is needed.

The goal of the work by Hassan *et al*. was to promote the growth of research in automated evaluation of brief discursive responses [3]. Text-to-text similarity, knowledge-based similarity that depends on a synonym dictionary, and corpus-based similarity that depends on a related corpus were the three main methodologies discussed in the linked publications. The study used an n-gram depending on resemblance and classification process to analyse three sets of Portuguese language responses to questions, two of which (Biology and Geography) were obtained from a higher education admissions process and the third (Philosophy) from a virtual learning environment. The applied method included a five-stage pipeline design with the following stages: corpus selection, preprocessing, variable formation, classification, and accuracy validation. Numerous similarity metrics and distances resulting from the combining of unigrams and bigrams were investigated in these three corpora. Two techniques were applied throughout the classification phase: multiple linear regression and K-Nearest Neighbours (KNN). Concurrently, certain research questions underwent revisions that produced insightful results. The results show advantages in terms of a more

straightforward procedure combined with high accuracy when compared to those from previous studies conducted by other methodologies.

Thus, the goal of the work by Sirotheau *et al*. is to promote further research in automated discursive answer evaluation [4]. Three primary approaches were provided in the linked works: corpus-based similarity, which relies on a related corpus, text-to-text similarity, and knowledge-based similarity, which depends on a synonym dictionary. Three sets of Portuguese language answers were used in their study: two sets (Biology and Geography) came from the admissions process for higher education, and the third set (Philosophy) came from a virtual learning environment. The researchers used an n-gram based similarity and a categorization process. The pipeline architecture used in the used method consisted of five stages: preprocessing, variable generation, classification, accuracy validation, and corpus selection. Numerous distances and similarity metrics arising from the combination of unigrams and bigrams were investigated in these three datasets. Two techniques were applied in the classification phase: K-Nearest Neighbours (KNN) and multiple linear regression. Three sets of Portuguese language answers were used in their study: two sets (Biology and Geography) came from the admissions process for higher education, and the third set (Philosophy) came from a virtual learning environment [5, 6].

The researchers used an n-gram based similarity and a categorization process. The pipeline architecture used in the used method consisted of five stages: preprocessing, variable generation, classification, accuracy validation, and corpus selection. Numerous distances and similarity metrics arising from the combination of unigrams and bigrams were investigated in these three datasets. Two techniques were applied in the classification phase: K-Nearest Neighbours (KNN) and multiple linear regression. Concurrently, several research inquiries were reformulated, yielding significant discoveries. Regarding the system efficiency, the results for the biology corpus showed an accuracy of 84.01 system vs. human as opposed to 93.85 human vs. human; for the geography corpus, the accuracy was 86.29 system vs. human as opposed to 84.93 human vs. human; and for the Philosophy corpus, the results showed an accuracy of 81.59 system vs. human respectively. These results show advantages in terms of a simpler procedure combined with good accuracy when compared with those from previous trials conducted by other methodologies.

The Technique

The cosine similarity metric is used to compare two comparable non-zero vectors, or the inner product of two vectors in space. Any angle in the range of (0) radians is less than 1, and 0° is equal to 1. Because the unit of measurement is the cosine of the angle across the two vectors, any angle that falls in that range is more than 1. Any two vectors in this situation, regardless of their size, will have a cosine similarity of 1, a similarity of 0 when they are oriented at 90°, and a similarity of –1 when they are opposed. The two vectors of the documents are text-matched using the Euclidean dot product in this method. The suggested system considers two input strings, the first of which is the user's response. and the second one is the database-stored model response. Following the vectorization of both strings, each vector is processed to produce a corresponding keyword score.

Assessment of the results of student learning has frequently utilised AES. Few studies, meanwhile, concurrently determine the levels and scores of the community academy's competency tests using Automated Essay Scoring (AES) [7, 8]. Using AES to evaluate essays on the competency certification test is the goal of this work. The AES can forecast exam results and categorise exam takers' competency levels. Back Propagation Neural Networks (BPNN) are utilised in the construction of AES. Because of its simplicity and convenience of model construction, BPNN was chosen.

PROPOSED SYSTEM

A computer-based method designed by Pisat *et al*. will produce objective-based questions that are more appropriate for academic student grading [9].

Architectural Design

We provide user information, user answer information, and the actual answer model as input to the system design as shown in Figure 1. From the actual answer model and user response, the information extraction module extracts the keywords and other question-specific information. The cosine similarity based on the feature-matches, score will be calculated after we apply grammatical checking in the weighing module (Figure 1).

Madhavi *et al*. examine a range of natural language processing methods using well-known datasets, including Microsoft Paraphrase Identification, the SICK dataset, and the STS benchmark [6]. Techniques for optical character recognition can be tested using datasets such as MNIST, EMNIST, and IAM.

Similar methods are employed in automated essay scoring (AES) systems to determine the grade for student responses. Over time, various approaches to computing similarity have surfaced. Only a few of them, nevertheless, have seen widespread application in the AES arena. The results of a 10-year study on similarity techniques used in AES systems were presented in the article by Ramnarain-Seetohul *et al*., together with information on the effectiveness and drawbacks of existing approaches [10]. 34 articles that were published between 2010 and 2020 were included in the final review. Also highlighted were the parameters employed to assess the effectiveness of the AES systems. Three research questions and a search strategy were created as part of the Kitchenham technique, which was used to conduct the review. One such algorithm and a relatively new method being used for numerous classification processes with little datasets is the capsule network [11, 12].

Yogish *et al*. created an intelligent system that can respond precisely to user questions in natural language [13].

Module Description (Figure 1)

Image 2 Text, User Answer extractor unit, Answer verifier unit, and Result set unit are the four modules that make up our system:

1. *Image 2 text:* In this module, the text from user-uploaded images is extracted using an OCR model.

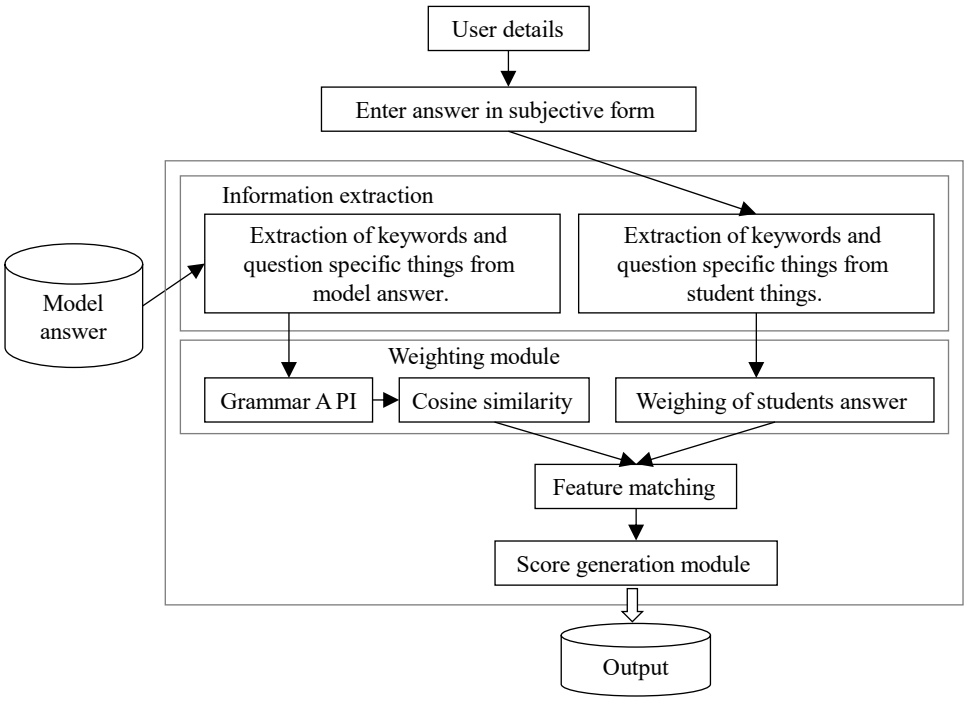

Figure 1. Answer verifier system.

2. *plUser answer extractor:* User Answer Extractor is a unit that retrieves the keywords from the default response together with their synonyms.
3. *Answer verifier unit:* The Cosine Similarity module and Text Gears Grammar API are the two distinct modules that make up this unit. The angle of similarity between two relationships between related strings is described by cosine similarity. Then, the retrieved keywords were cross-checked to the student's response's array. Any product can incorporate language processing technology thanks to Text Gears' grammar API. No matter if it is a sophisticated enterprise system or a mobile application. Depending on the input data, this API will either return a value of 0 or a value of 1. If the grammar is flawless, the API outputs 1, however if there are any errors in the statement, the API outputs 0.
4. *Result set unit:* The three properties "Keywords", "Grammar", and "QST" (Question Specific Terms) make up the Result Set Unit. The "keywords" and "QST" characteristics' values are so defined as:
    i. Keyword values range from 1 to 6, with 1 representing Excellent and 6 representing Very Poor.
    ii. The "grammar" attribute has values that range from 0 for improper to 1 for proper.
    iii. "Class" has a value between 1 and 9, with 1 representing the greatest and 9 the worst. The final output is produced following the results calculation and includes the final score as well as the grammar, keywords, and class values of each response.

CONCLUSION

Due to the subjective answer verification by applying the proper weights, the proposed work is robust. The method is also tested for adding new words, and weightage has no impact on it. Future system developments are also possible with this one. Therefore, any grammar checking can be modified in accordance with the standards. The second module is additionally created and put to the test for the "Cosine Similarity" method. Any other algorithm that can provide a more reliable and effective answer verifier and grading system can be tried to see how well it works.

REFERENCES

1. HM Balaha, MM Saafan. Automatic exam correction framework (AECF) for the MCQs, essays, and equations matching. IEEE Access. Jan 2021; 9: 32368–32389.
2. Suzen, *et al*. Automatic short answer grading and feedback using text mining methods. Procedia Comput Sci. 2020 Feb; 169: 726–743.
3. Hassan et al. Encouraging paragraph embeddings to remember sentence identity improves classification. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. 2019, July 28–August 2. pp. 6331–6338.
4. Sirotheau S, dos Santos Joao CA, Favero Eloi L, de Freitas Simone N. Automated Evaluation of Short Answers Using Text Similarity for the Portuguese Language. J Comput Sci. 2019 Mar; 15(11): 1669–1677.
5. Gusti Putu Asto Buditjahjanto I, Mohammad Idhom, Munoto Munoto, Muchlas Samani. An Automated Essay Scoring Based on Neural Networks to Predict and Classify Competence of Examinees in Community Academy. TEM J. 2022 Jan; 11(4): 1694–1701.
6. Desai Madhavi B, Desai Visarg D, Gupta Rahul S, Mevada Deep D, Mistry Yash S. A Survey on Automatic Subjective Answer Evaluation. Adv Appl Math Sci. 2021 Sep; 20(11): 2749–2765.
7. Merien Mathew, Ankit Chavan, Siddharth Baikar. Online Subjective Answer Checker. Int J Sci Eng Res. 2017; 8(2): 15–17.
8. Sakshi Berad, Prakash Jaybhaye, Sakshi Jawale. AI Answer Verifier. Int Res J Eng Technol. 2019; 06(01): 343–344.
9. Pisat P, Modi D, Rewagad S, Sawant G, Chaturvedi D. Question paper generator and answer verifier. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai. 2017; 1074–1077. doi: 10.1109/ICECDS.2017.8389603.

10. Ramnarain-Seetohul V, Bassoo V, Rosunally Y. Similarity measures in automated essay scoring systems: A ten-year review. Educ Inf Technol. 2022; 27(4): 5573–5604. https://doi.org/10.1007/s10639-021-10838-z.
11. Jeena Jacob I. Performance Evaluation of Caps-Net Based Multitask Learning Architecture for Text Classification. J Artif Intell Capsul Netw. 2020; 2(01): 1–10.
12. Vijayakumar T, Vinothkanna R. Capsule Network on Font Style Classification. J Artif Intell Capsul Netw. 2020; 2(02): 64–76.
13. Yogish Deepa, Manjunath TN, Hegadi Ravindra S. Survey on trends and methods of an intelligent answering system. In 2017 IEEE International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEEECCOT). 2017; 346–353.
14. Ashutosh Shinde, Nishit Nirbhavane C, Sharda Mahajan, Vikas Katkar, Supriya Chaudhary. AI Answer Verifier. Int Journal of Engineering Research in Computer Science and Engineering (IJERCSE). 2018 Mar; 5(3): 143–145.

