International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Rethinking Automated Essay Scoring Solutions for Frontline Academics

Kennedy A. Osakwe[1], Kunle Ola[2], Pete Omotosho[3]

[1]PhD, MAIOH, COH, FIIRSM, FRSPH, Senior Lecturer & Program Manager, PCPM, RMIT University, VIC 3000, Melbourne, Australia
Corresponding Author E-mail: *kennedy.osakwe[at]rmit.edu.au*
Orchid: 0000-0003-1415-6092

[2]Thomas More Law School, Australian Catholic University, St John Paul II (Building 212), 1100 Nudgee Road, Banyo Qld 4014

[3]RMIT University, School of Property, Construction and Project Management, 124 La Trobe St, Melbourne VIC 3000, Australia

Abstracts: ***With the surging digital revolution and increasing throughput in student numbers, the traditional methodsadopted in academia would need to give way to more efficient, effective and faster methods.One of such areas lagging in academia is the grading of student's scripts especially essay tasks which for ages had been done manually.The automated essay scoring (AES) system automatically scores and evaluate essay scripts and provide outcomes. Theobjective of this study is to identify safety nets requiredin the deployment of AES systemsin institutions of higher learning.To achieve this, a cross section survey of academics was undertaken to seek their opinion. It revealed that theAES system should integrate cyber safety, feedback, similarity scoring, re-configurability and artificial intelligence capabilities.***

Keywords: Safety nets, feedback, re-configuration, similarity scoring

1. Introduction

The term automated essay scoring (AES) is often used interchangeably with terms like 'automated essay grading' (AEG), 'automated writing evaluation' (AWE) and 'automated essay evaluation' (AEE)[1]. The different slants of the concept all convey the samemessage of a programable to automatically evaluate essay's and provide feedback. Ellis Batten Page has been credited with the pioneering works on automated essay scoring (AES) [3],[1], [2].

In"The Use of the Computer in Analyzing Student Essays" [4], he proposeda machine scoring technology program called Project Essay Grade [PEG]. The program (PEG) is a software powered by artificial Intelligence (AI) technology to score essays using trins and proxes[3]. The program reads, understands, processes and provides results. PEG as well as other AES systems address the burden associated with scoring high numbers of student essays and enables consistency in the grading process [4], [5].According to Valenti, the available automated assessment system in the commercial market are project Essay Grade (PEG), Intelligent Essay Assessor (IEA), Educational Testing Service, Electronic Essay Rater (E-rater), C-rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free text marking Engine and Automark [6].

The developments in statistical and linguistic strategies as well as in artificial intelligencewere instrumental to Page's researchwhich was first published in 1967 [7]. The programwas a culmination of his earlier research worksincluding"The imminence of grading essays by computer","Grading essays by computers" and "The analysis of essays by computer" [4]. Initial results from the program indicated an impressive prediction with human graders as it showed a high similarity between the program and two human graders[4]. PEG used a set of human graded essays and applied linear regression using a variety of automatically extractable textual features to predict the teachers' grades [4].The resultshowed a multiple R correlation with teachers' scores of 0.78—almost as strong as the 0.85 correlation between two or more teachers' [8].Further development and implementation of the program at the time was impracticable until the 1990's when access to the internet and computers became cheaper and available to the average person[9].

Drawing from PEG's inspiration, several other solutions have been developed for automated essay scoring including e-rater, IEA, IntelliMetric, Bookette, CRASE, Autoscore, Lexile, OzEgrader, Markit, SAGrader and several others.Today,AES isan integral part of the educational system and is widely used for scoring examinations such as Pearson Test of English (PTE), Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL), Graduate Management Admissions Test (GMAT), Cambridge Advanced English (CAE) and a host of others.In the 1960s when page commenced his research, automatically extractable features from text were limited to surface features [9].Predictable features in an essay identified by Page included number of words,length of words, uncommon words and punctuation[10]. Page described these features as intrinsic qualities of a competently written essay.He deployed indirect measures due to computational difficulties of implementing direct measures [4]. Although the program successfully predicted teachers' essay ratings, his earlier version of the program was not widely accepted within the educational circles because he usedindirect measures [10]. The argument was that using indirect measures left the system vulnerable to cheating;as students could trick the system by for example writing longer essays, thereby enhancing their chances for a higher score and ultimately manipulating the program.Another criticism of Page's program was the

inability of indirect measures to deal with essential issues such as content, organization, and style and that therefore the provision of instructional feedback to students would be impracticable [10].

The E-rater was the first AES system that addressed the gaps identified in the PEG [11]. In 1999 it was deployedas one of two assessors for the written segment of the Graduate Management Admissions Test (GMAT) [12].

Essay writing is typically part of the evaluation process in academia [12]. Based on the number of students being evaluated, it's pertinent to have an automated system that does not only reduce time but is also cheap for the level of value it creates. So many organisations have invested in AES systems to help achieve effective and sustainable outcomes. These systems effectively grade and assign scores essays without human intervention [13]. Several researches have gone into ensuring the efficacy of such systems and have been found to be up to standard. Typically, the scores from the use of AES systems are compared to that assigned by human graders. Review of essays and student feedback are crucial parts of effective essay grading. Obviously, providing response to papers is mostly draining for graders, especially where there are multiple assessments and hundreds of papers to grade. Situations like these make a huge case for the use of AES systems.

Top on the list of benefits is the ability of AES to provide quick feedback to the students. AES systems can provide instantaneous feedback to students after submission. This speed cannot be matched by human graders. Swift grading is important, and it is fast becoming best practice in academic and professional examinations. It is an indispensable consideration and may be directly linked to learners' enthusiasm to study. AES systems can grade and deliver feedback to students within seconds and has been found to be accurate and reliable[14]. In addition to the swiftness in providing feedback, one of the most valuable benefits is the consistency that comes with the use of an AES system. The same benchmark is used for all the students making grading less prone to human errors. Consequently, human bias is not an issue for the AES system.These days, training and assessment have moved beyond wall and mortar schools; online and distance education has become the order of the day. AES can be used not only to grade essays but to support the students with tutoring [15]. In the long run, using an AES system is less expensive and faster for both the tutor and educational organisations. Engaging human graders have been found to more expensive than human rater [16].Manual grading often requires engagingmore than one grader to evaluate large-scale assessments to uphold quality and decrease human bias. Human graders require training on an all-inclusive assessment rubric and the exercise is not cheap. The AES system comes as a cheaper option. [17].

It is possible to use similar words that are well-thought-out to compare to human ways of articulation. AES systemsare designed to accomplish repetitive functions without boredom and discrepancy. They are flexible and adaptable as tests can be carried out anytime according to any pre-configured algorithm. In a nutshell, it is possible to imitate the human selection of words and classifications [17]. The underlying idea can be summarized as:
"meaning of word 1 + meaning of word 2 + ... + meaning of word k = meaning of passage" (Landauer et al., 2003)

AES guarantees consistency by usings a standardised approach with the application of exactly the similar measures to all answers, thereby leading to consistency in grading. Automated responses are possible with time and date stamps. Other features such as plagiarism detector, style, error checks, validation and academic reference checks can be embedded in the system. AES systems do not only offer scholars with chances to write quality thesis, but also offer swift and precise feedback about grammar, content, organisation, errors, style, and referencing [17].The objective of this study is to identify key technical capacities and measures to mitigate the backdrops in the use of AES systems.

2. Material and Method

A cross sectional survey of scholars using questionnaires designed on survey monkey. The questionnaire was administered to twenty university lecturers and was designed using the outcome of previous published works by authors,'Contactless Academia – A scooping Review on Automated Essay Scoring (AES) System in Covid Pandemic[18]' and 'Use of Technology and Occupational Health Exposures Encountered by Academics in Institutions of Higher Learning – An Exploratory Study [19]'.Beside general background data on age and gender, the questions sought to know the opinion of respondents on the suitability and integration of the following capabilities in the design of AESsolutions, namely,cyber security, configurability of platforms, feedbacks, similarity scoring and assessment rubrics.Output of respondent answers were analysed using basic statistical analysis including averages and percentages.Preferred controls and technical capabilities to be integrated in the design of AES were analysed using weighted mean score on a Likert scale of 'strongly disagree', 'disagree', 'neutral', 'agree' and 'strongly agree'.

3. Findings

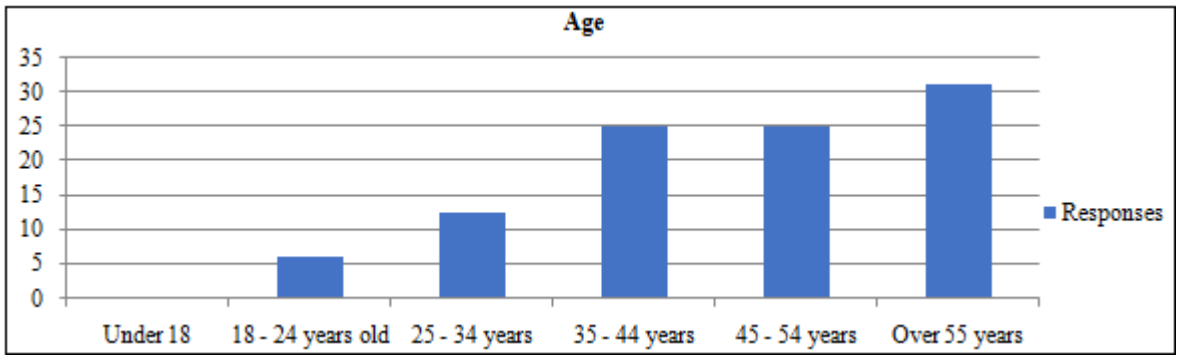

Figure 1: Age of the Respondents

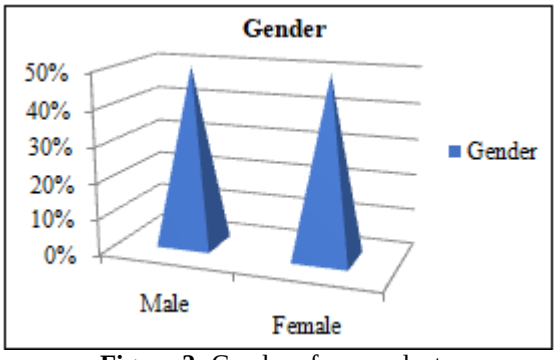

Figure 2: Gender of respondents

Controls for automated scoring

Table 1: Capabilities and Controls for Automated Essay Scoring Solutions

Controls	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Sum	Weighted Average	Rank
Cyber security	0(0.00%)	0(0.00%)	2(12.50%)	4(25.00%)	10(62.50%)	16	4.50	1st
Configurability of solutions	0(0.00%)	0(0.00%)	2(12.50%)	6(37.50%)	8(50.00%)	16	4.38	2nd
Feedback	0(0.00%)	0(0.00%)	2(12.50%)	7(43.75%)	7(43.75%)	16	4.31	3rd
Similarity scoring	1(6.25%)	0(0.00%)	2(12.50%)	4(25.00%)	9(56.25%)	16	4.25	4th
Standard computer workstation	0(0.00%)	0(0.00%)	3(18.75%)	6(37.50%)	7(43.75%)	16	4.25	4th
Customized rubrics	1(6.25%)	1(6.25%)	1(6.25%)	6(37.50%)	7(43.75%)	16	4.06	6th
Artificial intelligence	3(18.75%)	1(6.25%)	3(18.75%)	7(43.75%)	2(12.50%)	16	3.25	7th

P≤ 0.001 shows that the data was of high statistical significance, therefore valid.

4. Discussion

The analysis from the study in Figure 1 revealed that most of the respondents were in the "above 55 years age bracket". This bracket represents the elderly who are generally more skilled and experienced in marking manual and electronic essays.

It also buttresses the inference from the findings of Basseythat older people consistently outshine younger ones on all measures of wisdom, offering more thoughtful, sophisticated advices and assessment of students' performances [20]. Lecturers' age has been found to have influence on their assessment of students' learning activities and overall teaching effectiveness as reported in the study of Bassey [20]. This underscores the importance of the respondent age and supports the reliability of the conclusions.

This study further revealed equal distribution of male and female lecturers which does not differ in the global ratings from their students [21]. On a further analysis, statistically significant differences were found, where some students rated the male lecturers higher than their female counterparts on having positive impact on learning while female lecturers were rated higher on class participation than their male counterparts [21]. This study however stands on the global preconception that students' evaluation of male and female lecturers as professionals are not different and the result of the analysis of this study shows equal distribution of male and female. Notable also is the fact that other categories of gender identified in the analysis do not have any response attached to them.

Respondents acceded to the necessity of integrating safety nets in the design and use of AES butatvarying levels of importance. Cyber security was considered as the most important control by the respondents.Cyber hacking, information theft, corruption of question bank and more are the existential threats to databases which might impact the integrity of assessments and its process. Since AES is network dependent, it could be remotely accessed and potentially vulnerable to attack. This makes cyber security an important feature to integrate in the successful deployment of the AES system.Re-configuration of AES architecture to make product bespoke to by clients ranked second in the respondent'sopinion. It wouldenable the

useradapt variables, tasks, assessments and narratives, thereby making it useable by different institutions with different intranet environment. Survey outcome further revealed thatsimilarity scoringcapability is a sine-qua-non to achieving academic integrity and ethical compliance, thus should be a design specification for AES systems. Automated provision of feedback was unanimously accepted as an essential feature that the AES system should possess, thus corroborates Dikli' [22]. A surprising outcome was the low rating on artificial intelligence (AI) which might be because of poor understanding of what AI entails. Essentially, it encompasses all the choices previously rated acceptable to the respondent.

Strength and Weaknesses – The strength of this studies lies in the involvement of academics across several universities in the world and a rich mix of literature in benchmarking the study results. The limitation however lies in the limited number of participants in the study.

Further Work - Further work would be required to identify key technical capabilities an AES system should possess.

5. Conclusion

This study explored the opinion of academics on suitable safety nets that should be integrated in the use of AES.It was found that there should be cyber security, feedback, similarity scoring, re-configurability and AI capabilities.

6. Ethical Statement

The authors declare that there was alignment with high ethical standards during the study

7. Conflict of Interest

The authors declare that there are no conflicts of interest

References

[1] Zupanc K & Bosnic Z (2015) Advances in the Field of Automated Essay Evaluation. Informatica 39 (2015) 383–395 383.
[2] Bereiter C (2003) Automated essay scoring: a cross disciplinary approach. Lawrence Erlbaum Associates: Mahwah, NJ.
[3] Rudner L. M. & Gagne P (2000) An overview of three approaches to scoring written essays by computer. Practical Assessment, Research, and Evaluation 7(26), 1-5.
[4] Page EB (1994). New computer grading of student prose, using modern concepts and software. Journal of Experimental Education 62(2), 127-142.
[5] Page EB (1968) The Use of the Computer in Analyzing Student Essays. International Review of Education14(3), 253-263.
[6] Valenti S, Neri F and Cucchiarelli A (2003) An overview of current research onautomated essay grading. Journal of Information Technology Education 2:319 – 330
[7] Page EB (1967) Statistical and linguistic strategies in the computer grading of essays. In proceeding of the Conference on Computational Linguistics, COLING 1967, 1-13.
[8] Kukich K (1992)Techniques for automatically correcting words in text. ACM Comput. Surv., 24, 377-439.
[9] Norman J (2020) Ellis Batten Page Begins Automated Essay Scoring 1967 CE History of Information, http://www.historyofinformation.com/detail.php?id=3544
[10] Hearst M(2000)The debate on automated essay grading. IEEE Intelligent Systems 15(5):37
[11] Rudner LM, Garcia V, & Welch C(2006) An Evaluation of the IntelliMetric Essay Scoring System. The Journal of Technology, Learning and Assessment, 4(4), 3–20.
[12] Rudner LM & Liang T(2002) Automated Essay Scoring Using Bayes' Theorem, The Journal of Technology, Learning and Assessment, 1(2), 3–21.
[13] Taghipour K & Ng HT(2016) A Neural Approach to Automated Essay Scoring. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. doi: 10.18653/v1/d16-1193
[14] Oluade-Ibijola A, Wakama I, & Amadi JC (2012) An Expert System for Automated Essay coring (AES) in Computing using Shallow NLP Techniques for Inferencing. International Journal of Computer Applications, 51(10), 37–45. doi: 10.5120/8080-1480
[15] Chung KWK & O'Neil HF (1997) Methodological approaches to online scoring of essays (ERIC reproduction service no ED 418 101).
[16] 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. https://doi.org/10.1007/978-3-030-52237-7_44
[17] Burstein J, Kukick K, Wolff S, Lu C & Chodorow M(2001). Enriching automated essays coring using discourse marking. proceedings of the workshop on discourse relations and discourse marking, annual meeting of the association of computationallinguistics; Canada: Montreal.
[18] Osakwe K., Ola K & Omotosho P (2021)Contactless Academia – The Case for Automated Essay Scoring (AES) System in COVID 19 Pandemic. Current Journal of Applied Science and Technology, 40(4):17-29.
[19] Osakwe KA, Ola K & Omotosho P (2021)Use of Technology and Occupational Health Exposures Encountered by Academics in Institutions of Higher Learning – An Exploratory Study.Internal Journal of Science and Research (IJSR)
[20] Bassey BA (2016) Undergraduates' View of lecturers' age as a factor in their teaching effectiveness. Global Journal of Social Sciences, 15, 1 – 11
[21] Appiah SO & Agbelevor EA (2015) Impact of lecturers' gender on learning: Assessing University of Ghana students' views. Journal of Education and Practice, 6(28), 30 – 37
[22] Dikli, S. (2006). Automated Essay Scoring. Turkish Online Journal of Distance Education 7(1), 49 - 62