Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 07 July-2024, Page No.- 4382-4388

DOI: 10.47191/etj/v9i07.06, I.F. – 8.227

Artificial Intelligence Based Essay Grading System

Adeyanju Ibrahim A.[1*], Oderinde Kehinde Rachael[2*], Adedeji Oluyinka Titilayo[3*], Gbadamosi Omoniyi. Ajoke[4*], Makinde Bukola Oyeladun[5*], Falohun Adeleye Samuel[6*]

[1*]Department of Computer Engineering, Federal University, Oye-Ekiti, Ekiti State, Nigeria

[2*,6*]Department of Computer Engineering, Ladoke Akintola University of Technology, Ogbomoso. Nigeria.

[3*]Department of Information System Science, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

[4*]Olusegun Agagu University of Science and Technology, Okitipupa, Ondo State, Nigeria.

[5*]Department of Computer Science, Osun State College of Technology, Esa-Oke. Nigeria

Corresponding Author(s): ibrahim.adeyanju@fuoye.edu.ng, oderinde.okr@gmail.com, otadedeji@lautech,edu.ng, oa.gbadamosi@oaustech.edu.ng, bukolamakinde22@gmail.com, asfalohun@lautech.edu.ng

ABSTRACT: There are quite a number of challenges faced by teaching staff which if ameliorated can help reduce the quality of time spent on monotonous task associated with students essays grading More so, these days in Nigeria and other related parts of the world institutions deal with a large number of students. These however make grading time consuming and costly, therefore an automated system that can handle the tasks is highly desirable. The developed automated essay grading system in this research is restricted to grading of short and structured essay responses. Java programming language was used for the implementation and MySQL relational database system with WAMPSERVER as the back end. The results showed the user-friendly modules which enable users to navigate through various interfaces easily and work as required. The developed system has an Exact Agreement rate of 0.4 and a Pearson Correlation of 0.93 with human graders.

KEYWORDS: Essay Grading, Natural Language Processing, Feature Extraction, knowledge representation

1. INTRODUCTION

Williams (2001) informed that automated essay grading was proposed more than thirty years ago but only recently were its practical realizations getting popular. Checking the students' answers it is however time-consuming for the teachers which can be diverted to other areas of profitability. The Automatic evaluation of the students' answers should be checked. Assessment items can be broadly classified as selected responses (e.g., multiple-choice, true-false) as in the Computer-Based Test (CBT) or it can be constructed in different ways (e.g., short-answer, essay).(Uto M and Okano M, 2020) Free-text answers belong to the constructed response items which require the student to construct a response in natural language without the benefit of any prompts in the question. One of the benefits of Electronic-Assessment is instantaneous feedback delivery of grades. The automated essay type grading system will thus help to solve a lot of problems associated with grading of large number of students in tertiary and pre-tertiary institutions. (Tashu TM and Horváth T, 2020)

2. LITERATURE REVIEW

Cunningham (1998) asserts that education has gradually become one of the premier public policy issues in the world and that conventional tests can be divided into two main categories. Objective items that require students to choose answers from several choices such as true-false, multiple-choice, and matching exercises. The second is constructed response items for which the student must create and write out response which can be brief, as in the case with restricted response such as short-answer, completion, and fill-in-the-gap. Meanwhile a look at the existing grading techniques, though not limited to, is as elucidated hereafter.

2.1 Manual Grading System

The manual grading system involves the use of human being for grading students' performance. It involves manually marking, scoring and recording students' performance in an examination. However, with the advent of technological improvement and large number of students in various stages of learning, this method of grading is gradually becoming obsolete as it is a monotonous process which is oftentimes prone to errors due to fatigue, mood, inconsistency. The manual grading process is also a time consuming and resources draining process.

Conversely, the essay-type grading which can also be called Automated Essay Grading (AES) is the ability of computer technology to evaluate and score written prose (Shermis and Burstein, 2006). Examples of its implementation strategies utilized the following schemes.

2.1.1 Natural Language Processing (NLP)

NLP is an aspect in computer science which focuses on developing systems which makes computers to interact with

people using everyday language. It can also be referred to as Computational Linguistics that is concerned with the way computational methods can help understand human languages. (Catulay et al. 2021) The aim of the production and comprehension of natural language is communication. Language communication according to Emuoyibafarhe (2009) has seven component steps which can be sub-grouped into two; the communication for the speaker and for the hearer. (Zhu W and Sun Y, 2020)

2.1.2 Genetic Algorithm

Genetic algorithm (Xiaoping and Cao, 2002) proposed to look for the best distinguished parameter by using genetic evolution mechanisms and also to find the survival of the fittest in natural selection. Through this technique, misleading judgments are removed thereby leading to improvement in the accuracy of document classification.

2.1.3 Decision Tree

The decision tree rebuilds the manual categorization of training documents by constructing well-defined true/false-queries in the form of a tree structure. In a decision tree structure, leaves represent the corresponding category of documents and branches represent conjunctions of features that lead to those categories.

2.1.4 Artificial Neural Network

Artificial neural networks or connectionist systems are computing systems that are inspired by biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules

Which according to Ruiz and Srinivasan (1998) are constructed from a large number of elements with an input fan order of magnitudes larger than the computational elements of traditional architectures. These elements referred to as artificial neurons are interconnected into groups using a mathematical model for information processing.

2.2 Related Works

Page *et al.* (1994, 1996 and 2001) developed Project Essay Grader (PEG) that is one of the ancient implementations of an automated essay grading which primarily relies on style analysis of surface linguistic features of a block of texts. It operated on a school of thought that an essay is predominantly graded on the basis of writing quality, taking no account of content.

Hearst (2000) ; Jerrams-Smith *et.al.*(2001) reported the development of Intelligent Essay Assessor (IEA) in the late nineties which depends on the Latent Semantic Analysis (LSA) technique that was initially designed for indexing documents and text retrieval (Deerwester *et.al.*, 1990). And the test conducted by Valenti *et al* (2003) with the use of GMAT essays using the IEA system resulted in percentages for adjacent agreement with human graders between 85%-91%.

Burstein and Kaplan (1995) of the Educational Testing Service developed the Educational Testing Service (ETS I) in early nineties which worked only on a sentence fragment of between 15 and 20 words (Whittington and Hunt, 1999). The technique used lexical-semantic techniques to build a scoring system, based on small data sets. It used a domain-specific, concept-based lexicon and a concept grammar, both built from training data. The training data essays were parsed by Microsoft Natural Language Processing (MsNLP) tool where any suffix was removed by hand, and a list of stop words was also expunged.

Burstein *et.al.*,(1998) authored the development of Electronic Essay Rater (E-Rater) which used the MsNLP tool, the togetherness of statistical and NLP techniques for parsing all sentences in the essay to get linguistic features from the essays meant for grading.

Burstein *et al*. (2001) also reported on Conceptual Rater (C-Rater), also a NLP based prototype aimed at the assessment of short answers related to content-based questions. It adopted many of the some natural language processing tools and techniques developed for E-Rater but the former aimed at scoring a response as being either correct or not correct.

The Paperless School free-text Marking Engine (PS-ME) was however designed as an integrated component of a Web-based Learning Management System (Mason and Grove-Stephenson, 2002) but because of its processing criteria, the PSME does not grade essays in real-time.

Oduntan et.al. (2018) examined a comparative analysis of Euclidean Distance and Cosine Similarity measure in an Automated Essay-Type Grading System where result showed that cosine similarity measure has a higher positive correlation than the Euclidean distance.

Vijaya et al (2022) described and compared different methods based on machine learning, artificial intelligence and natural language processing that can be adopted to evaluate and score essays written by students

Suresh et al (2023) developed an AI-powered system for automated essay grading. The system utilized natural language processing and Graph based techniques to analyze, and grade written essays. It not only checked the syntax, semantics and grammar but also graded according to the similarity of sentences using a Graph based approach. The system will trained on a dataset of labelled essays and was able to accurately grade new essays based on their content and writing quality. The system was able to integrate with existing learning management systems. The system was able to provide a more efficient and accurate essay grading process, so the teachers can provide valuable feedback to students. The system was able to analyze, and grade written essays by using natural language processing and machine learning techniques. The system was trained on a dataset of labelled essays, which was used to teach the system to recognize patterns and characteristics of high-quality

writing. This enabled the system to accurately grade new essays based on their content and writing quality.

Summarily, the essay-grading system is still a work in progress as there is a continuous quest for finding a more excellent way to achieving an automatic means of doing students examination assessment, hence the need for this research.

3. DESIGN METHODOLOGY

In other to design an automated essay-type grading system design using artificial neural networks these five basic steps are highly necessary. These include document vector creation, comparison material creation, and feature extraction, scoring model creation, and scoring assignment. The steps involved in the design are discussed in the paragraphs below.

3.1.1 Document Vector Creation

The development of automated essay-type grading system using artificial neural networks system starts with the creation of the document vector. The document vector here is described as the questions which the examiner intends to ask the students and also the students' responses to these questions. This is to say that two fields must be created in the document vector which are; the question field and the response field.

The question field is created so that only the examiner also known as the administrator has access to this field. The examiner is given some privileges which are; determine number of questions to set, the number of questions each student is expected to provide answers to, assign mark for each questions set and also set time in which the student is expected to answer all questions in an examination.

The response field is created to be accessible to students which the examiner has registered for a particular examination. The registered students are able to see the following; the total time for the examination, the number of questions expected, access to go to the previous question, access to go to the next question and access to submit after the student had finished the examination.

This phase of the work is very important and strict attention must be given it most especially by the examiner. This is because any mistake made at this stage will have a great consequence on the overall accuracy of the system. It might reduce the accuracy and efficiency of the system and hence defeat its objective, hence strict attention and care was put into the design of the interfaces which will capture the information supplied at this stage.

3.1.2 Comparison Material Creation

The comparison material creation has to do with the creation of the marking guide by the examiner. This also is another important and vital aspect of the work. The comparison material created is used to compare with the response field by the student in the document vector that had been created. The examiner is the one given the privilege to access this stage. This is because the examiner is responsible for grading the students and each question and the accurate or precise solution to it is known to the examiner only.

3.1.3 Feature Extraction

The process of pre-processing is to ensure that the border of each language structure is cleared and also to remove as many factors that are language dependent, tokenization, stop words removal, and stemming. Feature extraction is the step of pre-processing which is used to present the text documents into clear word format.

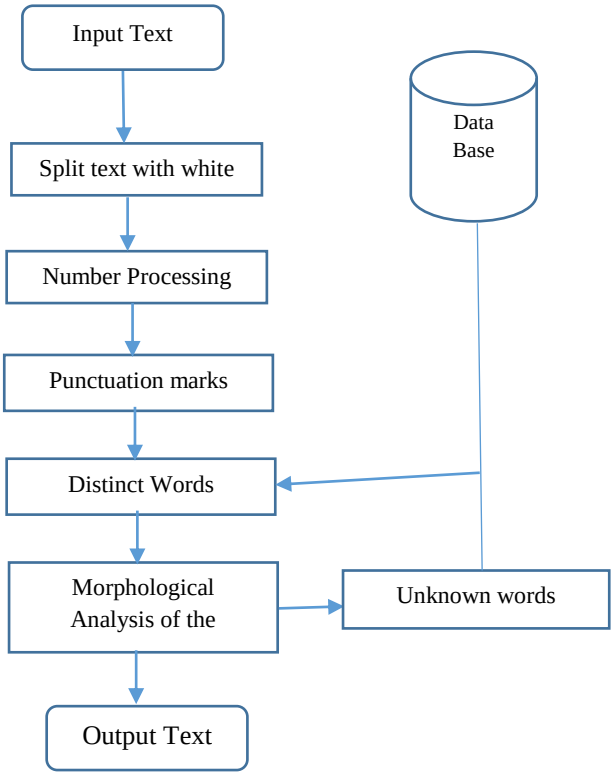

Figure 1: Feature Extraction Stages

3.1.4 Scoring Model Creation

This system used the content scoring models for assigning scores to student responses for each question. In other to compute the final essay score, a linear equation was used. For the final score for each examination to be computed, the sum of the question's mark is calculated and a percentage is calculated to arrive at the final score. This is due to the fact that the system generates questions randomly for students. This is to say that different students answer different questions which may have different marks assigned to them by the examiner. The equation for calculating the final score of student for each examination is described in equation (1)

$$Score = (\sum_{j=1}^{n} i(j) / \sum i) \times 100 \quad (1)$$

where i= mark obtainable for a question, j= question number and n= total number of questions

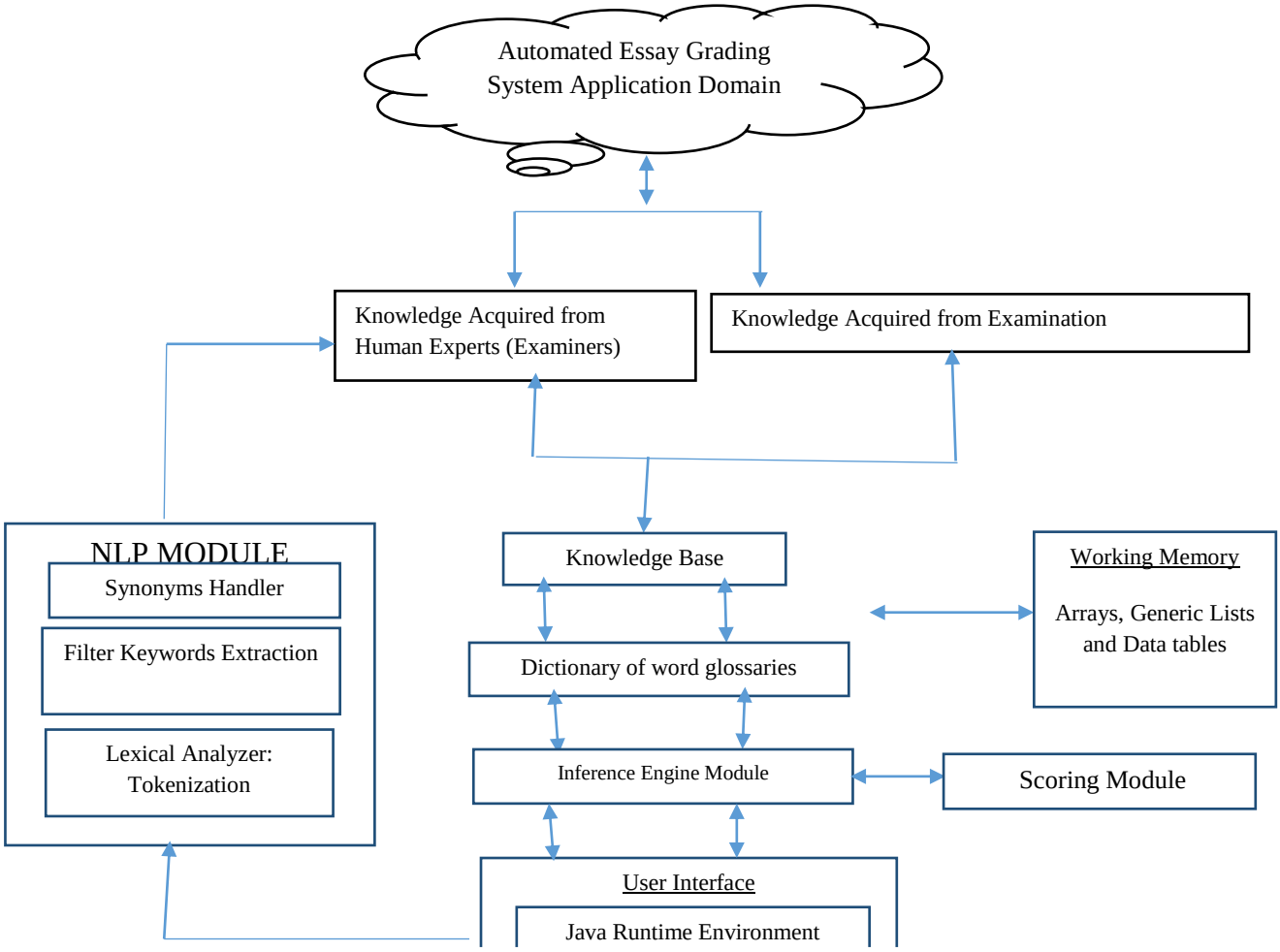

Figure 2: Architecture of the Developed System

3.1.5 **Scoring Assignment**

The scoring assignment phase of the system deals with calculating the total score of an examination in percentage and assigning the score a student has to his/her identity. The student is given the privilege to view the score he/she obtains in an examination while the examiner has the privilege to view the scores obtained by all the students who took the examination and the time in which the examination was taken for each student respectively.

3.2 Design Implementation Using Java and MySQL

The choice of JAVA programming language is because there exists a numerous libraries that works with natural language processing. Natural Language Toolkit (NLTK) was used and text mining for most NLP tasks. Also, integrating other libraries into the JAVA programming language is quite easy and less tedious. WordNet 2.1 was used for both the synonym handler and the dictionary glossaries imported during the implementation stage. My Structured Query Language (MySQL) Relational Database Management System (RDBMS) was used to represent the knowledge base which contains the scoring model, document vectors and comparison materials using WAMP Server as the back end.

3.2.1 System Architecture and Requirement

Figure 3.2 shows the System Architecture of the developed system.

The system was developed using NetBeans Integrated Development Environment (IDE) version 8.1 on a personal computer with the following specifications; RAM size 4GB, Processor speed 2.40GHz, Hard disk 500GB, Windows 8.1 Operating System(64-bit)

3.3 Performance Evaluation of the Automated Essay-Type Grading

Numerous metrics have been proposed, adopted and used in the reviewing of electronic essay assessors. Some notable metrics are: Measure of Exact Agreement, Adjacent Agreement or Reliability, the Pearson Correlation, Spearman or nonparametric Correlation, Mean and Standard Deviations, Kappa Measure and F-Score. It should be noted that exact agreement measures how frequently two or more evaluators assign the same rating (e.g., if both give a rating of "4" they are in agreement), and reliability measures the relative similarity between two or more sets of ratings. Therefore, two evaluators who have little to no agreement could still have high inter-rater reliability However, since the performance values of most AES system studied in this

work is available in Pearson Correlation Coefficient, it was adopted it in measuring the performance of the newly developed automatd essay grading system in comparative respect to the existing Electronic Essay Assessors.

A description of exact agreement measurement for the developed system is reflected in equation 2 where X is the developed systems' array of scores and Y represents the human graders' array of scores. Then,

$$Measure\ of\ Exact\ Agreement = (n(X,Y)) \times 0.1) \quad (2)$$

and n(X,Y) = number of exact rating of X and Y

Standard correlation was also used which measured the teachers' scores or true scores (Y) in relation with the systems' scores (X). It is appropriate when answers are being evaluated with a numerical score. At some point, the result of the average consensus of several teachers is the true score. The Pearson correlation was used for the determination of the accuracy of the developed automated essay grading system.

$$Correlation\ (X,Y) = \frac{Covariance(X,Y)}{StandardDev(X) \times StandardDev(Y)}$$

4. RESULTS AND DISCUSSIONS

4.1 User Interfaces

The figures below are snapshots of the various user interfaces involved in the implementation of the automated essay grading system developed.

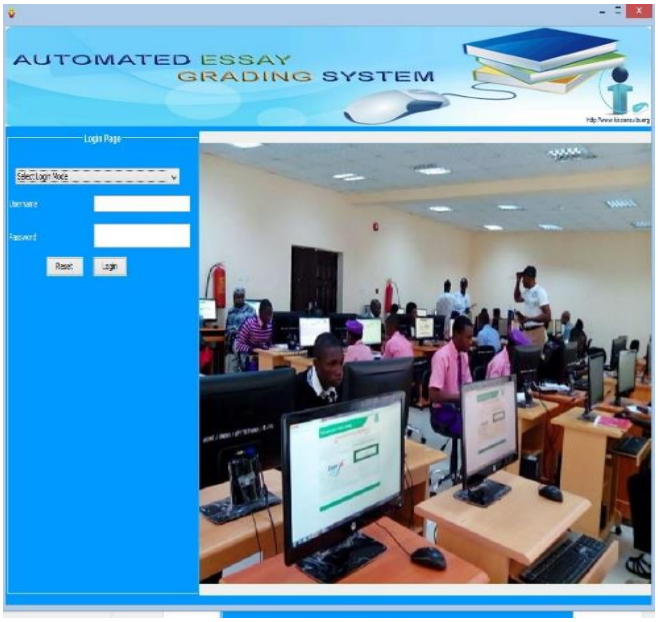

Figure 3: Login Module

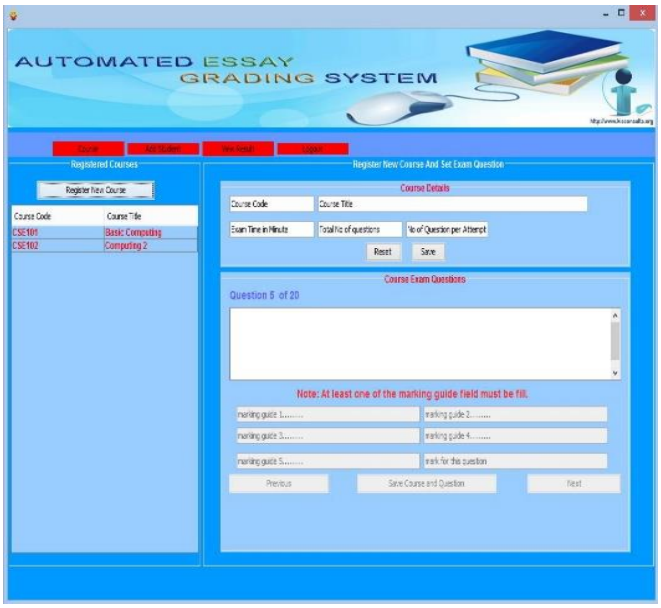

Figure 4: Knowledge Acquisition Module

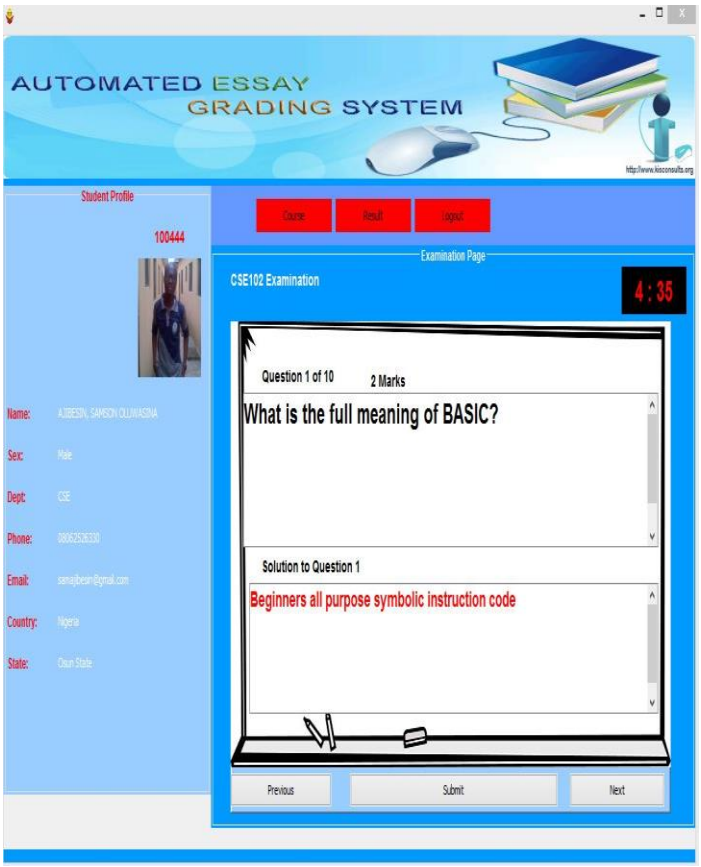

Figure 5: Examination in Progress Module

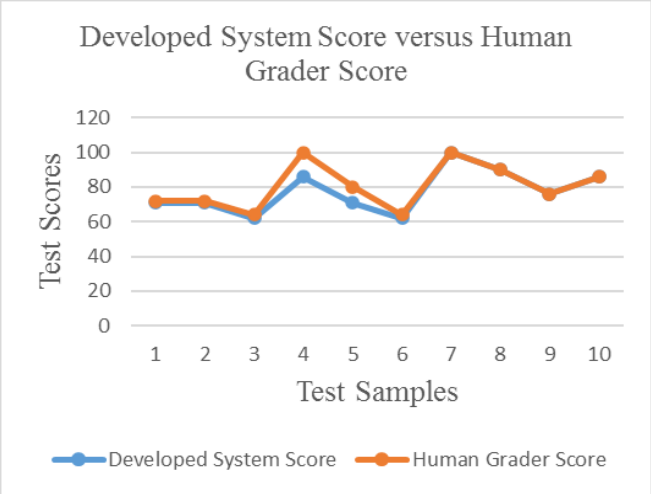

Figure 6: Explanation Module

Figure 3 shows the Login interface of the automated essay grading system developed. The Login Interface allows both the examiner to login with the default username and password while figure 4 depicts the knowledge acquisition page.

The 'Examination in Progress Page' is represented by figure 5 which has the time duration and other necessary instructions. The field for each question is non-editable while the field to provide the solution for each question is editable. Figure 6 is termed the explanation module, which shows the examiner's guide and the students' response to each question.

4.2 Performance Evaluation of the Developed System

The overall metric used for the performance evaluation of the developed automated essay grading system is the accuracy. The feedback (score for each test) provided by the system is compared with feedback provided by human grader to determine the accuracy of the developed system. Table 1 shows the feedback provided by the developed system against the feedback provided by human grader for 10 tests conducted to determine the accuracy of the developed system with each test having ten questions. The tests were carried out making use of ten distinctive students and the results are as shown below. The developed systems' score given below is a percentage score so also is the human grader score which was computed using the equation (1).

Table 1: Comparison of Feedback from the Developed System and Human Grader

TEST NUMBER	DEVELOPED SYSTEM SCORE	HUMAN GRADER SCORE
Test One	71	72
Test Two	71	72
Test Three	62	64
Test Four	86	100
Test Five	71	80
Test Six	62	64
Test Seven	100	100
Test Eight	90	90
Test Nine	76	76
Test Ten	86	86

The table shows that the developed system has an exact agreement rate of 0.4 with the human grader; this was computed using of equation (2). From the Pearson Correlation formula earlier stated, a correlation of 0.93 was obtained by the developed system against the human grader from the metric described in equation (3) and corroborated by the graph in figure 7.

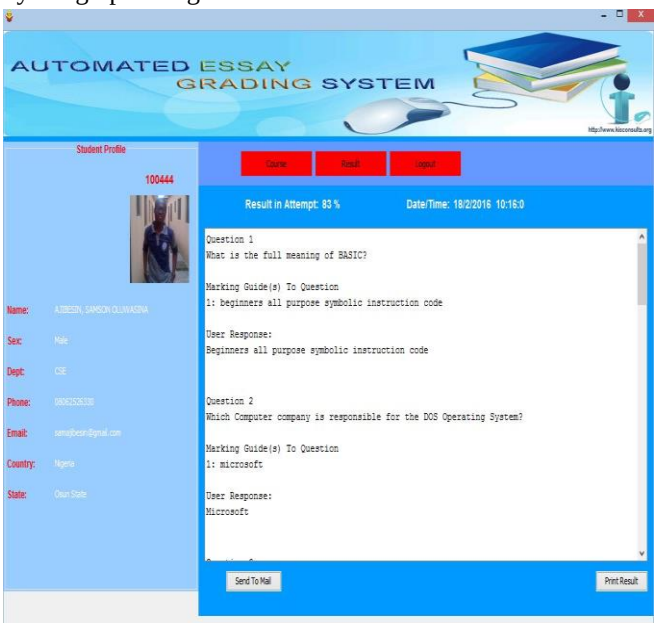

Figure 7: Correlation between the Developed Systems' Score and Human Grader Score using 10 Test Samples

5. CONCLUSION

This research examined the use of an essay grading system with user friendly facilities via Graphical user Interfaces (GUI) for both examiners and the students. The system was trained and tested usin features of computer Natural Language Processes of Artificial Intelligence. A correlation of 0.93 obtained revealed that there was not so much deviation of the auto-grading system from the human examiner. But the developed system can only assess short free-text responses of students and can't handle questions that involve calculations and diagrams. Grading systems that can handle the latter is hereby recommended for further research in this field.

REFERENCES

1. Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In A. Hermann and M.M. Kulski (eds). Expanding Horizons in Teaching and Learning.Proceedings of the 10th Annual Teaching and Learning Forum, Perth: Curtin University of Technology.
2. Cunningham, George K. (1998). Assessment in the classroom: Constructing and interpreting tests. London: Falmer Press. vii + 225 pages.

3. Application to computer in assessment and analysis of writing Mark D Shermis, Jil Burstein and Claudia Leacock (2006).
4. Emuoyibarhe O. Justice (2009), Introduction to Computational Intelligence Paradigms.
5. Xiaoping Cao,Feng Wang & Shu Guo A new convergent approach to Dendritic Macromolecules.Journal Synthetic Communication .An international Journal for rapid Communication of Synthetic Organic Chemistry Volume 32,2002 – Isuse 20
6. Miguel E. Ruiz , Padmini Srinivasan Hierarchical Neural Networks for Text Categorization (1999) . In poceedings of the 22 nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval
7. Hearst, M. (2000). The debate on automated essay grading. IEEE Intelligent Systems, 15(5), 22-37, IEEE CS Press.
8. S Deerwester, St Dunamis,Gw Furnas---Journal of the 1990-wiley-onloine library
9. Salvatore Valenti, Francesca Neri and Alessandro Cucchiarelli DIIGA(2002) - Universita' Politecnica delle Marche, Ancona, Italy. An Overview of Current Research on Automated Essay Grading.
10. Whittington, D., and H. Hunt. 2019. "Approaches to the Computerized Assessment of Free Text Responses". figshare. https://hdl.handle.net/2134/1775.
11. Jerrams-Smith, J., Soh, V., and Callear D. (2001). Bridging gaps in computerized assessment of texts. Proceedings of the International Conference on Advanced Learning Technologies, 139-140, IEEE.
12. Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman R. A. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.
13. Burstein,Kukich,Wolff,Chi and Chodorow(1998). Automated scoring using a hybrid feature identification technique.Proceeeding ACL '98/COLING '98 procedings of Annual Meeting of the Association for Computational Linguistics and 17th interational Conference on Computational Linguistics –Volume 1 pages 206-210
14. Burstein, J., Leacock, C., and Swartz, R. (2001). Automated evaluation of essay and short answers. In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.
15. Mason, O. and Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In M. Danson (Ed.), Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.
16. Odunayo E. Oduntan, Ibrahim A. Adeyanju, Adeleye S. Falohun and Olumide O. Obe (2018). A Comparative Analysis of Euclidean Distance and Cosine Similarity Measure for Automatic Essay-Type Grading. Journal of Engineering and Applied Sciences. 13(11): 4198-4204.
17. V. Suresh, R. Agasthiya, J. Ajay, A. A. Gold and D. Chandru, (2023)"AI based Automated Essay Grading System using NLP," *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2023, pp. 547-552, doi: 10.1109/ICICCS56967.2023.10142822.
18. Ramesh, D., Sanampudi, S.K. (2022) An automated essay scoring systems: a systematic literature review. *Artif Intell Rev* **55**, 2495–2527 (2022). https://doi.org/10.1007/s10462-021-10068-2
19. Zhu W, Sun Y (2020) Automated essay scoring system using multi-model Machine Learning, david c. wyld et al. (eds): mlnlp, bdiot, itccma, csity, dtmn, aifz, sigpro
20. Tashu TM, Horváth T (2020) Semantic-Based Feedback Recommendation for Automatic Essay Evaluation. In: Bi Y, Bhatia R, Kapoor S (eds) Intelligent Systems and Applications. IntelliSys 2019. Advances in Intelligent Systems and Computing, vol 1038. Springer, Cham
21. Uto M, Okano M (2020) Robust Neural Automated Essay Scoring Using Item Response Theory. In: Bittencourt I, Cukurova M, Muldner K, Luckin R, Millán E (eds) Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science, vol 12163. Springer, Cham
22. Vijaya Shetty, S., Guruvyas, K.R., Patil, P.P., Acharya, J.J. (2022). Essay Scoring Systems Using AI and Feature Extraction: A Review. In: Bindhu, V., Tavares, J.M.R.S., Du, KL. (eds) Proceedings of Third International Conference on Communication, Computing and Electronics Systems . Lecture Notes in Electrical Engineering, vol 844. Springer, Singapore. https://doi.org/10.1007/978-981-16-8862-1_4
23. Catulay J, Magsael M, Ancheta D and Costales J. (2021). Neural-Network Architecture Approach: An Automated Essay Scoring Using Bayesian Linear Ridge Regression Algorithm 2021 8th International Conference on Soft Computing & Machine Intelligence (ISCMI). 10.1109/ISCMI53840.2021.9654801. 978 -1-7281-8683-2. (196-200).