



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)

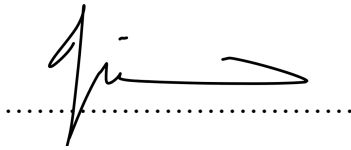
Faculty of Computer Science and Information Technology

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', 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 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.

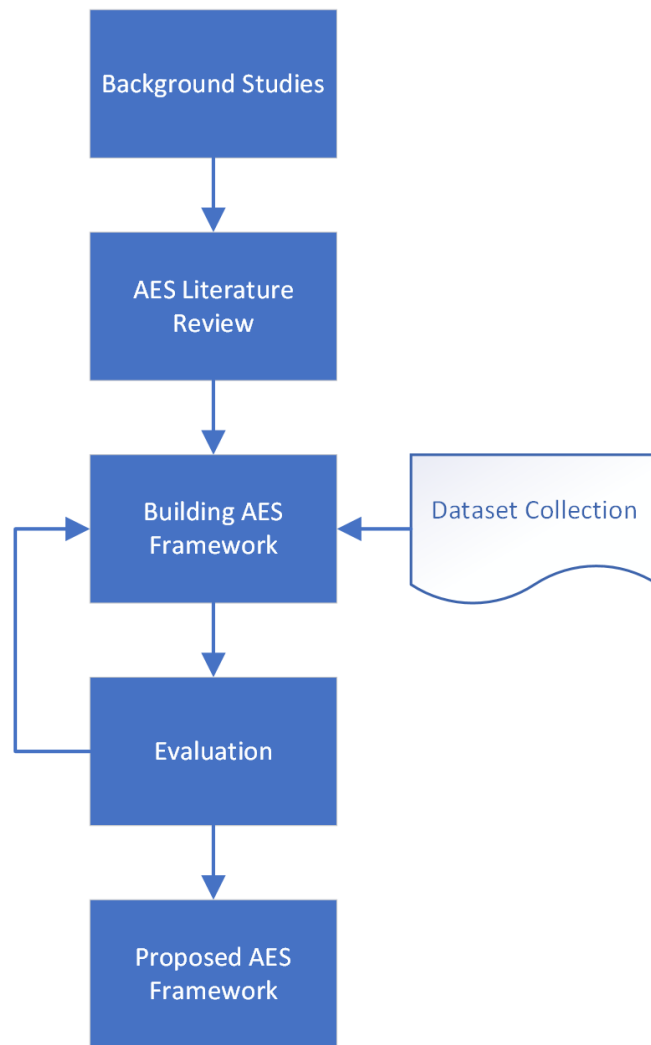


Figure 1.1: Steps in Methodology

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.

Additionally, cultural differences in writing style and expectations between Malaysia and the West may further complicate the development of an effective AES system. Therefore, understanding these challenges is critical to the development of a successful AES system for Malaysian society. The importance of this research will be emphasized in the context of the study, highlighting the potential impact of a successful AES system on education and grading efficiency in Malaysia.

AES Literature Review: This study focuses on reviewing past AES literature in Asia since many Asian countries, including Malaysia, learn English as a second language. By conducting a critical analysis of existing literature, this study emphasizes advancements in the Asia AES field. Based on the past literature review, significant findings and limitations of research are studied, and gaps in existing AES development are identified and analysed. By focusing on AES research in Asia, this study seeks to provide insights and recommendations for the development of a successful AES system in Malaysia. The analysis of past literature can help identify potential challenges and provide strategies for overcoming them, leading to more accurate and efficient grading of English writing in the Malaysian educational context.

Data Collection: Samples of English essays and their grades marked by human raters were collected from SMK Green Road and UNIMAS Pre-University to act as training data and testing data for our research. The collected essays were for the MUET, and they were graded according to bands set, which were from Band 1 to Band 6 (Band 1 being the lowest while Band 6 being the highest). These bands were still being used during the time of this study before the introduction of the new MUET format and bands in 2021. The collected English essays were converted to digital forms and the incomplete essays were removed to

make sure all training data were well-structured and contained complete content. Besides, the essays underwent several steps include word tokenization, stop-words removal, stemming and lemmatization to reduce the noise in data.

Building AES Framework: To develop an AES system for grading MUET English essays, a framework will be built using collected MUET English essays. Next, the features will be extracted from the pre-processed essays and act as input data for machine learning algorithms to predict the essay grade. The framework will use the machine learning algorithms to identify the relationship between the collected essays and their corresponding grades without human intervention. The combination of features and machine learning algorithms will be evaluated to identify the most effective framework composition for grading MUET essays.

Evaluation: In evaluating the effectiveness of the proposed AES framework, the predicted essay grades generated by the framework were compared to the grades given by human raters. Two commonly used evaluation metrics, Accuracy and Quadratic Weighted Kappa, were used to assess the performance of the framework. Additionally, a 10-fold cross-validation technique was applied to the experiment to ensure the reliability of the results. This technique involves dividing the data into ten subsets, training the model on nine subsets, and evaluating its performance on the remaining subset. The process is repeated ten times, with each subset serving as the evaluation set once to ensure that the evaluation results are not biased towards a particular subset of the data.

Proposed AES Framework: Based on the results from the evaluation, a proposed AES framework will be presented. This framework will include the most effective combination of features and machine learning algorithms for grading MUET essays accurately and

efficiently. Additionally, a use case of the proposed AES framework will be demonstrated to showcase the performance and accuracy of the system in real-life scenarios. The proposed AES framework is expected to contribute to the improvement of education and grading efficiency in Malaysia.

1.8 Chapters Overview

The thesis is presented in the following arrangement:

Chapter 2 describes several well-known Automated Essay Scoring (AES) systems, past literatures on AES in Asia and the mechanisms behind the systems that have been studied.

Chapter 3 discusses in detail about our research methodology and the proposed AES frameworks.

Chapter 4 describes the experimental results as well as explaining the performance of our proposed AES frameworks based on syntactic features, semantic features, and both.

Chapter 5 presents a use case study to demonstrate the effectiveness of our proposed AES framework for integrating the corresponding structure into an e-learning platform.

Chapter 6 concludes the study. We discuss the limitations and the suggestions for future improvement of our proposed AES framework.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter explores into several well-known Automated Essay Scoring (AES) systems, examining their mechanisms and drawbacks. Additionally, the methodologies utilized in designing and developing an AES system are reviewed by analysing past literature on AES in Asia. The AES research process is divided into multiple steps, which are outlined in this chapter. Section 2.3 discusses benchmark datasets commonly used in AES research, as well as datasets under the Malaysian University English Test Scoring Scheme. In section 2.4, various essay features are discussed and assessed for their applicability in essay assessment. Section 2.5 describes the learning algorithms used in AES frameworks and reviews the performance of related works that apply these algorithms in grading essays. Section 2.6 reviews resampling techniques utilized to address imbalanced grade distribution issue in datasets. Various evaluation metrics that are commonly used in AES research are reviewed in section 2.7. Finally, section 2.8 presents a summary of previous AES research in Asia, outlining their research details and findings.

2.2 Automated Essay Scoring Systems

This section gives an overview of AES systems that have been widely publicised and commercialised with significant success. The proprietary software used in these systems was mostly created in the Western nations. To describe each product with more specific information, we include the manufacturer or developer, main objective, essay features, grading system, and quantity of training examples needed for each product.

2.2.1 Project Essay Grader[®] (PEG)

The first AES system is Ellis Page's Project Essay Grader (PEG) (Page, 1966). By utilising trins and proxies, it focuses on assessing essays based on their writing style. PEG assumes that a person's writing style possesses intrinsic features known as trins that may be quantified or associated with observable elements known as proxies (Rudner & Gagne, 2001). For instance, there is a correlation between the vocabulary used (proxe) and the essay's fluency (trin). PEG enables statistical regression analysis to estimate essay scores using a training set of 100 to 400. PEG has created more than 500 trins so far to be utilised in scoring essays (Measurement Incorporated, 2020). The drawback of PEG, however, is that it emphasises writing style rather than content and text semantics (Srivastava et al., 2020). As a result, essays may earn good marks even if the material is off topic.

2.2.2 Intelligent Essay Assessor[™] (IEA)

Pearson Knowledge Technologies (PKT) has developed the Intelligent Essay Assessor (IEA) to evaluate the quality of essay contents (Foltz et al., 1999; Intelligent Essay Assessor (IEA)[™] Fact Sheet, 2010). Latent Semantic Analysis (LSA), a computational distribution model to evaluate the semantic similarity of texts, is used by IEA to evaluate essays (Landauer et al., 1998). LSA is performed on domain-specific corpora and the essays are represented by the multidimensional semantic space of the words they include, and the similarity is determined by comparing them to other essays' semantic representations (Foltz et al., 1999). It was claimed that the IEA's LSA scores correlate more closely with human raters than with grades derived from correlations of essay features (Landauer et al., 2000). Furthermore, IEA uses NLP strategies to extract paper features such as complex lexical usage, grammar, mechanics, style, and organizational features of writing (Zupanc & Bosnic, 2015). In contrast to other AES systems, the IEA requires only 100 samples of pre-scored

training papers to score a prompt-specific paper (Dikli S. , 2006). Due to the lack of consideration of word order by the IEA, since it is not considered a key component by the LSA (Srivastava et al., 2020), the grammar in the paper may be ignored in the analysis.

2.2.3 IntelliMetric®

Since 1998, Vantage Learning has introduced IntelliMetric as a proprietary AES system for grading essays (Vantage Learning, 2020). It is believed that IntelliMetric was the first AES system to replicate the scoring process using machine learning and artificial intelligence (AI) (Dikli, 2006; Hussein et al., 2019). More than 400 features, including semantics, syntactics, and discourses, are used by IntelliMetric to generate an essay score. These features fall into five categories under the IntelliMetric Feature Model: focus and unity (coherence), organisation, development, and elaboration, sentence structure, mechanics, and conventions. According to IntelliMetric, there are several automated scoring systems in use, and each one uses a different mathematical model to score essays such as linear analysis, Bayesian, and LSA (Vantage Learning, 2005; Vantage Learning, 2020). As opposed to previous scoring systems that use a single scoring engine, IntelliMetric's multiple scoring engine simulates the effects of a panel of judges to produce a final score that is deemed more accurate. The capability of IntelliMetric to grade essays in languages other than English is another standout quality (Elliott, 2003). However, IntelliMetric's requirement for at least 300 graded essays in order to function is one of its drawbacks (Zupanc & Bosnic, 2015).

2.2.4 E-rater®

The Educational Testing Service (ETS) has been using and developing E-rater since 1999 (Attali & Burstein, 2006). It uses unique NLP algorithms to extract linguistic aspects to assess the essay's style and content. E-rater 2.0 uses a topical, discourse, and syntactic

analysis module to evaluate essay elements (Dikli S. , 2006). Grammar mistakes, word usage mistakes, mechanics mistakes, style, segment organisation, and vocabulary content are among the aspects (Shermis et al., 2010). Writing quality (grammar, use, mechanics, style, organisation, development, word choice, average word length, appropriate prepositions), and collocation usage content or use of prompt-specific terminology are the two categories that the E-rater version now expands the essay grading characteristics into (Ramineni & Williamson, 2018). Regression modelling is used by E-rater to determine an essay's final grade. Additionally, the regression model needs a collection of around 250 training essay examples (Zupanc & Bosnic, 2015).

The comparison of these well-known AES systems is shown in Table 2.1.

Table 2.1: Summary of well-known AES

AES System	Vendor/ Developer	Main Focus	Essay Scoring Mechanism	Essay-Scoring Features	Training Samples Required
IntelliMetric	Vantage Learning	Style & Content	AI - cognitive processing, computational linguistics, and classification	Focus & Unity (Coherence) Organization Development & Elaboration Sentence Structure Mechanics & Conventions	300

Table 2.1 continued

e-rater	Educational Testing Service (ETS)	Style & Content	Regression Analysis	Grammatical Errors Word Usage Errors Mechanics Errors Style Organizational Segment Vocabulary Contents	250
PEG	Measurement Incorporated	Style	Statistical	Trins & Proxes	100 - 400
IEA	Pearson Knowledge Technologies	Content	LSA	Content Style Mechanics	100

2.3 Dataset

Based on previous literature reviews, many studies on (AES) conducted in Asia were carried out using the country's own language or ethnic language. The datasets used in AES research comprise collections of essays that could be sourced from tests, competitions, or daily writing assignments. The process of collecting these datasets is typically time-consuming and expensive since they usually comprise hundreds of human-written essays that must be graded by a human rater before being utilized in AES research. For AES research conducted in English, most researchers use the dataset provided by the Automatic

Student Assessment Program (ASAP) competition. In this section, we examined the benchmark datasets utilized in AES research sourced from ASAP, as well as the datasets utilized under the grading scheme of the Malaysian University English Test.

2.3.1 Malaysian University English Test (MUET)

According to the Malaysian Examinations Council (2019), it has been administering the Malaysian University English Test (MUET) since 1999. 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 and their ability to use and understand English in various situations in the higher education environment. The MUET examinations may also be used to assess the English proficiency of adult English learners, such as teachers and others who must use English in the workplace in the context of education. MUET is offered 3 times a year for local and international students who wish to pursue their first degree in public and private universities in Malaysia. The current MUET syllabus revision is the second since 1999 to ensure that the MUET maintains its status as a test of English proficiency. Based on MUET grading scheme in 2014, MUET tests candidates' four language skills of listening, speaking, reading, and writing, and uses an aggregated scores from 1 to 300 to assess a candidate's English proficiency, which corresponding to Band 1 to 6 (Malaysian Examinations Council, MUET Regulations and Test Specifications, 2014). However, the listening, speaking, reading, and writing test specifications have been revised in 2021 in order to align the MUET with the Common European Framework of Reference (CEFR). Each language skills test has a maximum score of 90, and uses an aggregated scores from 1 to 360 to assess a candidate's English proficiency, which corresponding to Band 1 to 5+ (Malaysian Examinations Council, MUET Specification Regulation, 2019).

Among the four-language skill tests, writing tasks are used to measure candidates' ability to communicate in English in a variety of situations (Malaysian Examinations Council, MUET Specification Regulation, 2019). The tasks are similar to what students should complete in formal and informal situations during university, and the content is appropriate for the purpose of communication. The purpose of the writing test is to stimulate and assess candidates' skills in completing a specific task in communicative behaviours that require the use of English. The writing test consists of two questions and the possible genres cover report, article, letter, and essay. The first question requires the candidate to write interpretation of information based on specific stimuli provided. The second question is our main focus in this study, which requires the candidate to write an essay based on a given topic and discuss ideas, evaluate an argument, or approach a problem and give examples and supporting information such as advice, reasons, opinions, and justifications. The scoring rubric of the MUET essay covers the aspect of accuracy, appropriacy, coherence and cohesion, use of language functions, and task fulfilment. Table 2.2 shows the specification of each scoring rubrics based on the MUET grading scheme 2014.

Table 2.2: Scoring Rubrics of MUET essay

Scoring Rubrics	Test Specifications
Use of Language Functions	Defining, describing, explaining Comparing and contrasting Classifying Giving reasons Giving opinions Expressing relationships Make suggestion and recommendations Expressing agreement and disagreement Persuading Interpreting information from non-linear texts Drawing conclusion Stating & justifying points of view Presenting an argument
Accuracy	Using correct spelling and mechanics Using correct grammar Using correct sentence structures
Appropriacy	Using varied vocabulary & expression Using clear varied sentences Using language appropriate for intended purpose and audience Observing conventions appropriate to a specific situation or text type

Table 2.2 continued

Coherence and Cohesion	Develop and organising ideas Using appropriate markers and linking devices Using anaphora appropriately together with cohesive devices
Task Fulfilment	Presenting relevant ideas Providing adequate content Show a mature treatment of topic

From Table 2.2, we can summarize the MUET scoring rubrics mainly in terms of syntactic and semantic dimension. Syntactic dimensions such as grammatical errors, spelling errors, and use of appropriate vocabulary can reflect the accuracy and appropriateness of the MUET scoring rubric. Based on the study of Wong (2021), the semantic dimension is highly related to the coherence and cohesion rubrics of the MUET scoring rubric and it refers to the ability to understand topic, developing ideas and elaborate ideas which is related to use of language functions and task fulfilment of MUET scoring rubrics. Figure 2.1 shows the mapping of syntactic and semantic dimension with MUET scoring rubrics.

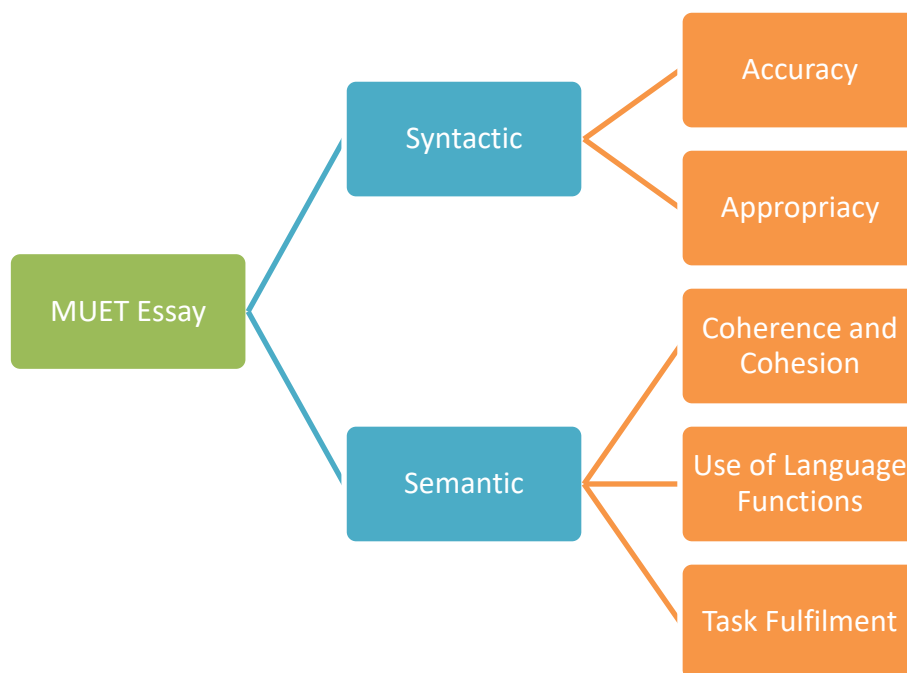


Figure 2.1: Mapping of Syntactic and Semantic Dimension with MUET Scoring Rubrics

2.3.2 Automatic Student Assessment Program (ASAP)

The Automatic Student Assessment Program (ASAP) is an event held in 2012 that brings together researchers and developers in the field of educational technology to showcase their innovative solutions for automatic student assessment. The primary objective of ASAP is to advance the current state of automated assessment methods and tools, while also comparing their efficacy and cost with that of human graders (The Hewlett Foundation: Automated Essay Scoring, 2012). The event also aims to highlight the capabilities of these solutions to state departments of education and other decision-makers interested in adopting them. Participants in the competition are challenged to develop solutions that can automatically grade a variety of essays, and are assessed based on their accuracy, speed, and scalability.

The competition features a total of eight essay sets, each of which was generated from a single prompt. The average length of selected articles is 150 to 550 words per response, with some being dependent on source information while others are not. All of the responses were written by students in grades 7 through 10 and were carefully hand-graded and double-scored. Each of the eight sets of data has its own distinct attributes, providing a diverse range of characteristics designed to challenge the capabilities of scoring engines to their limits. Table 2.3 shows the details of each ASAP datasets.

Table 2.3: Details of ASAP Datasets

Dataset	Essay Prompt	No. of Essays	Average Word Count	Score Range
1	The writers had to write a letter to their local newspaper in which they stated their opinion on the effects computers have on people.	1785	350	1 – 6
2	The writers had to write a persuasive essay reflecting their views on censorship in libraries.	1800	350	1 – 6
3	The writers had to read an extract from Rough Road Ahead: Do Not Exceed Posted Speed Limit by Joe Kurmaskie. They then had to explain how the features of the setting affected the cyclist.	1726	150	0 – 3

Table 2.3 continued

4	The writers had to read an extract from Winter Hibiscus by Minfong Ho. They then had to explain why the author concludes the story in the way that she did.	1772	150	0 – 3
5	The writers had to read an extract from Narciso Rodriguez by Narciso Rodriguez. They then had to describe the mood created by the author with supporting evidence from the extract.	1805	150	0 – 4
6	The writers had to read an extract from The Mooring Mast by Marcia Amidon Lusted. They then had to answer a question about the difficulties faced by the builders of the Empire State Building in allowing dirigibles to dock there.	1800	150	0 – 4
7	Write a story about a time when you, or someone you know, was patient.	1569	300	0 – 3
8	Write a story in which laughter plays a part.	723	650	1 - 6

Chen and He's (2013) study on the public English dataset ASAP demonstrated that their approach achieved a quadratic weighted Kappa of around 0.80 for prompt-specific rating and around 0.78 for generic rating, surpassing the state-of-the-art algorithms. Another research also found that the random forest regression model outperformed the linear regression and support vector regression models on grading ASAP datasets, with a minimum

absolute error of 1.22 (Ghanta, 2019). Moreover, Dong and Zhang's (2016) empirical investigation into a hierarchical CNN model for AES revealed that automatically learned features were competitive with discrete handcrafted features, indicating significant potential for deep learning in AES. Supporting this, Chen and Zhou's (2019) experimental results demonstrated that the automatic scoring model based on CNN+OR achieved high accuracy on the ASAP dataset when compared to the existing system based on manual extraction. Taghipour and Ng (2016) also discovered that long short-term memory networks were more effective than a strong baseline, improving quadratic weighted Kappa by 5.6%, and requiring no feature engineering.

2.4 Features

In this section, features extracted from the assessment of essays are discussed. In neural networks and other supervised machine learning algorithms, features are extremely important since a good set of features can improve the performance of AES systems (Ramesh & Sanampudi, 2022). The AES systems scores student essays based on three diverse types of features, which are syntactic, semantic and hybrid features.

2.4.1 Syntactic Features

Syntactic features, which include part-of-speech, stemmed words, word connectors, word counts, etc., are the essays' surface features. Spelling check, stemming, lemmatization, word segmentation (Loraksa & Peachavanish, 2007), n-gram (Islam & Hoque, 2013; Xu et al., 2016; Chen & Zhou, 2019), and normalisation (Taghipour & Ng, 2016; Ratna et al., 2019a), are prominent syntactic features employed in AES. These surface features of the writings are proven to be helpful in grading essays (Ong et al., 2011). Additionally, efforts to simplify the use of external knowledge bases like WordNet (Omar & Mezher, 2016;

Shehab et al., 2018) and ontology (Contreras et al., 2018) have been reported in order to increase the effectiveness of grading. Wong (2021) asserts that while grammatical and spelling errors are significantly negatively correlated with essay grades, essay syntactic features such as vocabulary count, advanced parts of speech, word count, and average words per paragraph are significantly positively correlated with essay grades. Additionally, previous research suggests that essay syntactic factors influence essay scores more than the semantic aspects (Wong & Bong, 2021).

2.4.2 Semantic Features

The meaning of a word, phrase, sentence, and text are known as semantic features. It is often believed that semantics features are employed to defend the semantic similarity of the entire or partial writings (Landauer et al., 2000; Ghosh, 2008; Islam & Hoque, 2013; Omar & Mezher, 2016; Sendra et al., 2016). AES that is entirely focused on semantic features is becoming popular as research shows that semantic factors produce more accurate grades or ratings. A semantic space in natural language modelling seeks to produce representations of the natural language that can capture the context. The first form of semantic space may be traced back to the Vector Space Model (VSM), which was applied to measure content similarity based on the co-occurrence of words in essays (Al-Jouie & Azmi, 2017).

2.4.2.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA), which is a prominent syntactic-blind semantics feature that is excellent at determining content analysis (Landauer et al., 1998). LSA is a distributional model which utilised to extract meaning from a text. According to Landauer (1998), LSA is a theory and method for extracting and capturing the contextual use meaning

of words using statistical calculations applied to a large corpus of text. Essays are represented by the LSA as a term-document matrix, which is then roughly reduced using singular value decomposition (SVD). The goal of the dimension reduction procedure in LSA is to identify words that are likely to be comparable to one another if they are ever used in other essays in a similar context. The findings of the experiment demonstrated that the inclusion of LSA over syntactic features enhances the scoring capability of the AES system (Omar & Mezher, 2016). LSA or any of its derivatives were frequently employed in contents similarity-based AES systems to derive grades or scores (Ong et al., 2011; Alghamdi et al., 2014; Contreras et al., 2018; Shehab et al., 2018; Amalia et al., 2019; Awaida et al., 2019).

2.4.2.2 Word2Vec

According to Elalfi (2019), Word2Vec is a Google-proposed neural network for processing text data. A new method to document representation has evolved as a result of Google's launch of word2Vec. This adds more semantic features to the text classification process. Word2Vec was also used to learn "word embedding," which is the process of learning word vector representations. Word2Vec is a collection of two learning models, Continuous Bag of Words (CBOW) and Skip-gram, rather than a single method. Skip-gram predicts the context provided by a word, whereas CBOW predicts the word given its context. Finally, by feeding the text corpus into a single learning model, Word2Vec generates the word vectors. Word2Vec learns the vector representations of each word after first building a vocabulary from the training text corpus.

Based on the study, an AES system that used Natural Language Processing, Word2Vec and support vector machine achieved higher level of accuracy which reaches high precision (94%) relative to human resident scores (Elalfi et al., 2019). Another research claims that by using a benchmark dataset of about 12,000 essays, a combination of 30-

manually extracted features, 300-Word2Vec representation, and 768-word embedding features extracted by another language representation model results up to 77.2 ± 1.7 of Kappa statistics for rescaled regression problem and 75.2 ± 1.0 of accuracy value for quantized classification problem (Beseiso & Alzahrani, 2020).

2.4.3 Hybrid Features

The combination of syntactic and semantic features extracted for the purpose of representing and evaluating essays is called hybrid features. The Japanese Grading System (JESS) provides an illustration of how to facilitate syntactic aspects on essay grades based on three syntactic categories: rhetoric, organization, and content (Ishioka & Kameda, 2006). However, they found that the quality of an essay cannot be judged solely on the use of syntax and style. Thus, JESS utilizes both syntactic and semantic features in its content analysis. On the other hand, a hybrid AES for Arabic essays had been proposed and which makes use of three syntactic features (word stemming, word count, and spelling error frequency) and LSA (Alghamdi et al., 2014). The experimental findings highlight the usefulness of LSA for Arabic essay assessment, particularly when paired with other linguistic variables. In accordance with the system, 96.72% of the test data were accurately rated, and the correlation between the automated and manual scores was 0.78, which was nearly as high as the interhuman correlation of 0.7 (Alghamdi et al., 2014).

2.5 Learning Algorithms

According to Merriam-Webster Online Dictionary (2022), an algorithm is defined as “a procedure for solving a mathematical problem (as of finding the greatest common divisor) in a finite number of steps that frequently involves repetition of an operation”. In this section,

the learning algorithms behind AES frameworks are described and the performance of related works on grading essays which involved corresponding algorithms are discussed.

2.5.1 Cosine similarity

Based on the explanation of Han (2012), Cosine similarity is a measure which calculates the similarity between two vectors of an inner product space. It detects if two vectors are pointing in the same general direction by measuring the cosine of the angle between them. Cosine similarity is a commonly used measure in information retrieval and related studies especially in determining document similarity in text analysis (Rahutomo et al., 2012). Cosine similarity is a similarity metric that can be used to compare texts, for example, to rank documents based on a vector of query words. However, Cosine similarity is still unable to fully grasp the text's semantic meaning (Rahutomo et al., 2012). Syntactically implementing cosine similarity measurement between two word-vectors sometimes can produce misleading results. The Cosine similarity of two vector x and y is calculated as (Equation 2.1):

$$sim(x, y) = \frac{x \cdot y}{||x|| ||y||} \quad \text{Equation 2.1}$$

Where $||x||$ is the Euclidean norm of vector $x = (x_1, x_2, \dots, x_p)$, defined as $\sqrt{x_1^2 + x_2^2 + \dots + x_p^2}$. Conceptually, it is the length of the vector. Likewise, the Euclidean norm of vector y is $||y||$. The cosine of the angle between vectors x and y is computed using the measure. A cosine value of 0 indicates that the two vectors are orthogonal which are 90 degrees to each other and do not match. The smaller the angle means the better the match between vectors and the cosine value will be closer to 1.

A study shows that AES systems with cosine similarity achieved 93.3% in accuracy when compared to teachers' manual assessment (Amalia et al., 2019). Besides, another research also found that their proposed AES system with cosine similarity called ABESS achieved 98% accuracy which obtained a higher level of accuracy as compared to human graders (Islam & Hoque, 2013). The study by Ratna et al. (2019c) shows the cosine similarity of their ASE system can reach an average accuracy by adjusting the parameters. Furthermore, a study used cosine similarity AES system to analyze and score the content of high school Filipino essays and achieved 0.863 in overall accuracy metrics (Ong et al., 2011).

2.5.2 Machine Learning

According to IBM Cloud Education (2020), machine learning is a “branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy”. Machine learning is a crucial part of the rapidly expanding field of data science. Algorithms are trained to generate classifications or predictions using statistical approaches, revealing crucial insights in data mining initiatives. Based on the study, supervised learning techniques are used in all AES systems created in the previous 10 years (Ramesh & Sanampudi, 2022). The AES system was considered as a regression or classification task by researchers who employed supervised approaches. The purpose of regression task is to predict the score of an essay while the purpose of classification task is to classify the essays belonging to their suitable grade. The concept of the neural network applied in the development of AES become a popular trend in the last three years (Ramesh & Sanampudi, 2022).

2.5.2.1 Random Forest

A supervised machine learning method built from decision tree techniques is known as a random forest (Mbaabu, 2020). A random forest is a machine learning approach for solving classification and regression issues. It makes use of ensemble learning, which is a technique for solving complicated problems by combining several classifiers. A random forest algorithm is made up of multiple decision trees that determine the outcomes based on the decision trees' predictions. It predicts by averaging the output from various trees, which means that the number of trees used enhances the precision of the result. A random forest method overcomes the drawbacks of a decision tree algorithm, which eventually reduces dataset overfitting and improves precision.

The study by Kobayashi (2016) showed that 60.11% of the predicted output were accurately classified as a result of random forests using the out-of-bag error estimate in assessing second language (L2) spoken English using automated scoring techniques. The random forests model increased prediction by 22.48 points over the baseline accuracy of the simplest algorithm of always picking the most common level which is 37.63%, and the correlation coefficient between predicted scores and human scoring was 0.85 (Kobayashi & Abe, 2016). Besides, the study conducted by Ghanta (2019) indicated that the random forest regression model outperformed the linear regression and support vector regression models when comparing the predictive performance of all three machine learning models with a benchmark score of the mean of Rater 1 and Rater 2 score values. Chen (2013) stated that the proposed Random Forest AES system outperforms state-of-the-art algorithms in scoring a publicly available English essay dataset, the Automatic Student Assessment Award (ASAP), and achieves 0.8014 in quadratic weighted Kappa. These studies offer valid results on the potential use of random forest approach in an AES system in elevating its efficacy.

2.5.3 Deep Learning

Deep learning is a new field of machine learning research, and its architecture has recently made significant progress in various fields which enhance the capabilities of artificial intelligence (Vargas et al., 2017). It is an artificial neural network consisting of several hidden layers. The application of deep learning in the field of AES has recently attracted attention as neural network methods can extract essay specific features from word sequence patterns and achieve high accuracy in scoring essays (Uto et al., 2020).

2.5.3.1 Neural Network (NN)

In terms of computer science, a neural network (NN) is a computational model based on the composition and operation of biological neural networks that operates as an artificial human nervous system for information reception, processing, and transmission (Saxena, 2017). As developers discover better methods to train and adjust neural nets and to eradicate drastically incorrect behaviours that they could produce, neural nets are being studied and used more widely in recent studies (Schwab, 2009). Based on the seminal review study of Lippmann (1987) on artificial neural networks, the current understanding of the biological nervous system serves as the foundation for artificial neural net architecture. Neural net models have the most promise in fields like voice and picture recognition where several hypotheses are investigated simultaneously, fast computation speeds are needed, and the best systems now available are still far from performing at a level with humans.

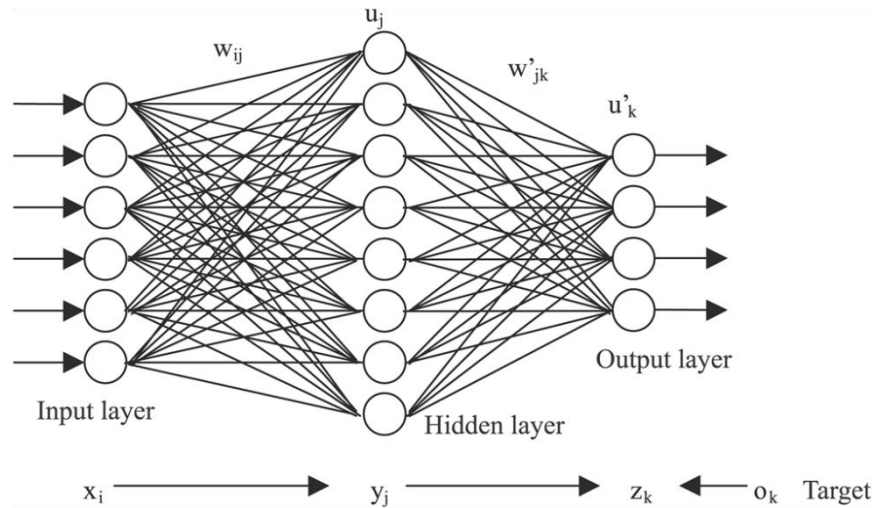


Figure 2.2: Diagram of Neural Network (Hruska, 2018)

Figure 2.2 shows the architecture of a Neural network. The three types of layers that make up a NN are generally input, hidden, and output layers. The purpose of a hidden network is to increase the capabilities of neural networks since without one, their output would simply be a function of their input (Wilczek, 2011). Each node in the input layer represents a neural network input, and each input has a corresponding weight. Utilizing the activation function, the outcome will be determined through the summation of multiplying each input value by its weight. The neural network will provide an output of true or false depending on whether the result is higher or lower than the activation value.

A network must be set up and trained using specifically prepared data that is used as inputs in order to create a neural network system. Structure, learning, and training data are only a few of the elements needed to construct the neural network architecture. The term “structure” describes the arrangement and kind of the neural network that will be created (Schwab, 2009). There are two types of neural networks: feed forward and recurrent. In a feed forward neural network, data in the network flows from the input layer to the output layer in a single direction, but in a recurrent neural network, data flows in two-way communication. In addition, how the neural network is organised pertains to how many

nodes and hidden layers it will employ. Unsupervised learning and supervised learning are the two different learning mechanisms that are used to train neural networks. An unsupervised learning mechanism involves having a programme automatically learn the neural network by statistically analysing the output and adjusting the weights accordingly. A supervised learning mechanism involves the developer providing numerous sets of training data that consist of input-output pairs to the neural network, and it will learn the patterns from the training data (Schwab, 2009).

Based on the study of Nguyen (2016), The 2-layer neural network model with the trained word vectors scored 0.9447875 on the quadratic weighted Kappa metric and outperformed the Long short-term memory neural network on the same dataset. This finding validates the potential of neural network architectures for solving natural language processing problems. The experimental results of another study show that the combination of LSA and ANN effectively simulates human raters under the experimental conditions, and the mean of errors between human and machine scores reaches 0.44 (Loraksa & Peachavanish, 2007).

2.5.3.2 Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is one of the deep learning algorithms which gets its name from the mathematical linear operation between matrices called convolution (Albawi et al., 2017). According to Albawi (2017), the CNN performs admirably in solving the machine learning problems and it obtained excellent result when deal with image data, computer vision, and in natural language processing (NLP). Moreover, when compared to other classification methods, the amount of pre-processing required by a CNN is significantly less (Saha, 2018). In fact, the core of CNN is a pattern-recognition machine which can use vast computational resources to find small, subtle visual patterns that would

otherwise go unnoticed by the human eye (Dickson, 2020). Although the CNN is good in recognizing the patterns of content, Dickson (2020) claims that the CNN performs poorly in understanding the meaning of content.

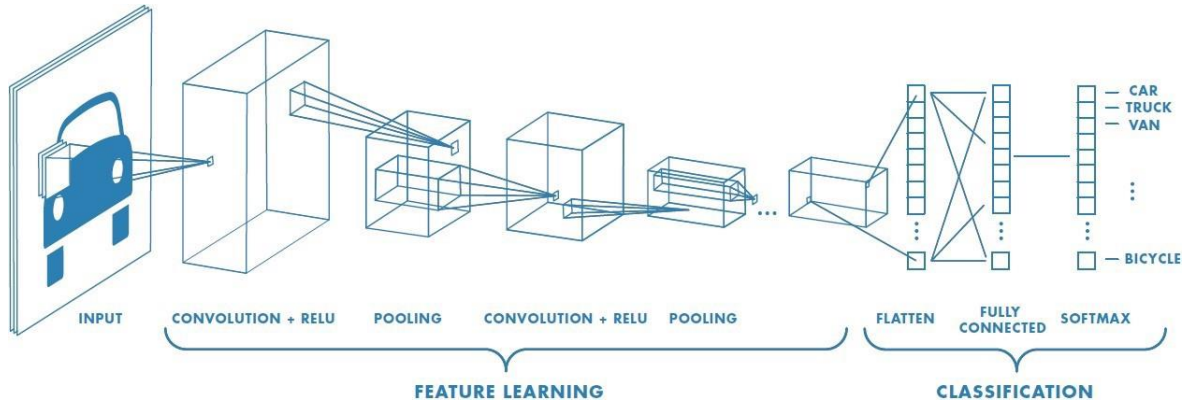


Figure 2.3: Architecture of CNN. Adapted from A Comprehensive Guide to Convolutional Neural Networks (Saha, 2018)

Figure 2.3 shows the architecture of CNN which consists of multiple layers, including convolutional layer, pooling layer, and fully connected layer. A CNN's fundamental building block is a convolutional layer. It contains a collection of filters or kernels whose parameters will be learnt during the training process (Mostafa & Wu, 2021). The filters' height and weight are smaller than input volume and each filter is convolved with the input volume to produce a neuron-based activation map. The filter is slid across the input's width and height, and the dot products between the input and filter are computed at each spatial point. The convolutional layer's output volume is calculated by stacking the activation maps of all filters along the depth dimension (Ke et al., 2018). In most cases, a pooling layer is sandwiched between two convolutional layers. By downsampling the representation, the pooling layer minimises the number of parameters and computations. The maximum or average pooling function can be applied in the CNN but max pooling is often utilised since it is more effective (Ke et al., 2018). After that, the output from the pooling and convolutional layer is a 3-dimensional matrix and it will be flattened to unroll all its

values into a vector (Arc, 2018). The vector will then feed into the fully connected layer which is a feed forward neural networks and form the last few layers in the network (Arc, 2018).

Taghipour's (2016) research suggests using a recurrent neural network to solve the problem of automated essay scoring. The approach does not rely on feature engineering but instead learns the necessary representations on its own. The best model is an ensemble of 10 CNN model instances and 10 LSTM model instances, with a QWK score of 0.761, which is 5.6% higher than the baseline system. Another research investigated a hierarchical CNN model for automatic essay scoring and found that automatically learnt features were competitive with discrete handmade features, with a quadratic weighted kappa score of 0.734 which show that deep learning has a lot of potential in AES (Dong & Zhang, 2016).

2.5.3.3 Long Short-Term Memory (LSTM)

A Long short-term memory (LSTM) network is a type of recurrent neural network (Brownlee, 2017). Unlike the standard feedforward neural networks, LSTM has feedback connections hence it not only processes single input data, but also the entire sequences of input data.

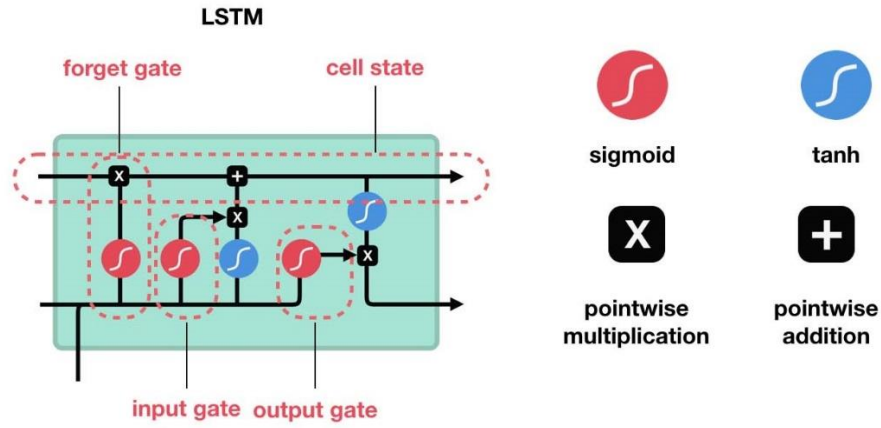


Figure 2.4: Architecture of LSTM (Phi, 2018)

Figure 2.4 shows the core concept of LSTM's are the cell state, forget gate, input gate and output gate (Phi, 2018). As stated by Phi (2018), the cell state is like a memory of the network, it can carry relevant information throughout the processing of the sequence chain. Hence, the networks represent a long-term memory characteristic which can capture information from the earlier time steps and use it in the later time steps. Next, the forget gate decides what information should be forget or keep in the network. Information from the previous hidden state and information from the current input is passed through the sigmoid function. The output of the sigmoid function will be between 0 and 1 which means the information will be forgotten if the output is closer to 0 or the information will be kept in the network if the output is closer to 1. The input gate is used to decide which information is important to be updated in the cell state. The information of the previous hidden state and current input is passed through the sigmoid function and tanh function, respectively. The output of tanh function will be between -1 and 1 to regulate the network and the sigmoid output will decide which tanh output information is important in order to update the cell state. The output gate decides what the next hidden state should be which contains information on previous inputs.

As highlighted by Gers (1999), the standard recurrent neural network fails to learn in the presence of time lags greater than 5 – 10 discrete time steps between relevant input events and target signals. This failure is because recurrent neural networks suffer from the vanishing gradient problem. If a gradient value becomes extremely small, then it will not be able to update a neural networks weight and therefore cannot contribute to learning. However, LSTM is not affected by this problem because it can learn to bridge minimal time lags in excess of 1000 discrete time steps by using mechanisms called gates. Since LSTMs are effective at capturing long-term temporal dependencies without suffering from the optimization hurdles that plague simple recurrent networks (SRNs), they have been used to advance the state of the art for many difficult problems such as language modelling and translation (Greff et al., 2017).

Liang (2018) developed a new model, Siamese Bidirectional Long Short-Term Memory Architecture (SBLSTMA), for autonomously learning text semantics and grading papers by including scoring criteria as part of the input into an AES system. The method outperforms the baseline by around 5% with a quadratic weight Kappa of 0.801. Another study used a benchmark dataset of about 12,000 essays and combined 30-manually extracted features, 300-word2vec representation, and 768-word embedding features with BERT model as the inputs for LSTM approach AES system which results up to 77.2 ± 1.7 of Kappa statistics for rescaled regression problem and 75.2 ± 1.0 of accuracy value for Quantized Classification problem (Beseiso & Alzahrani, 2020).

2.6 Resampling

In order to estimate a population parameter, resampling is a statistical technique that inexpensively creates repeated samples based on original data samples (Brownlee, 2018).

Resampling, in other words, is the process of choosing randomised cases and swapping them out for the created instances from the initial data sample in such a way that each number of the generated sample is similar to the original data sample.

As stipulated by Chawla (2002), Synthetic Minority Over-sampling Technique (SMOTE) is a method of over-sampling in which the minority class is over-sampled by creating synthetic instances rather than over-sampling using replacement, as stated in his article. It generates synthetic instances in a less application-specific manner since it operates in feature space as opposed to data space. The minority class is over-sampled by adding synthetic samples along the line segments connecting any of the k minority class nearest neighbours to each sample from that minority class. Depending on how much oversampling is required, random neighbours are chosen from the k nearest neighbours (Chawla et al., 2002). According to Blagus (2013), SMOTE is used to create a synthetically balanced, or roughly balanced, training set that is utilised to train the classifier. The SMOTE samples are created by combining two similar samples from the minority class (x and x^R) into a linear combination. are defined as (Equation 2.2):

$$s = x + u \cdot (x^R - x) \quad \text{Equation 2.2}$$

with $0 \leq u \leq 1$; x^R is picked randomly from the five minority classes which are closest to x . Most classifiers benefit from SMOTE when the data is low-dimensional since it reduces the bias of the majority class categorization (Blagus & Lusa, 2013).

Based on the experiment results of Filho (2021) in his research, the study found that using SMOTE can result in considerable increases in accuracy for true positives. Because class imbalance is one of the characteristics of essay grading corpora, this approach has shown to be effective for AES and has helped to achieve the requisite accuracies in this

study. He also stated that there are just a few research on the corpora balancing approach in the context of AES in the literature. Rajagede (2021) stated that the data imbalance issue is a challenge in AES system because the number of labels for correct responses is always significantly fewer than the number of labels for incorrect answers and therefore his study proposed SMOTE to solve the data imbalance issues in AES and enhance the model performance. However, the study of Filho (2019) revealed that the SMOTE have shown less impact than the others resampling methods although the balanced corpus applied by SMOTE also shown a considerable improvement in accuracy and average error when compared to unbalanced corpora.

2.7 Evaluation Metrics

In this section, various evaluation metrics that are prevalent in evaluating automated essay grading systems or research will be reviewed. The evaluation metrics are used to measure whether a proposed system can approach the level of human rater in grading essays.

2.7.1 Accuracy

According to past literature, accuracy is one of the popular evaluation metrics that researchers use to evaluate the performance of AES systems in their studies (Amalia, Gunawan, Fithri, & Aulia, 2019; Ratna, et al., 2018; Chen & Zhou, 2019; Wong & Bong, 2019; Contreras, Hilles, & Abubakar, 2018). Accuracy is used to examine the performance of different machine learning methods in predicting essay grades. This evaluation method compares the essay grades manually marked by the human raters against grades predicted by proposed system. The ratio of correct predictions over the total number of predicted essays was evaluated. The accuracy was calculated as (Equation 2.3):

$$Accuracy = \frac{\text{Number of Correct Prediction}}{\text{Total Number of Predicted Essays}} \quad \text{Equation 2.3}$$

2.7.2 Quadratic Weighted Kappa (QWK)

QWK scores were used to examine the degree of agreement between the human raters' grades and the system predicted grades. The output of QWK ranges between -1 and 1 while 1 indicates that the predicted grades are in complete agreement against human rater's, 0 indicates that there is a random agreement and -1 indicates that there is no agreement. In QWK, the result is regarded as bad if its value is lower than 0.4, fair to good if it is between 0.4 and 0.75, and excellent if it is higher than 0.75. (Cheon et al., 2015). Since this evaluation approach computes a confusion matrix between the predicted and actual values, it is appropriate for evaluating the multiclass classification model. These techniques often require a state-of-art scoring model as a baseline to assess the performance of the proposed model. As reiterated by the Hewlett Foundation: Automated Essay Scoring (2012), the quadratic weighted kappa is calculated as follows. First, an $N \times N$ histogram matrix, O is constructed, such that $O_{i,j}$ corresponds to the number of actual values i that received a predicted value j . An N -by- N matrix of weights, w , is calculated based on the difference between actual and predicted values as (Equation 2.4):

$$W_{i,j} = \frac{(i - j)^2}{(N - 1)^2} \quad \text{Equation 2.4}$$

An N -by- N histogram matrix of expected outcomes, E , was calculated assuming that there is no correlation between values. This was calculated as the outer product between the actual histogram vector of outcomes and the predicted histogram vector, normalized such that E and O have the same sum. From these three matrices, the quadratic weighted kappa was calculated as (Equation 2.5):

$$K = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \quad \text{Equation 2.5}$$

2.8 Summary of Related Works

To provide a foundation for this study, previous research on AES in Asia was analysed, and the results are divided into categories depending on the kind of attribute, methodology, prediction model, and findings. The type of attribute is the aspects evaluated by the proposed model in the research which include syntactic, semantic and hybrid. Syntactic attributes focus on linguistic features such as spelling mistake count, essay length and stop word count. Semantic attributes focus on verifying the correctness of content meaning and similarity between an essay with source text and the pre-graded essays. Methodology is the method used to extract the features from the essays. The prediction model is the technique or algorithms used to predict the score or grade of the essay. The results of the research are summarized in Table 2.4 in order to compare the performance of various proposed models.

Table 2.4: Summary of past literature in Asia

Type of Attribute	Methodology	Prediction Model	Dataset	Finding	Reference
Syntactic	Natural language processing (Linguistic features)	Nonparametric Weighted Feature Extraction, stepwise regression, and discriminant analysis	Self-Collected Chinese Dataset	Nonparametric Weighted Feature Extraction achieved best accuracy 51.3%.	(Yu et al., 2017)

Table 2.4 continued

		Linear regression	ASAP Dataset	Close to human rater	(Ramalingam et al., 2018)
		Random Forest	ASAP Dataset	0.8014 on Quadratic Weighted Kappa	(Chen & He, 2013)
			ASAP Dataset	Obtained a minimum absolute error of 1.22	(Ghanta, 2019)
			NICT JLE Corpus	60.11% accuracy and 0.85 correlation coefficient with human rater	(Kobayashi & Abe, 2016)
		Bayesian linear ridge regression	ASAP Dataset	0.784 on Quadratic Weighted Kappa	(Phandi et al., 2015)
Semantic	Latent Semantic Analysis	Neural Network	Self-Collected Thai Dataset	0.44 on mean of error	(Loraksa & Peachavani sh, 2007)
		Cosine similarity	Self-Collected Indonesia Dataset	69.80 % – 94.64 % accuracy for small class and 77.18 % - 98.42 % for medium class.	(Ratna et al., 2007)

Table 2.4 continued

			Self- Collected Indonesia Dataset	83.3% accuracy	(Amalia et al., 2019)
		Learning Vector Quantization	Self- Collected Indonesia Dataset	96.3% accuracy	(Ratna et al., 2018)
	Topic classified by Support Vector Machine (SVM) and assessed by LSA	Frobenius norm	Self- Collected Japanese Dataset	89.175% average accuracy (Japanese)	(Ratna et al., 2019a)
			Self- Collected Indonesia Dataset	72.01% accuracy (Bahasa Indonesia)	(Ratna et al., 2019b)
	GLSA	Cosine similarity	Self- Collected Bangla Dataset	0.98 for Precision, Recall and F1 scores	(Islam & Hoque, 2013)
	Latent Semantic Analysis, Disco2, Damera- levenshtein and N-gram	Similarity degree	Self- Collected Arabic Short Answers Dataset	N-gram achieved 0.82 correlation with rater	(Shehab et al., 2018)

Table 2.4 continued

	Latent Semantic Analysis and Winnowing algorithm	Cosine similarity	Self-Collected Japanese Short Answers Dataset	LSA 87.78% accuracy; Winnowing 86.72% accuracy	(Ratna, et al., 2019c)
	Latent Semantic Analysis with Multi-Level Keywords	Compared document vector	Self-Collected Indonesia Dataset	86% human raters' agreement	(Ratna et al., 2015)
	Modified LSA and syntactic features	Cosine similarity	Arabic Dataset	0.268 of Root Mean Square Error (RMSE)	(Omar & Mezher, 2016)
	Enhanced Latent Semantic Analysis	Cosine similarity	Self-Collected English Dataset	0.242 gap with rater	(Sendra et al., 2016)
	Arabic WordNet	Cosine similarity	Self-Collected Arabic Dataset	98% Pearson Corelation	(Awaida et al., 2019)
	Concept Indexing	Cosine similarity	Self-Collected Filipino Dataset	0.452 Exact Agreement Accuracy	(Ong et al., 2011)

Table 2.4 continued

	Contextualized Latent Semantic Indexing	SVM	Self-Collected Chinese Dataset	89.67 Rating Agreement	(Xu et al., 2016)
	Statistic (One-hot encoding)	CNN, LSTM and Bi-directional Long Short-Term Memory (BiLSTM)	ASAP Dataset	LSTM + CNN achieved 0.761 QWK score	(Taghipour & Ng, 2016)
	Statistic (Word embedding)	CNN	ASAP Dataset	0.734 on average Quadratic Weighted Kappa	(Dong & Zhang, 2016)
		CNN and ordinal regression (OR)	ASAP Dataset	82.6% accuracy	(Chen & Zhou, 2019)
		Siamese Bidirectional Long Short-Term Memory Architecture (SBLSTMA)	ASAP Dataset	0.801 on average Quadratic Weighted Kappa	(Liang et al., 2018)
	GloVe (Word embedding)	2-Layer Neural Networks	ASAP Dataset	0.9448 on Quadratic Weighted Kappa	(Nguyen, 2016)
	Word2Vec (Word embedding)	SVM	Self-Collected English Dataset	94% accuracy	(Elalfi et al., 2019)
	Hierarchical recurrent neural network	CNN, LSTM, BiLSTM	ASAP Dataset	0.773 on average Quadratic Weighted Kappa	(Chen & Li, 2018)

Table 2.4 continued

	Natural language processing (Unigram Language Model)	Machine Learning Classifier	Self-Collected MUET English Dataset	51.5% of mean accuracy	(Wong & Bong, 2019)
	Natural language processing (Word features, syntactic features, and dependency relation features)	Logistic regression and k-Nearest Neighbors	Self-Collected Korean Dataset	0.92 correlation with rater	(Cheon et al., 2015)
Hybrid	Natural language processing (Rhetoric, Organisation, Content)	Rule-based expert system	Self-Collected Japanese Dataset	0.57 correlation with rater	(Ishioka & Kameda, 2006)
	Vector Space Models (VSM)	Support Vector Regression (SVR)	Self-Collected MHK Chinese Dataset	Average correlation is 0.6107	(Peng et al., 2010)
	Natural language processing (Rhetoric, Organisation, Content and Length)	Rule-based expert system	Japanese Dataset from Taiyaku Database	0.562 Pearson's correlation coefficient	(Imaki & Ishihara, 2013)

Table 2.4 continued

	Latent Semantic Analysis, number of words, number of spelling mistakes, and word distance.	Linear regression	Self- Collected Arabic Dataset	96.72% accuracy and 0.78 correlation with rater	(Alghamdi et al., 2014)
	Latent semantic features	Support Vector Machine for Ranking	ASAP Dataset	Slightly increase the accuracy	(Jin et al., 2018)
	Latent Semantic Analysis, rhetorical structure theory (RST) and hand-crafted features	Rule-based expert system	Self- Collected Arabic Dataset	78.33% accuracy	(Al-Jouie & Azmi, 2017)
	Latent Semantic Analysis and Feature extraction	Linear Regression	Self- Collected English Dataset	47.16% accuracy	(Contreras et al., 2018)
	Linguistic features, Word2Vec representation and BERT word embedding	LSTM	ASAP Dataset	77.2 $\pm$ 1.7 of Kappa statistics and 75.2 $\pm$ 1.0 of accuracy	(Beseiso & Alzahrani, 2020)

2.9 Conclusion

This chapter has provided a comprehensive overview of the current state of research on Automated Essay Scoring (AES). The review has covered various topics, including the mechanisms and drawbacks of well-known commercial AES systems. This chapter also reviewed the benchmark datasets used in AES research obtained from ASAP, as well as the grading scheme of the Malaysian University English Test. In the previous 10 years, supervised learning techniques were used in all extant AES systems created because an AES system was considered as a regression or classification task by researchers who employing supervised approaches. The review of literature shows that the concept of the neural network applied in the development of AES systems become a popular trending since the last three years and indicating significant potential for deep learning in AES. Based on the study, the essay syntactic features have a greater impact on essay scores when compared to essay semantic features. The essay syntactic features such as vocabulary count, advanced part of speech, word count, and average words per paragraph are significantly positive-correlated with essay grades when compared to other features, whereas grammatical and spelling errors are significantly negative-correlated with essay grades. Furthermore, this chapter includes a review of various evaluation metrics commonly used in the assessment of AES systems or research. With regards to resampling methods, SMOTE seems promising in solving imbalanced dataset issue since many studies claimed that SMOTE can offer a considerable improvement in accuracy and enhance the performance of model after applied to AES system. This literature review will serve as a foundation for the development of an AES system tailored for the Malaysian University English Test.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter outline all relevant experimental, descriptive, theoretical and analytical techniques used in the research. The chapter begins with an overview of the proposed AES framework, followed by detailed explanations of each step in our methodology used for this study in corresponding sections. This chapter also provides a clear understanding of the methodology used, allowing other researcher to follow the process step by step, starting with the big picture and then narrowing down to the details.

3.2 Proposed AES Framework

This study attempts to develop a grading system for MUET essays and proposed three AES frameworks in grading essay primarily from two aspects, namely the syntactic and semantic dimensions in order to meet such wide-scope and occasionally ambiguous MUET essays scoring rubrics. One of the proposed frameworks called syntactic AES framework aims to evaluate accuracy and appropriacy of essay using syntactic features such as number of grammar errors, number of spelling errors, use of appropriate vocabulary. The other called semantic AES framework aims to evaluate coherence and cohesion, use of language functions, and task fulfilment of essay based on semantic features such as the ability to understand topic, developing ideas and elaborate ideas. Both of the frameworks can be combined as a new hybrid AES framework to evaluate essays with both syntactic and semantic features. This section provides an overview of the proposed AES framework to give a comprehensive understanding and a general idea of our approach. The subsequent sections in this chapter will detail each step in the framework.

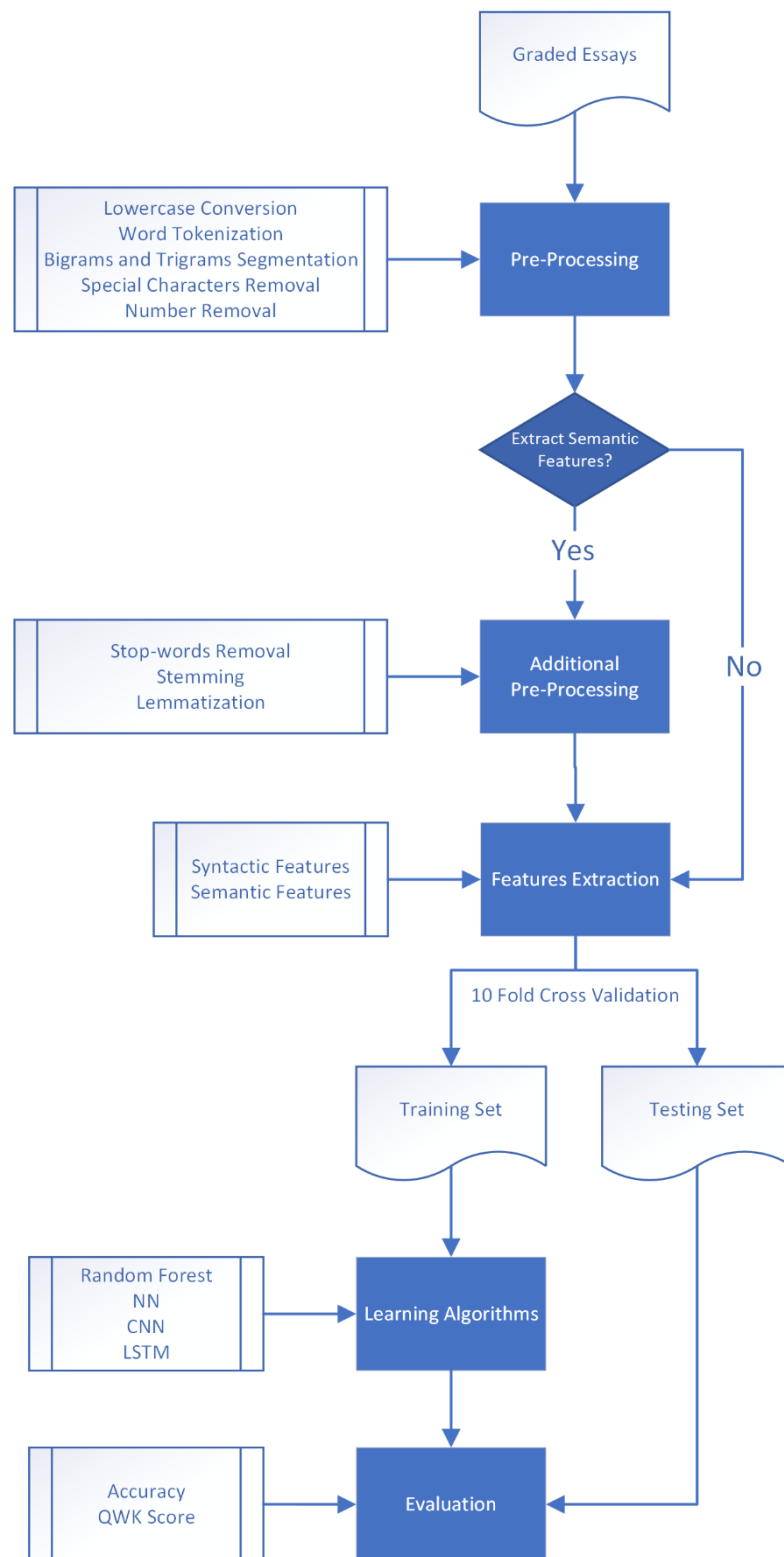


Figure 3.1: Proposed AES Framework Flow Chart

The flow chart of proposed AES framework is illustrated in Figure 3.1, providing a high-level overview of the process. Initially, the collected human-graded essays undergo pre-processing to remove noise. Additional pre-processing steps such as stop-word removal, stemming, and lemmatization are performed for the semantic framework. Next, the significant difference between each proposed framework is the feature extraction process. For the syntactic framework, 14 syntactic features of essay such as essay length and number of spelling mistakes are extracted, while the semantic framework extracts 300-dimension semantic features using LSA or Word2Vec methods. The hybrid framework extracts both syntactic and semantic features. After the feature extraction process, essays are separated into training and testing sets. The training set is utilized to train the machine learning model, and the testing set is used to evaluate the performance of the proposed framework. A 10-fold cross-validation is utilized to avoid overfitting. In the proposed AES frameworks, essay grading is treated as a multiclass classification problem in which each grade is treated as a separate class. The machine learning algorithms are trained using syntactic or semantic features as input to classify essays into corresponding grades. Finally, the results are evaluated using two popular metrics in AES, accuracy, and QWK scores. The details of each step in the framework will be described in each corresponding section in this chapter.

3.2.1 Syntactic AES Framework

The idea of this framework is to assign grades to new essays based on the syntactic essay features in order to fulfil the accuracy and appropriacy of the MUET grading scheme. The linguistic attributes of essay writing such as lexical richness, parts of speech, spelling error, and grammatical error are referred as syntactic features. It also had to do with the essay's length and count, such as the number of words, sentences, paragraphs, and average

words per phrase. These essays' surface features are found to be useful in grading essays (Ong et al., 2011).

The collected graded essays first underwent the pre-processing steps including lowercase conversion, word tokenization, bigrams and trigrams segmentation, special characters removal and number removal to reduce noises in data. Next, the pre-processed essays went through the feature selection process to extract the syntactic features as shown in section 3.5 which had been proven to have a significantly positive-correlated with essay grades (Wong & Bong, 2021). The syntactic features of essays then acted as input data of machine learning algorithms and used to predict the essay grade. Finally, the performance of the model is evaluated using accuracy and QWK scores, which are common metrics to assess classification models.

3.2.2 Semantic AES Framework

For semantic SES framework, the idea of this framework is to fulfil the coherence and cohesion, use of language functions, and task fulfilment of MUET grading scheme which assign grades to new essays based on how close the similarity of the essay to a gold standard: a collection of manually graded essays, covering all spectrum of grades on the respective topics. The meaning of a word, phrase, sentence, and text are referred to as semantic features. It is often believed that the semantic similarities between the entire or partial essays are justified using the semantic features.

The collected graded essays were subjected to pre-processing steps, including lowercase conversion, word tokenization, bigram and trigram segmentation, special character removal, and number removal to reduce data noise. To extract semantic features, additional pre-processing steps were carried out, including stop word removal, stemming,

and lemmatization. Following this, the pre-processed essays underwent feature selection to extract semantic features using either LSA or a word2Vec pretrained model. These semantic features then acted as input data for machine learning algorithms and were used to predict the essay grade. Finally, the model's performance was evaluated using accuracy and QWK scores to assess its essay grading capabilities.

3.2.3 Hybrid AES Framework

The hybrid framework combines both the goodness of syntactic and semantic framework which cover all scoring rubrics of MUET to derive essay grades or scores. Although the hybrid framework treats word representation and feature selection as the inputs of machine learning prediction model, both would contribute different weights on predicting essay scores. According to Peng (2010), Intelligent Essay Assessor (IEA) which was developed in the late 1990s, evaluates essays by measuring three types of features: content, style, and mechanics. However, it focuses more on content rather than other features because the content score is always given the greatest weight. The workflow of the hybrid framework is similar to syntactic framework, with an additional step where essays are transformed into semantic representation using the same feature extraction method as the semantic framework. These features are then fed into the machine learning algorithms to serve as supplementary semantic features. After that, both syntactic features and semantic features acted as input data of machine learning algorithms and used to predict the essay grade. Then, Accuracy and QWK scores are used to evaluate the performance of the model.

3.3 Data Collection

This study focused on proposing a new AES framework and evaluate its performance with different machine learning algorithms in grading MUET essays. Instead of using any

publicly available essays, MUET essays written by the students were collected and used in the training and testing processes. For our experimentation, we worked with SMK Green Road in Kuching, Sarawak, and Universiti Malaysia Sarawak to collect the essays. All the collected essays were written by Form Six students and pre-university students, and then graded by trained teachers to ensure the quality of essay meets the standard of MUET. Although the MUET grading scheme was revised in 2021, however, at the beginning of this study, the collected essay datasets and the conducted experiments were based on the old version of the MUET grading scheme, hence this study will focus on the MUET grading scheme in year 2014. The collected essays were graded according to six bands from Band 1 to Band 6 in which Band 1 is the lowest and Band 6 is the highest. These bands were used during the time the study was conducted before the introduction of the newer but equivalent bands (Band 1 to Band 5+) in 2021 by the Malaysian Examination Council (Malaysian Examinations Council, MUET Specification Regulation, 2019).

In this study, one of the datasets came from the Automated Student Assessment Prize (ASAP). ASAP is a series of competitions sponsored by William and Flora Hewlett Foundation (The Hewlett Foundation: Automated Essay Scoring, 2012), where competitors compete to demonstrate their ability to automate the essay grading process. The ASAP dataset was used to evaluate our proposed AES framework, as it is an internationally recognized dataset of essays and is frequently used in many studies related to AES (Mathias & Bhattacharyya, 2018). The high-quality grading of the ASAP dataset can also help us identify the differences between the international and local essay datasets and evaluate the performance of our proposed AES framework on both datasets. The average length of essays in the ASAP dataset is 350 words and the grade is between 1 to 6, similar to the collected MUET datasets, 1 is the lowest grade and 6 is the highest grade. The corresponding essay

titles with the grade distribution of the essay datasets are shown in Table 3.1, Figure 3.2, Figure 3.3 and Figure 3.4.

Table 3.1: Essay Title for Each Dataset

Dataset No.	Source	Topic
1	SMK Green Road	The imbalance between the number of boys and girls pursuing university education creates social problems. To what extent is this statement true? Discuss.
2	SMK Green Road	The trend today among youths is to own the latest technological gadgets such as mobile phones and other devices. Is this a healthy development among young people? Discuss.
3	ASAP	Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.

Table 3.1 shows the source and essay titles of datasets used in this study. Dataset 1 and Dataset 2 are local essay datasets collected from SMK Green Road Form Six students and all essays have been graded by their Secondary English teachers. Dataset 3, one of ASAP's essay sets, is an international recognition dataset that will be used to validate and examine our proposed AES framework's performance.

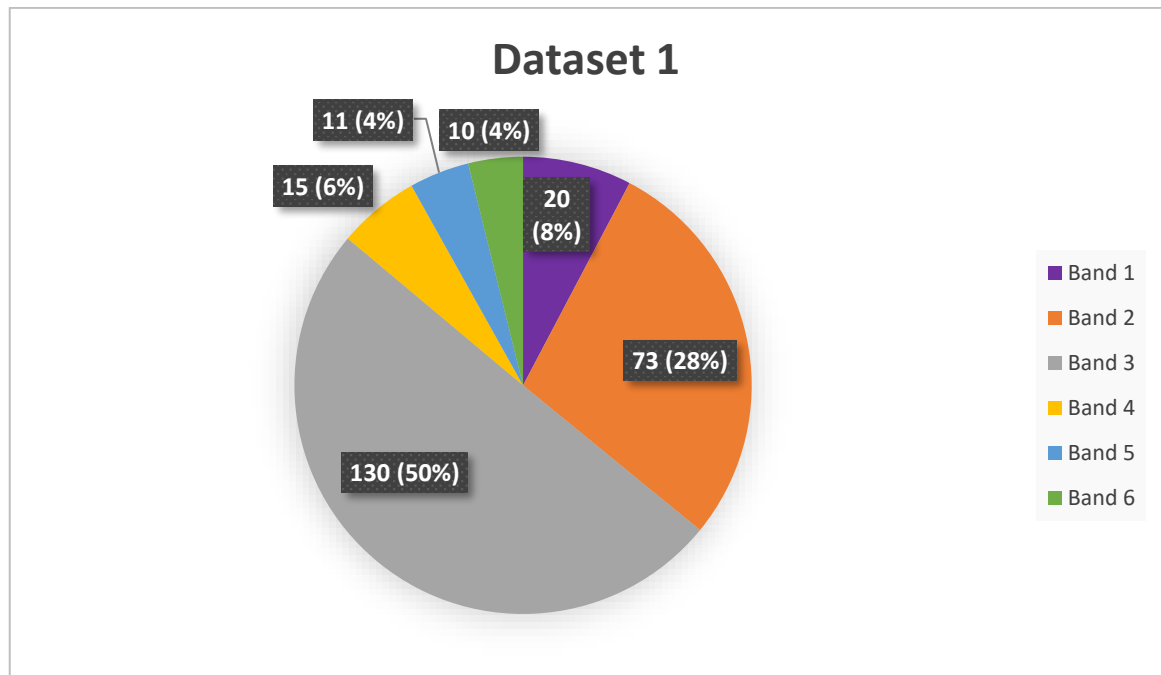


Figure 3.2: Pie Chart of Band Distribution in Dataset 1

Figure 3.2 shows the pie chart of band distribution in Dataset 1. It consists of 259 essays and 50% of them or 130 essays are of Band 3. The second highest frequency is Band 2 which are 28 % of them or 73 essays. Besides, Band 1, Band 4, Band 5, and Band 6 have almost similar proportions in this dataset which are 8%, 6%, 4% and 4% correspondingly. Furthermore, we can see that the students who participated in writing essays in this dataset generally had limited English proficiency, as 86% of the students achieved a grade of Band 3 or below.

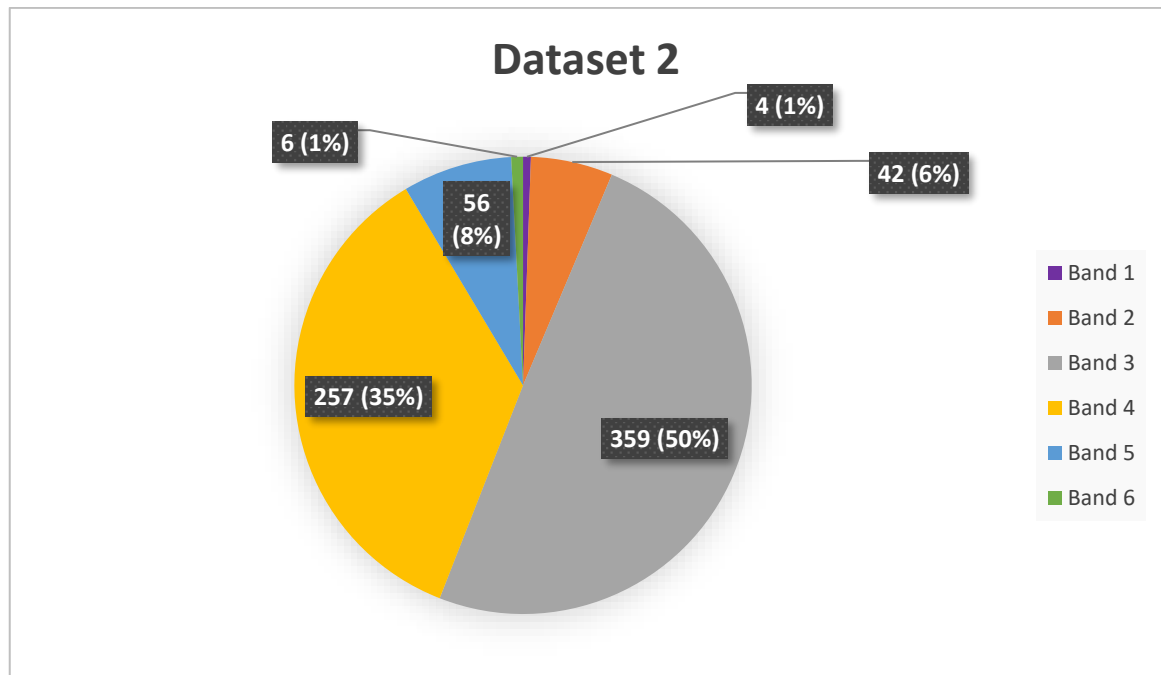


Figure 3.3: Pie Chart of Band Distribution in Dataset 2

Figure 3.3 shows the pie chart of band distribution in Dataset 2. It consists of 724 essays and 50% of them or 359 essays are Band 3. The second highest frequency grade is Band 4 which are 35 % of them or 257 essays. 42 essays were graded as Band 2, representing 6% of dataset while 56 essays were graded as Band 5 which represents 8% of dataset. From the figure, we can see that the English proficiency of the students who involved in this dataset is decent since most of the students' essay obtained Band 3 or Band 4. However, this dataset has the issue of imbalanced distribution of grades, as it only contains 1% of Band 1 and Band 6, so it may be a challenge for the proposed AES framework to correctly predict the full range of grades on essays due to the lack of sample of Band 1 and Band 6.

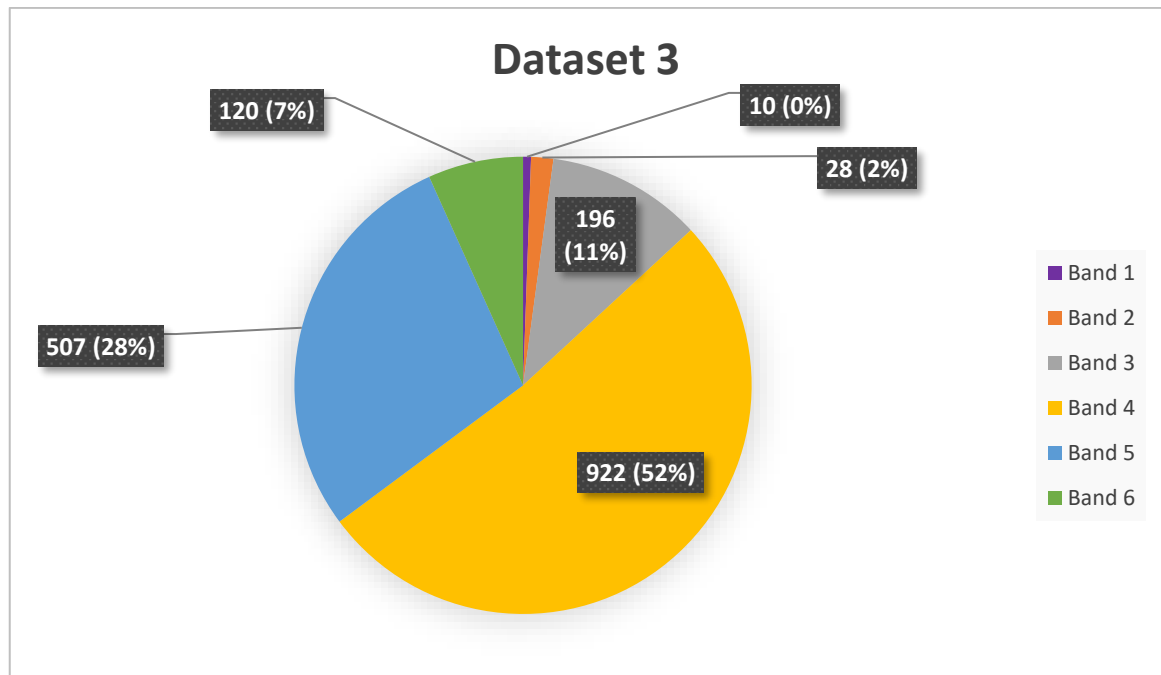


Figure 3.4: Pie Chart of Band Distribution in Dataset 3

Figure 3.4 shows the pie chart of band distribution in Dataset 3, which was obtained from the Automated Student Assessment Prize (ASAP) datasets. It consists of 1783 essays and 52% of them or 922 essays are Band 4. The second highest frequency is Band 5 which are 28 % of them or 507 essays. 196 essays were graded as Band 3 which accounted for 11% of the dataset while 120 essays were graded as Band 6 (7%). Furthermore, we can see that the students in this dataset generally had decent English proficiency, as 87% of the students achieved a grade higher than Band 3. From the pie chart, we can see that Dataset 3 has a more serious issue of imbalanced grade distribution than Dataset 2, as the dataset contains not more than 1% of Band 1 and only 2% of Band 1. Therefore, due to the lack of samples from Band 1 and Band 2, it may be a challenge for the proposed AES framework to correctly predict the grade of the essay.

Table 3.2: Details of Features for Each Dataset

Dataset No.	Word Court	Vocabulary Court	Lexical richness	Spelling Error	Grammar Error	Punctuation Types
1	415	156	0.39	13	16	4
2	444	184	0.43	25	15	4
3	369	162	0.45	26	17	6

Table 3.2 provides a summary of the average values of the features for each dataset. The word court feature indicates the average number of words per essay in each dataset, while the lexical richness feature is calculated by dividing the word court feature with the vocabulary court feature. The data in the table reveals significant differences between the local datasets (Dataset 1 and Dataset 2) and the international benchmark dataset which is Dataset 3. The average word count for Dataset 1 and 2 are 415 and 444 respectively, which are higher than the 369 words per essay for Dataset 3. However, the lexical richness of local datasets is lower than that of Dataset 3. Surprisingly, in terms of errors, the local datasets have fewer spelling errors and grammar errors compared to Dataset 3. Dataset 3 has the highest number of average punctuation types per essay, which is 6, compared to Dataset 1 and 2 which both have an average of 4 types of punctuation per essay.

Overall, the data suggests that there are differences in the characteristics of the local datasets compared to the international benchmark dataset. The local datasets have higher word court but lower lexical richness, as well as fewer spelling errors. The differences in the punctuation types and grammar errors suggest that the local datasets may have different writing styles or conventions compared to the international benchmark dataset.

3.4 Data Pre-Processing

The English essays collected from the SMK Green Road and UNIMAS were in handwritten form. To use the collected essays as our input for our proposed AES frameworks, we had to convert the essays into a digital form by retyping them in Microsoft Word. During this process, we cross checked the converted essay content to make sure it is same as the actual content including the grammatical errors and spelling errors. After that, essays acted as the input data and underwent several steps including lowercase conversion, word tokenization, bigrams and trigrams segmentation, special characters removal and number removal to reduce noises in data. For semantic framework, the essays went through additional pre-processing steps which are stop-words removal, stemming and lemmatization. These additional pre-processing steps can remove unnecessary or meaningless words in the essay, thereby reducing the noise in the essay and benefiting the extraction of semantic features because the semantic framework only focuses on the content or ideas discussed in the essay. However, syntactic frameworks cannot perform these additional pre-processing steps, as it would reduce the integrity of the essay and affect the extraction of syntactic features. For example, spelling and grammar errors are removed after these additional pre-processing steps. After pre-processing, the essay data was prepared and readied to be fed into our feature extraction process of proposed AES framework.

3.5 Features Extraction

Table 3.3 shows the selected syntactic features that are extracted from the essay. Based on Wong's (2021) study, it is demonstrated that the corresponding selected syntactic features are significantly highly correlated with the essay grades. From their research, they found that syntactic features such as number of vocabulary and number of parts of speech were positively correlated with essay grade, while syntactic features such as grammatical

error rate and spelling error rate were strongly negatively correlated with composition performance. Furthermore, the top five highly relevant syntactic features, such as vocabulary size, number of parts of speech, number of words, number of paragraphs, and grammatical error rate, produced better accuracy results on their prediction models. Therefore, the selected syntactic features will be used in our proposed AES framework as input to the deep learning method algorithms and the performance will be examined.

Table 3.3: Selected Syntactic Features

Features	Description
word_count	Number of Words
vocab_count	Count of Distinct Words
lexical_richness	Vocabulary Count divided by Word Count
sentence_count	Number of Sentences
word_sentence_ratio	Word Count divided by Sentence Count
spell_err	Number of Spelling Errors
spell_err_ratio	Spelling Error divided by Word Count
grammar_err	Number of Grammatical Errors
grammar_err_ratio	Grammatical Error divided by Word Count
adv_tag_count	Number of Advance Parts of Speech Tags Including Adjective, Adverb, Past Participle Verb and Gerund Verb
adv_tag_count_ratio	Advance Tag Count divided by Word Count
para_count	Number of Paragraphs
POS_count	Total Number of Parts of Speech in Essay
punctuation_count	Total number of punctuation types

To extract semantic features of the essays, the collected graded essays first underwent the pre-processing steps including additional pre-processing steps such as stop-words removal, stemming and lemmatization. Next, two methods were used to extract the semantic features separately from the essay, namely Latent Semantic Analysis (LSA) and Word2Vec. LSA is a topic modelling technique which was used to compute the content similarity of topic of each essay (Landauer et al., 1998). For this study, 300 topics will be generated from each dataset, and the content similarity score of 300 topics of essays then acted as input data of machine learning algorithms and used to predict the essay grades. For Word2vec, it is a Google-proposed neural network that used to process text data. The essay was then vectorized into multi-dimension semantic vector space when fed into one of the pre-trained Word2Vec called “word2vec-google-news-300” which contains 300-dimensional vectors for three million words and phrases. This step added more semantic features to the text classification process. The vector of essays then functioned as input data for machine learning algorithms and used to predict the essay grade.

3.6 Learning Algorithms

The proposed neural network model in this study consists of a specific number of nodes dedicated to the input features. For syntactic framework, the input layer comprises 14 nodes to represent 14 different syntactic features. For semantic framework, 300 input nodes are utilized to represent the 300-dimensional semantic features. The model also includes 4 hidden layers with 20 nodes in each layer, which enables the neural network to learn complex relationships among the input features. For the output layer, the model has 6 nodes, each corresponding to a grade of the essay. To classify the output, the activation function used in the output layer is softmax, which allows the model to predict the probability of each essay

grade. The loss function is categorical cross-entropy, and the Adam optimizer is used to optimize the weights and biases of the model.

For convolutional neural network (CNN) model was used, the first layer being a 1D convolution layer that creates a convolution kernel with a specific number of input layer dedicated to the features. The convolution layer has 128 output filters, and a kernel size of 5 is used to specify the length of the 1D convolution window. The activation function used in the convolution layer is relu. Following the convolution layer, a pooling layer is used to down-sample the input representation, using the Max pooling operation for 1D temporal data. A flatten layer is then used to flatten the input and send it to the output layer, which contains 6 nodes corresponding to each essay grade. To optimize the weights and biases of the model, the Adam optimizer was employed. The output layer utilized the softmax activation function and the categorical cross-entropy loss function used to calculate the difference between the predicted and actual values.

The Long Short-Term Memory neural network (LSTM) model in this study consists of an embedding layer for input of a specific number dedicated to the features. This is followed by a LSTM layer with 128 internal units, a Dropout layer with a frequency of 0.5 during training to prevent overfitting, a Dense layer with 100 units, and an output embedding layer of size 6 nodes representing each essay grade. The model is optimized using the Adam optimizer to adjust the weights and biases. The loss function used in the model is categorical cross-entropy, while the activation function used in the output layer is softmax.

3.7 Resampling

From the data collection section, we can see that the datasets used in this study have serious issues of imbalanced grade distribution due to the lack of samples from Band 1 and

Band 6. To overcome the imbalanced grade distribution, a resampling method called SMOTE was applied to the dataset to over-sample minority classes.

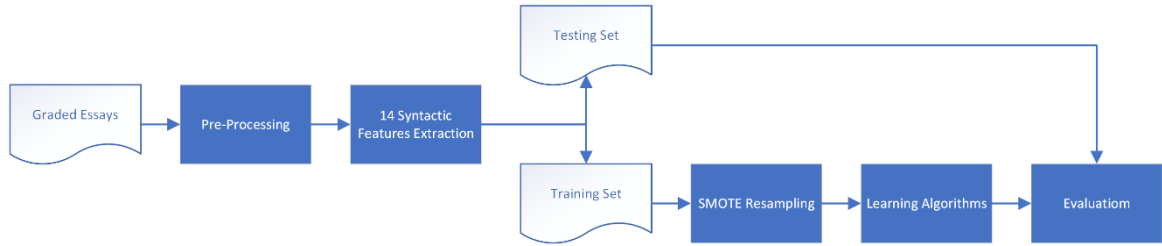


Figure 3.5: AES Framework based on Syntactic features with Resampling

Figure 3.5 shows the procedure of AES Framework based on Syntactic features with SMOTE. After the graded essays were pre-processed, the syntactic features were extracted from essays and classified into training set and testing set. Next, the SMOTE resampling method was applied to the training set to balance the grade distribution while the testing set remained as unseen data to evaluate the performance of prediction model after resampling. The 10-fold cross-validation was applied to evaluate the performance of the prediction model on the dataset.

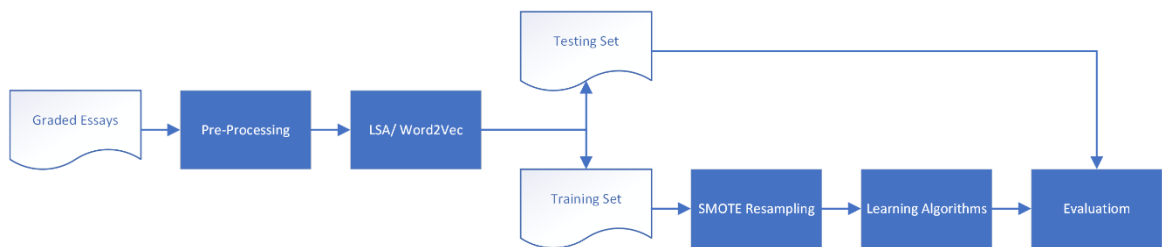


Figure 3.6: AES Framework based on Semantic features with Resampling

Figure 3.6 shows the procedure of AES Framework based on Semantic features with SMOTE resampling. To reduce the noise in essays, graded essays were pre-processed. The essays were then vectorized and classified into training and testing sets using LSA or Word2vec. The SMOTE resampling method was used only to the training set in order to balance the dataset's grade distribution, while the testing set remained as unseen data. The

resampled training set was then used to train machine learning prediction model and the 10-fold cross-validation was used to evaluate the prediction model's performance on the testing set.

3.8 Evaluation

In this study, Accuracy and QWK score were used to examine the experiment results. To ensure the reliability of the results, 10-fold cross validation was applied on both evaluation metrics and the performance of our proposed frameworks in predicting the grade for essay was evaluated. For the 10-fold cross-validation, 90% of the dataset is randomly selected as training data and the remaining 10% as test data for validation, and the process will be repeated a total of ten times in order to use the entire dataset in the validation process. Accuracy and QWK scores are popular evaluation metrics in the field of AES which used to assess the performance of a classification model. The accuracy metric simply measures the percentage of correctly classified essays from the total essays evaluated. On the other hand, QWK score measures the agreement between the model's predicted grades and human rater grades, taking into account the possibility of random agreement.

A high accuracy score indicates a good performance of the model in terms of correctly classifying the essays. However, accuracy alone may not always provide a complete picture of the model's performance as it doesn't consider the level of agreement between the predicted grades and the human rater grades. In contrast, a high QWK score indicates that the model's predicted grades are in agreement with human rater grades beyond what would be expected by chance. Therefore, both metrics are necessary to evaluate the performance of an AES model effectively.

3.9 Conclusion

This chapter explains our research methodology and the proposed AES frameworks. To evaluate performance of the proposed AES frameworks with different machine learning algorithms in grading MUET essays, three MUET essay datasets were collected. The datasets were then converted into the digital form. After the pre-processing steps to reduce the noise in dataset, the datasets were used as inputs for proposed AES framework. There are three proposed AES frameworks in this study which are syntactic, semantic and hybrid AES framework. To overcome the imbalanced grade distribution issue in the datasets, a resampling method called SMOTE was applied to the datasets to investigate the influence of resampling method toward the performance of the AES framework. The experiment results were examined using two evaluation metrics: Accuracy and QWK. The efficacy of our proposed framework in predicting the unseen essay grade was examined using the 10-fold cross validation on both evaluation measures to assure the reliability of the results.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Overview

This chapter presents the experimental results as well as explaining the performance of our proposed AES frameworks based on syntactic features, semantic features, and both. The applicability of evaluation metrics on AES to investigate which evaluation metric is able to provide more valuable information when evaluating performance of AES systems was also discussed. Moreover, we also compare the results of machine learning algorithms and deep learning algorithms to study which algorithm can perform better on AES task. Finally, we present the influence of dataset on performance of AES and the solutions to overcome the imbalanced grade distribution in dataset.

4.2 Predictive Performance of the Proposed AES System

In this subsection, we discuss the experimental findings and provide an explanation of the effectiveness of our suggested AES framework based on syntactic, semantic, and combined features. We also explore if the AES framework has the best chance of being applied in practise while comparing each one's performance.

4.2.1 Prediction Results based on Syntactic Feature

In this section, we evaluate the performance of our proposed syntactic framework with different machine learning algorithms using accuracy and QWK scores. The syntactic features show in Table 3.1 were derived from the datasets and acted as the input for the prediction model. The 10-fold cross-validation was used in the training process to estimate

the performance of the prediction model on new data. Table 4.1 shows our results of the AES system based on syntactic features.

Table 4.1: Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Syntactic Features

	Accuracy				QWK			
	1	2	3	Mean	1	2	3	Mean
Random Forest	0.61	0.45	0.67	0.58	0.73	0.05	0.71	0.50
NN	0.61	0.48	0.66	0.58	0.70	-0.02	0.70	0.46
CNN	0.57	0.45	0.65	0.56	0.71	0.03	0.68	0.47
LSTM	0.58	0.50	0.66	0.58	0.68	0.00	0.69	0.46

To examine the performance of each machine learning algorithm on different datasets in more detail, we will focus on the mean results of each machine learning algorithm because it can show the ability of each algorithm to handle different datasets and use them to predict essay grades. As shown in Table 4.1, Random Forest, NN and LSTM had the same highest accuracy of 0.58. When evaluated by QWK, Random Forest shows highest performance and achieved 0.50. Based on the result, the performance between machine learning algorithms did not show a significant difference. All machine learning algorithms performed well in datasets 1 and 3, with QWK results between 0.68 and 0.73, which can be considered fair to good as explained by Cheon (2015). It also shows that our proposed AES framework performed well in essay grading on both local and international essay datasets and demonstrates that the syntactic features extracted from essays were sufficient to represent and distinguish different essay grades. However, all machine learning algorithms performed poorly on dataset 2 with low QWK scores. When dealing with dataset 2, the

machine learning algorithms were noted to predict all essays to a majority grade (Band 3 and Band 4) in the dataset and in turn caused the accuracy for dataset 2 to be about 0.5 but the QWK score was close to zero.

Based on the review of literature presented in section 2.8, it can be concluded that Nguyen's (2016) research has demonstrated the highest performance in grading the ASAP dataset with an QWK score of 0.9448. However, the proposed syntactic framework utilized in this study has attained a QWK score of 0.71 when grading Dataset 3, which is the ASAP dataset. Although this score is slightly lower than the previous best performance, Cheon (2015) has explained that a QWK score of 0.71 can be considered as fair to good. The difference in scores may be attributed to the different techniques and methodologies utilized in each study and the proposed syntactic framework presented in this study may have the potential to improve with further refinement and optimization.

4.2.2 Prediction Results based on Semantic Features

In this section, we evaluated the performance of our proposed semantic framework with different machine learning algorithms using accuracy and QWK scores. LSA and Word2Vec were used to vectorize the essays representing semantic content. LSA is one of the most popular syntactic-blind semantics features which excels at deriving content analysis (Landauer et al., 1998). Word2Vec is a neural network proposed by Google that learn word associations from a large corpus of text and processes the text data into vector. The 10-fold cross-validation was used in the training process to estimate the performance of the prediction model on new data. Table 4.2 and Table 4.3 show the results of AES based on the LSA and Word2Vec separately.

Table 4.2: Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on LSA Semantic Features

	Accuracy				QWK			
	1	2	3	Mean	1	2	3	Mean
Cosine Similarity	0.34	0.39	0.35	0.36	0.27	0.02	-0.01	0.09
Random Forest	0.45	0.45	0.50	0.47	0.29	-0.05	0.13	0.12
NN	0.45	0.48	0.38	0.44	0.19	-0.01	0.04	0.07
CNN	0.38	0.46	0.42	0.39	0.22	-0.03	0.24	0.14
LSTM	0.41	0.38	0.38	0.39	0.24	0.01	0.09	0.11

From Table 4.2, Random Forest once again achieved the best performance yielding 0.47 of accuracy. When the performance was evaluated using QWK, CNN gained the highest score which was 0.14. When compared to the results of machine learning algorithms with syntactic features in Table 4.1, the QWK scores of all machine learning algorithms with LSA semantic features decreased significantly and the results can be considered as poor since all the results were below 0.4 (Cheon et al., 2015). Similar to the result of AES based on syntactic features, all machine learning algorithms performed poorly on dataset 2 and predicted randomly with semantic features since the QWK score was closest to zero. One of the reasons is the machine learning algorithms tend to predict all essays to the majority grade which is Band 3 and Band 4 when dealing with dataset 2 due to the lack of minority grade sample.

Table 4.3: Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Word2Vec Semantic Features

	Accuracy				QWK			
	1	2	3	Mean	1	2	3	Mean
Cosine Similarity	0.24	0.04	0.22	0.17	0.32	0.04	0.16	0.17
Random Forest	0.52	0.49	0.56	0.52	0.45	0.05	0.32	0.27
NN	0.39	0.35	0.46	0.40	0.43	0.04	0.38	0.28
CNN	0.38	0.42	0.31	0.37	0.30	0.00	0.35	0.22
LSTM	0.41	0.42	0.51	0.45	0.50	0.04	0.32	0.29

Based on Table 4.3, Random Forest achieved the best performance yielding 0.52 of accuracy. When the performance was evaluated using QWK, LSTM gained the highest score which was 0.29. From the QWK scores, the machine learning algorithms with Word2Vec semantic features performed better than LSA semantic features but still poorer than syntactic features. Same as the result of previous experiments, all machine learning algorithms performed poorly on dataset 2 and predicted randomly with semantic features since the QWK score was closest to zero. When dealing with dataset 2, all the machine learning algorithms assigned essays to majority grade in the dataset due to the lack of minority grade samples and therefore the QWK score was close to zero.

4.2.3 Prediction Results based on Hybrid Feature

In this section, syntactic features shown in Table 3.1 and semantic features vectorized by Word2Vec had been used as input data in AES hybrid framework. Based on section 4.3, we can see that the semantic features vectorized by Word2Vec achieved better performance than LSA and therefore been selected as input data. The accuracy and QWK scores were

used to assess the effectiveness of our proposed hybrid framework using various machine learning techniques. In the training procedure, the 10-fold cross-validation was utilised to measure the prediction model's performance on new data. The AES results based on hybrid features are shown in Table 4.4.

Table 4.4: Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Hybrid Features

	Accuracy				QWK			
	1	2	3	Mean	1	2	3	Mean
Random Forest	0.60	0.44	0.67	0.57	0.73	0.05	0.71	0.50
NN	0.23	0.21	0.54	0.33	0.27	0.01	0.57	0.28
CNN	0.32	0.32	0.27	0.30	0.36	0.04	0.28	0.23
LSTM	0.29	0.16	0.12	0.19	0.16	0.01	0.09	0.09

In Table 4.4, we can see that Random Forest achieved the best performance on AES hybrid framework of 0.57 in accuracy and 0.50 in QWK score. On the other hand, LSTM performed the worst with hybrid features yielding only 0.09 in QWK score. Like other proposed AES frameworks, the hybrid framework performs poorly on QWK scores on dataset 2, as it tends to only predict essays as the majority grade. Furthermore, all deep learning algorithms in the AES hybrid framework, such as NN, CNN, LSTM, perform worse than the syntactic feature-based AES framework except for random forest. Compared with the results of the AES syntactic framework in Table 4.1, the results of the deep learning algorithms in Table 4.4 are reduced to varying degrees.

4.2.4 Summary of the Prediction Results

The comparison of best performance between syntactic framework, semantic framework, and hybrid framework is summarized in Table 4.5.

Table 4.5: Best Mean Performance of Different AES Frameworks in Predicting Essay Grade

	Accuracy	QWK
Syntactic Framework	0.58	0.50
Semantic Framework	0.45	0.29
Hybrid Framework	0.57	0.50

As shown in Table 4.5, the semantic framework performed the worst in predicting essay grades while the syntactic framework and hybrid framework had similar performance. Hence, we can say that the syntactic features are more predictive on essay grades in hybrid framework since the semantic feature showing low accuracy and QWK scores. When comparing the feature extraction processes of both syntactic and hybrid framework, the hybrid framework took longer time on the feature extraction processes because it had to extract the semantic features. Therefore, we can conclude that our proposed syntactic framework achieved better result in predicting essay grades and consumed less time during features extraction process.

4.3 Prediction Results with Resampling Method

All the MUET datasets used in the experiments had the issue of imbalance grade distribution and therefore we found that most of the machine learning algorithms tend to predict the essay to majority grades to achieve higher accuracy. To solve this problem, a resampling method was applied to the datasets to balance the distribution of grades. In this

section, we evaluated the effect of SMOTE resampling method on the performance of proposed syntactic and semantic frameworks with different machine learning algorithms by using accuracy and QWK scores. The 10-fold cross-validation was used in the evaluation process to estimate the performance of the prediction model on unseen data.

4.3.1 Prediction Results Based on Syntactic features with Resampling Method

We evaluated the performance of our proposed syntactic framework with different machine learning algorithms using accuracy and QWK scores. The syntactic features show in Table 3.1 were derived from the datasets and acted as the input for prediction model. Next, the SMOTE resampling method was applied to dataset to balance the grade distribution. The 10-fold cross-validation was used in the training process to estimate the performance of the prediction model on new data. Table 4.6 shows the results of AES based on syntactic features with SMOTE.

Table 4.6: Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Syntactic Features with SMOTE

	Accuracy				QWK			
	1	2	3	Mean	1	2	3	Mean
Random Forest	0.52	0.40	0.62	0.52	0.71	0.18	0.71	0.53
NN	0.43	0.28	0.55	0.42	0.66	0.01	0.69	0.45
CNN	0.45	0.33	0.53	0.44	0.67	0.08	0.66	0.47
LSTM	0.34	0.19	0.54	0.36	0.62	0.02	0.71	0.45

Red colour in Table 4.6 shows a decrease in results compared to Table 4.1, while blue colour shows an increase and black colour shows no change. As clearly illustrated in Table 4.6, SMOTE resampling method not only could not improve the performance of the

prediction model but also decreased the accuracy and QWK score of prediction model in Table 4.1. Among the results, only the QWK score of Dataset 2 increased slightly after applying the SMOTE resampling method. The QWK score of Dataset 2 increased from 0.05 to 0.18 when using Random Forest, -0.02 increased to 0.01 when using NN, 0.03 increased to 0.08 when using CNN and 0.00 increased to 0.02 when using LSTM. However, considering the overall results, the SMOTE resampling method is not recommended for the AES syntactic framework since it cannot improve the performance of the prediction model.

4.3.2 Prediction Results Based on Semantic features with Resampling Method

We evaluated the performance of our proposed semantic framework with different machine learning algorithms using accuracy and QWK scores. LSA and Word2Vec were used to vectorize the essays representing semantic content. After that, the SMOTE resampling method was applied to the datasets to balance the grade distribution. The 10-fold cross-validation was used in the training process to estimate the performance of the prediction model on new data. Table 4.7 and Table 4.8 show the results of AES based on the LSA and Word2Vec with SMOTE separately.

Table 4.7: Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on LSA Semantic Features with SMOTE

	Accuracy				QWK			
	1	2	3	Mean	1	2	3	Mean
Random Forest	0.48	0.46	0.56	0.50	0.25	0.03	0.20	0.16
NN	0.36	0.36	0.33	0.35	0.17	-0.02	-0.05	0.03
CNN	0.36	0.37	0.42	0.38	0.27	0.01	0.02	0.10
LSTM	0.37	0.38	0.36	0.37	0.23	0.08	0.05	0.12

Red colour in Table 4.7 shows a decrease in results compared to Table 4.2, while blue colour shows an increase and black colour shows no change. When comparing the result between Table 4.7 and Table 4.2, only the performance of Random Forest with LSA increased slightly in which the mean accuracy increased from 0.47 to 0.50 and the mean QWK score increased from 0.12 to 0.16 after the SMOTE resampling method was applied. Besides, LSTM with LSA also marginally increased the mean QWK score from 0.11 to 0.12 after SMOTE resampling method was applied. However, the SMOTE resampling method decreased the performance of other machine learning algorithms. Hence, the SMOTE resampling method is not recommended to be applied in AES semantic framework since it cannot improve the performance of machine learning algorithms significantly.

Table 4.8: Performance of Different Machine Learning Algorithms in Predicting Essay Grade Based on Word2Vec Semantic Features with SMOTE

	Accuracy				QWK			
	1	2	3	Mean	1	2	3	Mean
Random Forest	0.49	0.49	0.51	0.50	0.52	0.08	0.17	0.26
NN	0.32	0.26	0.26	0.28	0.45	0.07	0.03	0.18
CNN	0.35	0.39	0.34	0.36	0.47	0.06	0.01	0.18
LSTM	0.34	0.25	0.27	0.29	0.49	0.01	0.13	0.21

When comparing the result between Table 4.8 and Table 4.3, the accuracy and QWK scores of all machine learning algorithms decreased when the SMOTE resampling method was used. Red colour in Table 4.8 shows a decrease in results compared to Table 4.3, while blue colour shows an increase and black colour shows no change. The QWK score of NN decreased the most from 0.28 to 0.18. Hence, the SMOTE resampling method is not

recommended to be applied in AES semantic framework since it cannot improve the performance of machine learning algorithms.

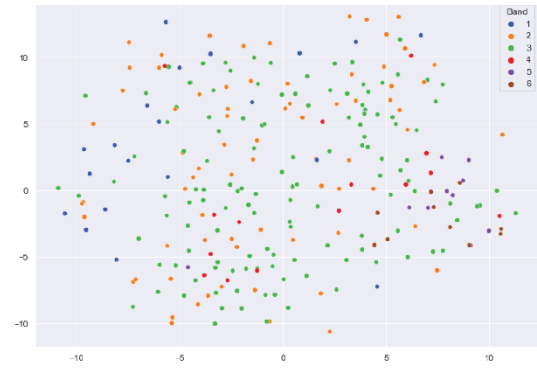
4.4 Comparison Between Syntactic Features and Semantic Features and Effect of SMOTE

In the previous section, it can be noted that the highest QWK scores of our proposed AES framework is Random Forest based on syntactic features with SMOTE which was only 0.53 and therefore the reason of unsatisfactory performance of prediction model needs to be investigated. In this section, a statistical method for visualizing high-dimensional data called t-distributed stochastic neighbour embedding (t-SNE) was used to reduce the high dimensions of both syntactic and semantic features and embedding the features for visualization in a two dimensions vector space to investigate how well the features represent the essay in the vector space. In the t-SNE visualized vectors space, each colour represents one essay grade and each vector point represent an essay. In an ideal situation, the same colour point should form a cluster in the space if the features represent the essay well. In addition, the shape and axis of t-SNE visualized vectors space are meaningless, what matters is the distance between each point, which indicates that the points have similar vectors and should form a cluster.

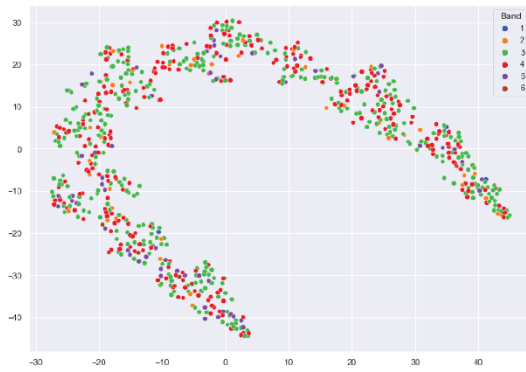
For instance, there are 14 syntactic features derived from the essay, please refer to Table 3.1 for more details. These 14 syntactic features were used to form a 14-dimensions vector to represent the essay, which were then reduced to 2-dimensional by t-SNE for visualization. For semantic features, the essays were converted to 300 dimensions vectors by using Word2Vec and then reduced to 2-dimensional by t-SNE for visualization. The visualized vectors space of syntactic and semantic features of dataset is shown in Figure 4.1.



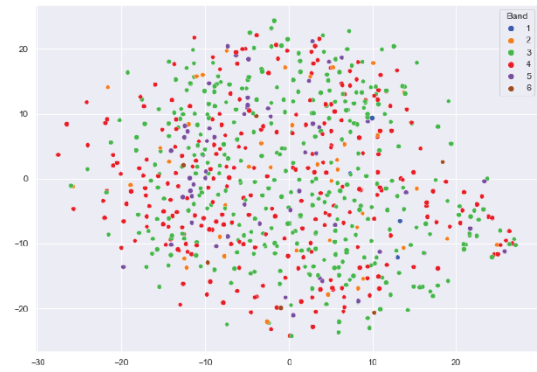
Dataset 1 Syntactic Features



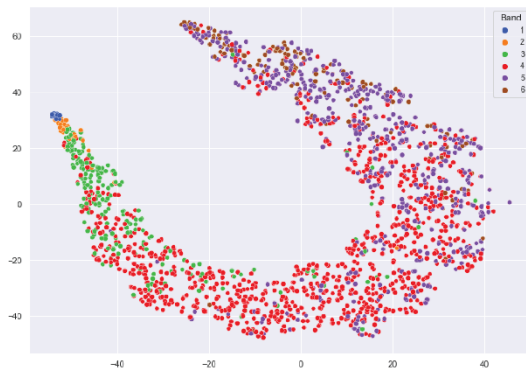
Dataset 1 Semantic Features



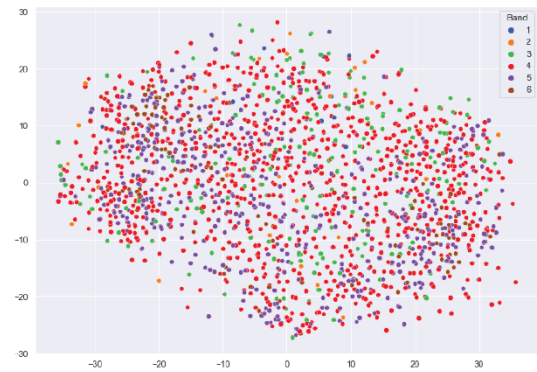
Dataset 2 Syntactic Features



Dataset 2 Semantic Features



Dataset 3 Syntactic Features

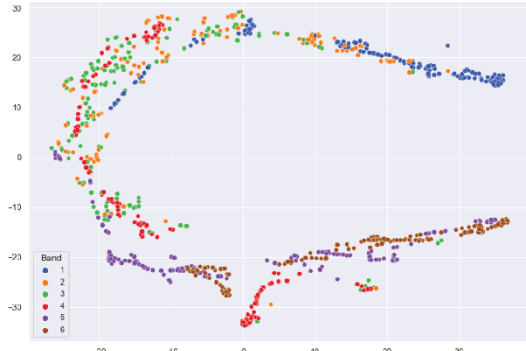


Dataset 3 Semantic Features

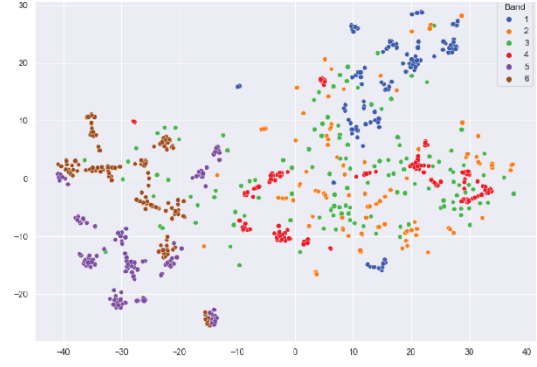
Figure 4.1: Visualized Vectors Space of Syntactic and Semantic Features of Dataset

In Figure 4.1, the distribution of vector point in Dataset 1 Syntactic Features and Dataset 3 Syntactic features is better than other datasets. Unlike other datasets whose vector points are randomly distributed throughout the vector space, the vector point distribution of Dataset 1 Syntactic Features exhibits more distinct clusters, with Band 1 grouped in the

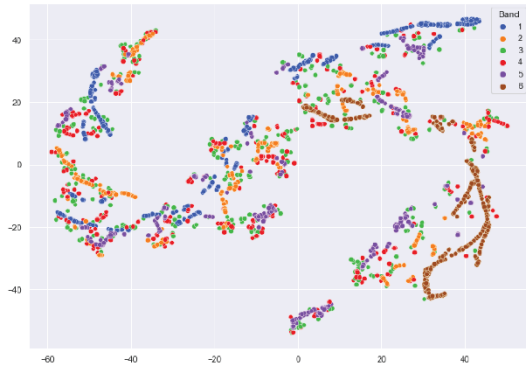
lower right vector space, while Band 5 and Band 6 are grouped in the upper left vector space. For Dataset 3 Syntactic Features, we can see that the distribution of vector points shows more significant distinct clusters, which Band 1, Band 2, and Band 3 grouped in upper left vector space while Band 5 and Band 6 grouped in upper right vector space. Therefore, the performance of predictive model based on Dataset 1 Syntactic Feature and Dataset 3 Syntactic Features are better than other datasets. As mentioned above, if the features represent essays well, it should be able to distinguish essay grade well in the vector space and therefore the same colour vector points should form a cluster in space. However, all the distribution of colour vector points on dataset with semantic features are randomly distributed in whole vector space without order. Therefore, we can say that the existing extracted syntactic features are more representative of essays and more appropriate to distinguish grades than semantic features, which is consistent with the experimental results obtained by Wong (2021) in his study.



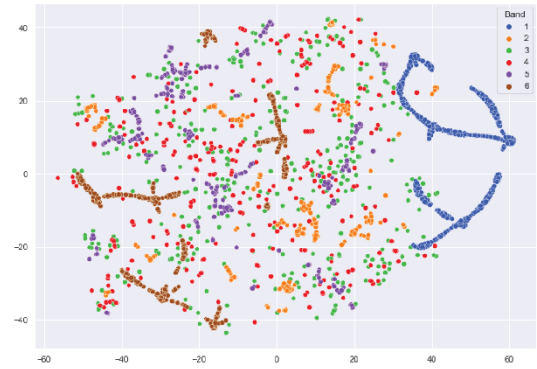
Dataset 1 Syntactic Features
After SMOTE



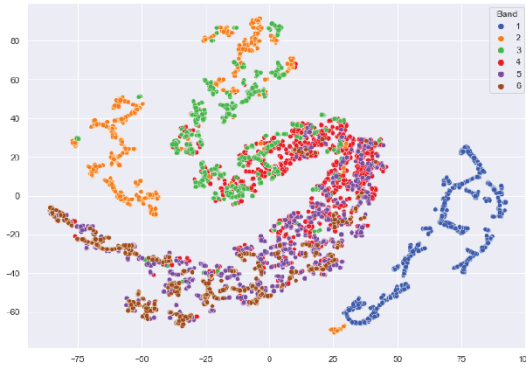
Dataset 1 Semantic Features
After SMOTE



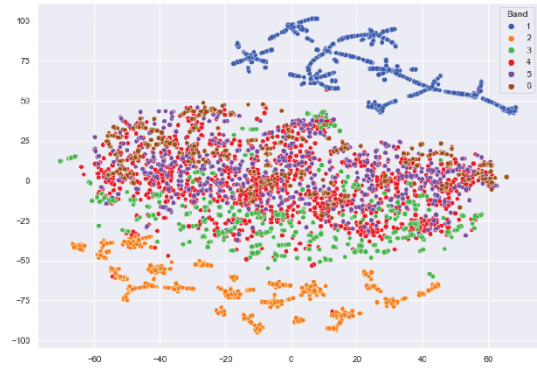
Dataset 2 Syntactic Features
After SMOTE



Dataset 2 Semantic Features
After SMOTE



Dataset 3 Syntactic Features
After SMOTE



Dataset 3 Semantic Features
After SMOTE

Figure 4.2: Visualized Vectors Space of Syntactic and Semantic Features of Dataset with SMOTE

Figure 4.2 shows the distribution of vector points in the vector space of syntactic and semantic features of the dataset after applying the SMOTE resampling method. The visualization illustrates that the SMOTE method has effectively increased the number of samples from minority classes such as Band 1 and Band 6 and grouped them into a cluster

in the vector space. However, it is essential to note that the shape of the distribution in the vector space is not significant, and what matters is the distance between each point, which indicates that the points have similar vectors and should form a cluster. Although the SMOTE method has performed well in resampling and clustering, as revealed by the visualization results, it is not recommended to be applied in the AES semantic framework based on the results in Section 4.3. The results in Section 4.3 shows that the SMOTE resampling method not only fails to improve the performance of machine learning algorithms but also decreases it. One reason for this situation could be the characteristic of QWK scores, which corrects the percent-agreement for the case of agreement that would be expected purely by chance. Therefore, even if SMOTE produces a large number of new minority class samples, the proposed framework will not make significant progress in QWK scores because it will increase the probability of the model successfully predicting the class of minority class samples and therefore reduce the QWK score.

Overall, the performance of machine learning algorithms in grading essay based on the syntactic features is better than the semantic features of essay in both accuracy metric and QWK metric. However, the results shows that only Syntactic AES Framework is sufficient to propose a method to be used in real life application while the performance of Semantic AES framework is not sufficient enough. The result also shows that the essay grade did not have a simple linear relationship with the syntactic features and semantic features since machine learning algorithms performed not efficiently in grading the essay based on those features. The majority of the AES systems currently in use presume a linear relationship between the essay's features and its grades. However, a study found that the correlation between the feature vector and essay grade is non-linear (Fazal et al., 2011). For instance, the majority of existing AES studies take essay length as a significant feature (Chen & He,

2013; Alghamdi et al., 2014; Wong & Bong, 2021) and utilise it to signify the quality of the essay, meaning the longer the essay, the better the essay score. However, these features are insufficient to describe the complex relationship between essay and its grades, as human evaluators would also give lower grade to a long essay if majority of the content is irrelevant.

4.5 Applicability of Evaluation Metrics on AES

In this study, two evaluation metrics which are accuracy and QWK were used to examine the performance of different machine learning algorithms in AES. Based on the results from Table 4.1 and Table 4.2, we can see that compared to QWK, the results of the three datasets in the experiment are not significantly different in the accuracy evaluation. The results are between 0.34 and 0.67. However, the QWK scores of each datasets have high variant which Dataset 2 scored approximate zero but Dataset 3 able to score 0.74 when predicted using syntactic features.

Table 4.9: Confusion matrix of result of dataset 2, random forest, syntactic features

Band	1	2	3	4	5	6
1	0	0	1	0	0	0
2	0	0	9	0	0	0
3	0	1	47	29	0	0
4	0	0	31	15	0	0
5	0	0	7	4	0	0
6	0	0	0	1	0	0

Table 4.9 shows the confusion matrix of the prediction result of Dataset 2 which predicted by random forest and based on syntactic features. From Table 4.9, the accuracy of prediction is 0.43 and the prediction model predicted most of the essays to be Band 3 and

Band 4. This is because Dataset 2 was a skewed dataset which consisted of only 4 samples from Band 1 or 0.01% of the dataset. On the other hand, it consisted of only 6 samples from Band 6 essays representing 0.01% of dataset, whereas Band 3 contained 359 samples or 50.0% of the whole dataset. The extremely imbalanced distribution affected the performance of machine learning algorithms as they were unable to grade the essays for Band 1 and Band 6. The result ended up predicting all the essays to be Band 3 since it had a greater chance to predict correctly and therefore obtain a higher accuracy. Hence, the accuracy of machine learning algorithms using Dataset 2 is around 0.50 but the QWK score is close to zero because it corrects the percent-agreement for the case of agreement that would be expected purely by chance. As reported in the study by Madasamy and Ramaswami (2017), if the data were in imbalanced ratio, the prediction process could be complicated as the pattern for each class is scarce. In term of the reliability, these data generally diminish the performance of prediction.

As a whole, the accuracy magnitude is greater than the QWK metric. By measuring the percentage of matched pairings with human raters, accuracy has been identified as the most often utilised assessment criterion in assessing the performance of an AES. However, when working with unbalanced datasets, the outcome could be distorted and biased toward the majority class (Tanha et al., 2020). Since the distribution of essay grades is typically approximated by a bell-shaped frequency distribution known as the normal distribution or the Gaussian distribution, it is typical to have an unbalanced percentage of grades in the real world. Hence, this study focuses on QWK score in the following section as a standard evaluation metric for AES since it provides more valuable information when dealing with imbalance ratio dataset, as QWK is belong to kappa-like family and kappa coefficient has been proven that it can provide valuable information on the reliability of ordinal scale data

(Sim & Wright, 2005). As emphasized by Wong and Bong (2019), the Kappa value corrects the percent-agreement for the case of agreement that would be expected purely by chance when estimating the degree of agreement between two raters. They also claimed that the Kappa value is a better measurement than simple percent agreement calculation in AES.

4.6 Comparison of Machine Learning and Deep Learning on AES Performance

As discussed in the previous section, we will focus on using QWK score to evaluate the performance of AES since it can show more valuable information than accuracy. As shown in Table 4.1, Random Forest achieved the highest mean QWK score which is 0.51 when predicting essay grade based on syntactic features. It shows that Random Forest can assign grade to corresponding essays and not just predict randomly. When predicting essay grade based on semantic features, LSTM with Word2Vec gain highest mean QWK score which is 0.29. Based on the experiment results, Random Forest performed better than other deep learning algorithms in predicting essay grade based on syntactic features and deep learning algorithm such as LSTM performed better in semantic features prediction. However, the difference between machine learning and deep learning algorithms in the experiments were not obvious and the performance of both algorithms cannot be considered as excellent since both of their QWK scores were not greater than 0.75 (Cheon et al., 2015).

4.7 Chapter Summary

In this chapter, the experimental results of our proposed AES frameworks based on syntactic features, semantic features, and hybrid features were explained and the reasons of unsatisfactory performance of prediction model were discussed. From the experiments, semantic framework performed the worst in predicting essay grade while syntactic framework and hybrid framework had the similar performance. Hence, we can conclude that

our proposed syntactic framework shows better performance in terms of time duration since it only extracts the syntactic features and achieved better result in predicting essay grades compared to another proposed AES framework. Furthermore, all the datasets used in the experiments had the issue of imbalance grade distribution and therefore a resampling method had been applied to dataset to balance the distribution of grade. From the result, we can see that the SMOTE resampling method not only cannot improve the performance of prediction model but also decreased the accuracy and QWK score of the prediction model. Therefore, the SMOTE resampling method is not recommended to be applied in AES framework. To investigate how well the features represent the essay in the vector space, both syntactic and semantic features of datasets were visualized into two dimensions vector space by t-SNE. However, the spatial distribution of vector points of the same colour did not form obvious clusters, it is distributed disorderly in the entire vector space except for Dataset 1 and Dataset 3 with syntactic features. This means that the existing extracted syntactic features are better than semantic features in representing essays with respect to paper grades, and this explained the reason that prediction models performed poorly with semantic features in grading paper grades. Most of the MUET datasets have imbalance grade distribution issue and therefore QWK score is found to be better than accuracy as a standard evaluation metric for AES since it provides more valuable information when dealing with imbalance ratio dataset. Based on experiment results, Random Forest performed better than other deep learning algorithms in predicting essay grade based on syntactic features and deep learning algorithm such as LSTM performed better in semantic features prediction. However, the difference between machine learning and deep learning algorithms in the experiments were not obvious and the performance of both algorithms cannot be considered as excellent since both of their QWK score were not greater than 0.75.

CHAPTER 5

USE CASE: PREDICTING MUET BAND AMONG PRE-UNIVERSITY STUDENTS

5.1 Overview

This chapter presents a use case study to demonstrate the effectiveness of our proposed AES framework for integrating the corresponding structure into a real-life e-learning platform. The syntactic AES framework has been used to develop an e-learning platform known as the UNIMAS DBRater which allows students to submit their essays and obtain a predicted grade or band immediately. This chapter also includes findings from the evaluation on whether our proposed framework can perform similarly as trained human raters.

5.2 Background

UNIMAS DBRater is an e-learning platform which is able to grade the essays automatically with high human rater agreement. One of the purposes of the UNIMAS DBRater is to provide a e-learning platform for students and teachers to let them can write and grade essay online. The essay written by student will be graded by teacher and then the graded essay will be collected from the e-learning platform and act as a new dataset to train our proposed AES framework. In the current stage, the DBRater system can assist human raters to reduce the cost of time and effort in grading essays by automatically grade each essay according to the past practices by using natural language processing and machine learning. Students can register themselves easily to the system and they are able to enrol to multiple classes and complete their essay assignment from each class. UNIMAS DBRater can grade their essays within a few minutes and help them to improve their writing skills.

UNIMAS DBRater allows students to stay focused on writing task as other functional features such as word count were done automatically. The system also provides comment function for students and teachers to let them communicate with each other and discuss the content of essay if they want to.

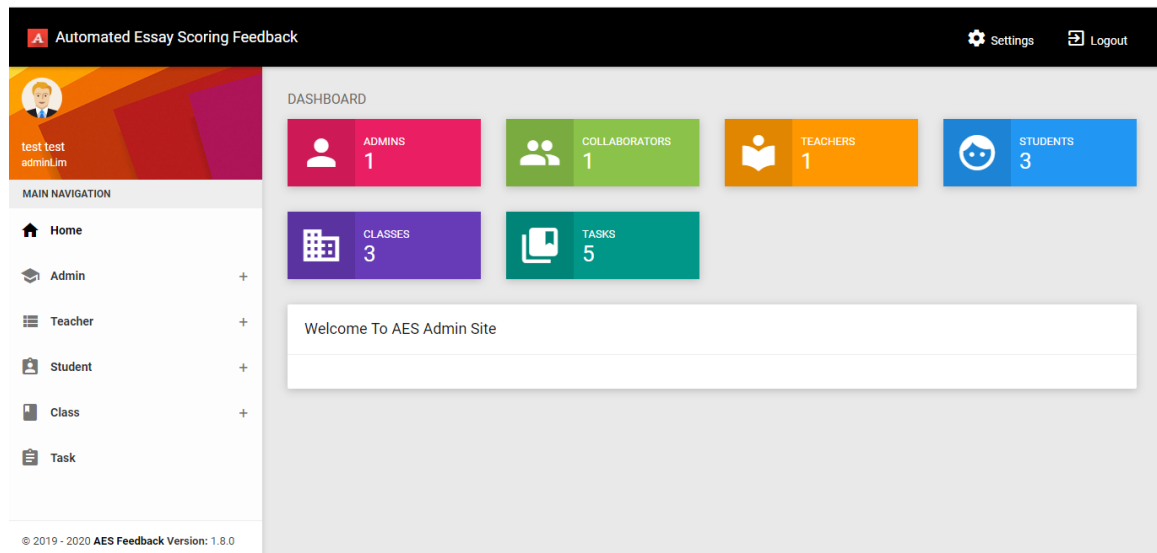


Figure 5.1: Dashboard of UNIMAS DBRater

Figure 5.1 shows the main page of DBRater once users logged in to the system. In the dashboard, users can view the details of system such as the number of teacher and student in this system and the number of current available tasks in system. The navigation panel on the right provides access to users to navigate to specific features or pages like create new class, new task, or new student.

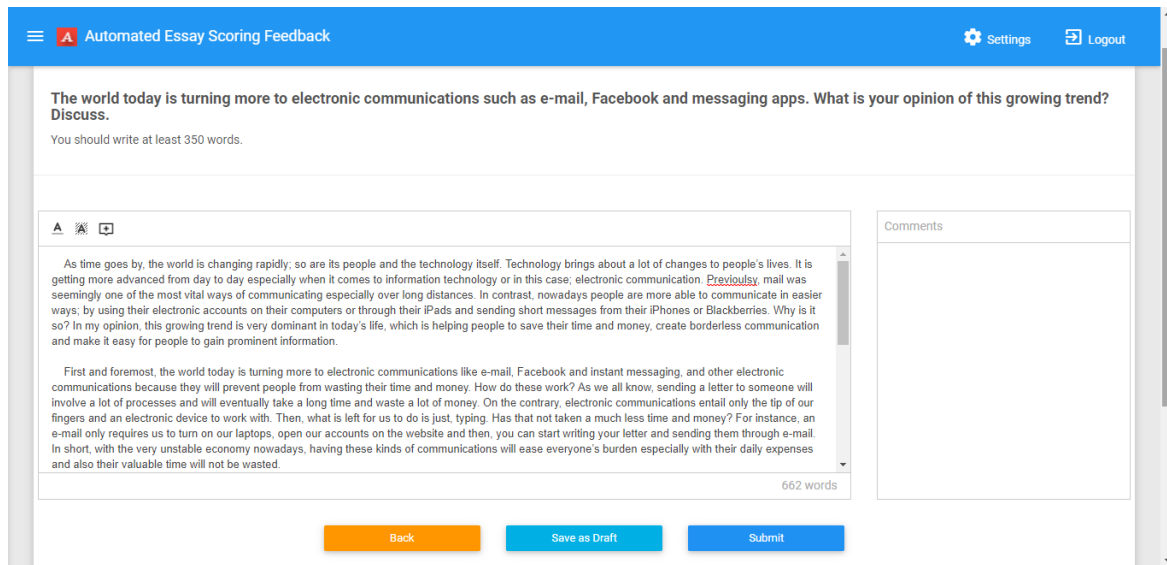


Figure 5.2: Writing Interface of UNIMAS DBRater

Figure 5.2 shows the essay writing interface for students and teachers in DBRater. The writing panel provides users with useful functional features such as automatic word count, word highlighting, and commenting functions for them to communicate with each other and discuss the substance of the essay.

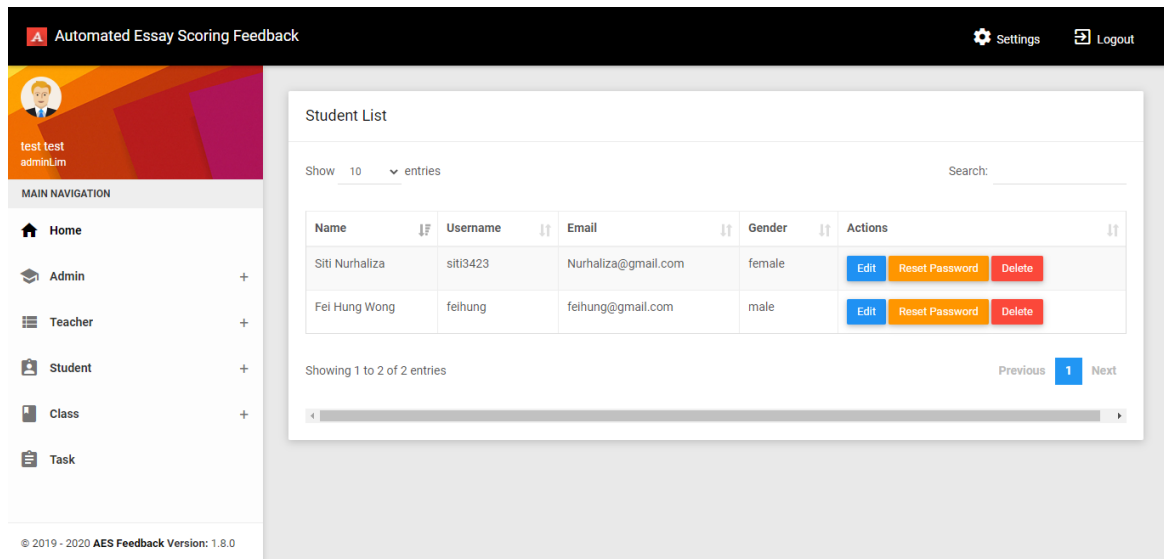


Figure 5.3: Interface of Student Management

Figure 5.3 show the student management interface of DBRater. School administrators are able to view and edit students' details, reset their passwords, and even delete their accounts from the system.

5.3 Task Description

To examine the performance of the proposed AES framework in real-life use cases, an experiment has been conducted to compare the results of the AES framework predicting student MUET essays with the results of teacher-graded student MUET essays. Based on Table 4.5 in Chapter 4, the proposed syntactic AES framework achieved highest scores in accuracy and QWK scores and therefore the syntactic framework was used in DBRater to predict essay grade. All the MUET essays used in the experiment were collected from DBRater and written by UNIMAS pre-university students, which were then graded by qualified graders. The grades ranged from Band 1 to Band 6 and the corresponding essay titles are shown in Table 5.1.

Table 5.1: Essay Titles for Each Dataset of DBRater

Dataset No.	Topic
A	Students should be given study leave before their major examinations. Do you agree? Illustrate your opinion with examples.
B	'Education comes not from books but from practical experience.' Do you agree? Support your answer with examples.
C	Meeting someone for the first time through the Internet is never a good idea. Do you agree?

From Table 5.1, there are three essay datasets which were collected from DBRater and used in the experiment. All essay datasets were argumentative essays, and students needed to use statistics, facts, and logic to support or oppose a particular position and convince readers of the merits of their position. The details and grade distribution of each essay datasets are shown in Figure 5.4, Figure 5.5 and Figure 5.6.

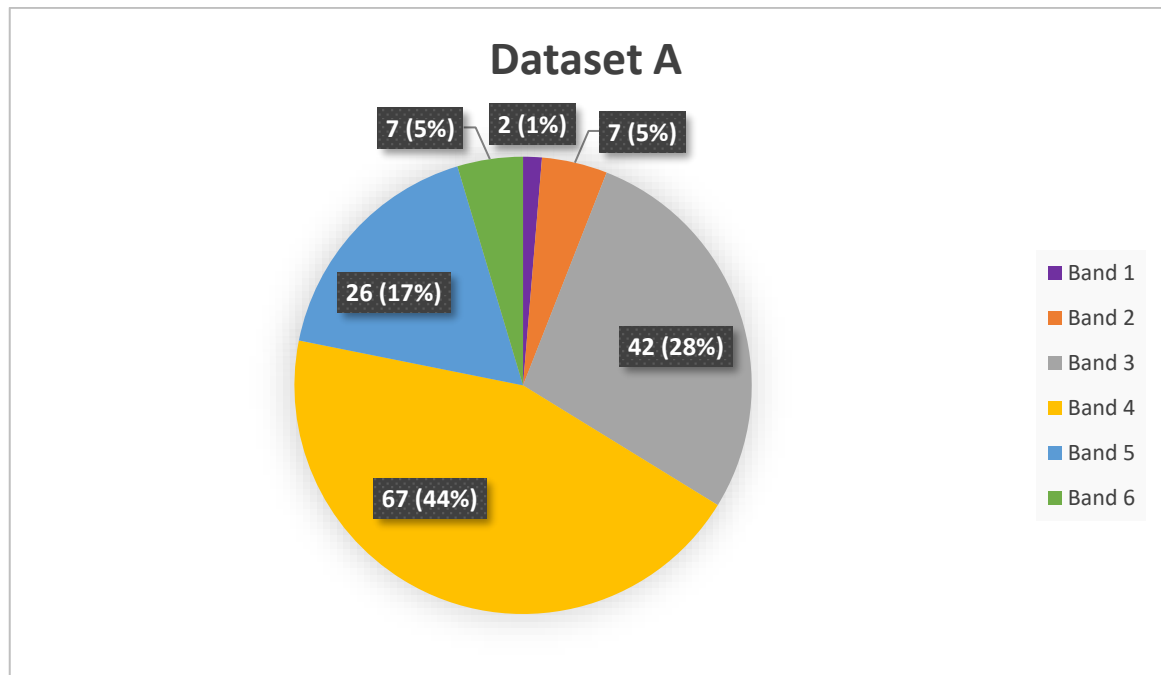


Figure 5.4: Pie Chart of Band Distribution in Dataset A

Figure 5.4 shows the pie chart of Band distribution in Dataset A and the dataset contains 151 essays. According to the pie chart, Band 4 is obtained by most of the students followed by Band 3 and Band 5. There are two Band 1 essays in the dataset which only consist 1% of dataset. Next, the Band 2 and Band 6 have the same number of essays which both contain 7 essays or 5% of dataset. For Band 3, a total number of 42 essays (28%) is obtained. Then, there are 67 essays in Band 4 or 44% of dataset. The total of Band 5 essays in the dataset is 26 or 17% of dataset.

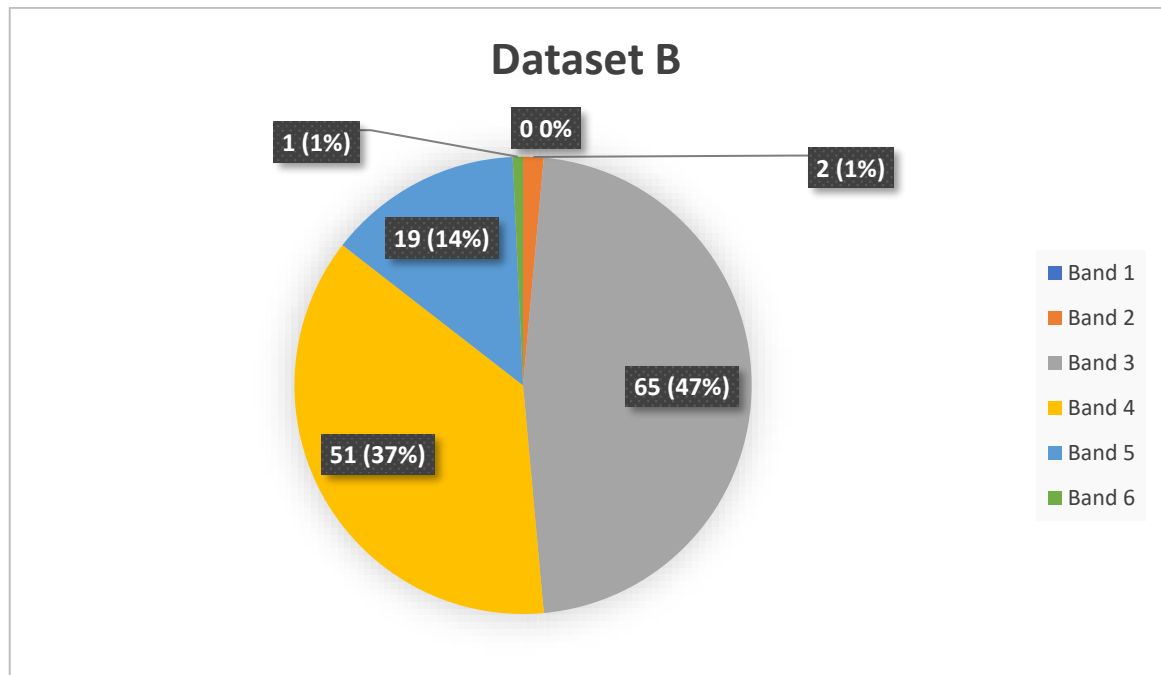


Figure 5.5: Pie Chart of Band Distribution in Dataset B

Figure 5.5 shows the pie chart of Band distribution in Dataset B and the dataset contains 138 essays. According to the pie chart, Band 3 is obtained by most of the students followed by Band 4 and Band 5. There are zero Band 1 essays in the dataset. Next, the Band 2 contains 2 essays or 1% of dataset. For Band 3, there are a total of 65 essays, which represents 47% of the dataset. Then, there are 51 essays in Band 4 or 37% of dataset. The total of Band 5 essays in the dataset is 19 or 17% of dataset. For Band 6, there is only 1 essay in the dataset which is 1% of dataset.

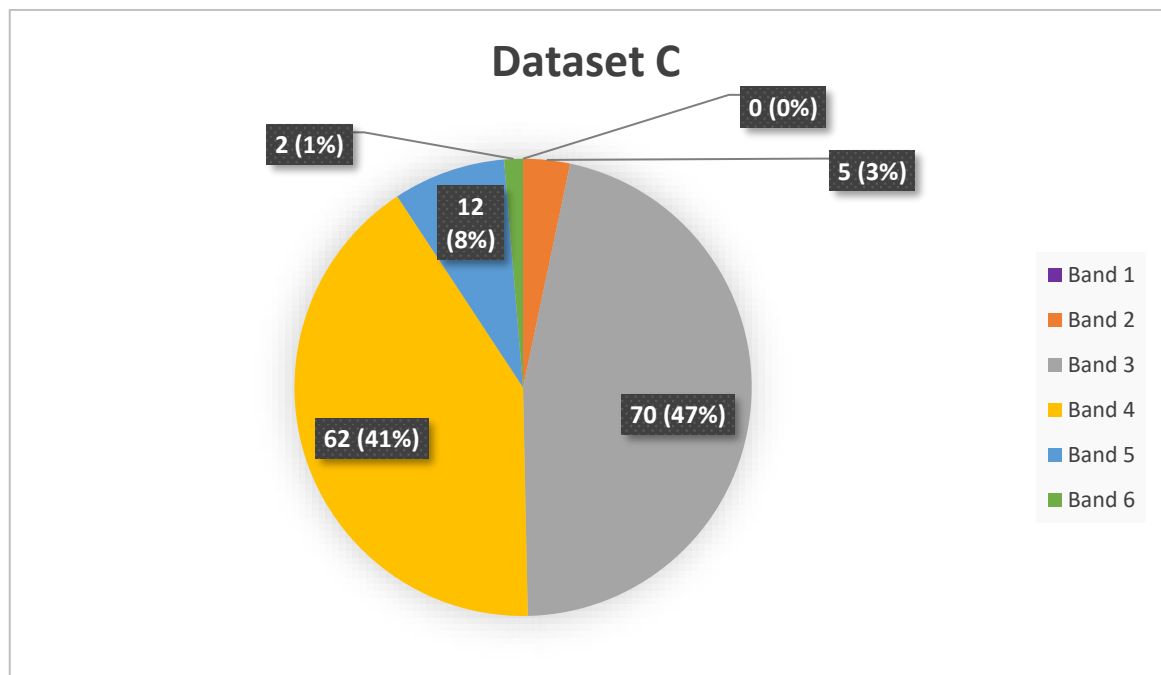


Figure 5.6: Pie Chart of Band Distribution in Dataset C

Figure 5.6 shows the pie chart of Band distribution in Dataset C and the dataset contains 151 essays. According to the pie chart, Band 3 is obtained by most of the students followed by Band 4 and Band 5. There are zero Band 1 essays in the dataset. Next, the Band 2 have 5 essays or 3% of dataset. Similar to Dataset B, a majority of the essays ($n=70$) in this dataset were graded as Band 3, accumulating 47% of the dataset. 62 essays were awarded Band 4 which occupied 41% of dataset. The total of Band 5 essays in the dataset is 12 or 8% of dataset. For Band 6, there are only 2 essays in the dataset which is 1% of dataset.

5.4 Results and Discussion for the Use Case Experiment

In this section, we evaluated the performance of our proposed syntactic framework with different machine learning algorithms by accuracy and QWK score. The syntactic features show in Table 3.2 were extracted from the dataset and acted as the input of prediction model. The 10-fold cross-validation was used in the training process to estimate

the performance of the prediction model on new data. Table 5.2 shows the results of AES based on syntactic features.

Table 5.2: Performance of Different Machine Learning Algorithms in Predicting DBRater Essay Grade Based on Syntactic Features

	Accuracy				QWK			
	A	B	C	Mean	A	B	C	Mean
Random Forest	0.51	0.47	0.46	0.48	0.33	0.08	0.21	0.21
NN	0.40	0.44	0.43	0.42	0.08	-0.02	0.15	0.07
CNN	0.47	0.42	0.44	0.44	0.30	0.16	0.26	0.24
LSTM	0.48	0.45	0.57	0.50	0.24	-0.01	0.31	0.18

In Table 5.2, LSTM has the highest accuracy mean of 0.50 followed by Random Forest and CNN. For the QWK score, CNN shows the highest mean performance and achieved 0.24. Based on the result, all machine learning algorithms performed poorly on dataset 2 with the QWK score closest to zero which means the prediction is similar to predict randomly.

Although the proposed AES framework have an unsatisfactory performance in the use case experiment, the specific machine learning AES framework shows higher correlation when compared to the correlation between two human raters on grading the three MUET dataset that used in this use case experiment. The accuracy and QWK score of each dataset between two human raters are evaluated and shown in Table 5.3.

Table 5.3: Accuracy and QWK of Dataset Between Two Human Raters

	Accuracy	QWK
Dataset A	0.43	0.28
Dataset B	0.41	0.16
Dataset C	0.56	0.28

From the result, dataset 3 has the highest correlation between two human raters which is 0.56 in accuracy and 0.28 in QWK scores while dataset 2 has the lowest correlation which is 0.41 in accuracy and 0.16 in QWK scores. When compared with the performance of our proposed AES frameworks in Table 5.2, our proposed AES frameworks performed better and had a higher correlation than two human raters in terms of accuracy and QWK score. For dataset A, the proposed AES framework with Random Forest and CNN achieved higher correlation which get 0.51 and 0.47 in accuracy, and 0.33 and 0.30 in QWK score. For dataset B, the proposed AES framework with CNN have the same correlation with two human raters. For dataset C, only proposed AES framework with LSTM shows better correlation than human rater which get 0.57 in accuracy, and 0.31 in QWK score. Hence, we can conclude that the proposed AES frameworks are able to capture and learn the behaviour of human raters on essay grading although the performance of proposed AES frameworks are not satisfactory enough which is below 80% in accuracy and QWK scores.

5.5 Chapter Summary

This chapter presents a use-case experiment to demonstrate the effectiveness of our proposed AES syntactic framework and integrated it into an e-learning platform called UNIMAS DBRater, which allows students to submit their essays and instantly predict their

grades. All the MUET essay datasets used in the experiment were collected from DBRater and written by UNIMAS pre-university students, which was then graded by qualified graders. Based on the experiment results, LSTM has the highest average accuracy at 0.50, while CNN has the highest average performance on QWK score at 0.24. As can be seen from the results, all machine learning algorithms performed poorly on dataset 2, with the QWK score closest to zero, which means the predictions are similar to random predictions. In terms of accuracy and QWK score, the correlation between the proposed AES framework and human raters is higher than the correlation between two human raters, hence we can conclude that the proposed AES framework is able to capture and learn the behaviour of human raters on essay grading, even though the proposed AES framework's performance is not satisfactory, with accuracy and QWK score below 80%.

CHAPTER 6

CONCLUSIONS

6.1 Conclusions of the Study

To conclude, the study's objectives as outlined in Chapter 1 have been met. The proposed AES framework is built successfully and have better correlation than two human raters, while syntactic AES framework achieve better performance than semantic and hybrid AES framework. According to the results of the experiment, Random Forest outperformed other deep learning algorithms in predicting essay grade based on syntactic data, whereas LSTM outperformed other deep learning algorithms in predicting semantic features. However, in the experiments, the differences between machine learning and deep learning algorithms were not obvious, and neither algorithm's performance can be considered exceptional because both methods' QWK scores were less than 0.75.

Instead of using any available public essay datasets, five MUET essay datasets were collected locally and used in training and testing processes of our proposed AES framework to evaluate the performance of AES framework. Most of the MUET datasets have imbalanced grade distribution. Therefore, the Quadratic Weighted Kappa (QWK) score was better than accuracy as a standard evaluation metric for AES since it provides more valuable information when dealing with imbalanced dataset. To overcome the imbalanced grade distribution, a resampling method called SMOTE was applied to datasets to investigate the influence of resampling method toward the performance of the AES framework. However, the SMOTE resampling method were not only unable to improve the performance of prediction model but also introduce more noises by decreasing the accuracy and QWK score of prediction models.

To demonstrate the applicability of our proposed method, this study presents a use case that integrates the proposed syntactic AES framework into an e-learning platform named DBRater to compare the correlation between the AES framework and human raters. The correlation between the proposed AES framework and human raters was found to be higher than the correlation between two human raters in terms of accuracy and QWK score. Hence, we can conclude that the proposed AES framework was able to capture and learn the behaviours of human raters when grading essays.

To investigate the reasons why our proposed AES framework is difficult to evaluate essay grades on some datasets, both syntactic and semantic features of the dataset were visualized into a two-dimensional vector space by t-SNE to examine how well these features represent essays in the vector space. From the visualized figure, the distribution of the same colour vector points did not form distinct clusters and they were scattered randomly throughout the vector space except for dataset 1 and dataset 3 with syntactic features. Hence, we can conclude that the extracted syntactic features outperformed the semantic features in terms of essay representation and were better at distinguishing essay grades. This finding explains why our proposed AES framework had difficulty in using semantic features to evaluate essay grades.

6.2 Limitation

There are several limitations in our study that need to be acknowledged. These limitations are as follows:

6.2.1 Limited Resources

There was a lack of available MUET essay datasets, so we were forced to collect ourselves through the collaboration with local schools. Also, the quality of the collected

datasets was not satisfactory, as the correlation between the two human raters was low, which meant that the datasets were highly contested on students' essay grades. In addition, the number of essays in the datasets was small (100~200 essays), which was not enough, and some minority grades were missing in the datasets.

6.2.2 Performance of Proposed AES Framework Is Not Satisfactory

Our proposed AES frameworks were not able to achieve over 80% in accuracy and QWK scores which was not satisfied enough to be viably used in real settings even though the proposed AES frameworks showed higher correlation as compared to the correlation between two human raters. Resampling methods such as SMOTE was applied to the dataset to improve the performance of the AES framework, but the results show that the improvement was not significant.

6.3 Contributions

This research has made significant contributions to related body of research concerning automation in educational evaluation using deep learning methods. Research has been conducted on the impact of deep learning and resampling methods on the performance of the AES framework and the results were analyzed. At present, there is a lack of research on AES with local English essays such as MUET, and this research has reasonably filled the gap. Additionally, the MUET essay dataset we collected from local schools was graded by two raters and can be used for future AES research, which we believe will be the starting point for the development of AES in Malaysia. Going a step further, this study developed an e-learning platform called UNIMAS DBRater, which is currently used by UNIMAS Pre-University English classes. We are also working with local educational institutions so that more teachers and students can join and benefit from this e-learning platform.

6.4 Future Work

Based on the limitations mentioned earlier in this chapter, there are several pertinent issues that warrant further investigation in future research. Firstly, using existing features to predict essay grades will be a challenge, so new semantic and syntactic features should be investigated to better represent essays. In particular, a more robust evaluation of the correlation between those features and essay grades is highly recommended as it can potentially enhance the accuracy of grading. Secondly, the optimal tuning of the predictive model should be investigated so that AES can perform better and achieve higher accuracy and QWK scores. It would be interesting to apply another machine learning algorithm in the AES framework and evaluate the performance. Thirdly, a thorough comparative study on the grades assigned by AES systems and trained human raters can be done through in order to effectively measure the differences in scoring.

REFERENCES

- Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology*, 1-6. doi:10.1109/ICEngTechnol.2017.8308186
- Alghamdi, M., Alkanhal, M., Al-Badrashiny, M., Al-Qabbany, A., Areshey, A., & Alharbi, A. (2014). A hybrid automatic scoring system for Arabic essays. *AI Communications*, 27(2), 103–111. doi:10.3233/aic-130586
- Algorithmia. (15 September, 2020). *A Comprehensive Overview of Natural Language Processing by the Algorithmia team*. Retrieved from AI Infrastructure Alliance: <https://ai-infrastructure.org/a-comprehensive-overview-of-natural-language-processing-by-the-algorithmia-team/>
- Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic Text Scoring Using Neural Networks.
- Al-Jouie, M., & Azmi, A. (2017). Automated Evaluation of School Children Essays in Arabic. *Procedia Computer Science*, 117, 19-22. doi:10.1016/j.procs.2017.10.089
- Amalia, A., Gunawan, D., Fithri, Y., & Aulia, I. (2019). Automated Bahasa Indonesia essay evaluation with latent semantic analysis. *Journal of Physics: Conference Series*, 1235, 012100. doi:10.1088/1742-6596/1235/1/012100
- Arc. (25 December, 2018). *Convolutional Neural Network*. Retrieved from Towards Data Science: <https://towardsdatascience.com/convolutional-neural-network-17fb77e76c05#:~:text=Fully%20Connected%20Layer%20is%20simply,into%20the%20fully%20connected%20layer>

- Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4(3).
- Awaida, S. a., Shargabi, B. A., & Rousan, T. A. (2019). Automated Arabic Essays Grading System based on F-Score and Arabic WordNet. *Jordanian Journal of Computers and Information Technology*, 5(3), 170-180. doi:10.5455/jjcit.71-1559909066
- Beseiso, M., & Alzahrani, S. (2020). An Empirical Analysis of BERT Embedding for Automated Essay Scoring. *International Journal of Advanced Computer Science and Applications*, 11. doi:10.14569/IJACSA.2020.0111027
- Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14, 106. doi:<https://doi.org/10.1186/1471-2105-14-106>
- Borade, J. G., & Netak, L. D. (2021). Automated Grading of Essays: A Review. *Intelligent human computer*, 12615, 238–249.
- Brownlee, J. (24 May, 2017). *A Gentle Introduction to Long Short-Term Memory Networks by the Experts*. Retrieved from Machine Learning Mastery: <https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1741-1752.

- Chen, M., & Li, X. (2018). Relevance-Based Automated Essay Scoring via Hierarchical Recurrent Model. *2018 International Conference on Asian Language Processing*. doi:10.1109/ialp.2018.8629256
- Chen, Z., & Zhou, Y. (2019). Research on Automatic Essay Scoring of Composition Based on CNN and OR. *2019 2nd International Conference on Artificial Intelligence and Big Data*. doi:10.1109/icaibd.2019.8837007
- Cheon, M., Seo, H.-W., Kim, J.-H., Noh, E.-H., Sung, K.-H., & Lim, E. (2015). An Automated Scoring Tool for Korean Short-Answer Questions Based on Semi-Supervised Learning. *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*. doi:10.18653/v1/w15-4409
- Contreras, J. O., Hilles, S., & Abubakar, Z. B. (2018). Automated Essay Scoring with Ontology based on Text Mining and NLTK tools. *2018 International Conference on Smart Computing and Electronic Enterprise*. doi:10.1109/icscee.2018.8538399
- Dickson, B. (6 January, 2020). *What are convolutional neural networks (CNN)?* Retrieved from BD Tech Talks: <https://bdtechtalks.com/2020/01/06/convolutional-neural-networks-cnn-convnets/>
- Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment*, 5(1), 1-35.
- Dikli, S. (2010). The Nature of Automated Essay Scoring Feedback. *CALICO Journal*, 28. doi:10.11139/cj.28.1.99-134
- Dong, F., & Zhang, Y. (2016). Automatic Features for Essay Scoring – An Empirical Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. doi:10.18653/v1/D16-1115

- Elalfi, A. E., Elgamal, A. F., & Amasha, N. A. (December, 2019). Automated Essay Scoring using Word2vec and Support Vector Machine. *International Journal of Computer Applications*, 177, 20-29. doi:10.5120/ijca2019919707
- Elliott, S. M. (2003). IntelliMetric: from here to validity. In *Automated essay scoring: a cross-disciplinary perspective* (pp. 71-86).
- Fazal, A., Dillon, T., & Chang, E. (2011). Noise Reduction in Essay Datasets for Automated Essay Grading. *On the Move to Meaningful Internet Systems: OTM 2011 Workshops Lecture Notes in Computer Science*, 484-493. doi:10.1007/978-3-642-25126-9_60
- Filho, A. H., Concatto, F., Nau, J., Prado, H. A., Imhof, D. O., & Ferneda, E. (2019). Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. *Procedia Computer Science*, 159, 764-773. doi:<https://doi.org/10.1016/j.procs.2019.09.235>
- Filho, A. H., Concatto, F., Prado, H. A., & Ferneda, E. (2021). Comparing Feature Engineering and Deep Learning Methods. *Proceedings of the 23rd International Conference on Enterprise Information Systems*, 1, 575-583. doi:10.5220/0010377505750583
- Foltz, P., Laham, D., & Landauer, T. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 27-31.
- Garbade, M. J. (15 October, 2018). *A Simple Introduction to Natural Language Processing*. Retrieved from <https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32>

- Gers, F., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: continual prediction with LSTM. *Ninth International Conference on Artificial Neural Networks*. 2, pp. 850-855. Edinburgh: Institution of Engineering and Technology. doi:10.1049/cp:19991218
- Ghanta, H. (2019). *Automated Essay Evaluation Using Natural Language Processing and Machine Learning*. Columbus State University, TSYS School of Computer Science. Columbus: Theses and Dissertations. Retrieved from https://csuepress.columbusstate.edu/theses_dissertations/327
- Ghosh, S. F. (2008). Design of an Automated Essay Grading (AEG) system in Indian context. *TENCON 2008 - 2008 IEEE Region 10 Conference*. doi:10.1109/tencon.2008.4766677
- Greene, P. (2 July, 2018). *Automated Essay Scoring Remains An Empty Dream*. Retrieved from Forbes: <https://www.forbes.com/sites/petergreene/2018/07/02/automated-essay-scoring-remains-an-empty-dream/#4474e4f74b91>
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28, 2222 - 2232. doi:10.1109/TNNLS.2016.2582924
- Han, J., Kamber, M., & Pei, J. (2012). 2 - Getting to Know Your Data. In J. Han, M. Kamber, & J. Pei, *Data Mining (Third Edition)* (pp. 39-82). Boston: Morgan Kaufmann. doi:<https://doi.org/10.1016/B978-0-12-381479-1.00002-2>
- Hruska, J. (15 February, 2018). *MIT Neural Network Processor Cuts Power Consumption by 95 Percent*. Retrieved from Extreme Tech:

<https://www.extremetech.com/computing/263951-mit-announces-new-neural-network-processor-cuts-power-consumption-95>

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated Language Essay Scoring Systems: A Literature Review. doi:10.7287/peerj.preprints.27715v1

IBM Cloud Education. (15 July, 2020). *IBM Cloud Learn Hub: What is Machine Learning?* Retrieved from IBM: <https://www.ibm.com/cloud/learn/machine-learning>

Imaki, J., & Ishihara, S. (2013). Experimenting with a Japanese automated essay scoring system in the L2 Japanese environment. *Papers in Language Testing and Assessment*, 2(2), 28-47.

Intelligent Essay Assessor (IEA)TM Fact Sheet. (2010). Retrieved 4 June, 2020, from Pearson Education: <https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf>

Ishioka, T., & Kameda, M. (July, 2006). Automated Japanese Essay Scoring System based on Articles Written by Experts. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 233–240. doi:10.3115/1220175.1220205

Islam, M. M., & Hoque, A. S. (2013). Automated Bangla essay scoring system: ABESS. *International Conference on Informatics, Electronics and Vision*. doi:10.1109/iciev.2013.6572694

Jin, C., He, B., Hui, K., & Sun, L. (2018). TDNN: A Two-stage Deep Neural Network for Prompt-independent Automated Essay Scoring. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. doi:10.18653/v1/p18-1100

- Ke, Q., Liu, J., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2018). Chapter 5 - Computer Vision for Human–Machine Interaction. In M. Leo, & G. M. Farinella (Eds.), *Computer Vision for Assistive Healthcare* (pp. 127-145). Academic Press. doi:<https://doi.org/10.1016/B978-0-12-813445-0.00005-8>
- Ke, Z., & Ng, V. (2019). Automated Essay Scoring: A Survey of the State of the Art. 6300-6308. doi:10.24963/ijcai.2019/879
- Knoema. (2018). *Sarawak - Number of students at government & government-aided secondary schools*. Retrieved from Knoema: <https://knoema.com/atlas/Malaysia/Sarawak/Number-of-students-at-government-secondary-schools>
- Knoema. (2018). *Sarawak - Number of teachers at government & government-aided secondary schools*. Retrieved from Knoema: <https://knoema.com/atlas/Malaysia/Sarawak/Number-of-teachers-at-government-secondary-schools>
- Kobayashi, Y., & Abe, M. (2016). Automated Scoring of L2 Spoken English with Random Forests. *Journal of Pan-Pacific Association of Applied Linguistics*, 20, 55-73.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284.
- Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. In M. A. Hearst (Ed.), *The debate on automated essay grading. IEEE Intelligent systems*, 27–31.

- Liang, G., On, B.-W., Jeong, D., Kim, H.-C., & Choi, G. S. (2018). Automated Essay Scoring: A Siamese Bidirectional LSTM Neural Network Architecture. *Symmetry*, 10, 682. doi:10.3390/sym10120682
- Lippmann, R. P. (April, 1987). An Introduction to Computing with Neural Nets. *IEEE Acoustics, Speech, and Signal Processing Society Magazine*, 4(2), 4-22.
- Loraksa, C., & Peachavanish, R. (2007). Automatic Thai-Language Essay Scoring Using Neural Network and Latent Semantic Analysis. *First Asia International Conference on Modelling & Simulation (AMS07)*. doi:10.1109/ams.2007.19
- Madasamy, K., & Ramaswami, M. (2017). Data Imbalance and Classifiers: Impact and Solutions. *International Journal of Computational Intelligence Research*, 13(9), 2267-2281. Retrieved from https://www.ripublication.com/ijcir17/ijcirv13n9_09.pdf
- Malaysian Examinations Council. (31 December, 2014). *MUET Regulations and Test Specifications*. Retrieved 15 March, 2021, from Malaysian Examination Councils Official Portal: https://www.mpm.edu.my/images/dokumen/calon-peperiksaan/muet/regulation/Regulations_Test_Specifications_Test_Format_and_Sample_Questions.pdf
- Malaysian Examinations Council. (18 July, 2019). *MUET Specification Regulation*. Retrieved from Malaysian Examination Councils Official Portal: https://www.mpm.edu.my/images/dokumen/calon-peperiksaan/muet/regulation/Test_Specification_Regulation.pdf
- Mathias, S., & Bhattacharyya, P. (May, 2018). ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

- Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from <https://aclanthology.org/L18-1187>
- Mbaabu, O. (11 December, 2020). *Introduction to Random Forest in Machine Learning*. Retrieved from Section: <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/>
- Measurement Incorporated*. (2020). Retrieved 4 June, 2020, from <https://www.measurementinc.com/products-services/automated-essay-scoring>
- Merriam-Webster. (n.d.). *Algorithm*. Retrieved 28 June, 2022, from Merriam-Webster.com dictionary: <https://www.merriam-webster.com/dictionary/algorithm>
- Mostafa, S., & Wu, F.-X. (2021). Chapter 3 - Diagnosis of autism spectrum disorder with convolutional autoencoder and structural MRI images. In A. S., El-Baz, & J. S. Suri (Eds.), *Neural Engineering Techniques for Autism Spectrum Disorder* (pp. 23-38). Academic Press. doi:<https://doi.org/10.1016/B978-0-12-822822-7.00003-X>
- Nguyen, H. (2016). *Neural Networks for Automated Essay Grading*.
- Omar, N., & Mezher, R. (2016). A Hybrid Method of Syntactic Feature and Latent Semantic Analysis for Automatic Arabic Essay Scoring. *Journal of Applied Sciences*, 16(5), 209–215. doi:10.3923/jas.2016.209.215
- Ong, D. A., Razon, A. R., Perigrino, J. M., Guevara, R. C., & Naval, P. C. (2011). *Automated Filipino Essay Grader with Concept-Indexing*.
- Page, E. B. (1966). The Imminence of... Grading Essays by Computer. *Phi Delta Kappan*, 47(5), 238–243.

- Peng, X., Ke, D., Chen, Z., & Xu, B. (2010). Automated Chinese Essay Scoring using Vector Space Models. *2010 4th International Universal Communication Symposium*, 149-153. doi:10.1109/IUCS.2010.5666229
- Phandi, P., Chai, K. M., & Ng, H. T. (2015). Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 431–439. doi:10.18653/v1/D15-1049
- Phi, M. (25 September, 2018). *Illustrated Guide to LSTM's and GRU's: A step by step explanation*. Retrieved from Towards Data Science: <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>
- PIM, L. H. (19 March, 2018). *Sarawak lacks well-trained teachers for three subjects*. Retrieved from The Borneo Post: <https://www.theborneopost.com/2018/03/19/sarawak-lacks-well-trained-teachers-for-three-subjects-manyin/>
- Rahutomo, F., Kitasuka, T., & Aritsugi, M. (2012). Semantic Cosine Similarity. *The 7th International Student Conference on Advanced Science and Technology ICAST 2012*.
- Rajagede, R. A., & Hastuti, R. P. (2021). Stacking Neural Network Models for Automatic. *IOP Conference Series: Materials Science and Engineering*, 1077, 012013. doi:10.1088/1757-899X/1077/1/012013

- Ramalingam, V. V., Pandian, A., Chetry, P., & Nigam, H. (2018). Automated Essay Grading using Machine Learning Algorithm. *Journal of Physics: Conference Series*, 1000, 012030. doi:10.1088/1742-6596/1000/1/012030
- Ramesh, D., & Sanampudi, S. K. (1 March, 2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. doi:10.1007/s10462-021-10068-2
- Ramineni, C., & Williamson, D. (2018). Understanding Mean Score Differences Between the e-rater® Automated Scoring Engine and Humans for Demographically Based Groups in the GRE® General Test. *ETS Research Report Series*, 1, 1-31. doi:10.1002/ets2.12192
- Ratna, A. A., Arbani, A., Ibrahim, I., Ekadiyanto, F., Bangun, K., & Purnamasari, P. (2018). Automatic Essay Grading System Based on Latent Semantic Analysis with Learning Vector Quantization and Word Similarity Enhancement. 120-126. doi:10.1145/3293663.3293684
- Ratna, A. A., Hartanto, D., & Budiardjo, B. (2007). SIMPLE: System Automatic Essay Assessment for Indonesian Language Subject Examination. *Makara Seri Teknologi*.
- Ratna, A. A., Kaltsum, A., Santiar, L., Khairunissa, H., Ibrahim, I., & Purnamasari, P. D. (2019a). Term Frequency-Inverse Document Frequency Answer Categorization with Support Vector Machine on Automatic Short Essay Grading System with Latent Semantic Analysis for Japanese Language. 293-298. doi:10.1109/ICECOS47637.2019.8984530
- Ratna, A. A., Khairunissa, H., Kaltsum, A., Ibrahim, I., & Purnamasari, P. D. (2019b). Automatic Essay Grading for Bahasa Indonesia with Support Vector Machine and

- Latent Semantic Analysis. *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 363-367.
- Ratna, A. A., Santiar, L., Ibrahim, I., Purnamasari, P., Luhurkinanti, D., & Larasati, A. (2019c). Latent Semantic Analysis and Winnowing Algorithm Based Automatic Japanese Short Essay Answer Grading System Comparative Performance. 1-7. doi:10.1109/ICA_wST.2019.8923226
- Ratna, A., Purnamasari, P., & Adhi, B. (2015). Simple-O, An Automated Essay Grading System for Indonesian Language Using the LSA Method with Multi-Level Keywords.
- Saha, S. (16 December, 2018). *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. Retrieved from Towards Data Science: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>
- Saxena, S. (26 October, 2017). *Artificial Neuron Networks(Basics) | Introduction to Neural Networks*. Retrieved from <https://becominghuman.ai/artificial-neuron-networks-basics-introduction-to-neural-networks-3082f1dcca8c>
- Schwab, B. (2009). *AI Game Engine Programming* (2nd ed.). Boston: Course Technology.
- Sendra, M., Sutrisno, R., Harianata, J., Suhartono, D., & Asmani, A. (2016). Enhanced Latent Semantic Analysis by considering mistyped words in automated essay scoring. 304-308. doi:10.1109/IAC.2016.7905734
- Shehab, A., Faroun, M., & Rashad, M. (2018). An Automatic Arabic Essay Grading System based on Text Similarity Algorithms. *International Journal of Advanced Computer Science and Applications*, 9. doi:10.14569/IJACSA.2018.090337

- Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated Essay Scoring: Writing Assessment and Instruction. *International Encyclopedia of Education*, 20-26. doi:10.1016/B978-0-08-044894-7.00233-5
- Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical Therapy*, 85(3), 257–268. doi:<https://doi.org/10.1093/ptj/85.3.257>
- Srivastava, K., Dhanda, N., & Shrivastava, A. (March, 2020). An Analysis of Automated Essay Grading Systems. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(6). doi:10.35940/ijrte.f9938.038620
- Taghipour, K., & Ng, H. T. (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
- Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data* 7, 70. Retrieved from <https://doi.org/10.1186/s40537-020-00349-y>
- The Hewlett Foundation: Automated Essay Scoring. (2 February, 2012). *The Hewlett Foundation: Automated Essay Scoring*. Retrieved from Kaggle: <https://www.kaggle.com/competitions/asap-aes/overview>
- Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2), 459–484. doi:10.1007/s41237-021-00142-y
- Uto, M., Xie, Y., & Ueno, M. (December, 2020). Neural Automated Essay Scoring Incorporating Handcrafted Features. *Proceedings of the 28th International*

- Conference on Computational Linguistics*, 6077–6088.
doi:10.18653/v1/2020.coling-main.535
- Vantage Learning*. (2005). Retrieved 4 June, 2020, from
http://www.vantagelearning.com/docs/intellimetric/IM_How_IntelliMetric_Works.pdf
- Vantage Learning*. (2020). Retrieved 4 June, 2020, from
<http://www.vantagelearning.com/products/intellimetric/faqs/#LongUsed>
- Vargas, R., Mosavi, A., & Ruiz, R. (June, 2017). DEEP LEARNING: A REVIEW. *Advances in Intelligent Systems and Computing*, 5.
- Wilczek, F. (2011). *What Scientific Concept Would Improve Everybodt's Cognitive Toolkit?*
Retrieved from <https://www.edge.org/response-detail/10351>
- Wong, W. S., & Bong, C. H. (2019). A Study for the Development of Automated Essay Scoring (AES) in Malaysian English Test Environment. *International Journal of Innovative Computing*, 9(1). doi:10.11113/ijic.v9n1.220
- Wong, W. S., & Bong, C. H. (April , 2021). Assessing Malaysian University English Test (MUET) Essay on Language and Semantic Features Using Intelligent Essay Grader (IEG). *Pertanika Journal of Science & Technology*, 29(2), 919-941.
- Xu, Y., Ke, D., & Su, K. (2016). Contextualized Latent Semantic Indexing: A New Approach to Automated Chinese Essay Scoring. *Journal of Intelligent Systems*, 26.
doi:10.1515/jisys-2015-0048
- Yu, J.-L., Kuo, B.-C., & Pai, K.-C. (2017). Developing Chinese Automated Essay Scoring Model to Assess College Students' Essay Quality.

Zupanc, K., & Bosnic, Z. (2015). Advances in the Field of Automated Essay Evaluation.
Informatica (Slovenia), 39.

APPENDIX

Appendix A: Journal Publications

1. **Lim, C. T.**, and Bong, C. H., and Wong, W. S., and Lee, N. K. (2021). A Comprehensive Review of Automated Essay Scoring (AES) Research and Development. *Pertanika Journal of Science & Technology*, 29(3), 1875-1899.