



Faculty of Computer Science and Information Technology

**A Study of Automated Essay Scoring Frameworks on Evaluating
Malaysian University English Test Essays Based on Syntactic and
Semantic Features**

Lim Chun Then

**Master of Science
2023**

A Study of Automated Essay Scoring Frameworks on Evaluating Malaysian
University English Test Essays Based on Syntactic and Semantic Features

Lim Chun Then

A thesis submitted

In fulfillment of the requirements for the degree of Master of Science

(Language Technologies)

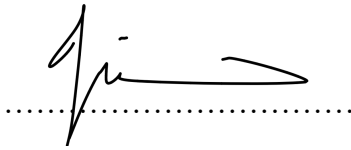
Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2023

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

A handwritten signature in black ink, appearing to read 'Lim Chun Then', is written over a horizontal dotted line.

Signature

Name: Lim Chun Then

Matric No.: 19020170

Faculty of Computer Science and Information Technology

Universiti Malaysia Sarawak

Date : 6/6/2023

ACKNOWLEDGEMENT

First of all, I would like to express my deepest appreciation to my supervisor, Dr. Bong Chih How and co-supervisor, Dr Lee Nung Kion who are abundantly helpful and offer invaluable assistance by stimulating suggestions and encouragement in performing this research. Without his help in the coordination, this research could not be done successfully. Besides, I would like to thank Malaysia Ministry of Higher Education and Universiti Malaysia Sarawak (UNIMAS) who funded this research study through the grant #F04/PRGS/1801/2019. I would also sincerely thank the faculty staff and Centre for Graduate Studies who gave the permissions to use all the required facilities and necessary material to complete this research. Finally, I wish to acknowledge with gratitude to those who have contributed directly or indirectly to this research.

ABSTRACT

An Automated Essay Scoring (AES) system can use a trained computational model to evaluate an essay as close to the grade that a human rater would assign. The purpose of this study is to examine the performance of different machine learning methods in predicting Malaysian University English Test (MUET) essay grade based on syntactic features and semantic features and generalize frameworks accordingly. Based on the results, we found that syntactic features of an essay have a higher effect than semantic features towards essay grades. Besides, we also found that the differences between machine learning and deep learning algorithms were not obvious, and neither algorithm's performance can be considered excellent because the quadratically weighted Kappa (QWK) scores were less than 0.75. Instead of using any available public essay datasets, five MUET essay datasets were collected locally for this study, and we found that all datasets suffer from imbalanced grade distribution. Therefore, QWK score is preferred over accuracy as the standard evaluation metric for AES because it provides more valuable information when dealing with imbalanced datasets. To overcome the problem of imbalanced grade distribution, a resampling method called Synthetic Minority Oversampling Technique (SMOTE) is applied to the dataset to study the impact of the resampling method on the performance of the AES framework. However, the SMOTE resampling method has been found to degrade predictive model accuracy and QWK scores. In addition, this study also developed an e-learning platform called UNIMAS DBRater, which is currently used by UNIMAS pre-university English classes, and more and more local educational institutions have expressed interest and willingness to join this e-learning platform.

Keywords: Automated essay scoring, MUET, machine learning, resampling

Pendekatan Pembelajaran Mendalam terhadap Pemarkahan Karangan Automatik

ABSTRAK

Sistem Pemarkahan Karangan Automatik (AES) merujuk kepada penggunaan model komputer yang terlatih untuk menilai karangan dengan mendapat permarkahan yang hampir sama dengan penilai manusia. Tujuan kajian ini adalah untuk mengkaji prestasi kaedah pembelajaran mesin yang berbeza dalam meramal markah karangan Malaysian University English Test (MUET) berdasarkan ciri sintaksis dan semantik serta membuat generalisasi rangka kerja yang sesuai. Berdasarkan hasil kajian, kami mendapati bahawa ciri-ciri sintaksis karangan mempunyai kesan yang lebih tinggi daripada ciri-ciri semantik terhadap gred karangan. Selain itu, kami juga mendapati bahawa perbezaan antara algoritma pembelajaran mesin dan pembelajaran mendalam tidak ketara, dan prestasi kedua-dua algoritma tidak boleh dianggap cemerlang kerana markah Quadratic Weighted Kappa (QWK) kurang daripada 0.75. Sebagai gantinya menggunakan set data karangan awam yang tersedia, lima set data karangan MUET dikumpulkan secara tempatan untuk kajian ini, dan kami mendapati bahawa semua set data mengalami masalah ketidakseimbangan pengedaran markah. Oleh itu, markah QWK lebih disukai daripada Accuracy sebagai metrik penilaian standard untuk AES kerana metrik ketepatan ini memberikan maklumat yang lebih bernilai apabila berurusan dengan dataset yang tidak seimbang. Untuk mengatasi masalah ketidakseimbangan pengedaran markah, satu kaedah persampelan yang dipanggil Synthetic Minority Oversampling Technique (SMOTE) digunakan pada dataset untuk mengkaji impak kaedah persampelan pada prestasi rangka kerja AES. Walau bagaimanapun, kaedah persampelan SMOTE telah didapati merosakkan ketepatan model prediksi dan markah QWK. Selain itu, kajian ini juga mengembangkan sebuah platform pembelajaran atas talian yang dipanggil UNIMAS DBRater, yang kini

digunakan oleh kelas-kelas bahasa Inggeris pra-universiti UNIMAS, dan banyak institusi pendidikan tempatan telah menyatakan minat dan keinginan untuk menyertai platform pembelajaran atas talian ini.

Kata kunci: *Pemarkahan karangan automatik, Ujian Bahasa Inggeris Universiti Malaysia, pembelajaran mesin, pensampelan semula*

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
ABSTRAK	iv
TABLE OF CONTENTS	vi
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xiii
CHAPTER 1: INTRODUCTION	1
1.1 Study Background	1
1.2 Problem Statement	2
1.2.1 Traditional Automated Essay Scoring Systems Reply on Manually Designed Features Which Require Manual Engineering by Human Experts	2
1.2.2 Most Existing AES Studies Are Trained and Evaluated Using Existing Western Essay Datasets Based on Western Test Scoring Scales, Their Impact on Self-collected Local Essay Datasets Is Remain Unclear	3
1.2.3 Malaysia Does Not Have a Mature and Usable AES System for Grading Essays, Making Grading Essays a Huge Burden for Teachers	4

1.3	Objectives	4
1.4	Hypothesis	5
1.5	Significance of the Study	5
1.6	Research Scope	6
1.7	Research Methodology	7
1.8	Chapters Overview	11
CHAPTER 2: LITERATURE REVIEW		12
2.1	Overview	12
2.2	Automated Essay Scoring Systems	12
2.2.1	Project Essay Grader [®] (PEG)	13
2.2.2	Intelligent Essay Assessor [™] (IEA)	13
2.2.3	IntelliMetric [®]	14
2.2.4	E-rater [®]	14
2.3	Dataset	16
2.3.1	Malaysian University English Test (MUET)	17
2.3.2	Automatic Student Assessment Program (ASAP)	21
2.4	Features	24
2.4.1	Syntactic Features	24
2.4.2	Semantic Features	25
2.4.3	Hybrid Features	27

2.5	Learning Algorithms	27
2.5.1	Cosine similarity	28
2.5.2	Machine Learning	29
2.5.3	Deep Learning	31
2.6	Resampling	37
2.7	Evaluation Metrics	39
2.7.1	Accuracy	39
2.7.2	Quadratic Weighted Kappa (QWK)	40
2.8	Summary of Related Works	41
2.9	Conclusion	48
	CHAPTER 3: METHODOLOGY	49
3.1	Overview	49
3.2	Proposed AES Framework	49
3.2.1	Syntactic AES Framework	51
3.2.2	Semantic AES Framework	52
3.2.3	Hybrid AES Framework	53
3.3	Data Collection	53
3.4	Data Pre-Processing	60
3.5	Features Extraction	60
3.6	Learning Algorithms	62

3.7	Resampling	63
3.8	Evaluation	65
3.9	Conclusion	66
CHAPTER 4: RESULTS AND DISCUSSION		67
4.1	Overview	67
4.2	Predictive Performance of the Proposed AES System	67
4.2.1	Prediction Results based on Syntactic Feature	67
4.2.2	Prediction Results based on Semantic Features	69
4.2.3	Prediction Results based on Hybrid Feature	71
4.2.4	Summary of the Prediction Results	73
4.3	Prediction Results with Resampling Method	73
4.3.1	Prediction Results Based on Syntactic features with Resampling Method	74
4.3.2	Prediction Results Based on Semantic features with Resampling Method	75
4.4	Comparison Between Syntactic Features and Semantic Features and Effect of SMOTE	77
4.5	Applicability of Evaluation Metrics on AES	82
4.6	Comparison of Machine Learning and Deep Learning on AES Performance	84
4.7	Chapter Summary	84
CHAPTER 5: USE CASE: PREDICTING MUET BAND AMONG PRE-UNIVERSITY STUDENTS		86
5.1	Overview	86

5.2	Background	86
5.3	Task Description	89
5.4	Results and Discussion for the Use Case Experiment	93
5.5	Chapter Summary	95
	CHAPTER 6: CONCLUSIONS	97
6.1	Conclusions of the Study	97
6.2	Limitation	98
6.2.1	Limited Resources	98
6.2.2	Performance of Proposed AES Framework Is Not Satisfactory	99
6.3	Contributions	99
6.4	Future Work	100
	REFERENCES	101
	APPENDIX	116

LIST OF TABLES

	Page
Table 2.1 Summary of well-known AES	15
Table 2.2 Scoring Rubrics of MUET essay	19
Table 2.3 Details of ASAP Datasets	22
Table 2.4 Summary of past literature in Asia	41
Table 3.1 Essay Title for Each Dataset	55
Table 3.2 Details of Features for Each Dataset	59
Table 4.1 Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Syntactic Features	68
Table 4.2 Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on LSA Semantic Features	70
Table 4.3 Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Word2Vec Semantic Features	71
Table 4.4 Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Hybrid Features	72
Table 4.5 Best Mean Performance of Different AES Frameworks in Predicting Essay Grade	73
Table 4.6 Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Syntactic Features with SMOTE	74
Table 4.7 Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on LSA Semantic Features with SMOTE	75
Table 4.8 Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Word2Vec Semantic Features with SMOTE	76
Table 4.9 Confusion matrix of result of dataset 2, random forest, syntactic features	82
Table 5.1 Essay Titles for Each Dataset of DBRater	90
Table 5.2 Performance of Different Machine Learning Algorithms in Predicting DBRater Essay Grade Based on Syntactic Features	94
Table 5.3 Accuracy and QWK of Dataset Between Two Human Raters	95

LIST OF FIGURES

	Page
Figure 1.1 Steps in Methodology	8
Figure 2.1 Mapping of Syntactic and Semantic Dimension with MUET Scoring Rubrics	21
Figure 3.1 Proposed AES Framework Flow Chart	50
Figure 3.2 Pie Chart of Band Distribution in Dataset 1	56
Figure 3.3 Pie Chart of Band Distribution in Dataset 2	57
Figure 3.4 Pie Chart of Band Distribution in Dataset 3	58
Figure 3.5 AES Framework based on Syntactic features with Resampling	64
Figure 3.6 AES Framework based on Semantic features with Resampling	64
Figure 4.1 Visualized Vectors Space of Syntactic and Semantic Features of Dataset	78
Figure 4.2 Visualized Vectors Space of Syntactic and Semantic Features of Dataset with SMOTE	80
Figure 5.1 Dashboard of UNIMAS DBRater	87
Figure 5.2 Writing Interface of UNIMAS DBRater	88
Figure 5.3 Interface of Student Management	89
Figure 5.4 Pie Chart of Band Distribution in Dataset A	91
Figure 5.5 Pie Chart of Band Distribution in Dataset B	92
Figure 5.6 Pie Chart of Band Distribution in Dataset C	93

LIST OF ABBREVIATIONS

AEG	Automated Essay Grading
AES	Automated Essay Scoring
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASAP	Automated Student Assessment Prize
BiLSTM	Bi-directional Long Short-Term Memory
CBOW	Continuous Bag of Words
CEFR	Common European Framework of Reference
CNN	Convolutional Neural Network
ETS	Educational Testing Service
IEA	Intelligent Essay Assessor™
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MUET	Malaysian University English Test
NLP	Natural Language Processing
NN	Neural Network
OR	Ordinal Regression
PEG	Project Essay Grader®
PKT	Pearson Knowledge Technologies
QWK	Quadratic Weighted Kappa
RMSE	Root Mean Square Error
SBLSTMA	Siamese Bidirectional Long Short-Term Memory Architecture

SMOTE	Synthetic Minority Over-sampling Technique
SRNs	Simple Recurrent Networks
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regression
t-SNE	t-distributed stochastic neighbour embedding
UNIMAS	Universiti Malaysia Sarawak
VSM	Vector Space Model

CHAPTER 1

INTRODUCTION

1.1 Study Background

Natural Language Processing (NLP) is a subfield of computer science, artificial intelligence and human language that deals with the interaction between computers and humans through natural languages. The goal of NLP is to aid computers to understand the human's natural language and giving computers the ability to generate meaningful human's natural language (Garbade, 2018). Nowadays, NLP is widely used in language translation, sentiment analysis, document summarization, automated essay scoring and other fields to lighten the burden on human works (Algorithmia, 2020). NLP contains two features, which are Natural Language Understanding and Natural Language Generation. The Natural Language Understanding can be used to analyse the content of essays and while Natural Language Generation can be used to generate example sentences for students to improve their writing skills.

One area where researchers have widely adopted NLP is the Automated System Scoring (AES) systems. An AES system aims to create a model that can automatically evaluate essays or reduce the involvement of human raters. It is a combination of NLP application and educational evaluation systems. An AES system may rely on grammar, spelling, and also more complex features such as semantics, discourse, and pragmatics (Chen & He, 2013). To model the scoring process used by human raters, a popular approach to AES is to use previously graded essays to train a scoring model. When given the same collection of essays for evaluation and a sufficient number of graded samples, AES systems generally achieve excellent agreement with skilled human raters.

The development of an effective AES system in Malaysia is challenging due to the unique linguistic and cultural context of the country, which differs significantly from Western educational models. Moreover, cultural differences in writing styles and expectations between Malaysia and the West may further complicate the development of an accurate AES system. Thus, it is crucial to comprehend these challenges to ensure the success of an AES system for Malaysian society. The significance of this research will be emphasized by highlighting its potential impact on education and grading efficiency in Malaysia. Therefore, this study will focus on examine the performance of AES system on grading Malaysian University English Test (MUET) essays. The MUET is a test of English language proficiency, largely used for university admissions in Malaysia. The goal of MUET is to assess candidates' English language proficiency in order to assist universities in making better judgments regarding prospective candidates' preparation for academic coursework. Furthermore, we applied deep learning in an AES system in order to increase grading accuracy and reduce the AES system development time. Deep learning is a subset of machine learning methods based on Artificial Neural Networks (ANN). With deep learning, the system itself is able to find the relationship between input data and desired results without any human interventions. Thus, it can prevent human bias when grading essays.

1.2 Problem Statement

We have identified three major problems in AES systems through an extensive review of previous studies in the area and have summarized them as follows:

1.2.1 Traditional Automated Essay Scoring Systems Reply on Manually Designed Features Which Require Manual Engineering by Human Experts

Traditional AES systems typically use handcrafted features manually created by human experts to predict scores and the development process usually requires a lot of time

and effort in feature engineering and tuning to achieve high grading accuracy in predicting essay scores (Alikaniotis et al., 2016; Uto, 2021). Besides, handcrafted features dramatically increase the number of tuning parameters, making the AES system training more challenging (Uto et al., 2020). As a result, the performance of an AES system tends to decrease due to the lack of human specialists. Although traditional AES models typically rely on manually designed features, deep learning based AES models that obviate the need for feature engineering have recently attracted increased attention (Uto, 2021) since the deep learning approach AES system does automatic feature extraction and finds unique, complex features (Borade & Netak, 2021). Therefore, more research is needed to compare the performance of traditional approach AES and the deep learning approach AES on grading essay.

1.2.2 Most Existing AES Studies Are Trained and Evaluated Using Existing Western Essay Datasets Based on Western Test Scoring Scales, Their Impact on Self-collected Local Essay Datasets Is Remain Unclear

Although AES research assumes that we can use enough graded essays for training, this assumption is often not met in practice because collecting graded essays is an expensive and time-consuming task (Ke & Ng, 2019). Hence, many studies use existing publicly available essay datasets as their training and test sets instead of collecting new datasets. According to research, 90% of AES systems use the Kaggle ASAP (2012) dataset as their training and testing dataset (Ramesh & Sanampudi, 2022), which were written by United States students from 7th grade to 10th grade. However, only a few AES studies can be found to train and test their AES systems based on Malaysian local essay datasets. Therefore, more research is needed to analyse the performance of AES systems on self-collected Malaysian local essay datasets.

1.2.3 Malaysia Does Not Have a Mature and Usable AES System for Grading Essays, Making Grading Essays a Huge Burden for Teachers

In the context of this study, there is a prevalent problem of insufficient trained English language teachers in Sarawak (PIM, 2018). This problem adds a huge burden on English teachers to evaluate all student's essays manually and give them timely feedback to improve their writing skills. Such burden could also potentially lead to inaccurate and inconsistent grading among the teachers as they must meet the deadlines. This issue may be solved by several well-known commercial AES adopted by western countries but most of the products are not related to the Malaysian English test context (Wong & Bong, 2021). As a result, those AES systems developed using essays written by native English speakers do not fully meet the needs of English as a Second Language (ESL) or English as a Foreign Language (EFL) students (Dikli S. , 2010). Therefore, more research is needed to develop an AES system that can grade essays in context of the Malaysian English Test.

1.3 Objectives

The main goal of the study is to propose a grading mechanism which can evaluate students' MUET essays with a satisfactory accuracy. The goal can be accumulatively achieved through accomplishing the following objectives:

- 1) To compare the performance of traditional approach AES and the deep learning approach AES on grading MUET essay.
- 2) To prepare and collect a corpus of annotated MUET essay datasets for evaluation to achieve objective 1.

- 3) To demonstrate a use case to show the efficacy of the proposed essay grading system which can automatically grade MUET essays against human raters' accuracy.

1.4 Hypothesis

This study will propose the following hypotheses:

- 1) The application of deep learning algorithm outperforms machine learning algorithm in MUET essay grading performance of AES system.
- 2) Syntactic features can better represent essays than semantic features and achieve better MUET essay grading accuracy.

1.5 Significance of the Study

The study focused on comparing the performance of traditional approach AES and the deep learning approach AES on grading MUET essay which benefits developers understand which approach is better for evaluating MUET essays when developing local AES systems. Besides, the study also provides a solution to reduce the dependence on human experts through an AES system that uses the deep learning approach instead of human handcrafted features. This can shorten the system development time because handcrafted features require human experts to spend a lot of time in various iteration and modification in order to achieve good performance. The deep leaning approach, on the other hand, can find out the relationship between essays and grades without human interventions.

Through this study, a few datasets of annotated MUET essays will be prepared and collected to evaluate the performance of proposed AES frameworks. These datasets can benefit researcher and fill the gap where there are no publicly available MUET datasets for

local AES research. The MUET essay dataset will be collected from local schools and graded by two raters which we believe will be the starting point for the development of AES in Malaysia.

Furthermore, this study will demonstrate a use case to show the efficacy of the proposed essay grading system which can automatically grade MUET essays against human raters' accuracy. This can benefit both students and teachers, as the proposed AES system can reduce the burden on teachers and quickly respond to students' essays with accurate grading. Most of the existing AES systems are developed for native English-speaking countries, but Malaysia is a multilingual country, and different languages will affect each other. This localization of English usage in Malaysia provides a unique opportunity for this study to contribute in terms of creating an AES system specifically for Malaysian English test settings.

1.6 Research Scope

This study was conducted as a quantitative study and the experimental results will be used to test the hypothesis. Moreover, this study focused on evaluating the performance of AES frameworks on grading English essays based on 2014 version of the MUET grading scheme. According to the scheme, the writing test are designed based on the accuracy, appropriacy, coherence and cohesion, use of language functions and task fulfilment of essay which build up knowledge of English and the ability to utilise it (Malaysian Examinations Council, MUET Regulations and Test Specifications, 2014). The results of this study may not be applicable to the new version of the MUET grading scheme, as the writing test specification was revised in 2021 to align MUET with the Common European Framework of Reference (CEFR).

Furthermore, this study also focuses on comparing the performance difference of machine learning and deep learning for grading MUET essays when integrated into an AES system. For machine learning methods, this study will focus on methods that perform well in grading essays based on past AES literature reviews. For deep learning, there are three types of neural networks, Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Through a review of past AES literature, this study will focus on one variant of each neural network that has been studied in the past with good essay grading performance and evaluate their performance in MUET essay grading.

1.7 Research Methodology

The methodology used in this research is described in this section and illustrated in Figure 1.1.

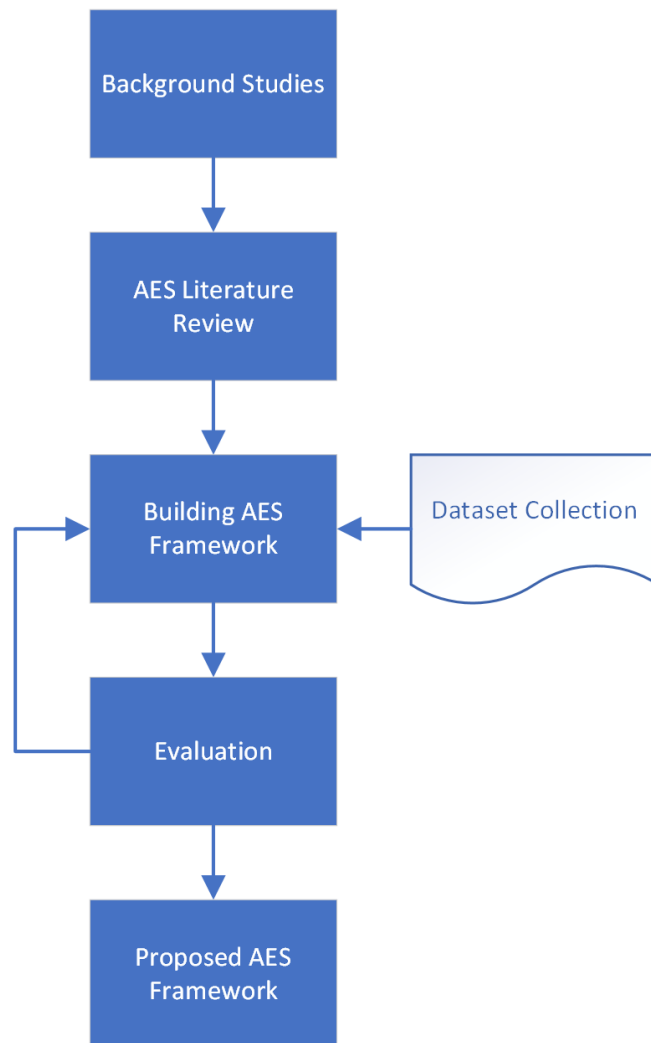


Figure 1.1: Steps in Methodology

Figure 1.1 shows the research methodology in this study. The research methodology involved data collection, building AES framework and evaluation.

Background Studies: Automated Essay Scoring (AES) systems have been developed in response to the increasing need for efficient essay grading. In this stage, the background and development of AES will be studied. Specifically, this research will investigate the challenges of developing an AES system in Malaysian society and the differences with existing human raters. Malaysia has a unique linguistic and cultural context that poses challenges for AES systems, which are often developed based on Western educational tests.