



Faculty of Computer Science and Information Technology

**Feature Engineering for Automated Essay Evaluator of Malaysian
University English Test (MUET) based on Linguistic Features**

Wong Wee Sian

**Doctor of Philosophy
2024**

Feature Engineering for Automated Essay Evaluator of Malaysian University
English Test (MUET) based on Linguistic Features

Wong Wee Sian

A thesis submitted

In fulfillment of the requirements for the degree of Doctor of Philosophy

(Computer Science)

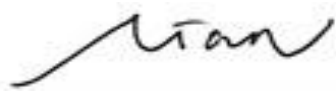
Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2024

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.



.....

Signature

Name: Wong Wee Sian

Matric No.: 14010085

Faculty of Computer Science and Information Technology

Universiti Malaysia Sarawak

Date: 18 March 2024

ACKNOWLEDGEMENT

This thesis is the result of extensive time and effort, representing a significant journey in my life.

First and foremost, my wholehearted gratitude goes to God. His grace and blessings have enabled me to complete the work. "The fear of the LORD is the beginning of wisdom." (The Bible).

I am truly indebted to Associate Professor Dr. Bong Chih How, my thesis supervisor, who not only imparted me the knowledge but also personal advice and encouragement, especially when I cannot meet the expectation. It can truly be said this work cannot be completed without his guidance.

I would like to express my gratitude to my beloved wife, Sharon, who never failed to be there for me, while I was occupied with my study.

My appreciation extends to all the lecturers, staff, and students in the Faculty of Computer Science and Information Technology and the Centre for Graduate Studies in UNIMAS who helped me during the journey.

A thesis dedicated to the memory of my late mother.

ABSTRACT

Automated Essay Scoring (AES) refers to the use of specialized computer programs to assess and score essays for overcoming time, cost, and reliability issues in an educational assessment context. It pertains to applications in the field of Natural Language Processing (NLP) and computational linguistics, which centres on the interactions between computer software and human languages. Several prominent proprietary AES systems are available in the commercial domain, and extensive academic research has been conducted to explore automated essay scoring. One of the issues in AES is its dependence on surface features (e.g., essay length) to score essays. These AES are often criticized because their scoring mechanisms are not associated with the rationale of how human raters typically score essays. Surface-level features from AES do not capture the linguistic aspects of an essay. To address the constraint of this “surface-level” assessment, several recent research have emerged, focusing on leveraging deep linguistic features, such as text cohesion and lexical diversity to assess essays. However, most of this research concentrates on specific linguistic dimensions – none of them provide comprehensive coverage of linguistic dimensions to score essays. Furthermore, AES systems, especially the commercial proprietary and deep neural network AES, exhibit a black-box nature. This non-transparent operation of the AES restrains the clear explanation and interpretation of essay features and scoring mechanisms employed for scoring essays. In response to these AES issues, this research is conducted to develop an AES system, namely the Automated Essay Evaluator (AEE), to score essays based on comprehensive deep linguistic features. It employed the Malaysian University English Test (MUET) essay as the case study for automated essay scoring. The research identified and categorized a total of 1,709 comprehensive linguistic feature indices into a taxonomy comprising eight distinct linguistic feature sets, and 43 linguistic feature

categories. These eight linguistic feature sets, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability, should be able to cover most if not all the linguistic features found in essays. A thorough correlation analysis between the linguistic features and the essay grades was conducted. Two feature selection schemes, namely the Correlation Rank and Minimum Redundancy Maximum Relevance (MRMR) Feature Selection have been formulated to select the optimized linguistic features that influence essay scoring. The overall performance of the selected linguistic features was evaluated using six different machine learning classifiers to score MUET essays. Lastly, an interpretation of the proposed linguistic feature set with the MUET essay scoring rubrics has been provided to explain how these linguistic features contribute to the overall essay score. According to the experiment result, this research found that readability, surface features, lexical diversity, and specific lexical sophistication are strong predictors of MUET essay scores. The linguistic features selected by the Correlation Rank and MRMR Feature Selection Scheme outperformed the baseline scheme, which consists of 50 randomly selected features. Furthermore, the linguistic-based automated scoring developed in this research demonstrated superior performance than the LigthSide AES vendor in scoring MUET essays. This linguistic-based essay scoring proposed can be used as the basis for developing a complete full-fledged local Malaysian AES by incorporating essay content features.

Keywords: Automated essay scoring, linguistic features, natural language processing, computational linguistic, machine learning

Kejuruteraan Ciri untuk Penilai Esei Automatik Malaysian University English Test (MUET) berdasarkan Ciri-Ciri Linguistik

ABSTRAK

Pemarkahan Esei Automatik atau “Automated Essay Scoring” (AES) merujuk kepada penggunaan program komputer khusus untuk menilai dan skor esei untuk mengatasi masalah masa, kos dan kebolehpercayaan dalam konteks penilaian pendidikan. Ia melibatkan aplikasi dalam bidang pemprosesan bahasa semulajadi dan linguistik komputasi, fokus kepada interaksi antara perisian komputer dan bahasa manusia. Beberapa sistem AES proprietari yang terkenal tersedia dalam domain komersial, dan penyelidikan akademik yang meluas telah dijalankan untuk meneroka pemarkahan esei automatik. Salah satu isu dalam AES ialah pergantungannya pada ciri dasar (cth., panjang esei) untuk skor esei. AES ini sering dikritik kerana mekanisme pemarkahan mereka tidak dikaitkan dengan rasional bagaimana penilai manusia biasanya skor esei. Ciri dasar daripada AES tidak menangkap aspek linguistik sesebuah esei. Untuk menangani penilaian “peringkat dasar” ini, beberapa kajian terbaru telah muncul, fokus kepada penggunaan ciri linguistik yang mendalam, seperti kohesi teks dan kepelbagaian leksikal untuk menilai esei. Walau bagaimanapun, kebanyakan penyelidikan ini memberi tumpuan kepada dimensi linguistik tertentu - tiada satu pun memberikan liputan komprehensif dimensi linguistik untuk skor esei. Tambahan pula, sistem AES, terutamanya proprietari komersial dan rangkaian neural AES, mempamerkan sifat kotak hitam. Operasi AES yang tidak telus ini menghalang penjelasan dan tafsiran yang jelas tentang ciri esei dan mekanisme pemarkahan esei. Memandangkan isu-isu tersebut, penyelidikan ini telah dijalankan untuk membangunkan satu sistem AES, iaitu “Automated Essay Evaluator” (AEE), untuk menilai pemarkahan esei automatik berdasarkan ciri linguistik mendalam yang komprehensif. Ia menggunakan esei

“Malaysian University English Test” (MUET) sebagai kajian kes untuk pemarkahan esei automatik. Penyelidikan ini mengenal pasti dan mengkategorikan sejumlah 1,709 indeks ciri linguistik komprehensif ke dalam taksonomi yang terdiri daripada lapan set ciri linguistik, dan 43 kategori ciri linguistik. Lapan set ciri linguistik ini, iaitu ciri dasar, kesilapan linguistik, kohesi teks, semantik, kepelbagaian leksikal, kecanggihan leksikal, kerumitan sintaksis, dan kebolehbacaan, seharusnya dapat merangkumi kebanyakan jika tidak semua ciri linguistik dalam esei. Analisis korelasi yang menyeluruh antara ciri linguistik dan skor esei telah dijalankan. Dua skema pemilihan ciri, iaitu “Correlation Rank” dan “Minimum Redundancy Maximum Relevance” (MRMR) telah digubal untuk memilih ciri linguistik yang optimum dalam pemarkahan esei. Prestasi keseluruhan ciri linguistik yang dipilih telah dinilai dengan enam pengelasan pembelajaran mesin untuk skor esei MUET. Akhir sekali, tafsiran bagi set ciri linguistik dengan rubrik pemarkahan esei MUET telah disediakan untuk menerangkan bagaimana ciri linguistik ini menyumbang kepada skor keseluruhan esei. Kajian ini mendapati bahawa kebolehbacaan, ciri dasar, kepelbagaian leksikal, dan kecanggihan leksikal tertentu adalah peramal yang kuat bagi skor esei MUET. Ciri linguistik yang dipilih oleh “Correlation Rank” dan MRMR mengatasi skema garis dasar, yang terdiri daripada 50 ciri yang dipilih secara rawak. Pemarkahan automatik berasaskan linguistik dalam kajian ini menunjukkan prestasi yang lebih baik daripada vendor LightSide AES dalam pemarkahan esei MUET. Pemarkahan esei berasaskan linguistik ini dapat menjadi asas untuk pembangunan AES tempatan Malaysia yang lengkap dengan integrasi ciri-ciri kandungan esei.

Kata kunci: *Pemarkahan esei automatik, ciri linguistik, pemprosesan bahasa semula jadi, linguistik komputasi, pembelajaran mesin*

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
ABSTRAK	v
TABLE OF CONTENTS	vii
LIST OF TABLES	xiv
LIST OF FIGURES	xvii
LIST OF ABBREVIATIONS	xix
CHAPTER 1 INTRODUCTION	1
1.1 Background of the Research	1
1.2 Automated Essay Scoring Process	2
1.3 Motivation of the Research	5
1.4 Problem Statement	7
1.4.1 Surface-level features from AES do not capture the linguistic aspects of an essay	8
1.4.2 Lack of comprehensive coverage of linguistic feature dimensions in AES research	9

1.4.3	The black box nature of AES (particularly the commercial proprietary and deep neural network AES)	10
1.4.4	The relationship between the problem statements and MUET essay scoring	11
1.5	Research Objectives and Research Questions	13
1.6	Significance of the Research	14
1.7	Scope of the Research	16
1.8	Thesis Organization	18
1.9	Chapter Summary	19
	CHAPTER 2 LITERATURE REVIEW	20
2.1	Automated Essay Scoring System	20
2.1.1	Project Essay Grader (PEG)	21
2.1.2	Intelligent Essay Assessor (IEA)	21
2.1.3	E-rater	22
2.1.4	IntelliMetric	23
2.1.5	Bayesian Essay Test Scoring sYstem (BETSY)	23
2.1.6	Coh-Metrix	24
2.1.7	LightSide	26
2.1.8	Deep Neural Network-based AES (DNN-AES)	26
2.1.9	Semantic Essay Grader for Essays (SAGE)	30
2.1.10	Comparison of the State-of-the-Art AES Systems	31

2.1.11	Summary of AES Systems	35
2.2	Essay Features for Automated Essay Scoring	37
2.3	Overview of Linguistic Features	41
2.3.1	Surface Features	42
2.3.2	Linguistic Errors	43
2.3.3	Text Cohesion	45
2.3.4	Semantics	46
2.3.5	Lexical Diversity	48
2.3.6	Lexical Sophistication	50
2.3.7	Syntactic Complexity	51
2.3.8	Readability	53
2.4	Overview of Content Features	54
2.4.1	Prompt Relevance / Adherence	55
2.4.2	Thesis Clarity	56
2.4.3	Organization and Development	56
2.4.4	Argumentation	57
2.5	Feature Selection	58
2.5.1	Feature Selection Techniques	59
2.5.2	Feature Selection in AES Research	63
2.6	Trends and Challenges in Automated Essay Scoring Research	66

2.6.1	Essay Features	66
2.6.2	Techniques of Essay Feature Extraction	67
2.6.3	Essay Dataset	68
2.6.4	Performance of AES Systems	69
2.7	Chapter Summary	72
CHAPTER 3 MATERIALS AND METHODS		76
3.1	Malaysian University English Test	76
3.1.1	Scoring Rubrics of Malaysian University English Test	77
3.2	Dataset for Essay Analysis and Scoring	79
3.3	Linguistic Features Specification	81
3.3.1	Taxonomy of Feature Set, Feature Categories, and Feature Indices	83
3.3.2	Surface Features	86
3.3.3	Linguistic Errors	88
3.3.4	Text Cohesion	90
3.3.5	Semantics	92
3.3.6	Lexical Diversity	95
3.3.7	Lexical Sophistication	98
3.3.8	Syntactic Complexity	103
3.3.9	Readability	106
3.4	Feature Selection of Linguistic Features	110

3.4.1	Correlation Rank Feature Selection	111
3.4.2	Minimum Redundancy Maximum Relevance (MRMR) Feature Selection	112
3.5	Chapter Summary	114
CHAPTER 4 EXPERIMENTS		116
4.1	Experiment Framework	116
4.2	Feature Extraction	118
4.3	Feature Evaluation	119
4.3.1	Spearman's Rank Correlation Coefficient	121
4.3.2	Kendall's Rank Correlation Coefficient	121
4.4	Feature Selection	122
4.4.1	Correlation Rank Feature Selection	122
4.4.2	Minimum Redundancy Maximum Relevance (MRMR) Feature Selection	124
4.5	Automated Essay Scoring Using Linguistic Features	124
4.5.1	Machine Learning Classifier	124
4.5.2	Automated Essay Evaluator (AEE) Models	127
4.5.3	Evaluation Metrics	129
4.6	Boxplot	134
4.7	Chapter Summary	136
CHAPTER 5 RESULTS AND DISCUSSION		137
5.1	Spearman's and Kendall's Rank Correlation Coefficient	138

5.2	Individual Linguistic Feature Set Evaluation	139
5.2.1	Surface Features	139
5.2.2	Linguistic Errors	141
5.2.3	Text Cohesion	142
5.2.4	Semantics	144
5.2.5	Lexical Diversity	147
5.2.6	Lexical Sophistication	148
5.2.7	Syntactic Complexity	151
5.2.8	Readability	152
5.3	Comprehensive Linguistic Feature Set Evaluation	154
5.4	Comprehensive Linguistic Feature Category Evaluation	155
5.5	Feature Selection	158
5.6	Performance of Automated Essay Scoring Using Linguistic Features	161
5.6.1	Effectiveness of Linguistic Features in Predicting Essay Score	161
5.6.2	Optimal Linguistic Features Selected by Feature Selection Schemes	166
5.6.3	Comparison of AEE Performance with LightSide AES	168
5.6.4	Comparison of AEE Models with the Highest and Lowest Performance	170
5.6.5	Evaluation of the Scoring of Different Essay Grades	176
5.7	Chapter Summary	182
	CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS	184

6.1	Summary of the Research	184
6.2	Summary and Discussion of Research Findings	185
6.2.1	Evaluation of Comprehensive Linguistic Features in Essay Scoring	185
6.2.2	Feature Selection of Optimal Linguistic Features in Essay Scoring	187
6.2.3	Performance of the Selected Linguistic Features in Essay Scoring	188
6.2.4	Mapping of MUET Scoring Rubrics with Linguistic Features	189
6.3	Contributions of the Research	195
6.3.1	Contribution to Body of Knowledge	195
6.3.2	Contribution to Practice	196
6.4	Implications of Linguistic-based Automated Essay Analysis and Scoring	197
6.4.1	Implication for Computer Science	197
6.4.2	Implication for Writing Assessment	197
6.4.3	Implication for Writing Instruction	198
6.5	Limitations and Future Works	199
6.5.1	The Constraint of Linguistic Feature Scoring	199
6.5.2	The Limitation of Essay Datasets	200
6.5.3	The Constraint of the Machine Learning Models	202
6.6	Concluding Remarks	204
	REFERENCES	206
	APPENDICES	235

LIST OF TABLES

	Page
Table 1.1: Comprehensive Linguistic Feature Sets	5
Table 2.1: Coh-Metrix Features	25
Table 2.2: A Comparison of the State-of-the-Art AES Systems	33
Table 2.3: Progress in AES Development	36
Table 2.4: An Overview of Essay Features for AES	39
Table 2.5: Feature Selection Techniques based on Filter Methods	60
Table 2.6: Feature Selection Techniques based on Wrapper Methods	61
Table 2.7: Feature Selection Techniques based on Embedded Methods	63
Table 2.8: Strengths and Weaknesses of Handcrafted Feature Engineering versus Neural-based Approaches in AES	68
Table 2.9: Performance of the State-of-the-Art AES Systems	70
Table 3.1: Scoring Rubrics of MUET Essay	78
Table 3.2: Grade and Score Distribution of MUET Essay Dataset	81
Table 3.3: Application of Linguistic Features in Assessing Essays	82
Table 3.4: Surface Feature Set	87
Table 3.5: Linguistic Error Feature Set	89
Table 3.6: Text Cohesion Feature Set	91
Table 3.7: Semantic Feature Set	93
Table 3.8: Lexical Diversity Feature Set	96
Table 3.9: Lexical Sophistication Feature Set	99
Table 3.10: Corpora used for Lexical Sophistication Features Indices Calculation	102
Table 3.11: Syntactic Complexity Feature Set (Phrasal Complexity, Clausal Complexity and Syntactic Sophistication)	105
Table 3.12: Syntactic Complexity Feature Set (Sentential Complexity)	106

Table 3.13: Readability Feature Set	108
Table 3.14: The Complete List of Linguistic Feature Sets and Categories Investigated in this Research	115
Table 4.1: Linguistic Analysis Tools	119
Table 4.2: Criteria for Correlation Rank Feature Selection	123
Table 4.3: Guideline for Interpreting the Size of a Correlation Coefficient	123
Table 4.4: Number of Selected Features by MRMR Feature Selection	124
Table 4.5: Machine Learning Algorithms Employed in the Experiment	125
Table 4.6: Hyperparameters Setting of the Machine Learning Classifier	126
Table 4.7: The 48 AEE Models Employed in the Experiment	128
Table 5.1: Standard Deviation of Feature Values by Different Semantic Analysis Techniques	147
Table 5.2: Number of Linguistic Feature Indices Used by Different Feature Selection Schemes	158
Table 5.3: Linguistic Feature Indices Selected by COR-07 Feature Selection	159
Table 5.4: Linguistic Feature Indices Selected by MRMR-10 Feature Selection	160
Table 5.5: Performance of Baseline, Correlation Rank, and MRMR Feature Selection Schemes	161
Table 5.6: Similar Features Indices Selected by COR-07, MRMR-10, and MRMR-25 Feature Selection	165
Table 5.7: Linguistic Feature Indices Frequently Selected by the COR-05, COR-07, MRMR-10, MRMR-25, and MRMR-50 Feature Selection	167
Table 5.8: Confusion Matrix - Support Vector Machine with MRMR-100 Features	173
Table 5.9: Confusion Matrix - Support Vector Machine with MRMR-10 Features	174
Table 5.10: Confusion Matrix - Random Forest with COR-03 Features	175
Table 5.11: Confusion Matrix - Random Forest with COR-07 Features	175
Table 5.12: Confusion Matrix – Random Forest with MRMR-100 Features	176
Table 5.13: Confusion Matrix - Generalized Linear Model with COR-03 Features	177
Table 5.14: Confusion Matrix - Generalized Linear Model with COR-05 Features	177

Table 5.15: Confusion Matrix - Generalized Linear Model with COR-07 Features	178
Table 5.16: Confusion Matrix - Generalized Linear Model with MRMR-10 Features	178
Table 5.17: Confusion Matrix - Generalized Linear Model with MRMR-25 Features	178
Table 5.18: Confusion Matrix - Generalized Linear Model with MRMR-50 Features	179
Table 5.19: Confusion Matrix - Generalized Linear Model with MRMR-100 Features	179
Table 5.20: Confusion Matrix - Generalized Linear Model with Baseline Features	179
Table 5.21: Class Precision and Recall of Each Essay Grade	180
Table 6.1: Optimal Linguistic Features Identified in this Research	188
Table 6.2: Detailed Mapping of MUET Rubric Specifications with AEE Linguistic Feature Set	191

LIST OF FIGURES

	Page
Figure 1.1: Illustration of an Automated Essay Scoring Process	3
Figure 2.1: Architecture of Convolutional Recurrent Neural Network	28
Figure 2.2: Classification of AES Systems	73
Figure 2.3: Essay Features for Automated Essay Scoring	74
Figure 2.4: Feature Selection Techniques based on Filter, Wrapper and Embedded Methods	75
Figure 3.1: The MUET Essay Prompt used in this Research	80
Figure 3.2: The Eight Linguistic Feature Sets in this Research	84
Figure 3.3: Taxonomy for Semantic Feature Set	85
Figure 4.1: Experiment Framework for Essay Analysis and Scoring	118
Figure 4.2: Components of a Boxplot	134
Figure 5.1: Comparison of Spearman's ρ and Kendall's τ Value	138
Figure 5.2: Correlation between Surface Feature Set and Essay Scores	140
Figure 5.3: Correlation between Linguistic Error Feature Set and Essay Scores	141
Figure 5.4: Correlation between Linguistic Error Feature Indices and Essay Scores	142
Figure 5.5: Correlation between Text Cohesion Feature Set and Essay Scores	143
Figure 5.6: Correlation between Semantic Feature Set and Essay Scores	145
Figure 5.7: Correlation between Applied Semantic Analysis Techniques and Essay Scores	146
Figure 5.8: Correlation between Lexical Diversity Feature Set and Essay Scores	148
Figure 5.9: Correlation between Lexical Sophistication Feature Set and Essay Scores	149
Figure 5.10: Correlation between Applied Word Form and Uniqueness with Essay Scores	150

Figure 5.11: Correlation between Syntactic Complexity Feature Set and Essay Scores	151
Figure 5.12: Correlation between Readability Feature Set and Essay Scores	153
Figure 5.13: Correlation between Readability Feature Indices and Essay Scores	153
Figure 5.14: Correlation between Comprehensive Linguistic Feature Set and Essay Scores	155
Figure 5.15: Correlation between Comprehensive Linguistic Feature Category and Essay Scores	157
Figure 5.16: QWK and Accuracy of Baseline, Correlation Rank and MRMR Feature Selection	163
Figure 5.17: Precision and Recall of Baseline, Correlation Rank and MRMR Feature Selection	164
Figure 5.18: F1 Score of Baseline, Correlation Rank, and MRMR Feature Selection	164
Figure 5.19: Comparison of QWK and Accuracy of Correlation Rank and MRMR Feature Selection with LightSide AES Features	170
Figure 5.20: AEE Models with High QWK and Accuracy	171
Figure 5.21: AEE Models with Low QWK and Accuracy	172
Figure 5.22: Class Precision and Recall of Each Essay Grade	181
Figure 6.1: High Level Mapping of MUET Rubrics with AEE Linguistic Feature Set	194

LIST OF ABBREVIATIONS

AEE	Automated Essay Evaluator
AES	Automated Essay Scoring
ASAP	Automated Student Assessment Prize
BETsY	Bayesian Essay Test Scoring sYstem
CAREC	Crowdsourced Algorithm of Reading Comprehension
CARES	Crowdsourced Algorithm of Reading Speed
CML2RI	Coh-Metrix L2 Readability Index
CNN	Convolution Neural Network
COCA	Corpus of Contemporary American English
DNN-AES	Deep Neural Network-based Automated Essay Scoring
ETS	Educational Testing Service
IEA	Intelligent Essay Assessor
L1	First Language
L2	Second Language
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory Network
MEC	Malaysian Examinations Council
MRMR	Minimum Redundancy Maximum Relevance
MUET	Malaysian University English Test
NLP	Natural Language Processing
PEG	Project Essay Grader

QWK	Quadratic Weighted Kappa
RNN	Recurrent Neural Network
SVD	Singular Value Decomposition
TOEFL	Test of English as a Foreign Language
TOEFL11	Corpus of Non-Native English TOEFL
TTR	Type-Token-Ratio

CHAPTER 1

INTRODUCTION

This chapter presents the introduction of this research which consists of nine sections. Section 1.1 presents the background of the research. This is followed by Section 1.2, which illustrates the automated essay scoring process. Section 1.3 states the motivation of this research, while Section 1.4 formulates the problem statement of the research. Section 1.5 outlines the research objectives with the research questions. Sections 1.6 and Section 1.7 present the significance and scope of the research accordingly. The sections are followed by a brief description of the organization of this thesis in Section 1.8 and a chapter summary in Section 1.9.

1.1 Background of the Research

Writing is considered to be one of the key 21st-century skills and has been incorporated as a critical component in many academic assessments (Foltz, 2016). The primary function of writing is no longer simply the conveyance of information but is perceived as the association of high-order cognitive capabilities such as critical thinking and reasoning. However, assessment of a student's writing or essay is by no means an easy task and can be time-consuming. It is undeniable that manual essay marking by human raters can be laborious and burdensome. The essays are difficult to score in an efficient, economical, and objective manner (Latifi, 2016). One of the fundamental issues in essay assessment is the time it takes to accomplish the scoring process, especially in the large-scale high-stakes language assessment environment. The vast number of essays needed to go through multiple human raters and adjudication of the scores where discrepancies occurred. To alleviate the workload and improve efficiency, extra staff may need to be hired for the marking process,

which can translate into additional costs. On the other hand, human markers can be inconsistent and subjective due to certain judgments and biases, and thus the same essay might have as many different grades as it does. Despite the unified grading rules, human graders can unintentionally introduce subjective bias into scores (Zupanc & Bosnić, 2018).

One feasible solution to address these problems is to automate the essay scoring process by incorporating the technology of Automated Essay Scoring (AES). AES is a computer-based assessment system that automatically scores or grades the student's essay by considering appropriate essay features (Ramesh & Sanampudi, 2021). Based on the extracted features that are deemed salient to the essay scores, AES constructs scoring models from pre-scored essays, using natural language processing (NLP) and machine learning approaches, and then employs these models to grade new sets of essays (Bennett & Zhang, 2016). AES offers many potential advantages for writing assessments, such as improving the quality of scoring, reducing time for score reporting, minimizing cost and coordination efforts for human raters, and the possibility of providing immediate feedback to students on their writing performance (Gierl et al., 2014; Foltz, 2016).

1.2 Automated Essay Scoring Process

AES is a multi-disciplinary field that incorporates research from computer science, linguistics, cognitive science, writing research, and education measurement (Shermis et al., 2013). Despite the availability of different AES systems, they all use similar fundamental processes to score written assessments. Figure 1.1 illustrates the end-to-end scoring process of an AES system.

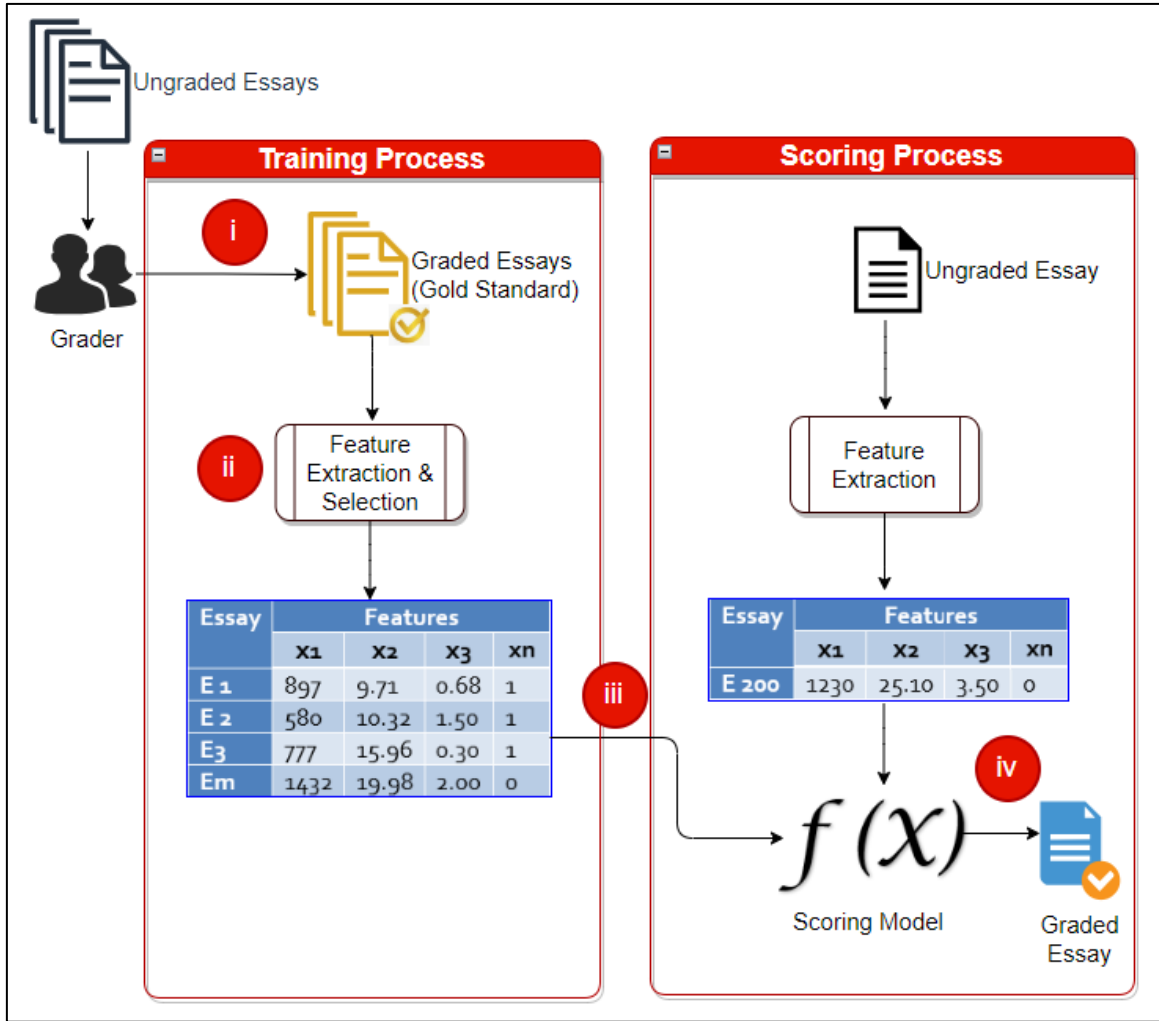


Figure 1.1: Illustration of an Automated Essay Scoring Process

The process can be decomposed into a few steps below:

i. Data Pre-processing

The initial step to building an AES is data pre-processing. Before the actual process automation, the expert human graders need to score a dataset of prompt-specific essays (i.e., essays on the same topic) manually. These graded essays will be used as the gold standard for the subsequent machine learning processes. Besides this human scoring process, the essays must be converted to the file format (e.g., text or csv file) that can be processed by the AES system.

ii. Feature Engineering

Once the data pre-processing is complete, the graded essays are inputted into the AES system for feature engineering. The purpose of feature engineering is to transform essays into a set of feature values. These essay features may include simple length-based features (e.g., number of words, number of sentences) and complex deep linguistic features (e.g., semantics and readability). The feature engineering phase involves two processes, namely the feature extraction to extract essay features and the feature selection to select key features, that are relevant or significant to the essay score. The selected features can be further transformed into reduced dimensions. Multiple reduction techniques are available to accomplish the task of feature reduction, e.g., Principal Component Analysis (PCA; Abdi & Williams, 2010), Latent Semantic Analysis (LSA; Landauer et al., 1998), Latent Dirichlet Allocation (LDA; Blei et al., 2003).

iii. Machine Learning Model Building and Evaluation

Once the essay features have been identified and extracted, the next step is the application of machine learning techniques to build the models. The AES scoring task is often addressed as a supervised machine learning problem, where a scoring model is used to predict the final score. The scoring model can either apply regression to predict the exact scores or adopt classification algorithms to classify discrete essay grades. The cross-validation sampling is then applied to evaluate and select the optimal machine learning model for scoring the essays.

iv. Essay Scoring

Once the machine learning model has been ready, it can be applied to analyse and score new ungraded essays. The AES system uses the data from the graded essays as

the gold standard to build the model that provides the foundations for scoring ungraded essays. The model follows the same scoring model to score new essays.

1.3 Motivation of the Research

The primary motivation and purpose of this research is to conduct a systematic and detailed investigation of linguistic features in assessing essay quality. Investigations on scoring writing quality should not concentrate solely on individual linguistic features (Crossley, 2020). Instead, linguistic research on writing ought to analyses various linguistic features to determine its overall quality. Therefore, this research is conducted to investigate multiple and comprehensive linguistic feature sets, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability in influencing essay scores.

The author attempts to exhaustively explore all potential linguistic features for scoring essays by reviewing numerous linguistic and AES-related literature. Based on the literature review, the author proposed the above eight linguistic feature sets as the comprehensive features in scoring essays. A vast number of comprehensive linguistic features can be derived from these feature sets. Table 1.1 provides the literature references corresponding to the linguistic feature sets proposed in this research. The detailed examination of the linguistic feature sets with their respective literature references will be further discussed in Section 2.3 and Section 3.3.

Table 1.1: Comprehensive Linguistic Feature Sets

Linguistic Feature Sets	References
Surface Features	(Perelman, 2014; Fleckenstein et al., 2020)
Linguistic Errors	(Allen et al., 2014; Crossley et al., 2019a)

Table 1.1 continued

Linguistic Feature Sets	References
Text Cohesion	(Crossley et al., 2016; Crossley, 2020)
Semantics	(Crossley et al., 2019b)
Lexical Diversity	(Zenker & Kyle, 2021; Crossley, 2020)
Lexical Sophistication	(Kyle et al., 2018; Crossley, 2020)
Syntactic Complexity	(Lu, 2010; Kyle, 2016)
Readability	(Crossley et al., 2011a; Crossley et al., 2019c)

Through the investigation of the essay's linguistic features, this research intends to formulate an automated essay scoring system, namely the Automated Essay Evaluator (AEE), for addressing the interrelated needs and purpose as stated below:

i. Formulation of a comprehensive linguistic-based AES System

This thesis argues that linguistic features are one of the essential aspects in assessing essay quality. Linguistic features exhibit their utility as domain-independent features for evaluating writing proficiency in terms of linguistic aspects, irrespective of the essay topics and domains. Numerous writing aspects, such as language errors, vocabulary usage, syntactic structure, etc., can be assessed through linguistic features. Therefore, the formulation of a scoring system based on linguistic criteria assessment is fundamental and yet significant in the field of AES.

ii. Formulation of a transparent AES System

This thesis argues that the formulation of a transparent AES system is vital for ensuring its reliability and validity as an educational measurement instrument. End users should be able to access and understand the respective AES scoring features.

This research fulfils this essential requirement by providing a comprehensive analysis of the essay quality from the perspective of its linguistic features. Based on the linguistic essay features, the underlying essay scoring features can be analysed, interpreted, and understood.

iii. Realization of a Malaysia-based AES System

There has been minimal research and development to create AES customized to the local context in Malaysia. All the well-known AES emerged and grew in the Western world, especially in the United States. If the Western AES software is adopted directly into the Malaysian context, it may cause assessment validity issues such as Construct Relevance and Representation (Williamson et al., 2012). Therefore, this work fills the gap of the lack of locally developed AES in Malaysia by proposing a Malaysia-based AES target in Malaysian English test environment. It utilizes the Malaysian University English Test (MUET) essay as a case study for automated essay scoring. The scoring of MUET essays hold considerable importance and substantial benefits given the high-stakes and large-scale nature of the MUET examination.

1.4 Problem Statement

An AES system takes a written essay as input and then assigns a numeric score to the essay reflecting its quality based on the content, grammar, and organization (Taghipour & Ng, 2016). The scoring process of the AES system is realized through machine learning algorithms with predefined essay features. Regardless of their wide usage and popularity, AES systems were often criticized and challenged in the aspects as elaborated below:

1.4.1 Surface-level features from AES do not capture the linguistic aspects of an essay

Some of the commonly selected features of AES to score essays are the surface features, which include non-linguistic characteristics, such as the total number of words per essay (essay length), number of sentences, and number of paragraphs, as well as basic linguistic characteristics, such as the total number of grammatical errors and the types of grammatical errors that appear in an essay (Kaplan et al., 1998). These surface features are superficial features that can be quantified easily by counting the basic elements in an essay. The use of surface features to assess essay quality can be found in works by Phandi et al. (2015), Amorim et al. (2018), Crossley et al. (2019a), and Uto et al. (2020). Although these surface features could be leveraged in the above research to attempt various complex algorithms to enhance scoring efficiency, they do not align with any specific aspects of linguistic theory.

Despite their potential in scoring essays, these AES systems relying on surface features are often criticized because their scoring mechanisms are not connected to the rationale of how human raters typically process and score essays. These AES systems have failed to identify and assess other essay features that define writing quality (Perelman, 2014). The assessment of writing quality using surface features, such as essay length, has been a subject of controversy, as highlighted by Fleckenstein et al. (2020). Wilson and Rodrigues (2020) highlighted the construct validity issue in AES due to its significant correlations between AES score and essay length and poor discrimination across various dimensions of writing quality. Crossley (2020) did not consider surface features as a construct-relevant aspect to assess writing quality but proclaimed deep linguistic constructs such as lexical, syntactic, and cohesion shall be adopted to assess writing proficiency. Furthermore, Li et al.

(2023) argued the traditional AES techniques concentrate on surface or shallow features of essays and overlook the potential deep linguistic information in the essay.

1.4.2 Lack of comprehensive coverage of linguistic feature dimensions in AES research

To address the limitation of “surface-level” assessment, several recent works have emerged to utilize deep linguistic features to assess essays. These deep linguistic characteristics, such as cohesion, lexical diversity, and lexical sophistication, are more theoretically connected to writing quality. For instance, Peng et al. (2023) predicted writing quality in argumentative writing by examining lexical diversity and lexical sophistication indices provided by Coh-Metrix (McNamara et al., 2014). Similarly, Shin and Epp (2023) explored the effect of text cohesion features on the writing clarity of academic papers. Other related research examining deep linguistic features on essay quality can be found in works of semantics (Crossley et al., 2019b) and lexical sophistication (Randy & Angel, 2023).

While these research focus on specific linguistic dimensions, they do not fully capture the comprehensive range of linguistic dimensions that contribute to essay quality. For example, some papers may solely focus on examining lexical diversity for scoring essays but ignore other significant linguistic dimensions, such as syntactic complexity. On the other hand, there may be inadequate representations of linguistic feature indices even within the linguistic dimensions that the research focuses on. For example, Coh-Metrix uses only four feature indices to represent lexical diversity and seven indices to measure syntactic complexity (Shin, 2018). To the best of the author's knowledge, none of the research attempts to comprehensively encompass the entire linguistic dimensions spectrum or delve deep into its vast number of fine-grained linguistic indices to score essays.

To bridge this gap, further investigation into a comprehensive spectrum of linguistic features is necessary. This comprehensiveness of linguistic dimensions might encompass all eight feature sets specified in Table 1.1, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability.

1.4.3 The black box nature of AES (particularly the commercial proprietary and deep neural network AES)

Recent advances in computational linguistics and NLP have catalysed the rapid innovation and growth of AES in the commercial proprietary domain and academic research. The top-rated commercial proprietary AES systems are Project Essay Grader (PEG; Measurement Incorporated, n.d.), Intelligent Essay Assessor (IEA; Pearson Education, 2010), e-Rater (ETS, 2023), and Intellimetric (Vantage Labs, 2023). As reported by Shermis (2014), these advanced AES systems can achieve high performance as represented by their Quadratic Weighted Kappa (QWK; Cohen, 1968) score: PEG (0.78), IEA (0.73), e-rater (0.76), Intellimetric (0.75). However, these AES vendors do not reveal the underlying technical implementation of the AES. Although certain vendors, such as e-rater and Intellimetric, provide an overview of their respective AES systems (Dikli, 2006), the documentation lacks a detailed description of the scoring mechanism and the specific fine-grained essay features used in the scoring process.

There are a significant number of recent research works that incorporate automatic feature extraction through deep neural network models. These Deep Neural Network-based AES (DNN-AES) can be found in works by Alikaniotis et al. (2016), Taghipour and Ng (2016), Dong et al. (2017), and Tay et al. (2018). While neural AES have demonstrated high effectiveness in scoring essays, their deep and complex architecture poses challenges in

interpreting and explaining the features used to generate essay scores. As highlighted by Ge and Chen (2019), the deep learning approach may improve AES performance, but it is difficult to provide profound justification for the essay scores produced based on writing theories.

In general, the above commercial and neural approach to AES systems demonstrates a non-transparent nature, resembling a black box system. Users can only input essays and observe the corresponding scores without having access to its underlying computational mechanism (Wong & Bong, 2019). The end users are unclear of the specific essay features that an AES system assesses and the corresponding weights it applies to those features during the essay scoring process. (Wilson & Andrada, 2016). In the occurrence of any discrepancies in the essay scores, the exact scoring algorithm cannot be examined. This black box operation of AES systems restrains the clear interpretation and explanation of the final scores, thus impacting the validity of the assessment.

1.4.4 The relationship between the problem statements and MUET essay scoring

To conclude this section, the author attempts to establish a connection between the three aforementioned problem statements and the nature of the MUET essay scoring. In the context of the MUET test, essays are assessed by adhering to the scoring rubrics outlined in the test specifications. These MUET rubrics cover various assessment criteria, namely, accuracy, appropriacy, coherence and cohesion, use of language functions, and task fulfilment (Malaysian Examination Council, 2014). Each rubric is further elaborated in its specification, providing clear guidelines for scoring MUET essays. Further discussion on the specific scoring rubrics used for MUET essays can be found in Section 3.1.1 and Table 3.1.

The relationships between the respective problem statements and the MUET essay scoring are described below:

- i. Considering the diverse scoring criteria outlined in the MUET scoring rubric, relying solely on surface-level features to score MUET essays may fall short of capturing the full spectrum of assessment criteria. As a result, employing surface features alone to score MUET essays could be considered inappropriate. Enhancing AES models to incorporate deeper constructs of linguistic features, such as text cohesion and lexical complexity becomes crucial to ensure the accurate assessment of MUET essays.
- ii. A comprehensive and fine-grained feature coverage should be incorporated in MUET essay scoring. This multifaceted and extensive approach facilitates a thorough examination of all elements outlined in the MUET scoring rubrics, reducing the likelihood of overlooking any scoring rubrics. Embracing such an approach enables a more holistic assessment aligned with the MUET scoring rubrics, leading to more reliable scores.
- iii. The formulation of a transparent AES system is required to avoid black box pitfalls when interpreting MUET essay scores. Through the development of an AES specifically customized to align with the MUET scoring rubric, the features employed to score MUET essays become transparent and interpretable, thus dispelling any doubts regarding the validity and reliability of the score. Educators, students, and stakeholders can then understand how the AES arrives at its scores, fostering trust and informed scoring of MUET essays.

1.5 Research Objectives and Research Questions

In response to the stated problem statements, this section outlines the specific research objectives and the corresponding research questions of this thesis. The research objective and research question are denoted by RO and RQ respectively in this section. Furthermore, each objective is accompanied by a brief explanation of how to measure the specific objective for assessing its achievement.

RO (1): To identify and evaluate comprehensive linguistic features for automated essay scoring.

RQ (1): What are the potential comprehensive linguistic features that can be employed to score essays?

This research will evaluate the linguistic features based on their correlation strength with the essay scores/grades. Furthermore, the comprehensiveness of the linguistic features specified in this objective will be demonstrated by the vast number of linguistic features derived from the eight feature sets (surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability) specified in Table 1.1.

RO (2): To select optimized linguistic features that influence automated essay scoring using various feature selection schemes.

RQ (2): Among the comprehensive linguistic features identified in RO (1), what are the optimized linguistic features that can be selected to score essays?

This research will employ various feature selection schemes to select the optimized linguistic features for scoring essays. The number and details of the selected linguistic features by each scheme will be discussed and outlined in tables.

RO (3): To assess the effectiveness of the selected linguistic features in predicting essay scores.

RQ (3): What is the performance of the selected linguistic features in predicting essay scores when applied to multiple AEE models?

This research involves experiments to assess how well the selected linguistic features perform in scoring essays when running on machine learning classifiers. The evaluation of these features' effectiveness will be measured by classification metrics such as QWK, accuracy, precision, recall, and F1 score.

RO (4): To establish an explanation of the linguistic feature set contributing to the MUET essay scoring rubrics.

RQ (4): How to map the MUET essay rubrics with linguistic feature sets for explaining the AEE's scoring in this research?

This research intends to address the AES black box nature, as stated in the problem statements, by providing an explanation of how the proposed AEE linguistic features play a role in the scoring of MUET essays. To achieve this, this thesis will theoretically map the MUET essay rubrics specification with the linguistic feature sets proposed, shedding light on their contribution to the scoring process.

1.6 Significance of the Research

Based on the research objectives stated above, this work implies its significance primarily to the analysis and scoring of essays, as well as the development of an AES customized for the Malaysian context. The significance of this research work is stated below:

i. Analysis of Essays

The formulated AES analyses comprehensive and multiple dimensions of linguistic features, including surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability in influencing essay scores. The compilation and application of such a large number of linguistic features for analysing essays can be considered the first of its kind in research. It provides most, if not all, feasible linguistic metrics to assess writing quality. Furthermore, this research classifies the linguistic features into their proper groups to provide a systematic structure for future investigation. Besides essay scoring, the result from the analysis can be used for other linguistic applications, such as second language acquisition and classroom pedagogy.

ii. Scoring of Essays

From the analysis results, this research identifies and selects optimal linguistic features influencing essay scores by using different feature selection schemes. Based on the selected features, this research attempts to construct the Automated Essay Evaluator (AEE) to score MUET essays. The AEE shows its uniqueness in scoring essays based on deep linguistic features instead of shallow surface features which are not construct-relevant to writing assessment. It leverages a diverse range of comprehensive linguistic features to score essays. Furthermore, the AEE demonstrates its explainability and interpretability in scoring essays by providing transparent and communicable linguistic scores associated with the MUET's official scoring rubrics. This AEE's characteristic addresses the black box nature of AES.

iii. Creation of a Malaysia-based AES System

There is limited research and development work on AES in Malaysia. To the best of the author's knowledge, there was only one AES-related research focusing on heuristic-based grammar checking of tenses in essays (Omar et al., 2009; Maasum et al., 2012). The work only addressed grammatical errors in essays without considering the other linguistic and content assessments. On the other hand, all the prominent AES software products have their origins in the Western English-speaking world. Their scoring mechanism might be closely associated with the scoring rubrics of a particular English test context. The direct adoption of such AES to evaluate essays in the Malaysian English test environment may raise reliability and validity issues (Wong & Bong, 2019). Therefore, it is essential to carry out this AES research to fill the existing gap in locally developed AES. The linguistic-based AEE developed in this research for MUET scoring can serve as the foundation for the realization of a full-scale Malaysia-based AES by incorporating other essay features.

1.7 Scope of the Research

The aim of this research, however, is not the development of a full-scale AES system but instead an investigation into the linguistic features that might be good predictors for essay scores. The development of a full-scale AES system that assesses every aspect of an essay is the work of established educational testing and assessment organizations, involving multiple teams of computer scientists, linguistic specialists, and educators, which are far beyond the ability and scope of this research. This research is thus limited to linguistic feature analysis and will not investigate other means of essay scoring, such as argumentation, organization, and development. Recognizing that writing quality is affected by other factors

besides linguistic attributes, the prediction score produced by the AEE may not be highly accurate but only reflects the linguistic aspect of the writing.

As the main objective of this thesis is to analyse and select linguistic features in scoring essays, the research will not focus on any specific machine learning model to achieve the highest feasible accuracy. The fine-tuning of the classifier hyperparameters and application of the wrapper method in feature selection (e.g., forward selection, backward elimination) to achieve the highest feasible accuracy may produce the subset of best features prone to overfitting and associated with the specific learning function. The selected best features may not be generalized to other classifiers. To focus on the analysis and explanation of the linguistic features in essay scoring, this research will not investigate the application of deep learning in scoring essays, despite many recent AES systems being neural-based. A big issue of deep learning is that its overall operation is a black-box structure, and it is usually difficult to find the principle and rationale of the given score (Ge & Chen, 2019). This may hinder the objective of this research to provide explainable influential linguistic features for essay scoring.

The MUET essay dataset used in this research consists of 724 essays written by sixth-form students and scored by qualified English teachers. Due to time, geographical, and financial constraints, the dataset collection is limited to schools within the Kuching district in Sarawak, Malaysia. Therefore, the analysis results derived from this research might only be generalized to the context of MUET students in Kuching, even though it is hoped to apply to a wider MUET population in Malaysia. On the other hand, there may be some minor discrepancy in the teacher's scoring despite briefing and training being conducted for

assessing the essays based on the MUET's essay scoring rubrics (Malaysian Examination Council, 2014). This discrepancy may affect the prediction accuracy of the AES classifier.

1.8 Thesis Organization

This thesis is organized into six chapters as follows.

i. Chapter One (Introduction)

This chapter presents the introduction of this research - discussing the research background and automated essay scoring process, the research's motivation, problem statements, research objectives, significance, and scope of the research.

ii. Chapter Two (Literature Review)

This chapter reviews related literature on automated essay scoring - state-of-the-art AES Systems, features for scoring essays, linguistic features, content features, and feature selection. This chapter concludes with its final section discussing the trends and challenges in automated essay scoring research.

iii. Chapter Three (Materials and Methods)

This chapter discusses the materials (MUET scoring rubrics and essay dataset), and methods (detailed linguistic features specification and feature selection schemes) used in this research.

iv. Chapter Four (Experiments)

This chapter presents the experimental framework for analysing and scoring essays - feature extraction, feature evaluation, feature selection, AEE models, and the boxplot diagram.

v. Chapter Five (Results and Discussion)

This chapter discusses detailed findings and results of this research - the correlation coefficient of linguistic features and essay scores, optimal linguistic features selected by different feature selection schemes, the effectiveness of the linguistic features, and the performance of the AEE models.

vi. Chapter Six (Conclusions and Recommendations)

This final chapter concludes the research by discussing the summary, the findings based on research objectives, the contributions and implications, and potential future works stemming from the limitations of this research.

To enhance readability, each chapter commenced with an introductory paragraph outlining the chapter's structure and concluded with a summary section highlighting key points in the chapter.

1.9 Chapter Summary

Automated Essay Scoring (AES) refers to the Artificial Intelligence (AI) application with the “intelligence” in assessing and scoring essays. These AES systems are often criticized because their scoring mechanisms are associated with shallow surface feature assessments. It does not capture the deep linguistic aspects of an essay. In Malaysia, the research and development of AES are scarce. This research intends to bridge the gap by conducting a systematic and detailed investigation of deep linguistic features in assessing essay quality. The thesis proposes comprehensive linguistic features, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability in influencing essay scores. Based on these deep linguistic features, this work attempts to construct an AES classifier, namely the Automated Essay Evaluator (AEE) to score MUET essays.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews some of the important literature related to automated essay scoring. Section 2.1 provides a complete literature review of the state-of-the-art AES systems, with a comparison and summary at the end of the section. Section 2.2 discusses the most important aspect of automated essay scoring, i.e., the use of multiple and diversified essay features to score essays. Based on the comprehensive review of related literature, Section 2.3 elaborates on the influence of multi-dimensional linguistic features on essay quality, while Section 2.4 explores the evaluation of essays based on content features. Section 2.5 reviews different feature selection techniques, as this research examines the selection of optimal linguistic features to score essays. Section 2.6 presents the author's comments on this AES review by highlighting trends, challenges, and opportunities in the AES research. Finally, Section 2.7 concludes this chapter by providing visual representations of AES systems, essay features, and feature selection techniques.

2.1 Automated Essay Scoring System

With the advancement of the Internet, machine learning, natural language processing tools, and e-learning systems, AES became a support technology in education. The Internet provides the common platform to submit digitized text for assessment, while the machine learning model formulates complex algorithms for scoring essays. Nowadays, the AES systems are used in combination with human graders in different large-scale high-stake assessments. In the following section, this thesis presents a comprehensive review of the state-of-the-art AES system, ranging from commercial proprietary software to research tools provided by the academic community.

2.1.1 Project Essay Grader (PEG)

PEG is a commercial proprietary AES system and was considered the first AES introduced in 1966 (Page, 1966). In 1998, a web interface was incorporated in response to the emergence of the World Wide Web (Shermis et al., 2001). PEG system scores the writing style of an essay by measuring "trins" and "proxes". "Trins" is defined as the essay's underlying intrinsic characteristics (e.g., fluency, grammar, and punctuation), whereas "proxes" denotes the approximation (correlation) of the "trins" (e.g., word count and error count) (Dikli, 2006). For example, the "trins" fluency is measured through the "proxes" amount of vocabulary. PEG employs regression analysis to score new essays based on a training set of 100 - 400 essays (Page, 1994). To date, PEG has developed more than 500 "trins" to score essays (Measurement Incorporated, n.d.).

2.1.2 Intelligent Essay Assessor (IEA)

In 1998, Pearson Knowledge Technologies released Intelligent Essay Assessor (IEA). IEA focuses on the assessment of the content-related features of essays (Dikli, 2006). It scores essays by using its Knowledge Analysis Technologies (KAT) engine, which is the proprietary implementation of Latent Semantic Analysis (LSA; Landauer et al., 1998), an approach to measure the semantic similarity of words and passages by analysing large bodies of relevant texts (Pearson Education, 2010). IEA relies on LSA to derive content, organization, and development-based attributes of essays. Besides LSA, IEA uses NLP techniques to extract essay attributes, measuring grammatical, mechanical, lexical sophistication, stylistic, and organization. The system employs around 60 attributes to score an essay based on its content (e.g., LSA essay semantic similarity), grammar (grammar errors), mechanics (e.g., spelling and punctuation), lexical sophistication (e.g., word maturity and word variety), style and organization (e.g., coherence) (Zupanc & Bosnić,

2015). Compared to other AES, IEA requires only a relatively small number of 100 – 300 pre-trained essay samples to score an essay (Dikli, 2006).

2.1.3 E-rater

E-rater is a commercial proprietary AES system developed by the Educational Testing Service (ETS) in 1998 (ETS, 2023). It relies on NLP techniques to extract essay features (syntactic, discourse, and topical) for evaluating the style and content of an essay (Dikli, 2006). Based on the literature by Burstein et al. (2004), Attali and Burstein (2005), and Zupanc and Bosnić (2015), E-rater 2.0 extracts multiple essay features that can be categorized into ten attribute classes to score an essay. The relevant attribute classes are listed below.

- i. Grammatical errors (e.g., subject-verb agreement errors)
- ii. Word usage errors (e.g., "their" vs. "there")
- iii. Mechanical errors (e.g., spelling errors)
- iv. Presence of essay-based discourse elements (e.g., thesis statement, topic sentences)
- v. Development of essay-based discourse elements
- vi. Writing style weaknesses (e.g., overly repetitious words)
- vii. Topical word usage based on Content Vector Analysis (CVA)
- viii. Content measure based on the word frequency in high score vs. low score essays
- ix. Lexical sophistication
- x. Correct usage of prepositions and collocations, and variety of sentence structure

The above essay features are then combined to produce a final essay score, with the weight of each feature determined by using multiple regression. A collection of

approximately 400 training essay samples is required for training the regression model (Hussein et al., 2019).

2.1.4 IntelliMetric

Intellimetric is another commercial proprietary AES system released by Vantage Learning to score essays operationally since 1997 (Vantage Labs, 2023). IntelliMetric claims itself as the very first AES system that relies on a combination of artificial intelligence, natural language processing (NLP), and statistical models to simulate the manual essay scoring process by human raters (Hussein et al., 2019). IntelliMetric leverages its proprietary CogniSearch and Quantum Reasoning technologies to evaluate over 300 semantic-, syntactic-, and discourse-related features, which can be categorized into five broad categories namely focus and unity (coherence), organization, development and elaboration, sentence structure, mechanics, and conventions (Dikli, 2006). As an advanced AI-based AES system, IntelliMetric uses multiple mathematical models, including linear analysis, Bayesian approach, and LSA to predict the final scores and combines the models into a single final essay score (Rudner et al., 2006). These multiple-scoring models emulate the equivalent of a panel of multiple judges for producing a more comprehensive and accurate final score. As such, IntelliMetric claims its advantage over other AES in having these multiple automated scoring systems at work. IntelliMetric requires a sample of at least 300 pre-scored essays to train its scoring model (Zupanc & Bosnić, 2015).

2.1.5 Bayesian Essay Test Scoring sYstem (BETSY)

In 2002, Lawrence Rudner developed BETSY (BETSY, n.d.) - an AES program based on the Bayesian theorem (Dikli, 2006). BETSY employs two Bayesian models, i.e., the multivariate Bernoulli model and multinomial model to classify essays into different

nominal or categorical scales (e.g., appropriate, partial, or inappropriate) based on content (e.g., unigrams and bigrams) and style attributes (e.g., sentence length) (Rudner & Liang, 2002). This essay classification is operated based on a large set of essay feature extraction and the assumption that each essay feature is independent of another. To score new essays, BETSY needs to be trained on a large number of human-graded essays (around 1000 samples) (Dikli, 2006).

Unlike the commercial proprietary AES products described above, BETSY was the first AES to be made available publicly (Rudner & Liang, 2002). Although BETSY has been designed for scoring essays, it can be used for any other text classification task (Hussein et al., 2019). However, BETSY can only be treated as a research tool rather than a full-scale product. The authors never continued with their work after their preliminary investigation (Zupanc & Bosnić, 2015).

2.1.6 Coh-Metrix

Coh-Metrix is a computational tool that computes language and discourse measures based on empirical and theoretical foundations of psycholinguistics methods, discourse processing, and computing algorithms (Graesser et al., 2004, 2011; McNamara et al., 2014). As the Coh-Metrix's name ("Coh" & "Metrix") implies, it focuses on measuring text cohesion and coherence based on linguistic metrics. Coh-Metrix defines text cohesion as the explicit features in text that guide the reader in interpreting the substantive ideas in the text, connecting ideas with other ideas, and connecting ideas to higher-level global units (e.g., topics and themes) (Graesser et al., 2004). These cohesive devices interact with the reader's background knowledge to form a coherent mental representation of the text in the reader's mind (hence the term of "coherence").

Coh-Metrix 3.0 (The University of Memphis, n.d.) provides 108 numeric indices of linguistic and discourse representations of a text. It provides a web interface for the user to analyse texts of less than 15,000 words. The Coh-Metrix indices values can be used in various text analysis tasks, such as assessing essay quality and distinguishing between writing in a First Language (L1) learner and a Second Language (L2) learner. These indices are categorized into 11 feature groups to provide a comprehensive cohesion measurement of a text (Shin, 2018). Table 2.1 shows the summary of Coh-Metrix's 11 feature groups with the corresponding number of indices in each group.

Table 2.1: Coh-Metrix Features

Feature Name	No. of Indices	Feature Name	No. of Indices
Descriptive	11	Situation Model	8
Text Easability Principal Component Scores	16	Syntactic Complexity	7
Referential Cohesion	12	Syntactic Pattern Density	8
LSA	8	Word Information	22
Lexical Diversity	4	Readability	3
Connectives	9	Total	108

Although Coh-Metrix was not designed particularly for essay scoring, it was included in this literature review because there exists a number of automated essay scoring research by leveraging Coh-Metrix as the instrument. Xu and Liu (2016) introduced an AES framework using the Coh-Metrix indices and eight additional deep language features to score Chinese students' writing. Jagaiah, (2017) relied on Coh-Metrics indices to investigate the syntactic complexity and its relationship to writing quality in essays. Shin (2018) formulated

a neural network approach to essay scoring by using Coh-Metrix indices. Furthermore, Latifi and Gierl (2021) evaluated the Coh-Metrix performance by extracting the deep linguistic feature from the 12,978 Automated Student Assessment Prize (ASAP) Essay Dataset (Kaggle, 2012) and predicting the essay score by Random Forest.

2.1.7 LightSide

In 2013, Mayfield and Rosé released LightSide, a user-friendly, GUI-based automated text evaluation system (Mayfield & Rosé, 2013). The system is designed as a tool for non-experts to leverage text mining technology for multiple purposes, including essay assessment. LightSide reached a significant milestone in the AES field as the first open-source AES. The LightSide program, namely, the LightSide Researcher's Workbench (LightSide, 2019), which consists of a machine-learning engine, a feature-extraction core, and a graphical user interface, is publicly available for download. LightSide streamlines the entire process of machine learning through six modules: extract features, restructure data, build models, explore results, compare models, and predict labels (Cunningham-Nelson et al., 2018). It allows users to select multiple text features and algorithms to construct a prediction model (e.g., Naïve Bayes, Logistic Regression, and Support Vector Machine) (Mayfield et al., 2014). The set of text features includes n-grams, Parts of Speech (POS) tags, and "count-occurrence" attributes. However, the system allows users to input the code for new attributes manually (Zupanc & Bosnić, 2015). To score an essay, LightSide requires a sample of 300 pre-scored essay samples to train its prediction model (Shermis, 2014).

2.1.8 Deep Neural Network-based AES (DNN-AES)

With the emergence of deep learning in the AI field, many latest AES systems are Deep Neural Network-based AES (DNN-AES). Compared to traditional work in AES that

focuses on feature engineering, an advantage of deep learning approaches is that they avoid the laborious task of constructing evaluation features manually (Ge & Chen, 2019). Furthermore, DNN-AES demonstrates its notable ability to achieve state-of-the-art performance (Hussein et al., 2019) when trained on large essay datasets.

2.1.8.1 Neural Approach AES

Taghipour and Ng (2016) proposed the first neural approach AES, which uses a Recurrent Neural Network (RNN; Medsker & Jain, 1999; Xiao & Zhou, 2020) to score essays without feature engineering. The system uses a Convolution Neural Network (CNN; LeCun et al., 1998; Li et al., 2021) as the first layer to extract n-gram level features based on the input of the one-hot-vectors of words in essays. These n-gram features (representing the local contextual dependencies of the words) are then passed to a recurrent Long Short-Term Memory Network (LSTM; Hochreiter & Schmidhuber, 1997; Staudemeyer & Morris, 2019) to output a vector that represents the long-distance relations of the words in the essay. These generated vectors at each time step are then combined by a pooling layer to form a vector, which is used as the input to a dense layer to produce the essay score. Figure 2.1 illustrates the Convolutional Recurrent Neural Network Architecture used in this AES. This neural-based AES system used the ASAP essay dataset provided by Kaggle in their experiment. A total of approximately 7,787 essays (60% of the 12,978 ASAP training dataset) were used to train the neural network.

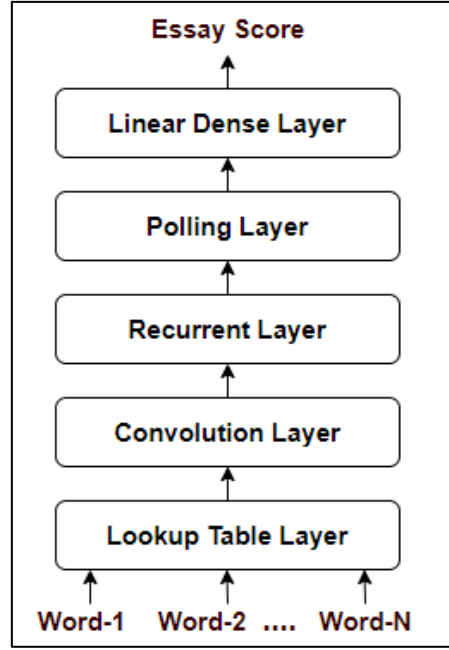


Figure 2.1: Architecture of Convolutional Recurrent Neural Network

2.1.8.2 Score-Specific Word Embeddings (SSWE) AES

This SSWE model is an extension of Taghipour and Ng (2016) neural-based AES. Instead of using one-hot word vectors to train the neural network, Alikaniotis et al. (2016) rely on word embedding to train the essay scoring model. A word embedding is a dense real-valued vector representation of a word, typically with several hundred dimensions. These vectors are trained from a large corpus so that words with similar meanings or contexts are closer together in the embedding space. Word embedding is considered a better model than sequential one-hot word vectors because it embeds the semantic aspect of the word. Based on the C&W Embedding Model (Collobert & Weston, 2008), Alikaniotis et al. (2016) propose score-specific word embeddings (SSWEs) by augmenting the C&W model with an additional output that corresponds to the score of the essay (Ke & Ng, 2019). These SSWE values are then used as features to train the AES neural network. The SSWE is claimed as a better representation of word embedding because it can better discriminate between

informative and under-informative words. Therefore, the model is expected to result in higher performance as it can filter out under-informative words (e.g., is, are, at) which have no or low impact on the essay scores. To construct this SSWE-based AES, around 8,306 ASAP essays (64% of the 12,978 ASAP training dataset) were used to train the neural network.

2.1.8.3 Attention-Based Hierarchical Model AES

Dong and Zhang (2016) enhanced Taghipour and Ng (2016) model by incorporating the hierarchical structure of an essay, i.e., its words and sentences. They hypothesized that an essay is to be written by combining first its words to construct the sentences and then combining sentences to create the essays. They model the hierarchical structure by using two convolution layers - a word-level convolution layer and a sentence-level convolution layer. The word-level convolution layer has each sentence in one-hot word vectors as input and extracts n-gram features from the sentence independently of other sentences (Lagakakis & Demetriadis, 2021). Then, after a pooling layer, the extracted n-gram features are then combined into a sentence vector. The sentence-level convolution layer then takes the sentence vectors from different sentences to generate the n-gram-level features over different sentences. These sentence-level features are then combined by another pooling layer to construct a vector for the input to a dense layer for generating the final essay score. To automatically identify important characters, words, and sentences in an essay, Dong et al. (2017) further enhanced the model to incorporate an attention mechanism. Instead of using simple pooling, the enhanced model leverages the attention pooling layer after the CNN and LSTM layer to focus on important features while suppressing irrelevant redundant information. This AES followed the same setting of the ASAP dataset used in Taghipour and Ng (2016) neural-based AES, i.e., around 7,787 essays.

2.1.8.4 SKIPFLOW AES

Tay et al. (2018) proposed SKIPFLOW AES - a novel AES that incorporates the capability of learning coherence features within the context of an end-to-end neural architecture. They hypothesize that modelling relationships between multiple positions of an essay enables the extraction of text coherence features. They called this feature the neural coherence feature. The model is an LSTM-based model with an addition of the SKIPFLOW mechanism to extract the neural coherence feature. The SKIPFLOW mechanism is realized by adding a neuron tensor layer after the LSTM layer for taking two positional outputs at different time steps from the LSTM to compute the neural coherence features between these two pairs of outputs. These neural coherence features are then concatenated with the output vector of the LSTM vector (which represents the local and long-distance relationship of words in an essay) to generate an augmented vector. This final vector is then fed into a dense neural layer to predict the essay score. Like Taghipour and Ng (2016) model, they used 7,787 essays (60% of the 12,978 ASAP training dataset) to train the SKIPFLOW model.

2.1.9 Semantic Essay Grader for Essays (SAGE)

The lack of consideration of text semantics is one of the main shortcomings of the existing AES system. To address this issue, Zupanc (2018) proposed a semantic-based AES, called Semantic Automated Grader for Essays (SAGE), to score essays. SAGE approaches essay semantics from aspects of essay coherence and semantic error detection. The essay coherence is represented by two groups of novel attributes, namely, the coherence attribute obtained in a highly dimensional semantic space and the coherence attribute based on sentence-similarity networks. To derive the first coherence attribute value, SAGE transforms sequential parts of an essay into the semantic space and measures changes between them to estimate the text coherence (Zupanc & Bosnić, 2017). For the sentence-similarity attribute,

SAGE constructs sentence-similarity networks based on the undirected, weighted graph to describe similarities among sentences in an essay. Any two nodes, representing distinct sentences A and B, are connected if their similarity is higher than a given threshold (Zupanc et al., 2017). Besides essay coherence, SAGE comprises an automated error detection function that detects semantic errors from the perspective of essay consistency. These semantic errors are detected via the implementation of entity recognition, coreference resolution, open information extraction, ontologies, and logic reasoners (Zupanc, 2018). In the research, SAGE relied on the ASAP training dataset (datasets 1, 3, and 8), with a total of 4,232 essay samples to train its scoring model.

2.1.10 Comparison of the State-of-the-Art AES Systems

This sub-section provides a comparison of the state-of-the systems. Table 2.2 shows the characteristics of these AES systems based on their technical specification and methodology employed in scoring essays. These AES characteristics are described by using the following criteria.

i. AES type

The AES type can be divided into four groups: commercial proprietary, publicly available, open-source, and research tool. Most of the full-scale functional AES are proprietary commercial software, except BETSY (publicly available) and LightSide (open source). The other AES systems are simply by-products from research that tackles a particular AES problem and may have issues to be operationalized as full-scale end products.

ii. Release year

The AES systems in Table 2.2 are listed in chronological order of their release year to show the development progress in the field. Except for PEG, all the well-established AES systems were released after 1998.

iii. Scoring feature

The scoring features of the AES systems can roughly be divided into two groups: style and content. Style features focus on the linguistic aspect of the essay, such as mechanics, grammar, lexical richness, and syntactic sophistication. Content features describe the semantics of an essay by comparing the similarity between the essay and the essay prompt or other scored essays.

iv. Feature extraction technique

Different AES systems use different approaches to extract features from essays. The widely used approach relies on NLP techniques to extract linguistic features of essays. AES systems focusing on content mostly use LSA to analyse concepts between the essays and the contained terms. Earlier AES systems, such as PEG, merely rely on statistical methods for feature extraction, while recent research work in AES usually leverages deep learning to facilitate automatic feature extraction. Semantic-based AES (e.g., SAGE) uses NLP and information extraction frameworks such as Open Information Extraction to generate structured and machine-readable representations of knowledge from the essay.

v. Scoring model

The AES scoring task is often addressed as a supervised machine learning problem, where a scoring model is used to predict the final score. Regression is the widely used machine learning algorithm to predict the exact essay score. Classification

methods are another approach for building classifiers to classify discrete essay grades. Recent AES research involves the use of multilayer neural network structures such as LSTM as the classifier for essay scoring.

vi. Number of essays required for training

The last criterion is the number of pre-scored essays to train the scoring model. The training data must be sufficiently large for the prediction model to be accurate. Most commercial proprietary AES systems specify a few hundred pre-scored essays to train the model. However, AES based on deep learning requires a large dataset of thousands of essay samples to deliver the desired result.

Table 2.2: A Comparison of the State-of-the-Art AES Systems

AES System	AES Type	Release Year	Scoring Feature	Feature Extraction Technique	Scoring Model	# Essays for Training
PEG	Commercial Proprietary	1966	Style	Statistical	Multiple Linear Regression	100 - 400
IEA	Commercial Proprietary	1998	Content	LSA (KAT Engine)	Machine Learning	100 - 300
E-rater	Commercial Proprietary	1998	Style & Content	NLP (extract language errors, discourse elements, topical words, lexical sophistication, etc.)	Linear Regression	~400
Intellimetric	Commercial Proprietary	1998	Style & Content	NLP (extract 300 semantic, syntactic, and	Multiple Mathematical Model	300

Table 2.2 continued

AES System	AES Type	Release Year	Scoring Feature	Feature Extraction Technique	Scoring Model	# Essays for Training
				discourse features)		
BETSY	Publicly Available	2002	Style & Content	Statistical	Bayesian Text Classification	1,000
Coh-Metrix	Research Tool	2004	Style	NLP (extract referential cohesion, lexical diversity, syntactic complexity, readability, etc.)	Random Forest	12,978
LightSide	Open Source	2013	Content	Statistical	Machine Learning	300
Neural Approach AES	Research Tool	2016	Style & Content	Deep Learning	LSTM	~7,787
SSWE AES	Research Tool	2016	Style & Content	Deep Learning	SSWE, Two-Layer Bi-LSTM	~8,306
Attention-Based Hierarchical Model AES	Research Tool	2017	Style & Content	Deep Learning	Attention Mechanism, Two-Layer CNN	~7,787
SAGE	Research Tool	2017	Style & Content	NLP (extract essay coherence and semantic errors), Open Information Extraction	Linear Regression, Neural Network, Random Forest, etc.	4,232

Table 2.2 continued

AES System	AES Type	Release Year	Scoring Feature	Feature Extraction Technique	Scoring Model	# Essays for Training
SKIPFLOW AES	Research Tool	2018	Style & Content	Deep Learning	SKIPFLO W, LSTM	~7,787

2.1.11 Summary of AES Systems

In summary, AES has undergone substantial development over a span of more than 60 years since the inception of the first AES system, PEG, in 1966. Among the AES systems available, four commercial proprietary systems, namely PEG, e-rater, IEA, and IntelliMetric, have emerged as notable solutions for evaluating written content. These systems claimed to have robust performance and demonstrated widespread adoption in practical application. Alongside these commercial software, several other AES systems have emerged since the 2000s for research purposes, each aiming to address specific challenges by exploring novel scoring approaches. However, their application is largely confined to testing within the ASAP essay dataset, and they have not yet been fully implemented into real-world applications.

Most AES systems operate by employing various feature extraction methods to extract a multitude of features from essays. A notable observation is that the four prominent commercial AES systems (PEG, IEA, e-Rater, IntelliMetric) depend on manual handcrafted features for feature extraction. This approach requires domain experts to manually identify and define the essay features before the feature extraction process can commence. As a result of recent progress in deep learning, AES research has seen a trend towards Deep Neural Network AES systems (DNN-AES). This trend is exemplified by various AES systems such

as Neural Approach AES, SSWE AES, Attention-Based Hierarchical Model AES, and SKIPFLOW AES presented in Table 2.2. These systems have gained significant attention due to their ability to automatically learn relevant features and representations from the raw essay text without explicit manual feature engineering.

To facilitate a summary of this comprehensive AES system review, the author seeks to conclude this section by presenting a chronological account of the progress and development of AES. Table 2.3 serves as a concise record highlighting notable achievements and significant milestones in the evolution of AES systems. Through this chronological narrative, one can gain deeper insights into the trajectory of AES development, understanding the key breakthroughs that have shaped the AES progression.

Table 2.3: Progress in AES Development

Period	Milestone of AES Development
1960s	An early attempt of automated essay scoring was proposed by Ellis Page.
1970s	Ellis Page implemented the first AES system, namely the Project Essay Grader (PEG). The PEG scores essays by using the statistical approach of measuring “trins” (intrinsic variable) and “proxes” (quantifiable variable).
1990s	Development of the predominance commercial proprietary AES systems, namely the IEA, e-rater, and Intellimetric. IEA uses LSA to score essays, while e-rater and Intellimetric rely on NLP and machine learning algorithms to produce scores.
2000s	Release of BETSY as the first publicly available AES system in 2002. BETSY employs the Bayesian model to score essays. Release of Coh-Metrix as the first linguistic tool to assess the deep linguistic features of writing in 2004.
2010s	The Hewlett Foundation sponsored the ASAP competition on Kaggle in 2012 to enhance the AES algorithm performance. The competition catalysed the development of more AES research works and created the ASAP dataset which is the most widely adopted dataset in AES research.

Table 2.3 continued

	Release of LightSide as the first open-source AES system in 2013. LightSide relies on the statistical text-mining approach to score essays.
2013 - current	Research of automated essay scoring of specific dimensions of essays instead of holistic scoring, such as thesis clarity, prompt relevance, argumentation, and semantics.
2016 - current	Release of multiple linguistic analysis tools (Kyle, n.d.) to explore the deep linguistic feature in writing. The tool covers a wider scope of linguistic indices compared with Coh-Metrix.
2016 -current	Taghipour and Ng published the first work on the deep learning approach toward automated essay scoring in 2016. With the rise of deep learning models, such as CNN, RNN, and LSTM, deep learning approaches gained prominence in AES. It enables automatic feature extraction and shows promising results in improving AES system performance.

2.2 Essay Features for Automated Essay Scoring

One of the main requirements to be considered in developing an AES is the essay-scoring features used in scoring the essays (Ramesh & Sanampudi, 2021). The underlying rationale is that an essay is considered a good essay (and vice versa) if certain attributes are present (or absent) in the essay. Such essay features, either explicit or implicit, indicate the writing quality, and thus the score of the essay. This research perceives the selection of these essay scoring features, as the utmost important factor in formulating the scoring mechanism for the Automated Essay Evaluator (AEE).

A detailed study has been carried out in reviewing AES-related literature to identify and analyse essay features employed in recent AES research, with the objective of adopting some of these features into the AEE. Based on the review paper written by Ke and Ng (2019), Table 2.4 attempts to outline most, if not all, of the essay features employed in AES research.

These essay features can be broadly categorised into three groups below. The author considers semantics as both linguistic and content features because semantics by itself is a subdiscipline in the linguistic field, but it is used as the underlying theory for evaluating the essay content too.

i. Surface feature

Length-based features

ii. Linguistic features

Mechanical, grammatical, cohesion, coherence, word usage (lexical), style (syntactic), semantics

iii. Content features

Semantics, prompt relevance/adherence, thesis clarity, organization & development, argumentation

Table 2.4 presents 12 distinct essay features for essay scoring. The first seven features belong to the surface and linguistic feature group, followed by the semantic feature which is classified as both a linguistic and content feature. The remaining four features are designated as content-based features. Each essay features along with the corresponding feature group, and the description/definition is outlined in the table. The comprehensive details of these features can be found in Section 2.3 for linguistic features and Section 2.4 for content features.

Table 2.4: An Overview of Essay Features for AES

Feature Group	Features	Description
Surface Feature	Length-Based	These simple but valuable features deal with essay length, sentence length, word length, etc. (Zesch et al, 2015). The rationale is that when writing essays in a test context, there is usually a specified time limit within which the writer must complete their tasks. As a result, the quantity of text generated within the specific timeframe can be used as an indicator of the writer's productivity and potentially the essay quality.
Linguistic Feature	Mechanical	These features assess the mechanical correctness of an essay, such as the correct usage of spelling, punctuation, and capitalization (Ke & Ng, 2019).
Linguistic Feature	Grammatical	These features assess the grammatical correctness of an essay, i.e., the correct usage of sentence structure that complies with the grammatical or syntactic rule (Crossley et al., 2019a).
Linguistic Feature	Cohesion	Cohesion assesses the appropriate use of transition phases or cohesive devices (e.g., firstly, as a result) (Ke & Ng, 2019). These explicit clues may enhance comprehension of the essay, and thus improve the essay quality. For example, the use of connective words (e.g., furthermore, however) can serve as an indication of a cohesive and well-organized essay.
Linguistic Feature	Coherence	Coherence assesses the appropriate transitions between ideas (Ke & Ng, 2019), which enhances the reader's comprehension and understanding of the essay. Cohesion and coherence are interrelated concepts - the cohesion in essays facilitates the coherence and understanding of the reader.
Linguistic Feature	Word Usage (Lexical)	Word usage denotes the essay quality based on its lexical dimension. Word usage or lexical item typically includes

Table 2.4 continued

Feature Group	Features	Description
		evaluation of its lexical diversity and lexical sophistication (Crossley, 2020). The words used in essays demonstrate the vocabulary the writers actively use and thus reflect their level of competency.
Linguistic Feature	Style (Syntactic)	Style denotes the essay quality from its syntactic complexity dimension. Syntactic complexity involves assessing the variety and sophistication of syntactic structure in an essay (Crossley, 2020). Variation in the syntactic structures indicates proficiency in writing.
Linguistic / Content Feature	Semantics	Semantic features capture the lexical semantic relationships among different words, sentences, and paragraphs within an essay (Ke & Ng, 2019). Furthermore, these features are employed to evaluate the degree of similarity between the essay content and the given essay prompt. Semantic features are usually computed based on NLP techniques such as Latent Semantic Analysis (LSA; Landauer et al., 1998), Latent Dirichlet Allocation (LDA; Blei et al., 2003), and Word2vec (Mikolov et al., 2013a, 2013b).
Content Feature	Prompt Relevance/ Adherence	Prompt relevance or adherence assesses the extent to which the essay contents relate to the given essay prompt (Persing & Ng, 2014). It assesses whether the essay stays on-topic and the degree to which it remains focused. A high prompt relevance score essay exhibits consistent alignment with the essay prompt and is free of irrelevant digressions.
Content Feature	Thesis Clarity	Thesis clarity refers to the level of clarity with which writers present the main thesis of their essays (Persing & Ng, 2013). A high thesis clarity score essay usually consists of clear and

Table 2.4 continued

Feature Group	Features	Description
		explicit thesis sentences that are easy for the reader to understand.
Content Feature	Organization and Development	Organization features reflect how well the essay is structured. It shows the transitions between functions of discourse structure (Persing et al., 2010) and facilitates the development of ideas in an essay. A high organization score demonstrates the writer's proficiency in introducing a topic, expressing his stance, providing substantial support, and concluding by reiterating his standpoint effectively.
Content Feature	Argumentation	Argumentation features are features derived from the argumentative structure of essays. It is undeniable one of the most important aspects to be evaluated when assessing essay quality, particularly in argumentative essays (Ke, 2019). A high argumentation score in an essay indicates a strong persuasive argument for its thesis and is capable of convincing most readers (Persing & Ng, 2015).

2.3 Overview of Linguistic Features

From the 12 essay scoring features outlined in Table 2.4, this section drills down to the surface features and deep linguistic features that are the subject of investigation in this research. The features to be reviewed include surface features, linguistic error, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability. Each of the features is detailed in the following subsections, with the description of the fundamental concept of the feature, the measurement of the feature, and the relevant research of their influence on essay scoring.

2.3.1 Surface Features

Within the scope of this research, the surface feature is considered one of the linguistic features. This categorization is based on the rationale that the surface feature set specified in this research comprises not only the basic length-based features but also several linguistically related features, such as parts of speech. As detailed in Table 3.4, the surface feature set in this research consists of feature categories related to length-based features, short paragraphs, parts of speech, and advanced parts of speech. Among these categories, parts of speech and advanced parts of speech are related to linguistic features. On the other hand, the length-based features are a subset of the surface feature set. It specifies the count of the very basic elements in an essay, such as the number of letters, words, sentences, paragraphs, etc.

Among the length-based feature category, the essay length (sometimes referred to as text length in some literature), is perhaps the most important feature adopted for scoring essays. However, it is still debatable whether the essay length or other surface-level features can be considered construct-relevant when it comes to assessing writing quality (Fleckenstein et al., 2020). Crossley (2020) argued that he does not consider text length as a linguistic feature while acknowledging that text length is likely the strongest predictor of writing development and quality. Perelman (2014) and Phandi et al. (2015) found that these length-related features are widely used for AES, and highly correlated with essay scores. In contrast, Powers (2005) claimed that text length is either completely or partly irrelevant to the assessment of essay quality and can be considered as a bias.

In this research, the author argues that essay length and other length-based features are important constructs to be considered for scoring essays. Certain text lengths are essential

for effectively developing a point of view in the writing. The ability to produce longer texts can be considered as an indicator of higher competence in writing skills either from the perspective of linguistic, content, or development. The average sentence length in words and word length in characters are the indicators of the degree of complexity a writer can master (Attali & Burstein, 2006; Östling et al., 2013; Zesch et al., 2015). Besides the theoretical arguments, empirical results show that essay length has always been a strong indicator of essay scores by trained human raters as well as AES (Chodorow & Burstein, 2004; Powers, 2005; Kobrin et al., 2011; Guo et al., 2013). Powers (2005) in his review observed that most research that address the relationship between text length and text quality found correlations between them are within the range of $r = 0.50$ to $r = 0.70$. Another research by Kobrin and Shaw (2007) reported the number of words explains 39% of the variance in the SAT essay scores.

2.3.2 Linguistic Errors

In the speech or writing of a second or foreign language learner, a linguistic error refers to the use of a linguistic item (e.g., a word, a grammatical item, and a speech act) in a way which a fluent or native speaker of the language regards as showing faulty or incomplete learning (Richards & Schmidt, n.d.). These linguistic errors reflect the incompetence of the language learner in the correct rules of the target language variety. Based on linguistic error analysis, the many errors that appear in a written text can be categorized into numerous error types. However, in the context of academic essay assessment, these linguistic errors are usually referred to the common errors such as spelling errors and grammatical errors. Spelling errors which can be referred to as mechanical errors are those of orthography, which arise from not following the conventional spelling system of a language. Grammatical errors

are structural errors in which sentence structure does not adhere to a language grammar or syntactic rules, such as incorrect subject-verb agreement, pronoun disagreement, etc.

The importance of linguistic errors in determining writing quality was described in many writing assessment literature. These linguistic errors pose one of the greatest challenges for English as a Second Language (ESL) learners (Xie, 2019). Kampookaew (2020) analysed the most common essay writing errors and identified that grammatical errors are major concerns for L2 learners, and their presence or prevalence can significantly undermine the writing quality. Crossley et al. (2019a) found that there are significant correlations between grammatical and mechanical errors to the human rating of TOEFL and Writing-Pal essay corpora. However, another research by Allen et al. (2014) shows that linguistic errors, especially grammatical errors have a small effect on the expert judgment of essay scores.

To automate the error detection process, several tools have been developed to assess spelling and grammatical errors in texts. Criterion (Burstein et al., 2004) which was marketed by ETS was the first full-scale commercially available grammar and spelling tool, which identifies errors in essay and provide feedback to users for revision and improvement. There are also open-source tools available for linguistic error analysis. A good example is LanguageTools (Naber, 2003) which automatically group linguistic errors into several categories such as grammar, spelling, typographical, duplication errors, etc. Besides, the Computational Natural Language Learning (CoNLL) Shared Task on Grammatical Error Correction, introduced by Ng et al. (2014) is another effort for automatic identification and correction of grammatical errors.

2.3.3 Text Cohesion

Text cohesion is the characteristic of writing in which text segments, i.e., words, phrases, sentences, and paragraphs are glued together to create a logical flow in structure and meaning. This “glue” of text consists of cohesive devices such as linking words, punctuation, pronouns, etc. The major work in text cohesion is the one conducted by Halliday and Hasan (1976) in their book *Cohesion in English*, which categorised five general types of cohesive devices in text, namely reference, ellipsis, substitution, lexical cohesion, and conjunction. Text cohesion can occur at the sentence level (i.e., local cohesion) or span across larger text segments at the paragraph level (i.e., global cohesion) (Crossley et al., 2016; Crossley et al., 2019b). The common method to measure text cohesion is by quantifying or counting the cohesive devices between text segments including reference of previous elements (generally through pronouns), overlapping of lexical items, substituting lexical items, the use of connectives (e.g., because and therefore), and conjunction (e.g., and, but), etc.

When a text is well-formed and contains a high number of cohesive cues, it increases the text’s comprehensibility and thus assists the reader in following the text flow and understanding the text. In the absence of explicit cohesion clues, it will be difficult for readers to identify the relationship between text segments, thereby at a certain point, a reader’s mental representation of the text may break thus affecting comprehension. This mental representation of a text that a reader constructs is referred to as coherence (Sanders & Maat, 2006). A text containing numerous cohesive elements might lead to a more coherent text. However, coherence, as opposed to cohesion, is a more high-level construct of a text and varies as a function not only of text cohesion features but also of readers’

background knowledge and language proficiency (McNamara et al., 1996). A coherent text sticks to the point of writing/reading and thus improves text comprehension.

Text cohesion has been an important element in assessing writing quality. Research on the effect of text cohesive features on essay scoring is mixed, especially in the aspect of local cohesion. MacArthur et al. (2019) reported that high-quality texts are more likely associated with a greater density of referential cohesion (e.g., links between words across sentences). However, the research by Crossley and McNamara (2016), and Tian et al. (2020) found that local cohesion features (e.g., the occurrence of conjunctions and semantic overlap between adjacent sentences) were negatively associated with the essay quality. In contrast, a significant number of research have shown a positive correlation between global cohesion and essay quality. For instance, McNamara et al. (2013) found that global cohesion, particularly the lexical and semantic overlap between paragraphs was positively correlated with essay quality. Furthermore, Crossley and McNamara (2016) observed that student modifications to the text in the aspects of its global cohesion could lead to increased essay scores.

2.3.4 Semantics

In natural language, semantic analysis is about understanding the meaning of what is written in a particular text (Goddard, 2011). The semantic dimension of a written text covers the meaning and relationship of the occurrences of each word and the formation of phrases, sentences, and larger units of discourse in the text. Semantic analysis of a text requires an understanding of lexical hierarchy, including hyponymy and hypernymy, meronymy, polysemy, synonyms, antonyms, and homonyms (Manning & Schutze, 1999).

Semantics, as related to the meaning and interpretation of words, sentences, and texts largely determine our reading comprehension and writing quality. In automated essay assessment, semantic similarities are useful for understanding the coherence of the essays and their relevance to the essay prompt (Zupanc & Bosnić, 2017). There were research works to evaluate essays based on their semantic features as described in the literature by Zupanc and Bosnić (2017), Crossley et al. (2019b), Janda et al. (2019), and Hoblos (2020). These works compute semantic similarities from essays using various NLP techniques as listed below:

i. Latent Semantic Analysis (LSA)

LSA (Landauer et al., 1998) is a mathematical optimization approach that constructs high-dimensional semantic space from corpora of text which in turn is applied to compute the similarity between words, sentences, and documents (Hoblos, 2020). The semantic space is created based on word co-occurrences within documents that establish relationships between concepts. Mathematically, LSA constructs a term-document matrix, which is then normalized by using log-entropy, followed by Singular Value Decomposition (SVD; Golub & Reinsch, 1970) and projected over the most representative k dimensions. In practice, LSA is applied to reduce the dimensionality of text data; and analyses related concepts between the documents, the terms, and the relationships between them, as well as compute the similarities between the documents (Zupanc & Bosnić, 2017).

ii. Latent Dirichlet Allocation (LDA)

LDA (Blei et al., 2003) is a topic modelling technique based on a statistical method that analyse groups of words together to capture and discover how the meaning of words depends on the broader context (Janda et al., 2019). LDA perceives documents

as a mixture of multiple topics. Documents and words are topic distributions drawn from Dirichlet distributions, and topics (i.e., the latent variables) are perceived as Dirichlet distributions over the input vocabulary (Kotz et al., 2004). One typical application of LDA is topic discovery, i.e., to discover topics in a collection of documents, and then classify each document based on its relevance to the discovered topics.

iii. Word2vec

Word2vec (Mikolov et al., 2013a, 2013b) is a commonly used word embedding technique, and it is often considered the de facto standard for pre-trained word embedding in text analysis. The Word2vec's vector represents words in a k -dimensional vector-space model by using a neural network model. Each word's embedding is computed using the context around it; and thus, words co-occurred in similar contexts are represented as being nearer, whereas words with dissimilar contexts are represented as being farther apart. The rationale of Word2vec is to group similar words together on a feature space, thereby enabling the computation of similarities between words mathematically feasible. The Word2vec model uses two predictive models, which are the continuous bag-of-words (CBOW) and continuous skip-gram (CSG). In the CBOW model, the neural network predicts a word in a text from a window of surrounding context words, while in contrast, the CSG model predicts the surrounding words by looking at the centre word.

2.3.5 Lexical Diversity

Lexical diversity refers to the variety of unique words in relation to the total number of words by tokens that occur in a text. Consider, let's say this sequence of text "eat sleep eating sleeping eat sleep". This text has four word types and six word tokens, and its lexical

diversity as expressed by Type-Token-Ratio (TTR) is 0.67 (types/tokens). A higher value of lexical diversity denotes the text is lexically diverse, while a lower value indicates more words are repeated multiple times across the text. When the number of word types is equal to the total number of tokens, all the words are different.

As the traditional lexical diversity measurements such as simple TTR (Johnson, 1944) are sensitive to text length, a number of revised indices have been attempted to show that their indices are less susceptible to this effect. For example, Root TTR (Guiraud, 1959) is calculated as the number of types divided by the square root of the number of tokens; and Log TTR (Chotlos, 1944; Herdan, 1960) is computed by dividing the logarithm of the number of word types by the logarithm of the number of word tokens. Furthermore, advanced linguistic computational tools, such as Coh-Metrix provide additional measures of lexical diversity that stabilize the effect of text length. They are known as Measure of Textual Lexical Diversity (MTLD) and vocd-D. The MTLD is calculated based on the average number of tokens it takes to reach a given TTR value (McCarthy & Jarvis, 2010); and the vocd-D is derived through a series of random text samplings, each with 100 samples of text strings with 35 to 50 tokens (McCarthy & Jarvis, 2010).

The measures of lexical diversity approximate the lexical richness of essays. The more advanced writer uses a wider range of lexical items. As such, indices of lexical diversity are often applied in L2 writing assessment research as a measure of lexical richness and complexity (Zenker & Kyle, 2021). Treffers-Daller et al. (2018) found the basic measure of lexical diversity such as TTR expresses variance in overall scores of the Pearson Test of English Academic essays. Another work done by Juanggo (2018), showed that the lexical

diversity index of students at the higher level of the Common European Framework of Reference for Languages (CEFR) was greater than that of students at the lower level.

2.3.6 Lexical Sophistication

Lexical sophistication is the proportion of relatively unusual or advanced words in the learner's text (Read, 2000). Although there was no formal definition of lexical sophistication, the construct of lexical sophistication provides the metrics in reflecting both the depth and breadth of lexical knowledge of the writers. As such, lexical sophistication is usually adopted as a measurement of language proficiency in speaking and writing assessments (Kyle & Crossley, 2015; Kyle et al., 2018)

Various constructs can be used to measure lexical sophistication, with word frequency as the most basic metric being adopted. Word frequency calculates how frequently a word occurs in general usage, as measured by a representative corpus such as the British National Corpus (BNC; BNC Consortium, 2007), and the Corpus of Contemporary American English (COCA; Davies, 2010). Words that occur frequently in a reference corpus (e.g., good and bad) are considered less sophisticated than less frequent words (e.g., pragmatic and substantiate). Besides word frequency, word range which refers to the number of texts in a corpus in which a particular word occurs is another common metric used to measure lexical sophistication.

In recent research, a great number of constructs have been adopted to quantify lexical sophistication. For example, sophisticated words have been associated with words that are more likely found in academic text such as the Academic Word List (AWL; Coxhead, 2000, 2011). Sophisticated words have been defined as words that are less concrete, less imageable, and less familiar (Salsbury et al., 2011); and words that have higher latencies in

word naming and lexical decision tasks (Balota et al., 2007). Furthermore, recent research has begun to shift focus from single lexical items to larger units such as n-gram frequencies (e.g., in addition, and as such) in determining lexical sophistication (Garner et al., 2020).

Lexical sophistication is one of the common linguistic features used to assess writing quality, such that higher-quality texts are represented by more sophisticated words. Research indicates that the use of more academic words (Douglas, 2013), longer words and less familiar words (Crossley et al., 2011b), and greater use of infrequent words (McNamara et al., 2010) is a strong predictor of higher quality academic writing. In the aspect of n-gram frequencies, more proficient L2 writers tend to produce more bigrams (Paquot, 2019) and trigrams (Kyle & Crossley, 2015; Kyle et al., 2018) than the lower proficiency writers. Garner et al. (2020) in their recent research found that both bigram and trigram features account for significant variance in human judgments of writing quality for Korean L2 writers of English.

2.3.7 Syntactic Complexity

Syntactic complexity has been an important construct that can influence a text's perceived quality. The component of syntax provides structures in which discrete units (e.g., words) can be combined to create meaningful utterances (e.g., sentences). Despite the lack of a clear definition of syntactic complexity, Ortega (2003) and Lu (2011) describe syntactic complexity as the sophistication of syntactic forms and the variety of syntactic forms produced. Text that comprises shorter sentences, with few words before the main verb and few words per noun phrase, is syntactically simple. Conversely, longer sentences with multiple clauses are structurally dense and syntactically complex.

Different linguistic features have been adopted as the approximation to measure syntactic complexity in writing. Traditional indices to measure syntactic complexity have involved calculating sentence length and T-unit count. A T-unit is defined as the shortest grammatically allowable sentences into which writing can be split, or as a minimally terminable unit (Hunt, 1965). Thus, a T-unit can consist of a main clause plus subordinated clauses, but not two independent clauses joined together (Crossley, 2020). Besides the mean length of T-unit which is used in almost 90% of the syntactic complexity research (Ortega, 2003), more fine-grained indices of T-unit (e.g., number of clauses per T-unit, and number of T-units per sentence) have been developed (Ortega, 2003; Lu, 2011). On the other hand, there has been the development of syntactic complexity indices that do not rely on T-units. For instance, the Coh-Metrix tool approaches syntactic complexity with numerous indices, such as the number of words before the main verb, the number of constituents in a sentence, sentence syntax similarity, etc (Graesser et al., 2004). In addition, Kyle (2016), in his recent works, provides the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASCC) which measures phrasal and clausal complexity features (Kyle & Crossley, 2018) along with features related to the usage-based characteristics of verb argument constructions (Kyle & Crossley, 2017) for measuring syntactic complexity.

Like lexical properties, syntactic complexity is another approach that can be used to assess essay quality. The underlying rationale is that more complex syntactic structures can be used as an indicator of more advanced writing skills. Numerous research works have been carried out to investigate the syntactic differences among different writing ability levels. Research based on T-units found that more proficient writers produce longer and more varied syntactic structures (Ortega, 2003; Lu, 2011). Other research reported that higher-quality essays consist of greater clausal subordination and passive structures (Grant & Ginther,

2000), higher numbers of dependent clause features (e.g., the incidence of infinitives and “that” verb complements) (Crossley & McNamara, 2014). Kyle and Crossley (2018) reported that fine-grained indices of phrasal complexity (related to the nominal subject, direct object, and prepositional object) explained a great variance in essay scores, while in another research, they found that novel complexity indices based on verb argument constructions were good predictors of writing quality (Kyle & Crossley, 2017).

2.3.8 Readability

In scoring an essay, the concept of text readability can act as a proxy for its quality. The notion of text readability is best defined as the ease with which a text can be read and understood in terms of the linguistic features found within a text (Dale & Chall, 1949). Pikulski (2002) further elaborates on readability as "the level of ease or difficulty with which text material can be understood by a particular reader who is reading that text for a specific purpose". In practice, readability is widely used to assess the difficulty of texts, with the purpose of matching appropriate reading material for language learners.

A common method to measure readability is the use of a readability formula, which typically consists of several variables related to text characteristics and constant weights. These text characteristics can be shallow text-based features (e.g., the number of words per sentence, the number of characters per word), or other advanced linguistic features such as lexical sophistication, syntactic complexity, text cohesion, semantics, and sentiment (Crossley et al., 2019c). These readability formulas are manifold, with well over 200 models emerging since the 1940s (Benjamin, 2012). Some of the widely used traditional readability formulas are the Flesch Reading Ease (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid et al., 1975), and Dale-Chall Readability (Dale & Chall, 1948). The formulas of these

readability indices are attached in Appendix A. Both the Flesch Reading Ease and Flesch-Kincaid Grade Level have been implemented as a feature in word processing software such as Microsoft Word.

Furthermore, advances in natural language processing and computational linguistics have facilitated improvements in traditional readability formulas. The Coh-Metrix tool provides the Coh-Metrix L2 Readability Index (CML2RI) based on psycholinguistic and cognitive models to measure English text readability for L2 readers (Crossley et al., 2011a). In recent research, Crossley et al. (2019c) used crowdsourcing techniques to develop novel readability formulas, namely the CAREC (Crowdsourced Algorithm of Reading Comprehension), and CARES (Crowdsourced Algorithm of Reading Speed) for text comprehension and reading speed. These readability formulas will be detailed in Chapter 3.

In the aspect of essay assessment, the readability formula which consists of numerous linguistic features (e.g., lexical sophistication and text cohesion) might play a role in grading essays. However, the relationship between text readability and text assessment is still a widely debated topic in the field of linguistics. Nigam (2017) observed that the Flesch-Kincaid Grade Level is one of the important features in automated essay scoring, while Nemati and Azizi (2013) reported that readability indexes, such as Flesch Reading Ease Flesch-Kincaid Grade Level, and Dale-Chall Readability appeared to be of almost no relationship (Correlation Coefficient, $r < 0.15$), with the TOEFL holistic scores.

2.4 Overview of Content Features

This section reviews the content features based on the 12 essay features outlined in Table 2.4. It complements the linguistic features discussed in Section 2.3, which are the primary focus of this research. This research reviewed five content-based features, with four

of them elaborated in this section. The fifth feature, semantics, is included in Section 2.3, as this research considers semantics as both a linguistic and content feature.

2.4.1 Prompt Relevance / Adherence

Prompt relevance or adherence refers to the aspect of essay quality that addresses the essay's alignment with the given topic. It deals with the evaluation of whether an essay appears to be on-topic, off-topic, or how well the essay is on-topic (Persing & Ng, 2014). When students are asked to write an essay in response to a particular prompt, one of the primary goals is to assess their ability to address the topic effectively, stay on point, and provide relevant arguments or explanations. Prompt adherence is essential because an essay that diverts from the topic or fails to address the key points of the prompt reveals the student's limited writing abilities or understanding of the subject matter. Prompt relevance or adherence is considered a content feature because the essay content needs to be analysed and compared with the prompt to determine its relevancy.

Scoring the essay in the dimension of prompt relevance involves treating the problem as a binary classification task, where an essay is evaluated as either on-topic or off-topic. Louis and Higgins (2010) developed a binary classifier to detect on-topic and off-topic essays by measuring semantic similarity between the essay content and expanded essay prompts. In addition, a fine-grained approach can be used to assess prompt relevance, wherein essays are rated on a numerical scale, reflecting the extent to which the essay adheres to the prompt. Persing and Ng (2014) approached prompt relevance by proposing feature-rich criteria, including semantic similarity, thesis clarity keywords, thesis clarity errors, prompt adherence keywords, and LDA topics to assess the essay content.

2.4.2 Thesis Clarity

Thesis clarity refers to the ability of the writer to explain the thesis statement of an essay clearly (Jong et al., 2023). It assesses whether the central argument or main idea of an essay is effectively communicated and easily understandable to the reader. A well-defined thesis statement should be clear, concise, and focused, serving as a roadmap for the development of essay content. Moreover, the thesis statement should be free of significant errors, such as excessive misspellings, that could hinder the reader's comprehension of the essay (Persing & Ng, 2013).

One of the pioneering research works in assessing essay thesis clarity was carried out by Persing and Ng (2013). They approached the assessment of thesis clarity by identifying types of errors that could reduce the clarity of a thesis. In the work, they proposed five novel error features, including confusing phrasing, incomplete prompt response, relevance to prompt, missing details, and writer position. They identified features associated with each of these defined errors and proceeded to train five binary classifiers. Each classifier was responsible for predicting the presence of a particular type of error within the essays.

2.4.3 Organization and Development

Organization and development refer to the overall structure of an essay (Persing et al., 2010). It assesses the logical flow of the ideas in an essay, by which the writer presents their main arguments and supporting evidence. An essay with a high organization score should have a clear introduction paragraph that introduces the main thesis, body paragraphs that each focus on a single main point with supporting evidence, and a conclusion paragraph that summarizes the standpoint effectively.

Research on assessing the organization and development of essays can be found in works by Attali and Burstein (2006) and Persing et al. (2010). During the development of e-rater V.2, Attali and Burstein (2006) evaluate essay organization with a focus on the five-paragraph essay structure. This five-paragraph essay typically comprises an introductory paragraph, a minimum of three body paragraphs - each containing two main points supported by evidence, and a final concluding paragraph. They then score the essay organization by measuring the difference between the discourse elements counted in an essay and the specified five-paragraph essay structure. On the other hand, Persing et al. (2010) proposed methods to identify the structure of the discourse in an essay through paragraph function labels (introduction, body, conclusion, rebuttal) and sentence function labels (e.g., thesis, main idea, elaboration, support, conclusion, rebuttal). Based on these paragraph and sentence labels as essay features, they adopted both heuristic and machine learning approaches to predict the organization score of an essay.

2.4.4 Argumentation

Argumentation features are the features related to the argumentative structure of essays and are only valid for argumentative or persuasive essays (Ke, 2019). An essay generally consists of a major claim that represents the writer's standpoint on the topic being discussed. Based on the major claim, there are subsidiary claims that either support or oppose the claim. Each of these subsidiary claims is followed by one or more premises, which serve as evidence to support or oppose the respective claims (Ke, 2019). The structure of an argumentative essay is often depicted as a tree-like structure, with nodes representing the components of the arguments (e.g., the major claim, subsidiary claims, premises) and edges encoding the relationship (e.g., supporting or opposition) between the argumentative components.

The evaluation of argumentation features in an essay usually involves argument mining to identify and extract argumentative components from an essay, and the assessment of the argument's strength or persuasiveness from these extracted argumentative components. Falakmasir et al. (2014) attempted to detect the thesis statement and its conclusion in essays on peer-reviewing processes. They incorporated 19 features in three feature categories, including positional, sentence-level, and essay-level features. Using various classifiers, they classified sentences as thesis statements, conclusions, or other statements based on the proposed features. Ong et al. (2014) constructed an ontology by tree structure for labelling the different components of an argumentative essay. They then developed eight heuristic rules to assign each sentence in the essay to its corresponding category in the ontology. Persing and Ng (2015) formulated a feature-rich approach to evaluate argument strength in essays. This approach integrated diverse features, including POS n-grams, transition phrases, coreference, and prompt agreement to assess the strength of the arguments presented in the essay.

2.5 Feature Selection

Feature selection is the process of selecting the most informative and discriminative features that contribute to the predictive capability of machine learning models. The characteristics of the features in a dataset can be broadly classified into four categories: (a) completely irrelevant and noisy features, (b) weakly relevant and redundant features, (c) weakly relevant and non-redundant features, and (d) strongly relevant features (Cai et al., 2018). Feature selection is crucial to eliminate irrelevant and redundant features, especially when dealing with high-dimensional datasets. The benefits of feature selection are numerous. It helps to mitigate the “curse of complexity” by reducing the number of features considered to construct the machine learning models. This reduced number of features

facilitates easier explanation and interpretation of the models, as well as reduces required computational resources and training time.

This section reviews different feature selection techniques, as one of the objectives of this research is to select optimized linguistic features that influence automated essay scoring from the extensive range of identified features. Section 2.5.1 discusses various common feature selection techniques by categorizing them into different groups, while Section 2.5.2 reviews feature selection applied particularly in AES research.

2.5.1 Feature Selection Techniques

Based on different selection strategies, feature selection techniques can be classified into three types, namely the filter, wrapper, and embedded methods (Venkatesh & Anuradha, 2019). The following section outlines and describes various feature selection techniques based on these three methods:

2.5.1.1 Filter Methods

Feature selection techniques based on filter methods evaluate the relevance of a feature based on statistical criteria and they operate independently of any specific classifier (Alelyani et al., 2018). These filter methods are generally robust to overfitting, as they focus on intrinsic characteristics of features without relying on a specific machine learning model. Furthermore, they are computationally efficient, which makes them suitable to operate on large datasets with high-dimensional features. However, filter methods are insensitive to machine learning algorithms which potentially restricts their capability to further enhance the performance of the particular learning model. According to the study conducted by Venkatesh and Anuradha (2019), Pearson's correlation and mutual information are two commonly used techniques in filter methods. Other popular filter methods include the Fisher

Score, Relief, and Minimum Redundancy Maximum Relevance (MRMR) feature selection. Table 2.5 describes these feature selection techniques and their corresponding references for further information.

Table 2.5: Feature Selection Techniques based on Filter Methods

Feature Selection	Description	Reference
Pearson's Correlation	Pearson's Correlation measures the strength and direction of the relationship between features and the target variables.	(Pearson, 1895)
Mutual Information	Mutual Information quantifies the amount of information that one feature provides about the target variables by assessing the reduction of entropy in the target variables.	(Vergara & Estévez, 2014)
Fisher Score	Fisher score assesses the discriminative power of features by comparing the ratio of between-class and within-class variances. A higher Fisher score indicates a more significant feature.	(Duda et al., 2000)
Relief	The Relief algorithm uses the nearest neighbors approach to evaluate the relevance of features by estimating how well they can distinguish instances that are similar to each other. Features with higher Relief scores are considered more important.	(Robnik-Šikonja & Kononenko, 2003)
Minimum Redundancy Maximum Relevance (MRMR)	MRMR aims to select features that are highly correlated with the target variable while having minimal redundancy with other features.	(Ding & Peng, 2005)

2.5.1.2 Wrapper Methods

In wrapper methods, the feature selection process is integrated with the model training process. They use dedicated machine learning algorithms to evaluate different feature subsets for selecting the optimal combination with the best model performance. A searching strategy is usually employed to identify the optimal feature subsets. By incorporating the machine learning algorithm in the feature selection process, wrapper methods demonstrate the strength of identifying the optimal feature subset tailored for a particular learning model. However, the wrapper methods can be extremely computationally expensive, where the search space for m features is $O(2^m)$. Furthermore, the wrapper methods are model-dependent and probably unable to be generalized well to other models. The wrapper method feature selection involves various searching strategies, such as forward selection, backward selection, hill climbing, and genetic algorithm. Table 2.6 outlines these feature selection techniques with their relevant references for detailed information.

Table 2.6: Feature Selection Techniques based on Wrapper Methods

Feature Selection	Description	Reference
Forward Selection	Forward selection starts with an empty set of features and iteratively adds one feature at a time until a stopping criterion specified for the model performance is met.	(Khaire & Dhanalakshmi, 2022)
Backward Selection	Backward selection starts with the full set of features and iteratively removes one feature at a time until a stopping criterion specified for the model performance is met.	(Khaire & Dhanalakshmi, 2022)
Hill Climbing	Hill climbing begins with a random set of features and iteratively searches for better feature subsets by adding or removing features based on feature	(Wan, 2019)

Table 2.6 continued

Feature Selection	Description	Reference
	performance. The stopping criterion is usually set by defining the maximum number of iterations.	
Genetic Algorithm	Genetic algorithms employ principles inspired by natural selection to iteratively search for an optimal feature subset. It involves evolutionary processes such as selection, mutation, and crossover to evolve the initial population over multiple generations to produce the optimal feature set.	(Pham & Raahemi, 2023)

2.5.1.3 Embedded Methods

Embedded methods were introduced to bridge the gap between the filter and wrapper methods by embedding feature selection as an inherent part of the model training process. The learning algorithms determine optimal features through their respective internal mechanisms. Embedded methods combine the benefits of both filter and wrapper methods - they are less computationally intensive than wrapper methods and they incorporate the interaction with the learning model to enhance predictive capabilities. However, one drawback of embedded feature selection is that they are constrained by the choice of available learning algorithms. The accuracy of predictions can be impacted if the chosen algorithm does not effectively perform on the dataset. The widely used embedded method feature selections involve various algorithms, such as Least Absolute Shrinkage and Selection Operator (LASSO) regression, Ridge regression, and ensemble learning using Random Forest and Gradient Boosted Tree. Table 2.7 provides an overview of feature selection techniques based on embedded methods with the corresponding references.

Table 2.7: Feature Selection Techniques based on Embedded Methods

Feature Selection	Description	Reference
LASSO Regression	LASSO regression uses L1 regularization to shrink the coefficients of irrelevant features to zero, effectively performing feature selection.	(Bouchlaghem et al., 2022)
Ridge Regression	Ridge regression uses L2 regularization to penalize large coefficients, encouraging the model to select features that have a smaller impact on the outcome.	(Bouchlaghem et al., 2022)
Random Forest	Random forest is an ensemble learning model based on bootstrap aggregation or bagging algorithm. It involves building multiple decision trees where each tree is trained on a randomly selected subset of data and features. The randomness of the data and features helps to reduce the probability of overfitting.	(Breiman, 2001)
Gradient Boosted Tree	Gradient Boosted Tree is an ensemble learning model based on the boosting algorithm. It builds a series of decision trees sequentially, where each tree correcting the errors of the previous tree to boost the model performance.	(Mason et al., 1999)

2.5.2 Feature Selection in AES Research

Numerous feature selection techniques have been utilized in the domain of AES to identify the optimal feature subsets for predicting essay scores. Based on the review of recent literature, the following section presents a summary of AES research that have incorporated feature selection in their scoring methodology.

i. Information Gain

Rudner and Liang (2002) employed the Bayesian model combined with information gain to construct a probabilistic model AES. By computing the information gain from the words, phrases, and arguments in the essays, the AES could achieve an accuracy of 0.8. Similarly, Li and Liu (2017) utilized information gain to measure the feature importance of the extracted unigrams, bigrams, and Coh-Metrix linguistic features from Chinese English Learners essays. Their model demonstrated that the selected features based on information gain produced impressive results even when they used only the top 5 and top 15 selected features.

ii. Relief and Correlation

Madala et al. (2018) investigated the effectiveness of the Relief and correlation-based feature selection in selecting the surface and linguistic features from ASAP essays. By using the Relief approach to rank the features and the correlation analysis to identify the most relevant feature subset, the AES system achieved a favourable accuracy ranging from 73% to 93% by using k-Nearest Neighbour, Linear Regression, and Support Vector Machine classifiers.

iii. Forward Selection

Harika et al. (2017) used the forward selection approach to select the optimal surface feature subset from one of the ASAP essay datasets. The forward selection approach starts with an empty set and iteratively adds one more feature at a time until an optimal set is reached. Likewise, Shiva and Ravibabu (2019) employed the forward selection method to select the bag of words and Parts of Speech (POS) features to score essays.

iv. Pearson Correlation, LASSO, and Ridge Regression

Kumar and Boulanger (2020) applied a thorough feature selection process that includes Pearson correlation, LASSO, and Ridge regression to reduce the 1592 linguistic indices that measure text cohesion, lexical diversity, lexical sophistication, and syntactic sophistication in the ASAP essay dataset. The initial step involved using Pearson correlation to identify and remove features that had an absolute correlation value exceeding 0.7 with each other. Next, LASSO regression was applied to further prune the features based on their contribution to the model performance. Lastly, Ridge regression was employed to address multicollinearity issues among the remaining features. The values of the selected features were utilized to provide an explanation and interpretation regarding the scoring of the developed DNN-AES model.

v. LASSO Regression and Gradient Boosted Tree

Filho et al. (2021) conducted a comparison between feature engineering and deep learning approaches for scoring essays in the Brazilian High School Examination. For the feature engineering approach, they adopted the LASSO and Gradient Boosted Tree to select the 623 textual features across five dimensions: lexical diversity and statistic, bag of words, textual cohesion and coherence, theme adherence, and argumentation. The LASSO feature selection was used to improve the accuracy of linear regression by removing lower-impact features. It identifies relevant features while simultaneously reducing the impact of irrelevant features through regularization. On the other hand, the Gradient Boosted Tree was used to improve the accuracy of the decision tree by combining multiple weak decision trees, where each subsequent tree corrects the errors made by the previous tree. Both the LASSO

and Gradient Boosted Tree approaches attained a significant increase in model performance, with increments of 56.7% and 28.8% respectively in the QWK score.

2.6 Trends and Challenges in Automated Essay Scoring Research

This section summarizes the reviewed works and present the research trends, possible research gaps and justify the significance of this research.

2.6.1 Essay Features

Essay features play a significant role in automated essay scoring. In general, the essay features used by AES can be categorized into three groups, i.e., the surface features, linguistic features, and content features. These “top-level” feature groups can be further decomposed into hierarchical structures of quantifiable fine-grained features to score essays. For example, the linguistic features group consists of text cohesion features which can be further drilled down to fine-grained features indices of lemma overlapping measure between sentences to score essays. Most of the AES research work picks some essay features of their interest and attempts various machine learning models or neuro approaches with the main purpose of improving scoring accuracy. The works may involve a broad spectrum of essay features ranging from surface features, and n-grams, to semantics and syntactic structure, but they often lack detailed explanations and justifications of the adopted features. Systematic work which encompasses a detailed investigation of features in a particular feature group to score essays is not found. The author believes such work is significant as it involves an in-depth analysis of the essay features in that specific feature group to provide the scoring interpretability and construct validity of the given score. Such AES with the scoring interpretability capability can be further extended in writing instruction applications

to provide feedback to students for improving their writing skills, rather than merely a scoring engine of giving holistic scores.

2.6.2 Techniques of Essay Feature Extraction

AES systems can be broadly classified into two categories based on the approaches they applied to extract essay features. The first category involves handcrafted feature engineering, where domain experts are required to manually define and select the essay features. The second category adopts neural-based approaches (e.g., CNN, RNN) to automatically learn relevant features and representations from the raw essay text without explicit manual feature engineering. Among the 12 AES systems reviewed in this chapter, eight AES utilize handcrafted feature engineering, while the remaining four newer DNN-AES leverage neural approaches to extract essay features. All the top-rated and widely adopted commercial AES systems (PEG, IEA, e-Rater, Intellimetric) are feature engineering based, while the neural approach AES systems have recently received much attention because of their remarkable performance in scoring essays when trained on large and diverse datasets.

Both handcrafted feature engineering and neural-based approaches face their challenges when employed to extract essay features for scoring essays. To facilitate comparison, the author attempts to contrast the strengths and weaknesses of both approaches, as outlined in Table 2.8 below.

Table 2.8: Strengths and Weaknesses of Handcrafted Feature Engineering versus Neural-based Approaches in AES

Approach	Strengths	Weaknesses
Handcrafted Feature Engineering	▪ Handcrafted feature engineering enables domain experts to have more control over the selected features, ensuring that the feature can be more explainable and interpretable. ▪ This approach can be effective when the features are well-defined and relevant to the scoring criteria.	▪ Handcrafted features are time-consuming and labour-intensive. It can be expensive especially if domain experts are required to annotate data. ▪ The handcrafted features may not generalize well to essays in other test domains with other scoring rubrics.
Neural-based Approaches	▪ The neural-based approach automatically learns feature representations from the essay text. It eliminates the need for feature engineering and potentially discovers hidden patterns that are difficult to define manually. ▪ This approach can be more adaptable to different types of essays, as the feature set is not strictly hard-coded.	▪ The neural-based approach is a data-driven method that requires a substantial number of essays for training to achieve optimal performance. ▪ The neural network's black-box nature makes it difficult to understand and interpret the scoring mechanism, leading to potential issues of operational transparency and essay score interpretability.

2.6.3 Essay Dataset

To build machine learning models to score essays, a considerable number of accurate pre-score essays are required to train the model. This necessity of large essay datasets is

more rigorously required for training AES based on deep learning models. However, there is a lack of publicly available datasets in this domain. The commercial proprietary AES systems mostly use essays from their respective educational institution and do not reveal their dataset for public use. Based on the review by Ramesh and Sanampudi (2021), 90% of AES research relied on the ASAP essay dataset to build their scoring model. The ASAP dataset consists of 21,450 essays (12,978 training sets, 4218 validation sets, and 4,254 test sets) across eight essay prompts making it large enough to build efficient machine learning models even with the data-intensive deep learning models. However, the author observes the dataset has limitations that may restrict its efficacy from at least two perspectives. Firstly, the dataset may not be as genuine as the real essay. All the essays have been pre-processed to contain only one paragraph, and the name entities and capitalized words have been replaced by labels. Secondly, the ASAP dataset was collected from the test of 7th to 10th grade of native English-speaking US students which may be associated with different assessment criteria. The dataset may not be context-relevant to the Malaysian English test environment, which makes it inappropriate to be used for scoring essays in the local context. Thus, a local Malaysian English essay dataset collection is essentially required for formulating Malaysian AES.

2.6.4 Performance of AES Systems

The fundamental criteria to determine whether an AES is capable of scoring essays relied upon its reliability measurement. A reliable AES shall produce system scores that are agreeable to the human scores. Various evaluation metrics such as exact agreement or classification accuracy, Cohen's Kappa statistics (Cohen, 1960), its variance Quadratic Weighted Kappa (QWK; Cohen, 1968), and Pearson's product-moment correlation coefficient (Pearson, 1895) have been used to measure AES reliability. To understand the

performance of state-of-the-art AES systems, the author compiled the AES performance reported in several works in Table 2.9. Besides the AES results by the respective AES vendors/developers, Table 2.9 shows the AES system performance obtained from three large-scale independent third-party studies. The most prominent third-party study was conducted by Shermis (2014) who evaluated the performance of the four dominant AES vendors (PEG, IEA, e-rater, and Intellimetric) and LightSide using 21,450 ASAP essays dataset. Rupp et al. (2019) carried out another comprehensive study by evaluating the e-rater's performance using 9,628 TOEFL six-point rubrics essays from high school students in Germany and Switzerland. The third large-scale study was conducted by Latifi and Gierl (2021) who evaluated the Coh-Metrix performance by using the 12,978 ASAP training dataset.

Table 2.9: Performance of the State-of-the-Art AES Systems

AES System	Exact Agreement Rate / Accuracy	Quadratic Weighted Kappa (QWK)
PEG	0.66* (Shermis, 2014)	0.78* (Shermis, 2014)
IEA	0.61* (Shermis, 2014)	0.73* (Shermis, 2014)
e-rater	0.52 (Attali & Burstein, 2006) 0.63* (Shermis, 2014) 0.38* (Rupp et al., 2019)	0.76* (Shermis, 2014) 0.77* (Rupp et al., 2019)
Intellimetric	0.60 (Rudner et al., 2006) 0.66* (Shermis, 2014)	0.75* (Shermis, 2014)
BETSY	0.80 (Rudner et al., 2006)	N/A
Coh-Metrix	0.67* (Latifi & Gierl, 2021)	0.70* (Latifi & Gierl, 2021)
LightSide	0.62* (Shermis, 2014)	0.74* (Shermis, 2014)
Neural Approach AES	N/A	0.76 (Taghipour & Ng, 2016)

Table 2.9 continued

AES System	Exact Agreement Rate / Accuracy	Quadratic Weighted Kappa (QWK)
SSWE AES	N/A	0.94 – 0.96 (Alikaniotis et al., 2016)
Attention-Based Hierarchical Model AES	N/A	0.76 (Dong et al., 2017)
SKIPFLOW AES	N/A	0.76 (Tay et al., 2018)
SAGE	0.77 (Zupanc, 2018)	0.83 (Zupanc, 2018)

* Performance of AES systems reported by third-party studies

As shown in Table 2.9, the author finds that:

- i. QWK is the most adopted evaluation metric for assessing AES system performance. In general, most State-of-the-Art AES Systems achieve exact agreement rates/accuracies surpassing 0.60, and QWK scores exceeding 0.70.
- ii. Several AES vendors and researchers did not disclose their AES systems' exact agreement rate or accuracy. Instead, the AES systems' accuracy metrics were mostly reported through evaluations conducted by independent third-party studies.
- iii. While many AES systems reported high QWK scores, the exact agreement rates or accuracy of these systems always fall below their associated QWK values. These low accuracy values are unlikely to stem from the large rubric scale range, as the essay datasets mostly scored on moderate three to six-point scores.

Ifenthaler (2022) also observed a similar phenomenon as described in (ii) and (iii). He commented the exact agreement rates of AES systems with human raters are always below the desired level. Furthermore, he highlighted that many AES systems presented their results based on the adjacent agreement rates between human raters and AES systems rather than those of exact agreement rates.

2.7 Chapter Summary

The advancement of information technology, namely the Internet and Artificial Intelligence facilitates the growing interest in AES systems. However, most of the established AES systems are commercial proprietary software (e.g., PEG, IEA, e-rater, and Intellimetric). The other AES systems are simply by-products of research addressing a particular AES problem. Of all the AES systems, one of the main differences is the essay features employed in scoring essays. Despite the multiple and various features adopted in different AES systems, they can be broadly organized into three groups, i.e., the surface features, the linguistic features, and the content features. This chapter describes the 12 commonly adopted essay features in scoring essays. It focuses on the linguistic feature group to provide a comprehensive literature review of the eight linguistic feature dimensions, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability in affecting the essay score. Besides linguistic features, an additional four content features were reviewed, including prompt relevance/adherence, thesis clarity, organization and development, and argumentation.

Finally, the author created visual representations of the primary topics in hierarchical diagrams below to facilitate comprehension of the vast information presented in this chapter. Figure 2.2, Figure 2.3, and Figure 2.4 provide an overview of the AES systems, essay features, and feature selection techniques, as discussed in this chapter.

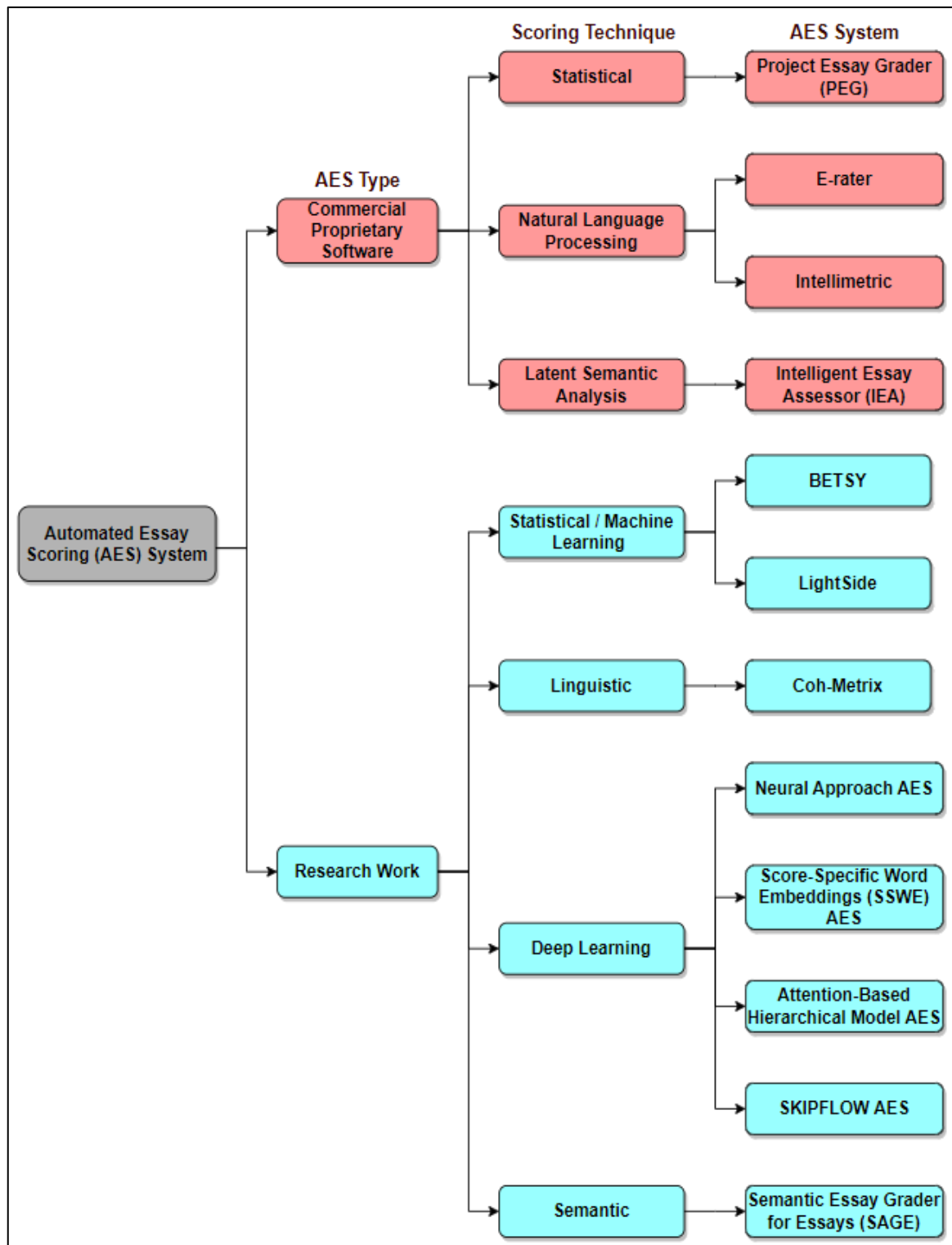


Figure 2.2: Classification of AES Systems

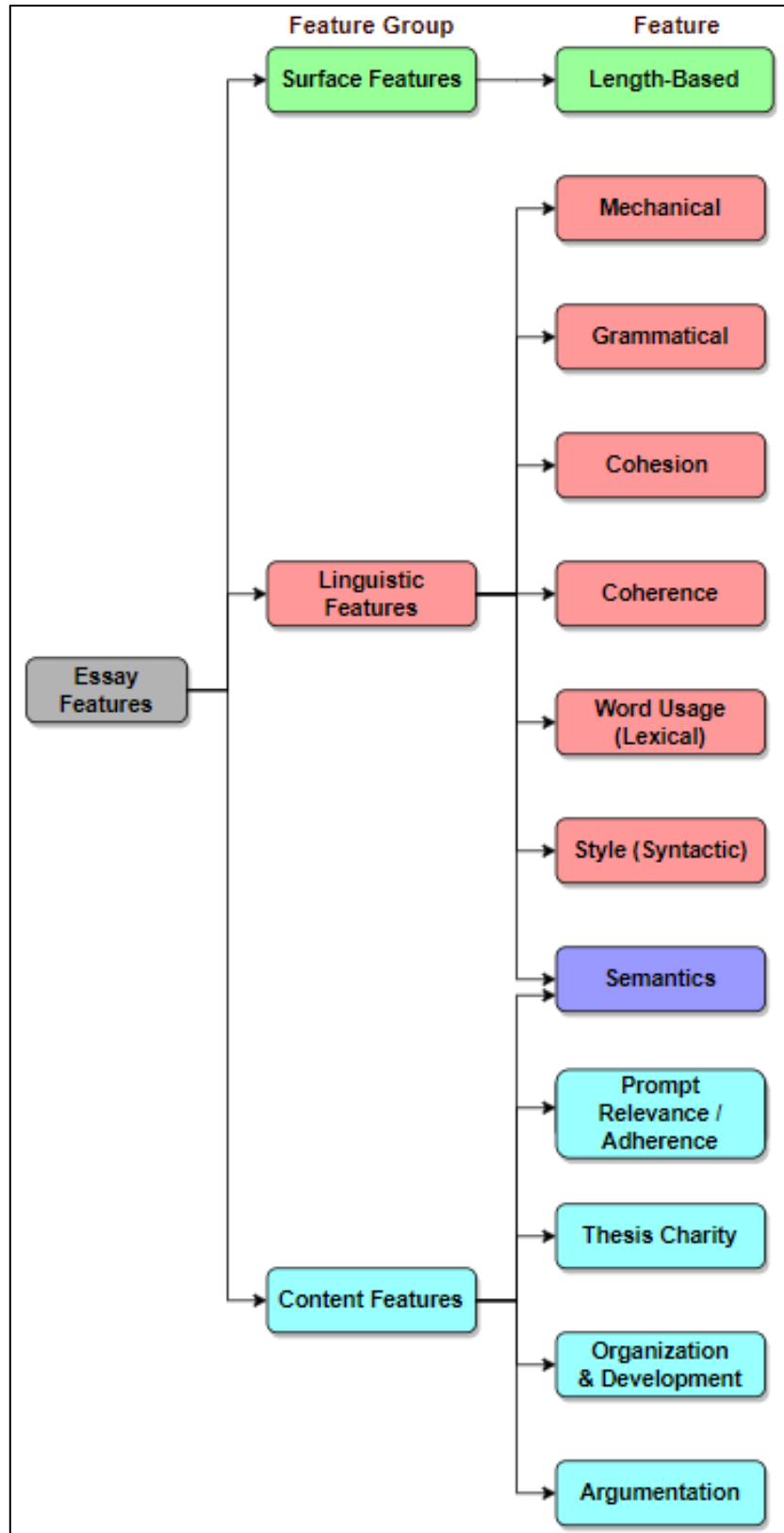


Figure 2.3: Essay Features for Automated Essay Scoring

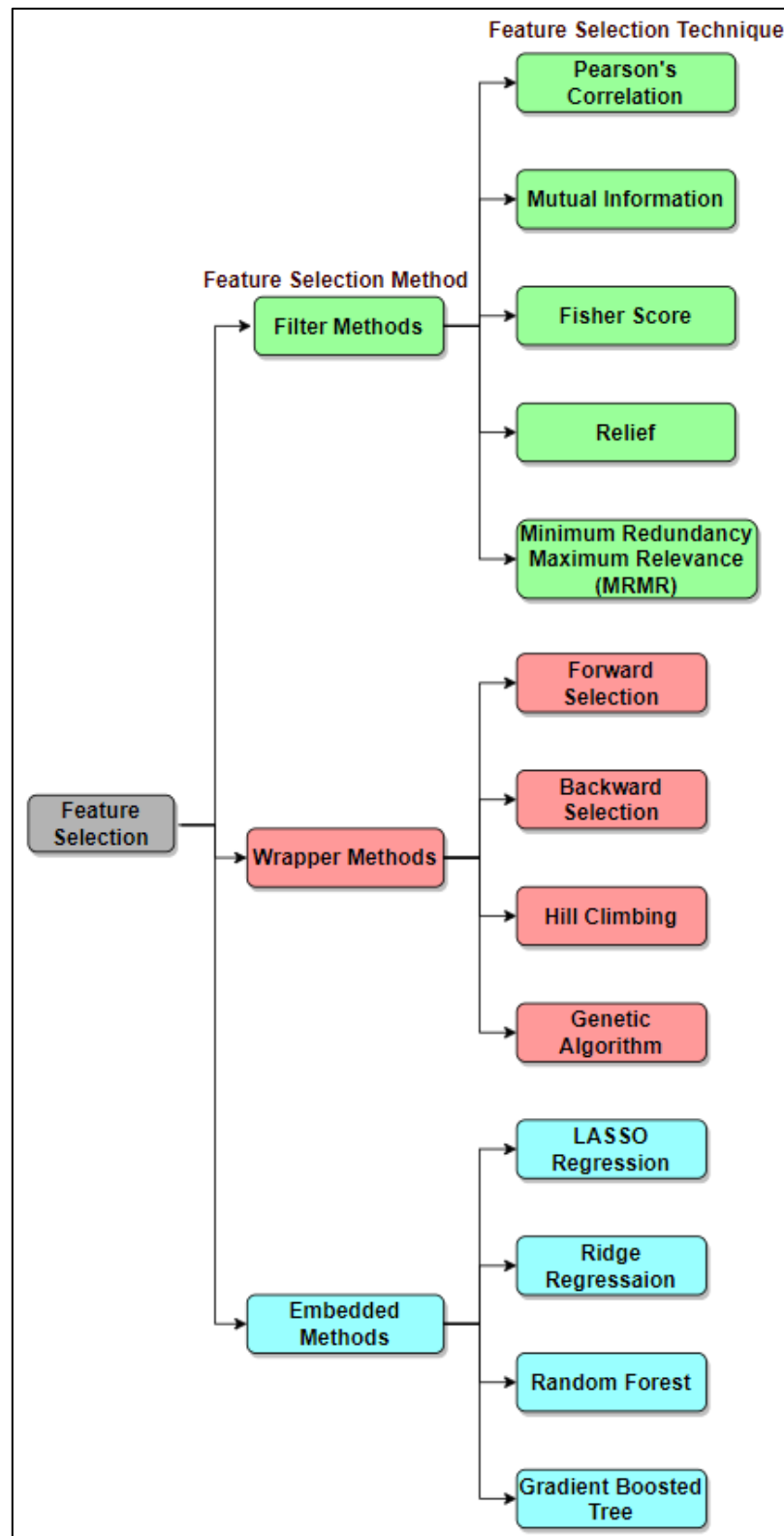


Figure 2.4: Feature Selection Techniques based on Filter, Wrapper and Embedded Methods

CHAPTER 3

MATERIALS AND METHODS

This chapter presents the material and methods used for the analysis and scoring of MUET essays in this research. Section 3.1 discusses the Malaysian University English Test (MUET) and the MUET essay scoring rubrics specified by the Malaysian Examination Council (MEC). The following Section 3.2 specifies the MUET essay dataset used in this research. Section 3.3 is the longest in this chapter, outlining and elaborating on the linguistic feature sets, feature categories, and feature indices used in this research. Section 3.4 details the two feature selection schemes used in this research, namely the Correlation Rank and MRMR Feature Selection. Lastly, Section 3.5 concludes the chapter with a complete list of investigated linguistic feature sets and feature categories in a table.

3.1 Malaysian University English Test

To formulate a linguistic-based automated essay scoring for the Malaysian test environment, this research used the Malaysian University English Test (MUET) essays as the specific case study for our investigation. The analysis and scoring of MUET essays are significant and yet beneficial due to the nature of MUET as high stakes (significant to test-takers) and large scales (involving a large number of test-takers).

MUET is an English test administrated by the Malaysian Examination Council to assess the English language competency of Malaysian students seeking admission to tertiary education. It evaluates the proficiency level of candidates in four language skills: listening, speaking, reading, and writing. MUET's marking scheme relies on aggregated score scales from zero to 300, which corresponds with a band-level system spanning from Band 1 to

Band 6 (Malaysian Examination Council, 2014). As described by the aggregated scores, Band 1 is the limited English user, while Band 6 is the highly proficient user. The MUET writing test consists of two writing tasks. The first task involves writing a text based on a non-linear source, such as graphs, charts, tables, etc. The second task requires candidates to write an essay in response to a prompt that may include various genres, such as descriptive, persuasive, analytical, and argumentative. This research focuses on analysing and scoring the essays written in the second task of the MUET writing assessment.

3.1.1 Scoring Rubrics of Malaysian University English Test

A typical approach to identifying which features to use for automated scoring is by referencing the scoring rubrics provided in the corresponding test specification (Briscoe et al., 2010). Interpreting the actual meaning in the scoring rubric can pose challenges, but it enables more construct-relevant assessment than the shallow surface features employed in numerous AES systems. It improves the validity of the scores. Therefore, the proposed linguistic features in this research must be associated with the MUET scoring rubrics to justify their validity, i.e., they score what they claim to score. As provided in the “MUET Regulations, Test Specifications, Test Format, and Sample Questions” document (Malaysian Examination Council, 2014), the scoring scheme of the MUET essay encompasses five specific rubrics, namely accuracy, appropriacy, coherence and cohesion, use of language functions, and task fulfilment. Table 3.1 provides the exact descriptions of these rubrics based on the MUET Test Specifications document.

Table 3.1: Scoring Rubrics of MUET Essay

Scoring Rubrics	Test Specifications
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
Coherence and Cohesion	▪ Develop and organising ideas ▪ Using appropriate markers and linking devices ▪ Using anaphora appropriately together with cohesive devices
Use of Language Functions	▪ Defining, describing, explaining ▪ Comparing and contrasting ▪ Classifying ▪ Giving reasons ▪ Giving opinions ▪ Expressing relationships ▪ Make suggestions and recommendations ▪ Expressing agreement and disagreement ▪ Persuading ▪ Interpreting information from non-linear texts ▪ Drawing conclusion ▪ Stating & justifying points of view ▪ Presenting an argument
Task Fulfilment	▪ Presenting relevant ideas ▪ Providing adequate content ▪ Show a mature treatment of topic

This research attempts to formulate a linguistic-feature scoring scheme not only to score MUET essays but also to fulfil the specifications of MUET scoring rubrics. Since the

objective of MUET is to measure English language proficiency (Malaysian Examination Council, 2014), this thesis hypothesized that linguistic features should be one of its main assessment criteria. The formulated Automated Essay Evaluator (AEE) should achieve a certain level of accuracy since it assesses the linguistic features in the MUET essays. Based on the AEE's performance, this research will attempt to explain the relationship between the MUET scoring rubrics and the proposed linguistic feature sets in the last chapter of the thesis.

3.2 Dataset for Essay Analysis and Scoring

Like other natural language processing research, this research faces the same challenges of lacking or even absence of relevant datasets. The commercial proprietary AES vendors, such as Measurement Incorporated, Pearson Education, Education Testing Services, and Vantage Learning do not make their respective essay datasets available for public use. Referring to the literature by Ramesh and Sanampudi (2021), 90% of the AES research relied on the Automated Student Assessment Prize (ASAP) essay dataset. This ASAP essay dataset is significant as it is the only freely available dataset in a large volume. However, this ASAP dataset uses different scoring rubrics for essay assessments and lacks paragraph information, making it unsuitable for developing the AES system tailored to MUET.

To ensure the AES scoring validity for the Malaysian English test context, this research relied upon MUET datasets collected by Ng (2018) in her work of the impact of AES with feedback for essay writing. The dataset consists of 724 MUET essays collected from sixth-form students at schools within the Kuching district in Sarawak, Malaysia. The sixth-form level of education is a pre-university stage where students need to undergo the MUET examination to be admitted to the local public university and colleagues in Malaysia.

The essays were collected from five schools, namely SMK Green Road, SMK St. Thomas, SMK St. Joseph, SMK Batu Lintang, and SMK Datuk Patinggi Haji Abdul (DPHA) Gapor.

The dataset was built using authentic handwritten students' essays scored by independent essay makers who were well-versed with the MUET essay marking scheme. Furthermore, the essay markers were briefed and trained to assess the essays based on the MUET essay scoring rubric specified by the MEC. The practice is to ensure unbiased grading to increase score reliability and validity. Figure 3.1 shows the MUET essay prompt used in this research. The essay question was based on the MUET 2014 March paper, where the validity of the question is assured by the MEC. Table 3.2 shows the grade and score distribution of the essay samples used in this research. The essay scores assigned by the essay markers are holistic scores that assess the overall quality of the essay. Another aspect of scoring which comprises analytic scores to provide an in-depth analysis assessment of each scoring rubric was not available in the dataset.

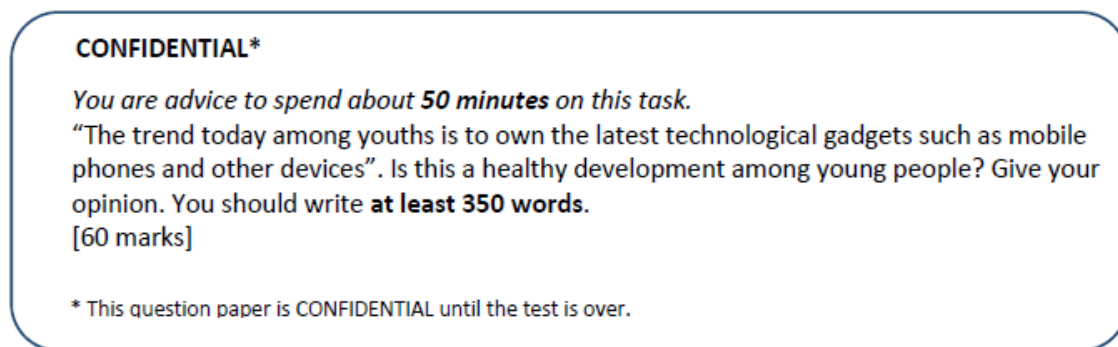


Figure 3.1: The MUET Essay Prompt used in this Research

Table 3.2: Grade and Score Distribution of MUET Essay Dataset

Grade	Score (Range)	Number of Essays	Percentage
Band 1	0 - 19	4	0.6%
Band 2	20 - 27	41	5.7%
Band 3	28 - 35	360	49.7%
Band 4	36 - 43	257	35.5%
Band 5	44 - 51	56	7.7%
Band 6	52 - 60	6	0.8%
Total Samples		724	

3.3 Linguistic Features Specification

Multi-dimensional linguistic features, namely, the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability have been investigated in this research for assessing essay scores. This section focuses mainly on outlining and elaborating the relevant linguistic features. The fundamental concept of the linguistic dimensions and the related research of their influence on essay scoring have been explained in Chapter 2, while the tools and experiments for extracting the relevant features will be further discussed in Chapter 4.

As stated above, this research encompasses multi-dimensional linguistic features to evaluate essay scores. These linguistic features demonstrate their strengths as domain-independent features for assessing general writing skills based on linguistic aspects, irrespective of the essay topics and domains. Many aspects of writing, such as language errors, vocabulary usage, sentence structure, cohesion, etc., can be assessed through

linguistic features. To exemplify how these linguistic features can be used, Table 3.3 presents the application of the eight linguistic feature sets proposed in this research to evaluate various aspects of writing skills. In summary, the linguistic features suggested in this research exhibit domain independence and provide a standardized and consistent approach to evaluating essays across different subjects and topics.

Table 3.3: Application of Linguistic Features in Assessing Essays

Linguistic Feature	Writing Skill
Surface Features	Length count of word, sentence, paragraph, etc.
Linguistic Errors	Spelling errors, grammatical errors
Text Cohesion	Text coherence of essays
Semantics	Semantic coherence of essays
Lexical Diversity	Vocabulary usage
Lexical Sophistication	Word choice, vocabulary depth
Syntactic Complexity	Sentence structure, sentence variety
Readability	Text clarity and comprehensibility

However, these linguistic features alone may have a limitation in assessing essays based on their content. The linguistic-based AES is insufficient to grasp the contextual understanding of the essay prompt, thus unable to determine how well the prompt is being addressed. For example, the AEE proposed in this research cannot assess the relevance of the essay content to the essay prompt to determine whether the essays are on-topic or off-topic. Furthermore, content-based assessment may require a deeper understanding of the subject matter. It involves the integration of domain-specific knowledge that is beyond the

scope of purely linguistic analysis. A full-fledged AES combining linguistic and content-based assessment capabilities is essential to ensure a holistic scoring of essays.

3.3.1 Taxonomy of Feature Set, Feature Categories, and Feature Indices

To facilitate a systematic approach to analysing and scoring the great numbers of linguistic features investigated in this research, this section attempts to formulate a taxonomy to organize the proposed features into a hierarchical structure. The taxonomy is described below.

- i. This research specifies eight linguistic dimensions, which are the primary investigated subjects in this work as feature sets. Feature sets are the highest tier in the structure. Figure 3.2 depicts the eight feature sets in this research, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability as discussed in Chapter 2.
- ii. Each feature set consists of one or many feature categories which are the middle tier in the structure. For example, Figure 3.3 represents the semantic feature set consisting of the global semantic and local semantic feature categories.
- iii. Each feature category can be drilled down to one or more feature indices, which are the fine-grained features defined in machine learning. Each of the feature indices is quantified by a numerical or categorical value. Figure 3.3 represents the fine-grained feature indices (e.g., Wordnet synonym count between sentences, LSA similarity between sentences) in the local semantic feature category.

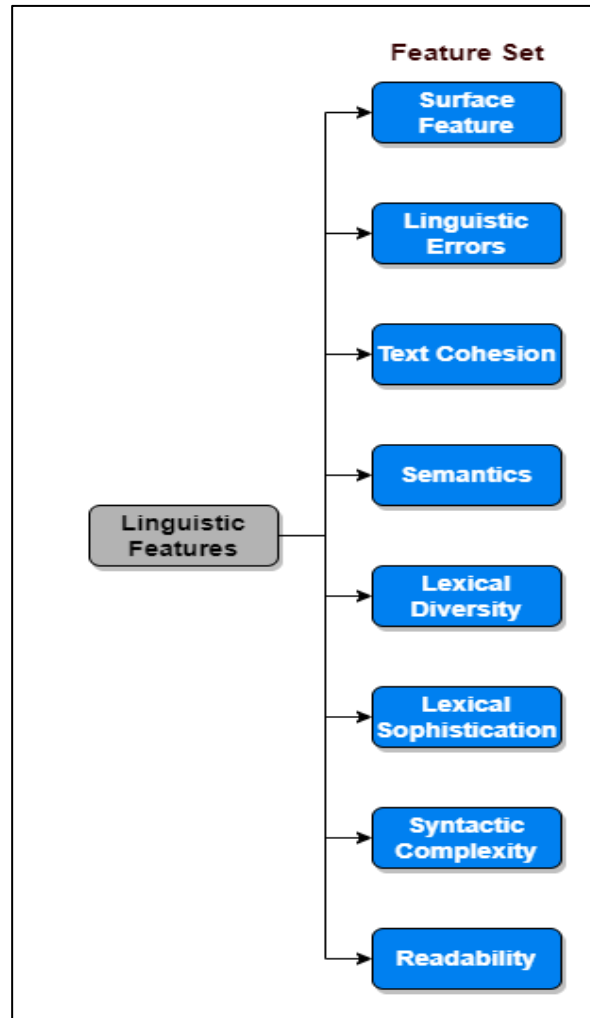


Figure 3.2: The Eight Linguistic Feature Sets in this Research

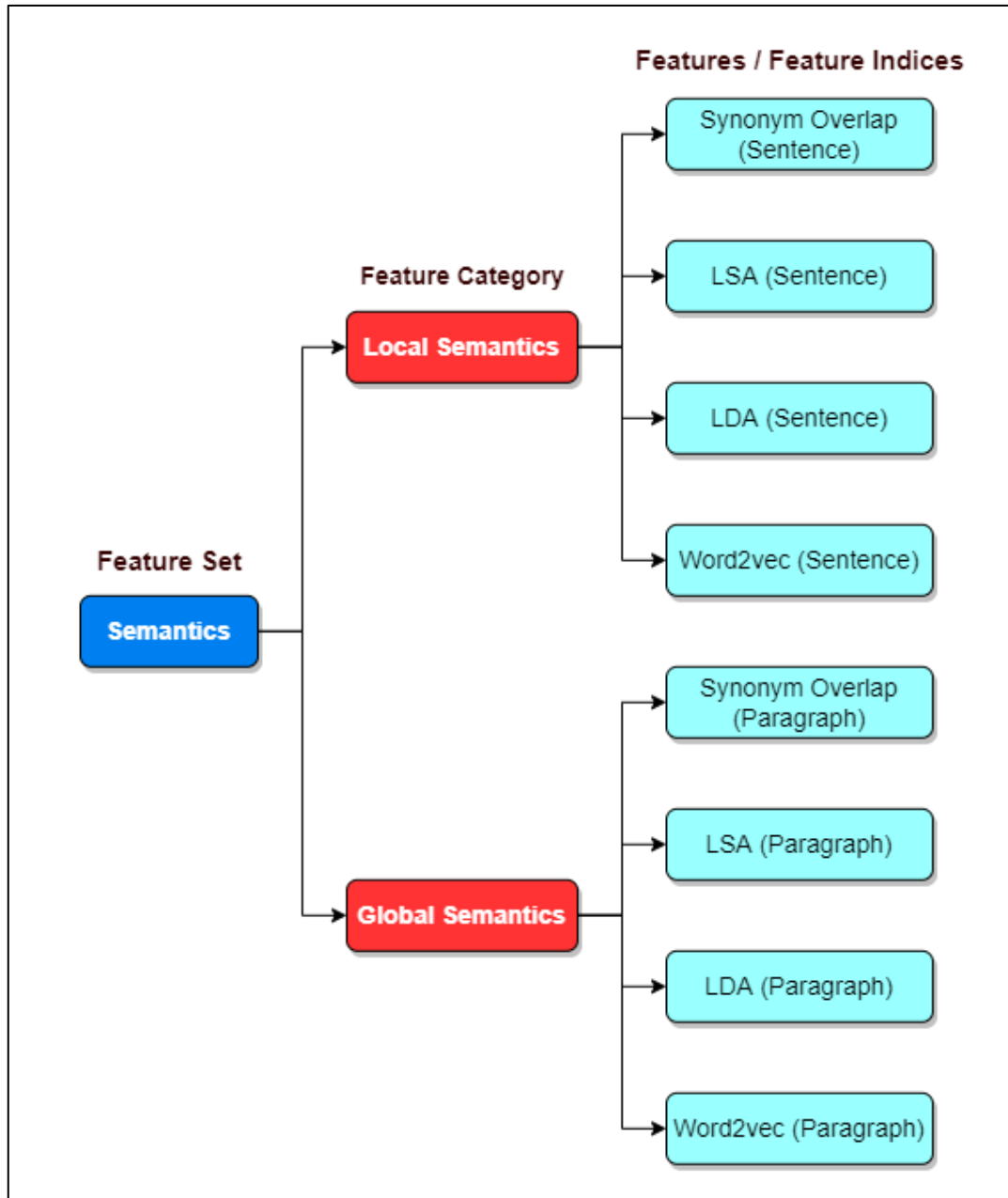


Figure 3.3: Taxonomy for Semantic Feature Set

The above-mentioned linguistic feature sets, feature categories, and feature indices are grouped based on literature in linguistics and writing assessment, except for the surface features and linguistic errors that are organized based on the author's logical understanding.

It is hoped that this proposed structure may ease other related future research using the same features.

In the following section, a detailed description of eight linguistic feature sets used in this research is presented. The relevant feature categories and feature indices are provided for surface features, linguistic errors, and readability feature sets. However, only feature categories are included in the text cohesion, semantics, lexical diversity, lexical sophistication, and syntactic complexity feature set. The feature indices in these feature sets are excluded due to their excessive length extending beyond the pages of this thesis. A comprehensive specification of all the feature sets, feature categories, and feature indices can be found in “1_Features_Specification.xlsx”, as provided in Appendix B, with a hyperlink to the directory containing the file.

3.3.2 Surface Features

Despite the controversy of whether the essay surface feature is construct-relevant to essay assessment (Fleckenstein et al., 2020), this feature is adopted as one of the primary factors for scoring essays in this research. As stated in Section 2.3.1, the author argues that surface features are important constructs to be considered for scoring essays. Unlike the other linguistic features, this surface feature may not be considered a deep linguistic feature, but a superficial feature based on the mere count of the element in the text. The features derived are mostly not grounded on any specific linguistic theory.

These surface or length-based features specify the count of the elements in the essay, ranging from the fundamental element of letters, words, sentences, and paragraphs to the count of the Parts of Speech (POS), such as the occurrence of determiner, conjunction, adjective, etc. Table 3.4 comprises the complete list of the fine-grained indices of the surface

features grouped by the feature categories of the length-based, short paragraph, parts of speech. and advanced parts of speech.

Table 3.4: Surface Feature Set

Feature Category	Feature Indices	Description
Length-Based	Word Count	This feature calculates the number of words in the text.
Length-Based	Sentence Count	This feature calculates the number of sentences in the text.
Length-Based	Paragraph Count	This feature calculates the number of paragraphs in the text.
Length-Based	Letter per Word	This feature computes the average number of letters in each word.
Length-Based	Words per Sentence	This feature computes the average number of words in each sentence.
Length-Based	Words per Paragraph	This feature computes the average number of words in each paragraph.
Length-Based	Sentences per Paragraph	This feature computes the average number of sentences in each paragraph.
Short Paragraph	Short Paragraph	This feature calculates the occurrence of short paragraphs (e.g., the number of paragraphs less than four sentences, the number of paragraphs less than five sentences).
Parts of Speech	Determiner Count	This feature calculates the incidence of determiners (e.g., a, an, the) in the text.
Parts of Speech	Demonstrative Count	This feature calculates the incidence of demonstratives (e.g., this, that) in the text.

Table 3.4 continued

Feature Category	Feature Indices	Description
Parts of Speech	Pronoun Count	This feature calculates the incidence of first-, second-, and third-person pronouns (e.g., I, me, you, he, his, they, them) in the text.
Parts of Speech	Conjunction Count	This feature calculates the incidence of conjunctions (e.g., and, for, but, because) in the text.
Parts of Speech	Connective Count	This feature calculates the incidence of connectives (e.g., however, in addition, as a result) in the text.
Parts of Speech	Negation Count	This feature calculates the incidence of negations (e.g., no, none) in the text.
Parts of Speech	Future-Word Count	This feature calculates the incidence of future-words (e.g., could, may) in the text.
Advanced Parts of Speech	Advanced Parts of Speech	This feature calculates the occurrence of advanced parts of speech (POS), i.e., the adjective, adverb, past participle tense, and present participle tense in the text.
Advanced Parts of Speech Rate	Advanced Parts of Speech Rate	This feature is derived from dividing the advanced parts of speech count by the word count.

3.3.3 Linguistic Errors

Effective writing depends on knowledge of the language system to include spelling and punctuation rules (i.e., mechanical rules), and grammatical rules (i.e., structural rules). Therefore, spelling and grammatical errors can be seen clearly as one of the traditional constructs in explaining essay quality. The count of the linguistic error that occurred in an

essay can be used as an explicit approximator for essay scoring. The higher incidence of linguistic errors in an essay reflects the inferior language ability of the writer and thus lower essay scores.

Table 3.5 specifies the fine-grained indices of the linguistic error features used in this research. Besides the raw count of spelling errors and grammatical errors, the error rates are included due to the sensitivity of the error count with the essay length. A higher score essay tends to be longer in text and may produce higher raw error counts. The error rates are adopted to mitigate this text length effect. In this research, the linguistic error features are categorized into linguistic error count and linguistic error rate.

Table 3.5: Linguistic Error Feature Set

Feature Category	Feature Indices	Description
Linguistic Error Count	Spelling Error Count	This feature calculates the number of spelling errors (mechanical errors) that occurred in the text.
Linguistic Error Count	Grammatical Error Count	This feature calculates the number of grammatical errors (structural errors) that occurred in the text.
Linguistic Error Count	Total Error Count	The sum of spelling errors and grammatical errors in the text.
Linguistic Error Rate	Spelling Error Rate	This feature is derived from dividing the spelling error count by the word count, for mitigating the text length effect.
Linguistic Error Rate	Grammatical Error Rate	This feature is derived from dividing the grammatical error count by the word count, for mitigating the text length effect.
Linguistic Error Rate	Total Error Rate	This feature is derived from dividing the total error count by the word count, for mitigating the text length effect.

3.3.4 Text Cohesion

Text cohesion is the first deep linguistic feature set included in this research. It is generally assumed that text cohesion is important to writing quality. The cohesive nature, i.e., the presence or absence of explicit cues in essays facilitates the human rater to make the logical connection between the text. Such logical connection creates coherence in the mind of the human rater and thus enhances their understanding of the information conveyed in the essay.

Following the work done by Crossley et al. (2016), this research provides various feature indices to measure the text cohesion of an essay. Based on the scope of the text segment it measures, the text cohesion features indices can be grouped into three primary feature categories below.

i. Local cohesion

These indices measure the cohesion at the sentence level, denoted by the overlapping of lexical items between adjacent sentences.

ii. Global cohesion

These indices measure the cohesion at the paragraph level, denoted by the overlapping of lexical items between adjacent paragraphs.

iii. Overall cohesion

These indices quantify the cohesion attribute in the overall text level, by measuring the percentage of a particular cohesive element in the text, such as the occurrence of connectives and the text-givenness features.

Table 3.6 details the text cohesion feature set as discussed above. Each feature category in the table can be further decomposed into its fine-grained feature indices for

measuring text cohesion. For example, the local cohesion feature category consists of feature indices for calculating lexical overlap between adjacent one sentence, adjacent two sentences, etc. Besides the “primary” feature categories, another set of alternate feature categories is specified in Table 3.6 merely to provide another perspective of feature grouping. This research focuses mainly on the primary feature categories only.

Table 3.6: Text Cohesion Feature Set

Feature Category (Primary)	Feature Category (Alternate)	Description
Local Cohesion	Lexical Overlap (Sentence)	This group of features calculates the average number of lemmas (all words, content words, function words, nouns, pronouns, arguments, verbs, adjectives, and adverbs) that are repeated between adjacent sentences. The span of the adjacent sentences can be one or two consecutive sentences.
Global Cohesion	Lexical Overlap (Paragraph)	This group of features calculates the average number of lemmas (all words, content words, function words, nouns, pronouns, arguments, verbs, adjectives, adverbs) that are repeated between adjacent paragraphs. The span of the adjacent paragraph can be one or two consecutive paragraphs.
Overall Cohesion	Connectives	These features provide the percentage of connective devices in the text, based on a predefined connectives list. The connective devices may include conjunction (and), disjunction (or), casual (because, therefore), opposition (but, however), temporal (after, before), etc.

Table 3.6 continued

Feature Category (Primary)	Feature Category (Alternate)	Description
Overall Cohesion	Givenness	Givenness indices approximate the proportion of given information in the text by examining pronoun density, pronoun-to-noun ratios, and repeated content lemmas.

3.3.5 Semantics

Semantics, as related to the study of meaning conveyed in a text is an important aspect of assessing essay quality. Existing AES systems use a variety of attributes to describe essay features (e.g., essay length, spelling and grammatical errors, lexical diversity). However, most systems lack an appropriate evaluation of text semantics (Zupanc & Bosnić, 2018). This research attempts to approach the semantic features of essays by measuring the semantic similarities within the text segment of sentences and paragraphs.

Based on the work by Crossley et al. (2019b), this research specifies numerous indices to measure the semantic features of an essay. Four types of semantic similarity feature grounded on semantic analysis theories, namely, the WordNet Synsets (Miller, 1995), Latent Semantic Analysis (LSA; Landauer et al., 1998), Latent Dirichlet Allocation (LDA; Blei et al., 2003), and Word2vec (Mikolov et al., 2013a, 2013b) have been extracted for measuring semantic similarity within an essay.

Similar to the text cohesion features, the semantic similarity indices can be grouped into two primary feature categories based on the context of the text segment they measure. These two primary categories are listed below.

i. Local semantics

These indices measure the semantic features at the sentence level, denoted by the average count of overlapping WordNet synonyms and the average similarity scores computed by the respective LSA, LDA, and Word2vec between adjacent sentences.

ii. Global semantics

These indices measure the semantic features at the paragraph segment, denoted by the average count of overlapping WordNet synonyms and the average similarity scores computed by the respective LSA, LDA, and Word2vec between adjacent paragraphs.

Table 3.7 presents the feature categories in the semantic feature set. Besides the “primary” feature categories, another set of alternate feature categories is specified based on the semantic analysis technique used to compute the feature indices. The LSA, LDA, and Word2vec models were trained on the newspaper and magazine sections of the Corpus of Contemporary American English (COCA; Davies, 2010).

Table 3.7: Semantic Feature Set

Feature Category (Primary)	Feature Category (Alternate)	Description
Local Semantics	WordNet Synonym Overlap (Sentence)	This group of features calculates the similarity between sentences by counting the mean overlapping of synonyms (nouns and verbs) using the WordNet Synonym Set (Synset).
Local Semantics	LSA Cosine Similarity (Sentence)	These LSA indices calculate the mean cosine similarity between neighbouring sentences. The neighbouring text segments span one or two adjacent sentences.

Table 3.7 continued

Feature Category (Primary)	Feature Category (Alternate)	Description
Local Semantics	LDA Divergence (Sentence)	These indices provide the mean similarity score between neighbouring sentences based on LDA. The neighbouring text segment comprises one or two adjacent sentences.
Local Semantics	Word2vec Similarity (Sentence)	These indices provide the mean similarity score between neighbouring sentences based on Word2vec. The neighbouring text segment comprises one or two adjacent sentences.
Global Semantics	WordNet Synonym Overlap (Paragraph)	This group of features calculates the similarity between paragraphs by counting the mean overlapping of synonyms (nouns and verbs) using the WordNet Synonym Set (Synset).
Global Semantics	LSA Cosine Similarity (Paragraph)	These LSA indices calculate the mean cosine similarity between neighbouring paragraphs. The neighbouring text segments span one or two adjacent paragraphs.
Global Semantics	LDA Divergence (Paragraph)	These indices provide the mean similarity score between neighbouring paragraphs based on LDA. The neighbouring text segment comprises one or two adjacent paragraphs.
Global Semantics	Word2vec Similarity (Paragraph)	These indices provide the mean similarity score between neighbouring paragraphs based on Word2vec. The neighbouring text segment comprises one or two adjacent paragraphs.

3.3.6 Lexical Diversity

Lexical features are perhaps the most valid linguistic construct for scoring essays. An essay with higher scores shall render both the depth and breadth of its lexical items. This research distinguishes two sets of lexical dimensions, namely, lexical diversity and lexical sophistication. Lexical diversity is often used as the measure of lexical richness/complexity in L2 writing assessment (Zenker & Kyle, 2021). The underlying rationale is that the lexical diversity score represents a writer's overall lexical knowledge - advanced learners use a wide range of lexical items.

The most basic and widely known lexical diversity measure is the Type-Token Ratio (TTR: Johnson, 1944), which is simply the number of unique words (types) divided by the number of running words (tokens). However, the TTR value is sensitive to text length, with longer text tending to produce a lower lexical diversity score. This text length effect is a serious concern, especially in the context of automated essay scoring because the essay lengths produced by students tend to be varied, even in a timed test environment. Several revised lexical diversity indices have been proposed over the years to mitigate this text length effect. With reference to the literature by Zenker and Kyle (2021), Table 3.8 outlines the lexical diversity feature set, accompanied by citations for each feature category. The table comprises several formula descriptions, ranging from the simple TTR calculation to the advanced TTR formula based on mathematical transformation such as the HDD, MTLD, MTLD-MA-BI, etc. Each of the feature categories consists of multiple feature indices. For example, the Simple TTR feature categories consist of TTR for all words, TTR for content words, and TTR for function words.

Table 3.8: Lexical Diversity Feature Set

Feature Category	Description	Reference
Lexical Raw Count	This group of features calculates the number of lexical items in the text, including the number of content word tokens, function word tokens, total word tokens, content word types, function word types, and total word types.	N/A
Lexical Density	Lexical Density denotes the percentage of text (token or type) that is content words (nouns, verbs, adjectives, adverbs).	(Crossley, 2020)
Simple TTR	Simple TTR is calculated as the number of unique words (types) divided by the number of running words (tokens) in the text.	(Johnson, 1944)
Root TTR	Root TTR is computed as the number of types divided by the square root of the number of tokens.	(Guiraud, 1959)
Log TTR	Log TTR is derived by dividing the logarithm of the number of word types by the logarithm of the number of word tokens.	(Chotlos, 1944; Herdan, 1960)
Maas' Index	Maas' Index is a more complex transformation of TTR, as denoted by Equation 3.1 below. $Mass = \frac{Log(ntokens) - Log(ntypes)}{Log(ntokens)^2}$ Equation 3.1	(Mass, 1972)
Moving-Average TTR (MATTR)	MATTR is calculated based on the moving average approach, whereby the mean TTR values of all text segments (with predefined length) are obtained. For example, TTR is computed on tokens 1–50, 2–51, 3–52, etc., and the corresponding TTR values are averaged to generate the final MATTR score.	(Covington & McFall, 2010)

Table 3.8 continued

Feature Category	Description	Reference
Hypergeometric Distribution Diversity Index (HDD)	For every word type in a text, HDD uses the hypergeometric distribution to compute the probability of encountering one of its tokens in a random sample of 42 tokens. These probabilities are then summed up to generate the final HDD value.	(McCarthy & Jarvis, 2007)
Measure of Textual Lexical Diversity (MTLD)	MTLD is computed based on the average number of tokens it takes to reach a predetermined TTR value (e.g., 0.66). MTLD moves through a text and computes the TTR on incremental segments (one token, two tokens, three tokens, etc.) until reaching a segment of at least 10 tokens that meets the predefined TTR value. Then, it proceeds to the adjacent word and repeats the process. All the corresponding segments (referred to as factors) with their lengths are recorded. The remaining segment whereby the predefined TTR cannot be obtained is counted as a partial factor. All these factor lengths are then averaged to produce the final MTLD score.	(McCarthy & Jarvis, 2010)
Moving-Average Bidirectional MTLD (MTLD-MA-BI)	MTLD-MA-BI is the revised version of MTLD based on the moving average. In addition to moving forward to obtain the MTLD factor length, MTLD-MA-BI moves backward starting with the last word in the text when an MTLD factor cannot be completed. The final MTLD value comprises the average factor length for all factors.	(McCarthy & Jarvis, 2010)
Moving-Average Wrapped MTLD (MTLD-MA-Wrap)	Like MTLD-MA-BI, MTLD-MA-Wrap takes a moving-average approach to calculate the TTR value. However, instead of navigating the text bi-	(McCarthy & Jarvis, 2010)

Table 3.8 continued

Feature Category	Description	Reference
	directionally, MTLD-MA-Wrap avoids partial factors by looping back around to the beginning of the text.	

3.3.7 Lexical Sophistication

Lexical sophistication is perhaps another valid linguistic construct to assess the quality of an essay. The presence of sophisticated words in essays indicates superior lexical knowledge and thus advanced writing skills. Word frequency features are probably the most common measure of lexical sophistication. The underlying assumption is that infrequent words are usually regarded as sophisticated.

In addition to the word frequency, lexical sophistication research have evolved over the years such that sophistication can encompass a vast number of other word properties. As a result, lexical sophistication tends to provide the largest group of features in assessing essay quality. Based on the work by Kyle and Crossley (2015) and Kyle et al. (2018), Table 3.9 details multiple aspects of lexical sophistication features used in this research. Besides the basic categories of word frequency and word range, this research includes various advanced categories derived from psycholinguistic and cognitive science studies (e.g., psycholinguistic norms, word recognition norms, and word neighbour information).

Table 3.9: Lexical Sophistication Feature Set

Feature Category	Description
Word Frequency	Word frequency refers to the number of times a word occurs in a reference corpus. The word frequency of a target text is derived by dividing the sum of the frequency score of every word by the total number of words with frequency scores. This group of features provides manifold word frequency indices based on different reference corpora, types of words (all words, content words, function words), raw-word/lemma, word token/word type, and raw-count/logarithm-based.
Word Range	Word range refers to the number of texts in a reference corpus in which a particular word occurs. For a particular text, the mean word range is calculated by dividing the sum of the word range score of every word by the total number of words with word range scores. Like the word frequency, various word range indices are provided in this feature category.
N-gram Frequency	N-gram frequency resembles word frequency, except the target unit is a combination of n number of words (e.g., bigram and trigram) instead of a single word.
N-gram Range	N-gram range resembles word range, except the target unit is a combination of n number of words (e.g., bigram and trigram) instead of a single word.
N-gram Association Strength	N-gram Association Strength measures the conditional probability that words will occur together. For example, the bigram “optimistic about” has a stronger association strength than the bigram “in the”.
Character Bigram Frequency	A character bigram is an association of two characters (e.g., "TH", "HE", "IN", "AN"). Their frequency of appearance makes it possible to obtain information from the text.
Academic Language	Academic language comprises words that occur frequently in academic texts. The indices measure the raw count or proportion of academic words in the target text based on several academic corpora.

Table 3.9 continued

Feature Category	Description
Age of Acquisition / Exposure	Age of acquisition (AoA) indices are derived based on human judgments of the age at which a particular word is learned. These indices of a particular text are computed by dividing the sum of AoA scores of every word by the number of words with AoA scores.
Contextual Distinctness	Contextual distinctiveness measures the diversity of contexts in which a word is encountered. These metrics are operationalized through free association tasks (tasks whereby participants are given a stimulus word and asked to produce the associated words that come to mind) or corpus-driven approach (probability of a word co-occurring with the other words in the corpus). Words with a greater number of associations or co-occurrence with other words are assumed to be less contextually distinct and thus less sophisticated.
Word Neighbour Information	Word Neighbourhood refers to the number of words that share orthographic, phonographic, and/or phonological similarities with a particular word (Peereman & Content, 1997). The number of a word's neighbourhoods has been shown to affect the word-naming and recognition tasks (Adelman & Brown, 2007).
Word Recognition Norms	Word recognition features provide several indices related to the average response latencies for a target word in lexical-decision tasks (tasks whereby participants need to decide whether a given word is a real English word) or word-naming tasks (tasks whereby participants begin to read a word aloud). These observed response latencies may reflect the ease or difficulty of processing a given word (Balota et al., 2004) and thus determine its sophistication level.
Hypernymy (Semantic Network)	Hypernymy refers to the number of superordinate terms a particular word has. A word such as "dog" has a more superordinate term than "animal" and thus yields a higher hypernymy score. Quality writing usually involves more specific lexical items, i.e., words with a higher hypernymy score (Guo et al., 2013). Several mean hypernymy scores

Table 3.9 continued

Feature Category	Description
	(nouns, verbs) of the target text are employed to measure these features.
Polysemy (Semantic Network)	Polysemy refers to the number of related senses or co-existing meanings a particular word has. A word such as "make" has more senses than "construct" and thus produces a higher polysemy score. Quality writing comprises words with less sense, or lower polysemy scores (Guo et al., 2013). Various mean polysemy scores based on the content words, nouns, verbs, adjectives, and adverbs of the target text are adopted to measure these features.
Psycholinguistic Norms	Psycholinguistic norms are the word information features derived from the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981). In this measurement, every word in the target text is scored from four psycholinguistic properties, namely: i. Familiarity: How familiar does a word seem to an adult? ii. Concreteness: How concrete or non-abstract a word is? iii. Imageability: How easy is it to construct a mental image of the word? iv. Meaningfulness: How high is the association of the word to other words?

Unlike the other linguistic features which compute their indices value from the target text directly, this lexical sophistication feature set relied upon multiple publicly available corpora as the reference to calculate its indices. Table 3.10 lists all the corpora employed to calculate the lexical sophistication indices in this research. The second column labelled "Application Scope" refers to the feature category of lexical sophistication within which the corpora have been used to compute lexical sophistication indices in this research.

Table 3.10: Corpora used for Lexical Sophistication Features Indices Calculation

Corpus / Database	Application Scope	Reference
Academic Corpus	Academic Language	(Coxhead, 2000, 2011)
Age of acquisition norms collection	Age of Acquisition/Exposure	(Kuperman et al., 2012)
British National Corpus (BNC)	Word Frequency, Word Range, N-gram Frequency, Contextual Distinctiveness	(BNC Consortium, 2007)
Brown Corpus	Word Frequency, Word Range	(Kučera & Francis, 1967)
Concreteness norms	Psycholinguistic Norms	(Brysbaert et al., 2014)
Corpus of Contemporary American English (COCA)	Word Frequency, Word Range, N-gram Frequency, N-gram Range, N-gram Association Strength	(Davies, 2010)
Four Corpora	Academic Language	(Simpson-Vlach & Ellis, 2010)
Hyperspace Analogue to Language (HAL)	Word Frequency, Character Bigram Frequency, Word Neighbour Information,	(Balota et al., 2007)
London-Lund Corpus	Word Frequency	(G. D. A. Brown, 1984)
Lorge's corpus	Word Frequency	(Thorndike & Lorge, 1944)
Medical Research Council (MRC) Database	Psycholinguistic Norms	(Coltheart, 1981)
SUBTLEXus	Word Frequency, Word Range	(Brysbaert & New, 2009)
Touchstone Applied Sciences Associates (TAS)	Age of Acquisition/Exposure, Contextual Distinctiveness	(Landauer et al., 1998; Dascalu et al., 2016)

Table 3.10 continued

Corpus / Database	Application Scope	Reference
TOEFL11 Corpus	Word Frequency, Word Range, N-gram Frequency, N-gram Range	(Blanchard et al., 2013)
WordNet	Hypernymy and Polysemy	(Miller, 1995)

3.3.8 Syntactic Complexity

Another common approach to analyse the quality of an essay is to examine the structural properties of the writing. Syntactic complexity suitably fits into this criterion by assessing the variety and sophisticated syntactic form produced in the essay. The underlying argument is that the ability to produce complex syntactic structures is an indicator of advanced writing skills.

Syntactic complexity reflects the difficulty of syntactic structures on a formal, grammatical basis. It mainly entails quantitative measures to pinpoint syntactic difficulty. With reference to the work by Kyle (2016) and Lu (2010), this research approaches the syntactic complexity assessment of an essay from four categories described below.

i. Phrasal Complexity (Noun Phrase Complexity)

Phrasal complexity (or noun phrase complexity in this context) involves the frequency count of dependents in a noun phrase. Different phrase types and dependent types are defined to facilitate the computation of these indices.

ii. Clausal Complexity

Clausal complexity measures the occurrence of different structures/types within the context of a clause. Different structures (e.g., adjective complement and conjunction) are defined in this category.

iii. Syntactic Sophistication

Syntactic sophistication refers to the usage-based perspective to measure complexity based on corpus data concerning frequencies of the construction structure. Syntactic sophistication adopts the same concepts as lexical sophistication - infrequent syntactic structures are considered more sophisticated and thus indicate a superior level of writing skills. Based on the work by Kyle (2016), this research focuses on Verb-Argument Constructions (VAC) which consists of a verb slot and the arguments it takes, as the unit measurement of syntactic sophistication. The statistic of this VAC usage is derived from the COCA corpus.

iv. Sentential Complexity

Compared to the phrasal and clausal complexity, sentential complexity (Bulté & Housen, 2014) is the simplest metric, which focuses on measuring the length of the production unit, i.e., the length of the clause, T-unit, and sentence. Fourteen sentential complexity indices based on the Syntactic Complexity Analyzer (SCA) constructed by Lu (2010) have been used in this research.

Table 3.11 specifies the phrasal complexity, clausal complexity, and syntactic sophistication features categories. Due to the vast number of the involved complexity indices, the complexity feature categories listed in the table can be further decomposed into fine-grained indices for measuring syntactic complexity. Another category of syntactic complexity, i.e., the sentential complexity is illustrated by the 14 fine-grained indices specified in Table 3.12.

Table 3.11: Syntactic Complexity Feature Set (Phrasal Complexity, Clausal Complexity and Syntactic Sophistication)

Feature Category	Description
Phrasal Complexity (Noun Phrase Complexity)	Phrasal complexity measures the number of dependents in a noun phrase. Several phrase types (e.g., nominal subject, direct object, agent), and dependent types (e.g., determiner, adjective modifier, preposition phrase) are identified for generating these phrase complexity indices. Three types of phrasal complexity indices are computed, namely: i. The average number of dependents in a particular phrase type (e.g., number of dependents per nominal subject) ii. The average occurrence of a particular dependent type regardless of phrase type (e.g., number of determiners) iii. The average occurrence of a particular dependent type in a particular phrase type (e.g., number of adjective modifiers per nominal subject)
Clausal Complexity	Clause complexity measures the occurrence of a particular structure in a clause. Numerous fine-grained structures (e.g., adjective complement, modal auxiliary, negation, conjunction, etc.) are defined and used for computing the clause complexity indices, such as adjective complement per clause, conjunction per clause, etc.
Syntactic Sophistication	Syntactic Sophistication measures the syntactic structures in a text with reference to their usage statistic in the COCA corpus. These feature categories focus on the computing of the VAC-related indices, which provide manifold indices of VAC Frequency, VAC Range, VAC Association Strength, and VAC TTR.

Table 3.12: Syntactic Complexity Feature Set (Sentential Complexity)

Feature Indices	Description
Length of Production Unit	
Mean length of clause (MLC)	Number of words per clause
Mean length of sentence (MLS)	Number of words per sentence
Mean length of T-unit (MLT)	Number of words per T-unit
Sentence Complexity	
Sentence complexity ratio (C/S)	Number of clauses per sentence
Subordination	
T-unit complexity ratio (C/T)	Number of clauses per T-unit
Complex T-unit ratio (CT/T)	Number of complex T-units divided by T-units
Dependent clause ratio (DC/C)	Number of dependent clauses per clause
Dependent clauses per T-unit (DC/T)	Number of dependent clauses per T-unit
Coordination	
Coordinate phrases per clause (CP/C)	Number of coordinate phrases per clause
Coordinate phrases per T-unit (CP/T)	Number of coordinate phrases per T-unit
Sentence coordination ratio (T/S)	Number of T-units per sentence
Particular Structures	
Complex nominals per clause (CN/C)	Number of complex nominals per clause
Complex nominals per T-unit (CN/T)	Number of complex nominals per T-unit
Verb phrases per T-unit (VP/T)	Number of verb phrases per T-unit

3.3.9 Readability

Despite the lack of literature investigating the relationship of readability with essay scoring, readability metrics may be adopted as a novel idea in assessing essays. The

readability formula, originally designed to assess the difficulty of reading material, can be leveraged to measure the quality of a writing text. This thesis claims that readability might be a valid construct to score essays with the two arguments below.

- i. Similar to matching appropriate reading materials to different "grades" of readers, readability can be used as a proxy to match the writing quality in essays to the corresponding grades or scores. A good quality essay shall be readable and understandable (i.e., the readability measurement) by the human raters, who are the readers in this context.
- ii. Unlike other linguistic features, readability is a "composite" attribute in which the readability formula may consist of variables related to multiple linguistic feature sets, ranging from the basic length-based features to advanced features such as text cohesion, semantics, lexical sophistication, syntactic complexity, etc. These multi-dimensional linguistic features incorporated in the readability metrics can facilitate a more comprehensive assessment of essay quality and thus might be a good predictor of essay scores.

In this research, several readability formulas have been reviewed and used in the experiment for selecting the appropriate readability metrics to formulate the AEE. The research includes the classic readability metrics and the latest readability formulas based on natural language processing and crowdsourcing. Table 3.13 presents the formulas of the readability indices, focusing on describing the linguistic features (e.g., lexical diversity and lexical sophistication) used in each formula. Furthermore, citations for the corresponding readability indices are provided in the last column. The readability feature indices can be categorized into three categories, namely Classic Readability, Crowdsourced Readability,

and Coh-Metrix Readability. The exact formula to compute these readability indices is attached in Appendix A of this thesis.

Table 3.13: Readability Feature Set

Feature Category	Feature Indices	Description	References
Classic Readability	Flesch Reading Ease	Flesch Reading Ease was one of the earliest readability formulas proposed by Flesch in 1948. The formula was derived based on superficial length-based features, i.e., the number of words per sentence (sentence length) and the number of syllables per word (word length).	(Flesch, 1948)
Classic Readability	Flesch Kincaid Grade Level	Flesch Kincaid Grade Level was the readability formula provided by Kincaid et al. in 1975. Like the Flesch Reading Ease, Flesch Kincaid Grade Level was calculated from the number of words per sentence and the number of syllables per word, but with different constant and coefficients.	(Kincaid et al., 1975)
Classic Readability	Automated Readability Index	Automated Readability Index was the readability formula provided by Senter et al. in 1967. This formula computes the text readability based on the number of words per sentence and the number of characters per word. The underlying idea is that the more characters that are present in a word, the harder it is to read.	(Senter & Smith, 1967)
Classic Readability	SMOG Grading	SMOG (an acronym for Simple Measure of Gobbledygook) Grading was formulated by Henry Mc Laughlin in 1969.	(Mc Laughlin, 1969)

Table 3.13 continued

Feature Category	Feature Indices	Description	References
		The formula was derived from the square root of the number of three or more syllables words per 30 sentences.	
Classic Readability	New Dale-Chall Readability	The New Dale-Chall Readability (Chall & Dale, 1995) was an improved version of the original Dale-Chall Readability Formula (Dale & Chall, 1948). Instead of merely based on sentence length, the formula relied upon the percentage of difficult words, which are not in the list of 3,000 familiar words compiled by Chall and Dale.	(Chall & Dale, 1995)
Crowdsourced Readability	Crowdsourced Algorithm of Reading Comprehension (CAREC)	CAREC was formulated based on multi-dimensional linguistic features of lexical sophistication (e.g., age of acquisition, bigram range, and word frequency), text cohesion (e.g., temporal connectives), lexical diversity (e.g., type-token-ratio), and sentiments (e.g., positive adjective scores). The formula was derived from measuring text comprehension collected by the crowdsourcing technique.	(Crossley et al., 2019c)
Crowdsourced Readability	Crowdsourced Algorithm of Reading Speed (CARES)	CARES was formulated based on multi-dimensional linguistic features of lexical sophistication (e.g., word naming response time and word concreteness score), lexical diversity (e.g., number of content lemmas and number of function words), syntactic complexity (e.g., complex nominals per T-	(Crossley et al., 2019c)

Table 3.13 continued

Feature Category	Feature Indices	Description	References
		unit), and length-based features (e.g., number of sentences per paragraph). The formula was derived from measuring reading speed collected by the crowdsourcing technique.	
Coh-Metrix Readability	Coh-Metrix L2 Readability Index (CML2RI)	CML2RI was the readability formula provided by the Coh-Metrix tool based on psycholinguistic and cognitive models of reading. The formula was constructed by the indices of the number of sentences, content word frequency, and proportion of noun and pronoun lemma.	(Crossley & McNamara, 2008; Crossley et al., 2011a)

3.4 Feature Selection of Linguistic Features

As specified in Section 3.3, there are a large number of linguistic features proposed to score essays. To handle this large number of features effectively, feature selection is highly required to reduce the dimensionality of features by identifying the most informative and relevant features. The selection of optimal features is significant to enhance the scoring performance of the proposed AEE and reduce the required computational resources to construct the scoring model. Furthermore, it facilitates a clear explanation and interpretation of the scoring criteria, which is one of the main objectives of this research. This section describes two feature selection approaches used in this research to select the optimal essay features for automated essay scoring.

3.4.1 Correlation Rank Feature Selection

This research proposed Correlation Rank Feature Selection as the feature selection approach to identify optimal linguistic features for scoring essays. This Correlation Rank Feature Selection relies on the filter method. The filter method (Chandrashekar & Sahin, 2014) uses the feature scores in various statistical tests (e.g., correlation, ANOVA, Chi-Square) to determine the relevancy of the features with the dependent variable. The features are then subsequently ranked based on their score, and a threshold is defined to filter out the irrelevant low-score features. This filter method operates independently of machine learning algorithms, ensuring its robustness against overfitting, and being computationally efficient.

The Correlation Rank Feature Selection was chosen as one of the feature selection approaches in this research due to several considerations, as detailed by the following rationales:

- i. Intuitive: The Correlation technique is easy to understand and interpret. Correlation values can be easily represented through plots to visualize the relationship between essay features and essay scores.
- ii. Robustness: Correlation-based feature selection exhibits robustness against overfitting. It operates on the feature values, independent of any machine learning algorithm.
- iii. Efficiency: Calculating correlations is computationally efficient, making it suitable for handling large datasets with high dimensions.
- iv. Usage in AES Research: Correlation-based feature selection has been applied in AES research, as evidenced by the works of Madala et al. (2018), and Kumar and Boulanger (2020).

The selection of optimal linguistic features for scoring essays in this Correlation Rank Feature Selection approach involves the following steps:

i. Calculation of correlation coefficient

The correlation coefficient value between each essay feature and essay score was computed. Spearman's rank correlation coefficient (Spearman, 1907) and Kendall's rank correlation coefficient (Kendall, 1938) were adopted as the correlation coefficient in this research. Further details of these Spearman's and Kendall's correlation coefficients are described in Section 4.3.

ii. Ranking of essay features

The essay features were then ranked according to their absolute correlation coefficient value. Features with higher correlation coefficients indicate a stronger association with the essay score and thus will be ranked higher and prioritized to be selected first.

iii. Defining correlation thresholds

In this step, correlation thresholds were defined to determine the significance of the correlation between each feature and the essay scores. This research proposed various correlation schemes, including COR-03, COR-05, and COR-07, each based on different correlation thresholds. For instance, under the COR-03 scheme, all essay features with correlation coefficient values equal to or greater than 0.3 would be selected. The configuration and description of these correlation thresholds can be found in Section 4.4.1.

3.4.2 Minimum Redundancy Maximum Relevance (MRMR) Feature Selection

One of the challenges in feature selection is the multicollinearity between the extracted features, where multiple features are correlated and redundant. The analysis and

processing of data under such conditions are time-consuming and tends to produce inaccurate prediction model, especially in the case of high-dimensional features. It is of utmost importance to balance the number of selected features. Removing too many might result in the loss of information while keeping too many might produce undesired results with expensive computation costs.

In the context of essay scoring, there may exist collinearity relationships between the essay features. For example, the word count and the sentence count may be highly correlated because a lengthy essay in word count usually indicates a higher number of sentences. To address this issue, i.e., to minimize the redundancy and maximize the relevance of feature selection, this research relied upon the Minimum Redundancy Maximum Relevance (MRMR) Feature Selection (Ding & Peng, 2005) as another feature selection approach to select the optimal linguistic features for scoring essays. The MRMR algorithm attempts to optimize two main criteria, i.e., maximum relevance and minimum redundancy. Maximum relevance evaluates how well each feature is individually correlated with the target variable. Minimum redundancy measures the multicollinearity between selected features, ensuring that the selected features are distinct and not redundant. It works by performing a sequential forward search so that it iteratively adds features that are most relevant to the label and least redundant to previously selected features.

This research proposed several MRMR schemes, including MRMR-10, MRMR-25, MRMR-50, and MRMR-100 based on the desired number of linguistic features to be selected. For example, the MRMR-10 scheme predefines the selection of 10 essay features with minimum redundancy and maximum relevance, whereas the MRMR-25 scheme

specifies the selection of 25 features with minimum redundancy and maximum relevance. The configuration of these MRMR schemes is detailed in Section 4.4.2.

3.5 Chapter Summary

This research reviewed comprehensive and multi-dimensional linguistic features to analyse and score essays. To facilitate the systematic investigation, analysis, and scoring of these linguistic features, this research attempted to organize them into a taxonomy of feature sets, feature categories, and feature indices. This chapter elaborates on the proposed eight linguistic feature sets, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability. The specification of each feature set is presented with their corresponding feature categories and feature indices. Table 3.14 summarizes the complete list of eight feature sets and 43 feature categories specified in this research. Furthermore, this chapter specifies two feature selection approaches, namely the Correlation Rank and MRMR Feature Selection Scheme to select the optimal essay features for automated essay scoring in this research.

Table 3.14: The Complete List of Linguistic Feature Sets and Categories Investigated in this Research

Surface Feature	Linguistic Error	Text Cohesion	Semantics	Lexical Diversity	Lexical Sophistication	Syntactic Complexity	Readability
Length-Based	Linguistic Error Count	Local Cohesion	Local Semantics	Lexical Raw Count	Word Frequency	Noun Phrase Complexity	Classic Readability
Short Paragraph	Linguistic Error Rate	Global Cohesion	Global Semantics	Lexical Density	Word Range	Clausal Complexity	Crowdsourced Readability
Parts of Speech		Overall Cohesion		Simple TTR	N-gram Frequency	Syntactic Sophistication	Coh-Metrix Readability
Advanced Parts of Speech				Root TTR	N-gram Range	Sentential Complexity	
				Log TTR	N-gram Association Strength		
				Maas' Index	Character Bigram Frequency		
				MATTR	Academic Language		
				HDD	Age of Acquisition / Exposure		
				MTLD	Contextual Distinctiveness		
				MTLD-MA-BI	Word Neighbor Information		
				MTLD-MA-Wrap	Word Recognition Norms		
					Psycholinguistic Norms		
					Hypernymy		
					Polysemy		

CHAPTER 4

EXPERIMENTS

This chapter discusses the experimental setup for the analysis and scoring of MUET essays in this research. Section 4.1 illustrates an experiment framework consisting of the five sequential processes to conduct this experiment. All the subsequent sections in this chapter are dedicated to describing the processes. Section 4.2 describes the linguistic feature extraction from the MUET dataset, while Section 4.3 discusses the feature evaluation using Spearman's and Kendall's rank coefficients. Section 4.4 specifies the Correlation Rank and MRMR Feature Selection used in this research. Section 4.5 discusses the automated essay scoring experiment in this research, describing the six machine learning models, evaluation of feature selection schemes, and evaluation metrics. The section is followed by a brief discussion of the result presentation format via the boxplot diagram in Section 4.6 and a chapter summary in Section 4.7.

4.1 Experiment Framework

The experiment in this research was set up with two primary purposes, i.e., to perform a detailed analysis of the MUET essay features and attempt machine learning models to evaluate the selected linguistic features in scoring essays. Figure 4.1 illustrates the experiment framework of this research, summarized as processes below.

i. Feature Extraction

The linguistic features of the MUET dataset were extracted by using multiple linguistic analysis tools and Python programming.

ii. Feature Evaluation

The extracted features were evaluated by correlation statistic tests of the essay features and essay scores.

iii. Feature Selection

The optimal essay features were selected based on different feature selection schemes.

iv. Automated Essay Scoring using Linguistic Features

The selected essay features were evaluated using various supervised learning classifiers to predict the essay grade.

v. Findings and Results

The analysis and scoring results were presented via boxplot diagrams and other charts in Chapter 5.

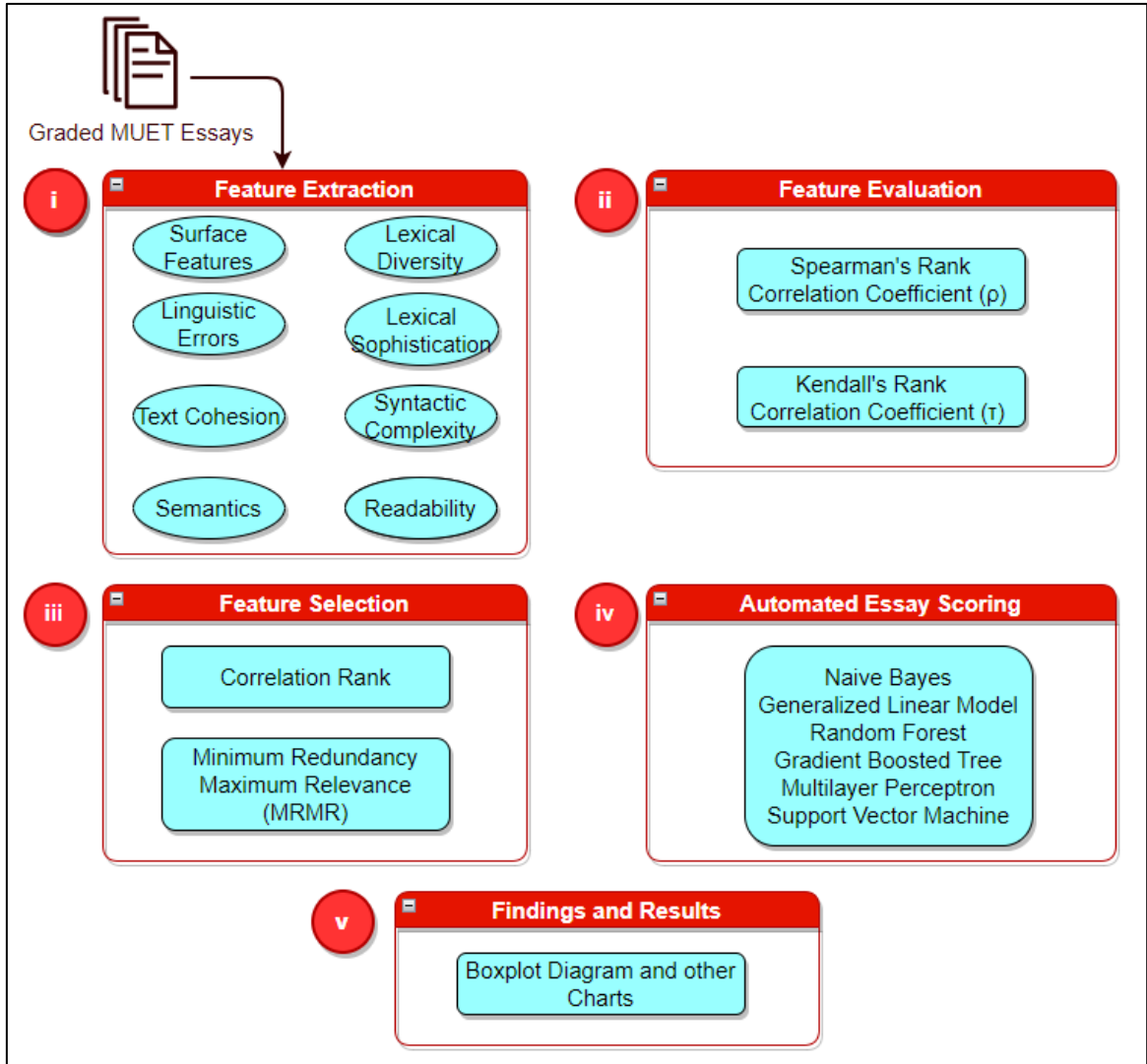


Figure 4.1: Experiment Framework for Essay Analysis and Scoring

4.2 Feature Extraction

This is the first process of the experiment. Comprehensive linguistic features were extracted from the 724 sets of the MUET essay. Multiple linguistic analysis tools (Kyle, n.d.), and Python Natural Language Toolkit (NLTK, n.d.) were used to extract the linguistic features. Table 4.1 specifies all the linguistic analysis tools used to extract different linguistic feature sets employed in this experiment. The linguistic analysis tools provide graphical user interfaces to extract features from the essays stored in plain text files encoded in UTF-8

format and produce the feature scores in comma-separated value (.csv) spreadsheets. Overall, a total of 1,709 linguistic feature indices were extracted, which is far beyond other linguistic analysis tools, such as 108 feature indices extracted by Coh-Metrix in Table 2.1.

Table 4.1: Linguistic Analysis Tools

Feature Set	Tools	Feature Indices
Surface Features	Simple Natural Language Processing Tool (SiNLP), Python Natural Language Toolkit (NLTK)	27
Linguistic Errors	Grammar and Mechanics Error Tool (GAMET)	6
Text Cohesion	Tool for the Automatic Analysis of Cohesion (TAACO)	137
Semantics	Tool for the Automatic Analysis of Cohesion (TAACO)	16
Lexical Diversity	Tool for the Automatic Analysis of Lexical Diversity (TAALED)	35
Lexical Sophistication	Tool for the Automatic Analysis of Lexical Sophistication (TAALES)	1,103
Syntactic Complexity	Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC)	377
Readability	Automatic Readability Tool for English (ARTE)	8
	Total Feature Indices Extracted	1,709

4.3 Feature Evaluation

Based on the 1,709 linguistic feature indices extracted (as specified in Table 4.1), correlation analysis has been performed to examine the strength of the association between the feature value with the essay scores. The significant essay features shall render a higher magnitude of a correlation value with the essay score. Instead of relying on the commonly

used Pearson's product-moment correlation coefficient (Pearson, 1895), this research adopts Spearman's rank correlation coefficient (Spearman, 1907) and Kendall's rank correlation coefficient (Kendall, 1938) as the metrics to evaluate the correlation of the essay features and essay scores. Both Spearman's and Kendall's correlation coefficients were computed by using the IBM SPSS Statistics (IBM, n.d.). A detailed comparison of Pearson's, Spearman's, and Kendall's correlation coefficients can be referred to literature written by Hauke and Kossowski (2011). The rationale of adopting both Spearman's and Kendall's correlation coefficients in this research is explained below:

i. Non-parametric Statistic

Spearman's and Kendall's correlation coefficients are non-parametric (distribution-free) statistics. They do not assume anything about the underlying distribution of the data (e.g., the data comes from a normal distribution). Therefore, these correlation coefficients are more appropriate to be applied in this research, as they do not require to assumption of the normality of the MUET essay score.

ii. Monotonic Rank Correlation

Spearman's and Kendall's correlation coefficients are monotonic rank statistics. They do not assume the relationship between the variables is linear, nor do they require the variables to be measured on interval scales; they can be used for variables measured at the ordinal level. As such, these correlation coefficients are more appropriate to be applied in this research, as they do not assume the linear relationship between essay features and essay scores. In addition, these correlation coefficients can be extended to measure the relationship between the essay feature and essay grades (which are on ordinal scales).

4.3.1 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient (Spearman, 1907) is a non-parametric rank-based statistic formulated by Charles Spearman. Spearman's coefficient (denoted as Spearman's rho or Spearman's ρ) assesses how well an arbitrary monotonic function can describe a relationship between two variables, without making assumptions about any fixed set of parameters for the probability distribution of the variables. Apart from the interval scale variables, it can be used to assess variables at ordinal scales. Equation 4.1 represents the formula to compute Spearman's ρ value.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad \text{Equation 4.1}$$

where:

ρ = Spearman's rank correlation coefficient

d_i = Difference between the two ranks of each observation

n = Number of observations

The Spearman's ρ value ranged from +1 to -1. A ρ value of +1 denotes a perfect association of ranks, a ρ value of zero denotes no association between ranks, and a ρ value of -1 denotes a perfect negative association of ranks. The closer of ρ value is to zero, the weaker the monotonical association between the ranks of the variables.

4.3.2 Kendall's Rank Correlation Coefficient

Kendall's rank correlation coefficient (Kendall, 1938), proposed by Maurice Kendall is a correlation coefficient that can be used as an alternative to Spearman's rank correlation coefficient for assessing the ordinal association between two variables. Kendall's coefficient (denoted as Kendall's tau or Kendall's τ) is equivalent to Spearman's ρ in terms of all the

underlying assumptions, but they are not identical in magnitude, as their underlying computational logic and formula are different. Kendall's τ usually renders a lower value than Spearman's ρ (Hauke & Kossowski, 2011). Equation 4.2 denotes Kendall's coefficient formula by using the concordant and discordant pairs between the observed variables to compute the τ value.

$$\tau = (C - D)/(C + D) \quad \text{Equation 4.2}$$

where:

τ = Kendall's rank correlation coefficient

C = Number of concordant pairs

D = Number of discordant pairs

4.4 Feature Selection

From the 1,709 linguistic feature indices extracted from the MUET dataset, features that were not statistically significant ($p\text{-value} > 0.05$) on the correlation test were discarded. Furthermore, any negligibly small effect features that did not reach an absolute correlation threshold of Spearman's $\rho \geq 0.1$ were removed. The remaining features were then passed to the two feature selection schemes, namely the Correlation Rank and MRMR Feature Selection described below. The theoretical explanation of Correlation Rank and MRMR Feature Selection can be found in Section 3.4

4.4.1 Correlation Rank Feature Selection

In this experiment setting, several Correlation-Rank Feature Selection Schemes, namely the COR-03, COR-05, and COR-07 scheme has been proposed based on different correlation threshold. Table 4.2 shows the relevant schemes with their feature selection

criteria. Spearman's ρ was used as the selection metric in the scheme to include more features for analysis, as Kendall's τ usually renders a lower correlation value.

Table 4.2: Criteria for Correlation Rank Feature Selection

Scheme	Criteria	Interpretation
COR-03	Features with Spearman's $\rho \geq 0.3$	Low-correlated features
COR-05	Features with Spearman's $\rho \geq 0.5$	Moderate-correlated features
COR-07	Features with Spearman's $\rho \geq 0.7$	High-correlated features

This research adopted the cut-off thresholds at 0.3, 0.5, and 0.7 based on Mukaka's Rule of Thumb (Mukaka, 2012) for interpreting the correlation coefficient. Table 4.3 specifies the guideline for correlation coefficient interpretation, where the values at 0.3, 0.5, and 0.7 denote the starting point of low, moderate, and high correlation.

Table 4.3: Guideline for Interpreting the Size of a Correlation Coefficient

Size of Correlation	Interpretation
.90 to 1.00 (-.90 to 1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to -.30)	Negligible correlation

Source: Mukaka (2012)

4.4.2 Minimum Redundancy Maximum Relevance (MRMR) Feature Selection

In this experiment, several MRMR Feature Selection Schemes have been formulated based on the desired number of features to be selected. Table 4.4 specifies the 10, 25, 50, and 100 features selected in this experiment, which are denoted by the scheme name of MRMR-10, MRMR-25, MRMR-50, and MRMR-100 respectively. This experiment used RapidMiner (RapidMiner, n.d.) to implement the MRMR algorithm. The number of linguistic features to be selected can be configured through the MRMR operator parameter.

Table 4.4: Number of Selected Features by MRMR Feature Selection

Scheme	Number of Selected Features
MRMR-10	10
MRMR-25	25
MRMR-50	50
MRMR-100	100

4.5 Automated Essay Scoring Using Linguistic Features

4.5.1 Machine Learning Classifier

To investigate the optimal linguistic features in scoring MUET essays, this research attempted various machine learning algorithms to predict the essay grade based on the different sets of feature indices selected by the Correlation Rank and MRMR Feature Selection Scheme. The experiment treated the essay scoring process as a multiclass supervised learning task by using the essay grades (discrete value) instead of the essay scores (continuous value) as the target class for prediction. Multiple and diversified machine learning algorithms were employed in the experiment to examine the overall performance of

the selected features in different classifiers, such that the selected features are generally "good" enough for essay scoring without the preference of a particular classifier.

This experiment attempted six supervised learning algorithms, namely Naive Bayes, Generalized Linear Model, Random Forest, Gradient Boosted Tree, Multilayer Perceptron, and Support Vector Machine for scoring the essays. Table 4.5 presents the six machine learning algorithms with the reasons for including them in the experiment. Reference to influential literature on the respective algorithms is also included in the table.

Table 4.5: Machine Learning Algorithms Employed in the Experiment

Machine Learning Algorithm	Abbreviation used in this Research	Reference	Reasons for Including the Algorithm in the Experiment
Naïve Bayes	NB	(Lewis, 1998)	One of the most simple and effective algorithms based on Bayes Theorem.
Generalized Linear Model	GLM	(Nelder & Wedderburn, 1972)	Alternative to conventional linear regression to handle multiclass classification and non-normality data.
Random Forest	RF	(Breiman, 2001)	Effective ensemble learner based on bagging on decision trees and considered one of the top performers in text classification tasks.
Gradient Boosted Tree	GBT	(Mason et al., 1999)	Effective ensemble learner based on gradient boosting on decision trees for reducing bias and thus avoiding overfitting.
Multilayer Perceptron	MP	(Rumelhart et al., 1985)	Alternative to conventional machine learning technique by leveraging

Table 4.5 continued

Machine Learning Algorithm	Abbreviation used in this Research	Reference	Reasons for Including the Algorithm in the Experiment
			neural network model to detect complex nonlinear relationships.
Support Vector Machine	SVM	(Chang & Lin, 2011)	Kernel function to handle non-linearly separable classes by mapping non-linear data into higher-dimensional space.

Table 4.6 specifies the basic hyperparameter setting of the classifiers employed in the experiment. Further hyperparameter tunings of the classifier were not performed so that the resulting outcome is generally associated with the selected set of features, but not tuned towards a particular classifier. This experiment used RapidMiner (RapidMiner, n.d.) as the tool to perform all the classification tasks. The classification results were obtained through the 10-fold cross-validation (Sammut & Webb, 2011) for each classifier.

Table 4.6: Hyperparameters Setting of the Machine Learning Classifier

Machine Learning Classifier	Hyperparameters
Naïve Bayes	Laplace Correction: Yes
Generalized Linear Model	Regularization: Yes
Random Forest	Number of Trees: 100 Criterion: Gain Ratio Maximal Depth: 10
Gradient Boosted Tree	Number of Trees: 50 Maximal Depth: 5 Minimum Rows: 10

Table 4.6 continued

Machine Learning Classifier	Hyperparameters
	Minimum Split Improvement: 1.0E-5 Number of Bins: 20
Multilayer Perceptron	Number of Hidden Layers: 1 Training Cycles: 200 Learning Rate: 0.01 Momentum: 0.9
Support Vector Machine	SVM Type: C-SVC Kernel Type: Linear Gamma: 0.0 C: 0.0 Cache Size: 80 Epsilon: 0.001

4.5.2 Automated Essay Evaluator (AEE) Models

This research conducted experiments to evaluate the proposed Correlation Rank and MRMR Feature Selection Schemes based on the six machine learning classifiers specified in Section 4.5.1. All the Correlation Rank schemes (COR-03, COR-05, COR-07) and MRMR schemes (MRMR-10, MRMR-25, MRMR-75, MRMR-100) were used to predict the MUET essay grades. To assess how well the proposed Correlation Rank and MRMR schemes are in scoring essays, this research compared their results with two benchmark schemes - a baseline scheme, and schemes provided by the open-source AES LightSide. The baseline scheme consists of 50 features which were randomly selected from the overall extracted 1,709 linguistic features using the RapidMiner "Select by Random" operator.

Based on the combination of the eight feature selection schemes (including the baseline scheme) and six machine learning classifiers, a total of 48 AEE models were built

to score essays. Table 4.7 presents an overview of the 48 AEE models employed in this research.

Table 4.7: The 48 AEE Models Employed in the Experiment

Feature Selection Scheme	Machine Learning Classifier					
	NB	GLM	RF	GBT	MP	SVM
COR-03	COR-03_NB	COR-03_GLM	COR-03_RF	COR-03_GBT	COR-03_MP	COR-03_SVM
COR-05	COR-05_NB	COR-05_GLM	COR-05_RF	COR-05_GBT	COR-05_MP	COR-05_SVM
COR-07	COR-07_NB	COR-07_GLM	COR-07_RF	COR-07_GBT	COR-07_MP	COR-07_SVM
MRMR-10	MRMR-10_NB	MRMR-10_GLM	MRMR-10_RF	MRMR-10_GBT	MRMR-10_MP	MRMR-10_SVM
MRMR-25	MRMR-25_NB	MRMR-25_GLM	MRMR-25_RF	MRMR-25_GBT	MRMR-25_MP	MRMR-25_SVM
MRMR-50	MRMR-50_NB	MRMR-50_GLM	MRMR-50_RF	MRMR-50_GBT	MRMR-50_MP	MRMR-50_SVM
MRMR-100	MRMR-100_NB	MRMR-100_GLM	MRMR-100_RF	MRMR-100_GBT	MRMR-100_MP	MRMR-100_SVM
Baseline	Baseline_NB	Baseline_GLM	Baseline_RF	Baseline_GBT	Baseline_MP	Baseline_SVM

Besides the baseline scheme, LightSide AES has been selected as the benchmark model for evaluating the proposed AEE with existing AES systems. The choice of LightSide as the benchmarking AES system is primarily based on its open-source nature and the accessibility of both its compiled and source codebase to the public. For this research, the author downloaded the LightSide Researcher's Workbench executable (LightSide, 2019) to score the MUET essays. Other AES systems like e-Rater, IEA, and Intellimetric cannot be

utilized due to their commercial proprietary nature, making it impossible to acquire the software without cost.

Furthermore, LightSide's extensive adoption in AES research makes it a preferable choice for comparison and benchmarking. LightSide was employed as one of the primary AES systems in the renowned large-scale AES evaluation study conducted by Shermis (2014). Moreover, Zupanc (2018), in her doctoral research, used LightSide as one of the benchmarking models to compare her developed semantic-based AES - Semantic Automated Grader for Essays (SAGE). Recent research, such as those by Cunningham-Nelson et al. (2018) and Uzun (2018), continue to utilize LightSide, demonstrating its ongoing relevance in the field.

In the experiment, two essay feature schemes, named LightSide_Unigram scheme and LightSide_Composite scheme, were used to compare against the proposed Correlation Rank and MRMR schemes. The LightSide_Unigram scheme is the basic unigram features extracted by LightSide. On the other hand, the LightSide_Composite scheme is the combination of basic unigram features, English parse features, and stretchy pattern features extracted by LightSide. This LightSide_Composite scheme is proposed with reference to the work by (Uzun, 2018), which uses the same LightSide features to score essays. In both the LightSide Unigram and LightSide_Composite scheme, the machine learning classifier provided by LightSide, namely the Naive Bayes, Logistic Regression, Support Vector Machine, Decision Trees, and Random Forest were used to score the MUET essays.

4.5.3 Evaluation Metrics

In this research, the performance of the proposed AEE (Automated Essay Evaluator) was evaluated using several metrics. Aligning with the standard practices in the field of

automated essay scoring, this research adopted the Quadratic Weighted Kappa (QWK) and classification accuracy/exact agreement as the primary evaluation metrics. Furthermore, supplementary evaluation metrics such as precision, recall, and F1 score were provided when deemed necessary. The details and calculation formulas for all the evaluation metrics used are described below.

4.5.3.1 Quadratic Weighted Kappa

The Quadratic Weighted Kappa (QWK; Cohen, 1968) is a statistical measure used to examine the degree of agreement between two raters. Compared with classification accuracy, the QWK is a superior evaluation metric as it accounts for agreement by chance (Doewes et al., 2023). Furthermore, it considers the severity of disagreement by imposing greater penalties for larger disagreements than minor ones. It is frequently employed as a standard evaluation metric in the field of AES (Doewes et al., 2023). Furthermore, QWK was adopted as the official evaluation metric of the Automated Student Assessment Prize competition (Kaggle, 2012). The QWK value ranges from -1 (perfect disagreement between two raters) to 1 (perfect agreement between two raters). A QWE value of 0 suggests that the observed agreement is no better than the expected agreement by chance.

QWK can be calculated according to the steps below:

- i. A weight matrix ($W_{i,j}$) is constructed based on Equation 4.3, such that i and j correspond to the reference rating (assigned by the human rater) and the hypothesis rating (predicted by the AES classifier), respectively. N is the number of possible ratings.

$$W_{i,j} = \frac{(i - j)^2}{(N - 1)^2} \quad \text{Equation 4.3}$$

where:

W = Weight matrix

i = Reference rating (assigned by human rater)

j = Hypothesis rating (predicted by AES)

N = Number of possible ratings

- ii. An observed-agreement matrix ($O_{i,j}$) is calculated, such that $O_{i,j}$ represents the number of essays that have a rating i by the human rater and a rating j by the AES classifier.
- iii. An expected-agreement matrix ($E_{i,j}$) is computed based on the outer product of two histogram vectors - $histogram_r$ and $histogram_h$ of the reference rating and hypothesis rating, respectively. The matrix E is then normalized to produce the final matrix E (as denoted by Equation 4.4).

$$E = Normalize(histgram_r \otimes histgram_h) \quad \text{Equation 4.4}$$

where:

E = Expected-agreement matrix

$histgram_r$ = Histogram of reference rating (assigned by human rater)

$histgram_h$ = Histogram of hypothesis rating (predicted by AES)

- iv. Based on the constructed matrix W, O, and E, the QWK value can be calculated according to Equation 4.5.

$$k = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \quad \text{Equation 4.5}$$

where:

k = Quadratic Weighted Kappa

W = Weight matrix

O = Observed-agreement matrix

E = Expected-agreement matrix

4.5.3.2 Accuracy / Exact Agreement

The accuracy or exact agreement measures the overall correctness of an AES classifier by calculating the ratio of correctly predicted essays to the total number of essays.

Equation 4.6 denotes the formula of accuracy.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad \text{Equation 4.6}$$

4.5.3.3 Precision

Precision measures the proportion of correctly predicted instances of a specific class out of all instances predicted as that class. It focuses on the correctness of positive predictions. Precision is derived by dividing the true positive predictions for a class by the total of true positive and false positive predictions for that class. Precision score focuses on the correctness of positive predictions. Equation 4.7 represents the formula of precision.

$$Precision = \frac{TP}{(TP + FP)} \quad \text{Equation 4.7}$$

where:

TP = True positives for the class

FP = False positives for the class

4.5.3.4 Recall

Recall measures the proportion of correctly predicted instances of a specific class out of all actual instances of that class. It emphasizes the model's ability to identify positive instances. The recall value is computed by dividing the true positive predictions for a class by the sum of true positive and false negative predictions for that class. Equation 4.8 denotes the formula of recall.

$$Recall = \frac{TP}{(TP + FN)} \quad \text{Equation 4.8}$$

where:

TP = True positives for the class

FN = False negatives for the class

4.5.3.5 F1 Score

The F1 score is the harmonic mean of precision and recall for a specific class. It provides a balanced single measure that considers both precision and recall. The metric is particularly useful when dealing with imbalanced datasets. The F1 value ranges from 0 (either precision or recall is zero) to 1 (perfect precision and recall). Equation 4.9 denotes the formula for calculating the F1 score for a particular class.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{Equation 4.9}$$

4.6 Boxplot

This experiment used the boxplot or box-and-whisker plot (Williamson et al., 1989) and other charts (e.g., column charts and combo charts) to present the results in this research. The boxplot diagram is a visualization method to describe the variability or dispersion of numerical through their four quartiles. The boxplot diagram comprises several components: first quartile (Q1), median, third quartile (Q3), minimum value, maximum value, and outliers. Figure 4.2 shows the key elements of a boxplot.

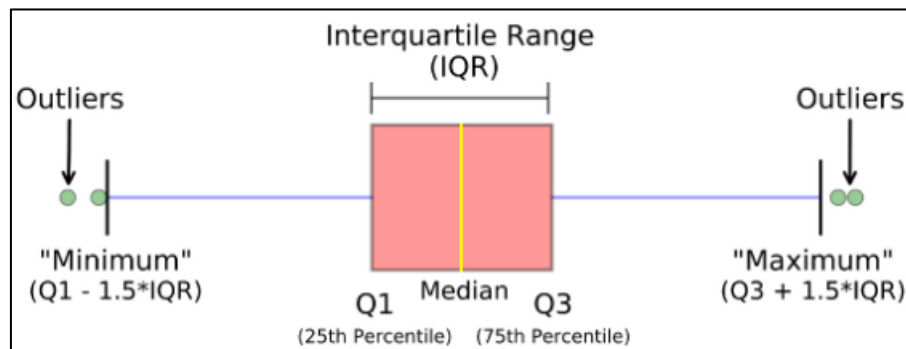


Figure 4.2: Components of a Boxplot

As depicted in Figure 4.2, the boxplot consists of:

i. First Quartile (Q1)

This value separates the lowest 25% of the dataset from the rest. It denotes the boundary of the 25th percentile of the data when arranged in ascending order.

ii. Median

The median divides the dataset into two halves, with 50% of the data falling below the median and 50% above it.

iii. Third Quartile (Q3)

This value divides the upper 25% of the dataset from the lower 75%. It denotes the boundary of the 75th percentile when the data is sorted in ascending order.

iv. Minimum

The minimum value represents the lowest data point observed in the data set. It is located at the bottom end of the whisker in the box plot. As denoted in Equation 4.10, the minimum value is derived by subtracting 1.5 times the Interquartile Range (IQR) from the First Quartile (Q1).

$$\text{Minimum} = Q1 - 1.5 \times IQR \quad \text{Equation 4.10}$$

v. Maximum

Conversely, the maximum value represents the highest data point observed in the dataset. It is located at the top end of the whisker. As denoted by Equation 4.11, the maximum value is calculated by adding 1.5 times the Interquartile Range (IQR) to the Third Quartile (Q3).

$$\text{Maximum} = Q3 + 1.5 \times IQR \quad \text{Equation 4.11}$$

vi. Outlier

Outliers are data points that significantly deviate from the rest of the dataset. They are typically located outside the whiskers of the boxplot. The outlier may indicate a true variability or an indication of errors in data collection.

These boxplot components collectively provide valuable insights into the distribution and variability of the dataset. It facilitates the visual representation of the dataset pattern, such as its symmetry and skewness, and the detection of outliers. The length of the box representing the Interquartile Range (IQR), indicates the spread of the middle 50% of the data. The whiskers extend from the box to the minimum and maximum values, showing the

extent of the data range, excluding the outliers. Outliers, if present, are plotted individually beyond the whiskers, aiding the detection and investigation of abnormalities.

Due to the large number of feature indices in this research, the correlation value of each feature index and essay scores are not well-suited to be rendered individually via bar charts or line charts. Therefore, the boxplot diagrams were primarily used for rendering the feature evaluation results in this experiment. The correlation values distribution of manifold feature indices within a particular feature category and feature set can be graphically visualized via the boxplots. The detailed results and analysis via the boxplots are reported in Chapter 5.

4.7 Chapter Summary

In this research, experiments were designed to analyse and score MUET essays from the linguistic dimension. The first step of the experiment is the linguistic features extraction from the MUET essay dataset by using multiple linguistic analysis tools. The correlations of the extracted essay features with the essay scores were evaluated using Spearman's and Kendall's rank correlation coefficient. The significant essay features affecting essay scores were then selected using the Correlation Rank and MRMR Feature Selection Schemes. Finally, the selected linguistic features were input into various machine learning classifiers to predict the essay grade. The results were presented through a variety of graphs in Chapter 5.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter is dedicated to presenting the findings and results of this research, which consists of two main parts, i.e., the evaluation of linguistic features and the automated scoring of MUET essays. Section 5.1 illustrates Spearman's and Kendall's rank coefficient values in a line graph. Section 5.2 discusses the correlation value between the individual linguistic feature set and the essay scores via boxplot diagrams. Section 5.3 and Section 5.4 present the key findings of this research by presenting the correlation value of the comprehensive linguistic feature sets and feature categories with the essay scores depicted through boxplot diagrams. Section 5.5 discusses the selected linguistic feature indices using the Correlation Rank and MRMR Feature Selection. Section 5.6 presents the performance of the linguistic-based AEE in this research by examining the effectiveness of the selected features, identifying the optimal linguistic features, comparing with a third-party AES vendor, an overview of high and low-performance AEE models, and evaluating the scoring of different essay grade via confusion matrices. Lastly, Section 5.7 summarizes the main findings and results in this chapter.

To facilitate further understanding of the findings presented in this chapter and to support future research, Appendix B includes a hyperlink to a directory containing the relevant schema and data in this experiment. The appendix specifies the hyperlink to the linguistic feature specification used in the research, the feature value extracted from the MUET essays, the correlation value of every feature with essay score/essay band, and the selected features based on different feature selection schemes.

5.1 Spearman's and Kendall's Rank Correlation Coefficient

Figure 5.1 exhibits Spearman's and Kendall's rank correlation coefficient of the first 500 linguistic features and the MUET essay scores. The plot in the line graph shows that both values are approximately similar, with Kendall's τ rendering a lower value than Spearman's ρ . This observation aligns with the theoretical aspect of the values (Hauke & Kossowski, 2011). Since the two values are similar, this chapter will use Spearman's ρ as the correlation metric to simplify the discussion and reporting of the findings in this research. The complete Spearman's and Kendall's correlation values of all the 1,709 linguistic features with the MUET essay score in this research can be referred in the online spreadsheet titled "3_Features Correlation.xlsx." Appendix B contains the hyperlink to access this spreadsheet.

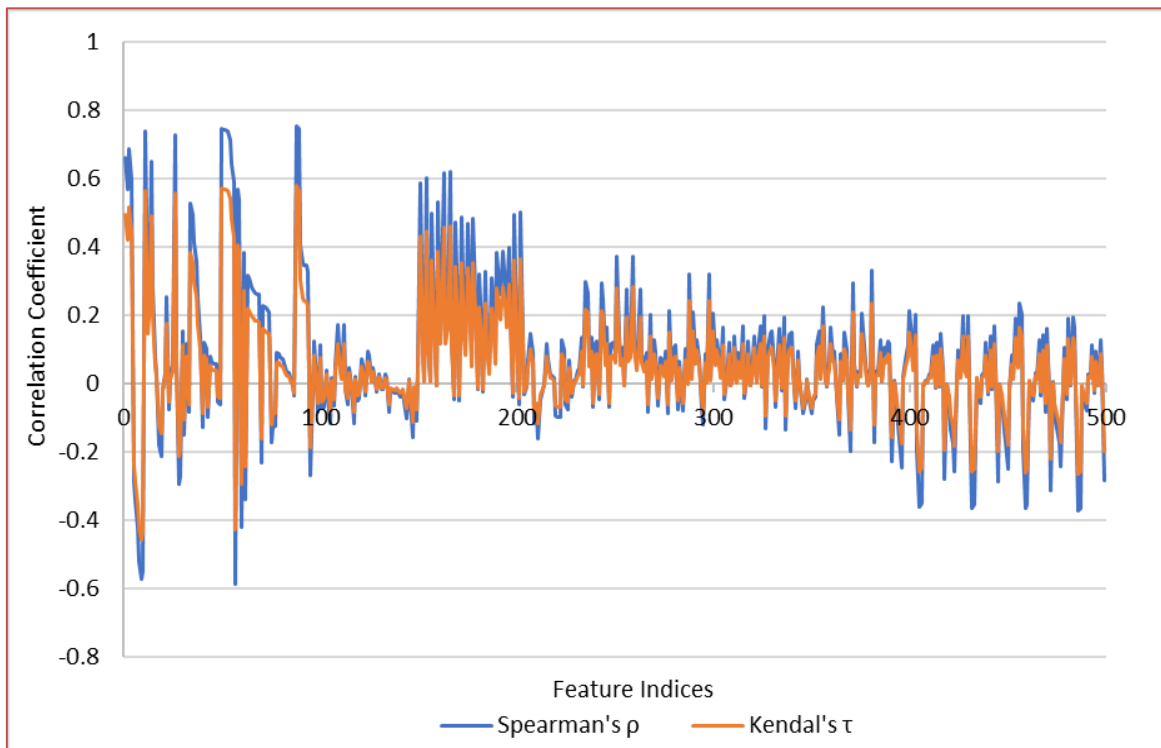


Figure 5.1: Comparison of Spearman's ρ and Kendall's τ Value

5.2 Individual Linguistic Feature Set Evaluation

This section discusses the detailed result of the individual linguistic feature set evaluation. In every sub-section, the boxplot diagram is illustrated to exhibit the correlation distribution between the respective feature set and essay scores.

5.2.1 Surface Features

Figure 5.2 shows the distribution of Spearman's ρ value based on different surface feature categories in the surface feature set. The boxplot indicates:

- i. The surface feature set overall shows a high positive correlation with essay scores.
- ii. Among the four surface feature categories, length-based features demonstrated the highest correlation with essay scores.
- iii. The advanced parts of speech, i.e., the occurrence of the adjective, adverb, past participle tense, and present participle tense tend to correlate with higher scores in essays. This finding is one of the novel results of this research.

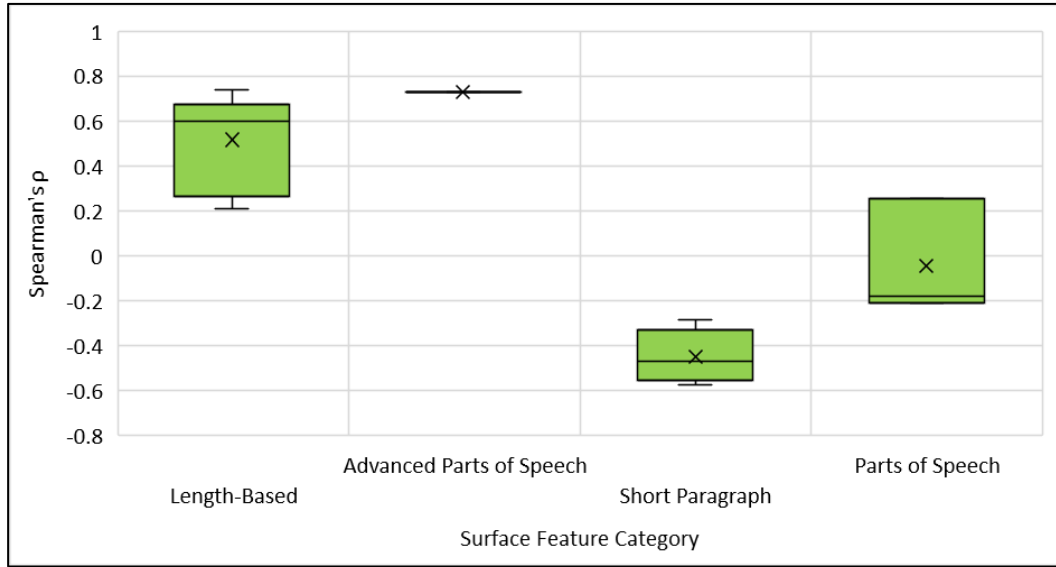


Figure 5.2: Correlation between Surface Feature Set and Essay Scores

The finding of the strong association between length-based features and essay scores as outlined in point (ii) above, implies their potential significance in predicting essay scores. This observation aligns with the literature by Perelman (2014) and Phandi et al. (2015). Perelman (2014) examined the evaluation report by Shermis (2014), which assessed the state-of-the-art AES systems including PEG, IEA, e-rater, and Intellimetric. He revealed that these AES systems primarily rely on the word-counting approach to predict essay scores with a considerable level of accuracy. On the other hand, Phandi et al. (2015) employed length-based features such as word count, word length, sentence count, and POS count as core features to construct an AES system with domain adaptation capabilities. By leveraging these length-based features, the AES achieved favourable results in comparison to the conventional domain adaptation approaches.

5.2.2 Linguistic Errors

Figure 5.3 shows the distribution of correlation coefficients between the linguistic error feature set and essay scores. Besides the boxplot, another column chart is plotted in Figure 5.4 to illustrate the correlation value between the fine-grained linguistic error feature indices and the essay scores. As exhibited in both plots:

- i. The linguistic error feature set shows a very small or negligible correlation with essay scores.
- ii. Compared with the raw error count, the error rate is a more reliable and consistent indicator of essay scores. The error rate shows a slightly higher correlation value and consistently negatively correlates with essay scores. Some of the raw error counts (e.g., spelling error count and total error count) show a positive correlation with essay scores. This result may be due to the text-length effect - a higher score essay tends to be longer in text and may produce higher raw error counts.

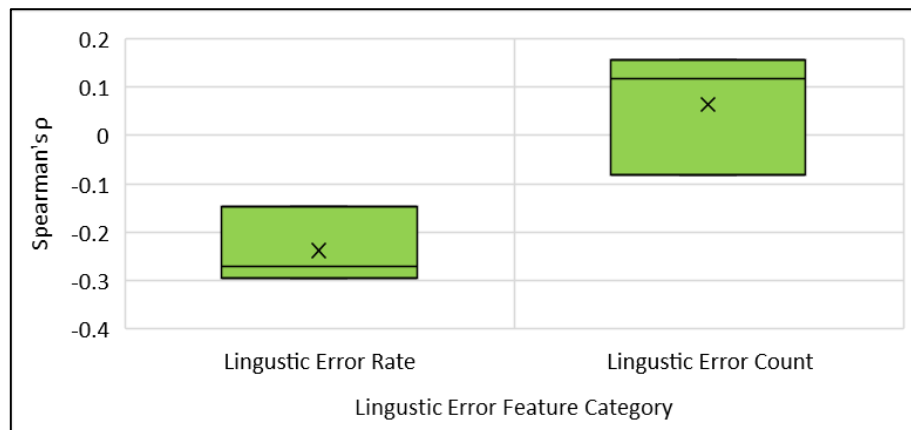


Figure 5.3: Correlation between Linguistic Error Feature Set and Essay Scores

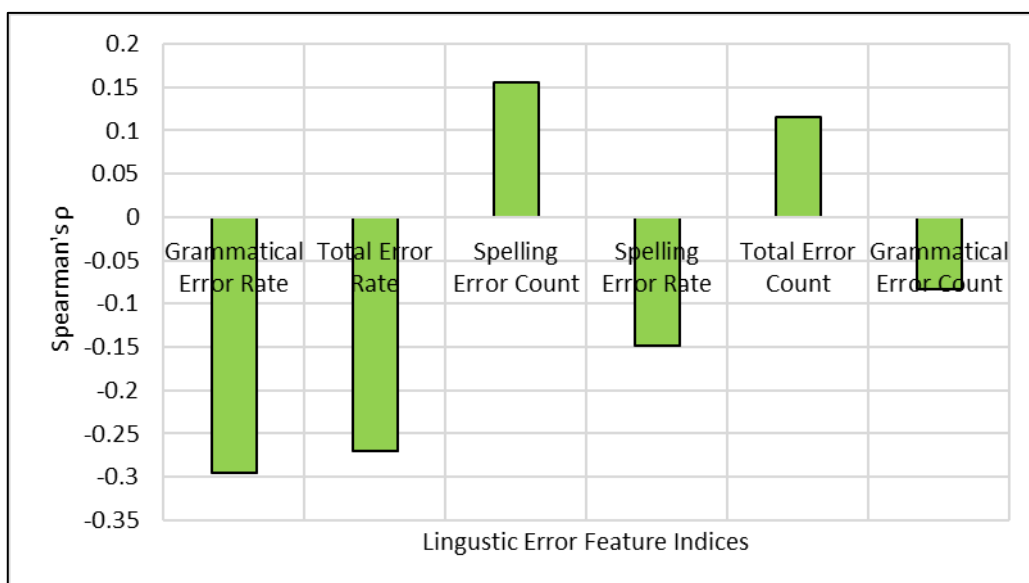


Figure 5.4: Correlation between Linguistic Error Feature Indices and Essay Scores

Although the finding in point (i) concerning the low correlation between linguistic errors and essay scores may appear contradictory to our common understanding, a similar result of the weak relation between linguistic errors and essay scores can be found in the literature by Allen et al. (2014). In the research, Allen et al. (2014) found a weak relationship between grammatical errors and holistic essay scores after examining linguistic error counts in essays written by high school students using the Writing Pal Intelligence Tutoring System.

5.2.3 Text Cohesion

Figure 5.5 shows the dispersion of correlation coefficients between the text cohesion feature set and essay scores. The boxplot reveals:

- i. The global cohesion feature category, i.e., the lexical items overlapping between adjacent paragraphs shows a moderate positive correlation with essay scores. The result denotes that the higher repetition of the same word between paragraphs is likely to produce higher essay scores.

- ii. The local cohesion feature category, i.e., the lexical items overlapping in the scope of adjacent sentences shows a weak correlation (within the range of Spearman ρ of 0.1 to -0.1) with essay scores. The result denotes that the higher repetition of the same word between sentences tends to have no impact or even produces lower essay scores.
- iii. The overall cohesion feature category, i.e., the usage of connectives and the text-givenness features in essays shows a negligible association (either positive or negative correlations) on essay scores.

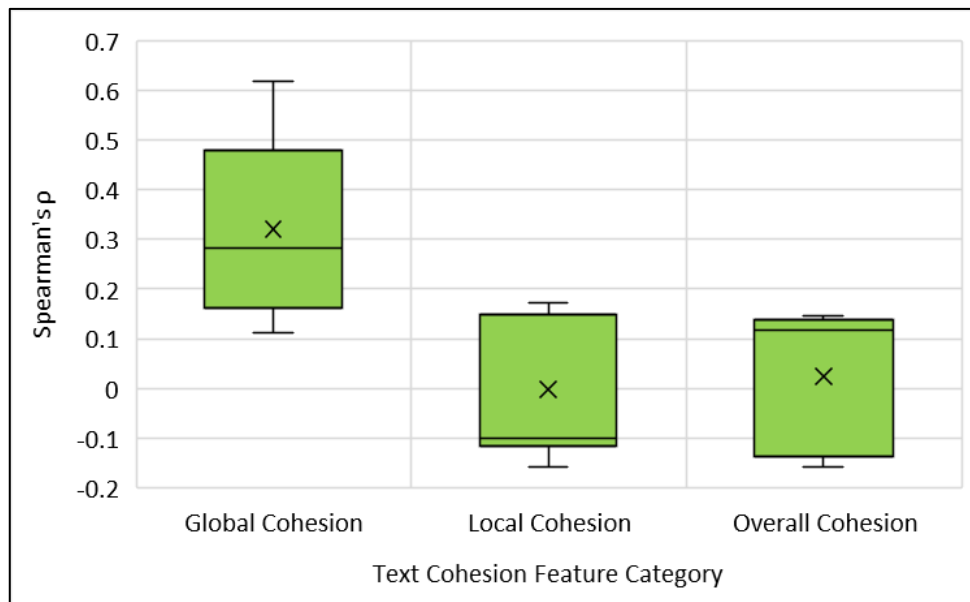


Figure 5.5: Correlation between Text Cohesion Feature Set and Essay Scores

The findings stated in points (i) and (ii) above, concerning the relationship between global and local cohesion with essay scores are consistent with the results reported by Crossley & McNamara (2016). Crossley & McNamara (2016) investigated the relationship between text cohesion with overall essay quality. They modified original essays to increase both local and global text cohesion elements, such as lemma overlap between adjacent

sentences or paragraphs. Their result showed a stronger positive correlation between essay quality and global cohesion than local cohesion, denoting that global cohesion is potentially a more significant feature for essay quality than local cohesion. The local cohesion indices reported a small effect size on the holistic score of essays.

5.2.4 Semantics

Figure 5.6 exhibits the dispersion of the correlation coefficients between the semantic feature set and essay scores. The plot shows a similar trend with the text cohesion feature set. Instead of measuring the overlapping of lexical items, these semantic features calculate the overlapping of WordNet synonyms and the average similarity scores computed by the LSA, LDA, and Word2vec between segments of sentences or paragraphs. The boxplot indicates:

- i. The global semantics, i.e., the semantic items overlapping between adjacent paragraphs show a moderate positive correlation with essay scores. The result implies that the higher similarity between paragraphs may produce a more coherent text and thus improve text comprehension and essay scores.
- ii. The local semantics, i.e., the semantic items overlapping in the scope of adjacent sentences show a weak correlation (either positive or negative correlation) with essay scores. The result implies that the higher repetition of the same semantic item between sentences may produce the same content and thus yield a lower essay score.

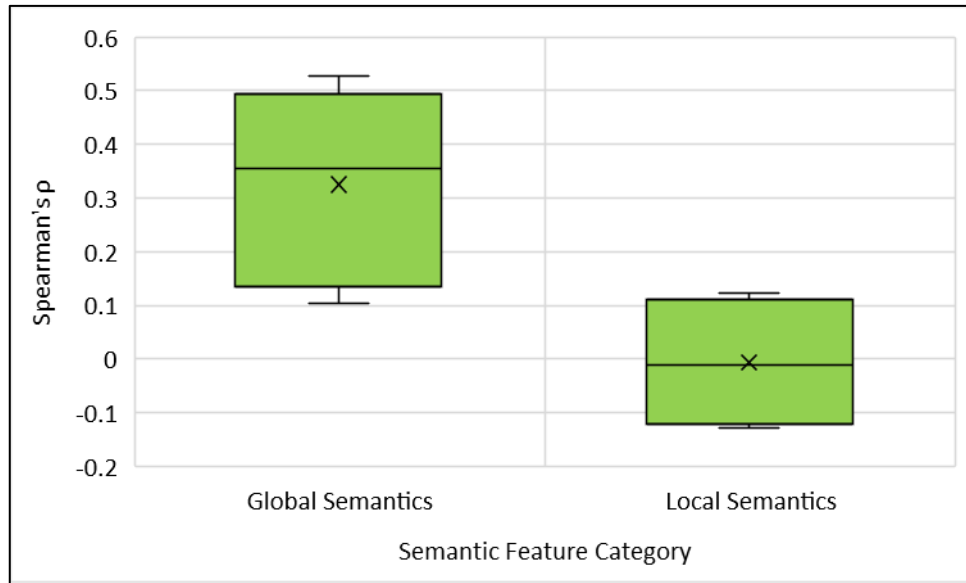


Figure 5.6: Correlation between Semantic Feature Set and Essay Scores

The above findings regarding the relationship between global and local semantics with essay scores in this experiment support the outcome reported by Crossley et al. (2019b). Crossley et al. (2019b) used an automated cohesion tool to assess the semantic similarities at the local and global level provided by LSA, LDA, and word2vec. Their findings indicated that the global semantic similarities (especially those measured by word2vec) were a significant predictor of human scoring of essay coherence, while local semantic variables demonstrated a weak correlation with the essay coherence score.

Besides Figure 5.6, another boxplot is plotted in Figure 5.7 to investigate the effectiveness of semantic analysis techniques, i.e., the WordNet Synonym Overlap, LSA, LDA, and Word2vec in predicting essay scores. In addition to utilizing techniques such as LSA, LDA, and Word2vec, this research incorporates WordNet Synonym Overlap as a method for semantic analysis. WordNet's lexical database is leveraged to assess the semantic similarity within essays. Both local and global semantic similarities within essays are

determined by measuring the overlap between individual words and their corresponding WordNet Synonym Sets (Synsets) across sentences and paragraphs.

As depicted in Figure 5.7, the WordNet Synonym Overlap approach emerges as a more effective essay score predictor, as it exhibits a wider range of correlation magnitude, followed by the LSA, Word2vec, and LDA. The underlying rationale is that an effective approach should exhibit a broader range of correlation magnitudes with essay scores. The finding is rational because human raters tend to make more sense and intuitive to the overlapping of WordNet synonyms, compared with similar words computed by complex algorithms such as LDA. This finding is one of the novel results of this research.

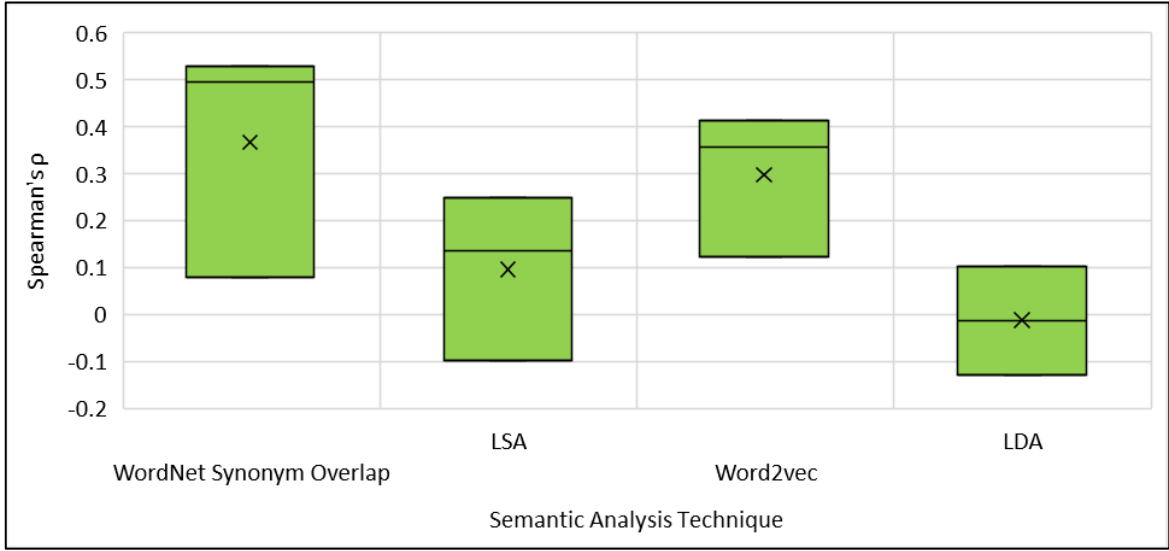


Figure 5.7: Correlation between Applied Semantic Analysis Techniques and Essay Scores

A similar pattern is observed in the standard deviation of the feature value. Table 5.1 presents the standard deviation of feature values for each semantic analysis technique. The values are derived by considering every local and global semantic feature in the respective

technique. Notably, the WordNet Synonym Overlap approach demonstrates a higher standard deviation value, followed by LSA, Word2vec, and LDA. WordNet synonym features with higher standard deviations tend to have more significant variability in their values and potentially capture more distinctions between data points. This higher variability may enhance the ability to effectively differentiate between different essay grades, making WordNet Synonym Overlap features a potentially good predictor of essay grade provided that they contain relevant information for essay grades.

Table 5.1: Standard Deviation of Feature Values by Different Semantic Analysis Techniques

Semantic Analysis Technique	Standard Deviation of Feature Values
WordNet Synonym Overlap	4.89
LSA	0.19
Word2vec	0.06
LDA	0.05

5.2.5 Lexical Diversity

Figure 5.8 shows the distribution of correlation coefficients between the lexical diversity feature set and essay scores. The boxplot indicates:

- i. The lexical diversity feature set generally exhibits a strong correlation with essay scores.
- ii. The lexical raw count exhibits a high positive correlation with essay scores, followed by the Root TTR and lexical density which shows a moderate positive correlation.

- iii. The Simple TTR and Log TTR feature categories are negatively correlated with essay scores. These results demonstrate the text length effect, i.e., a higher score essay is usually longer in text and tends to produce a lower TTR score.
- iv. The advanced lexical diversity indices (e.g., MTLD, HDD, and MATTR) show a low to negligible positive correlation with essay scores. The positive correlation value denotes the text length effect is mostly mitigated in these advanced TTR indices.

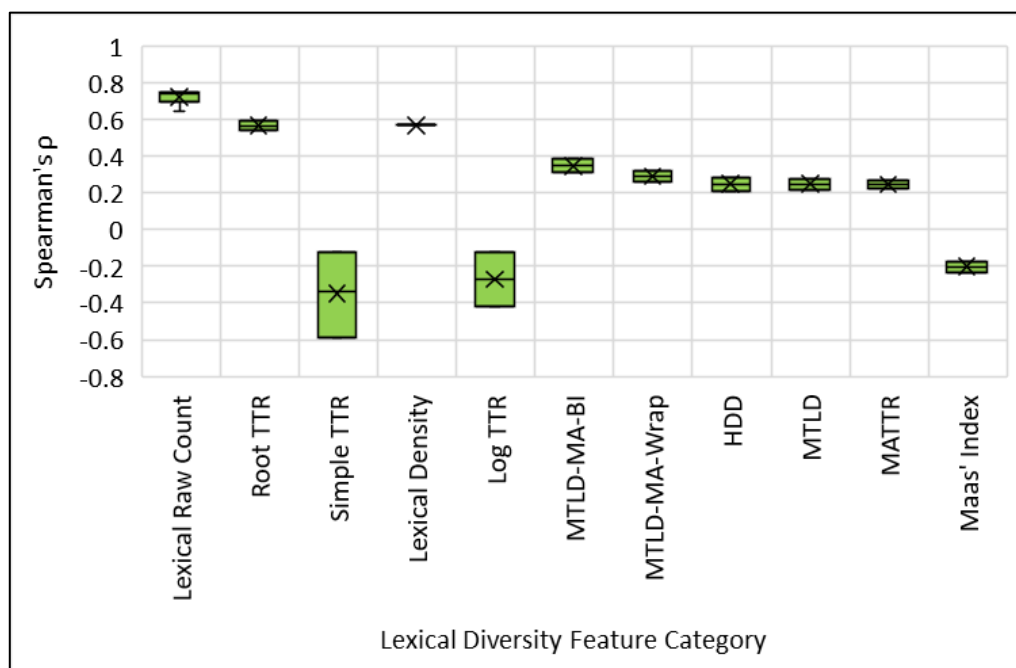


Figure 5.8: Correlation between Lexical Diversity Feature Set and Essay Scores

5.2.6 Lexical Sophistication

Figure 5.9 depicts the dispersion of Spearman's ρ value between the lexical sophistication feature set and essay scores. This lexical sophistication feature set consists of 14 feature categories and is the largest feature set investigated in this research. The boxplot denotes:

- i. Most of the feature category in the lexical sophistication feature set shows a low to negligible correlation with essay scores.
- ii. The word frequency and word range feature indices derived from COCA and TOEFL11 corpus show a high negative correlation with essay scores. The result is logical as a higher score essay tends to consist of sophisticated words which less frequently presented in the reference corpus.
- iii. The age of acquisition/exposure features computed from Kuperman's Age of Acquisition Norms Collection show a moderate positive correlation with essay scores. A higher score essay tends to consist of sophisticated words acquired at an older age.

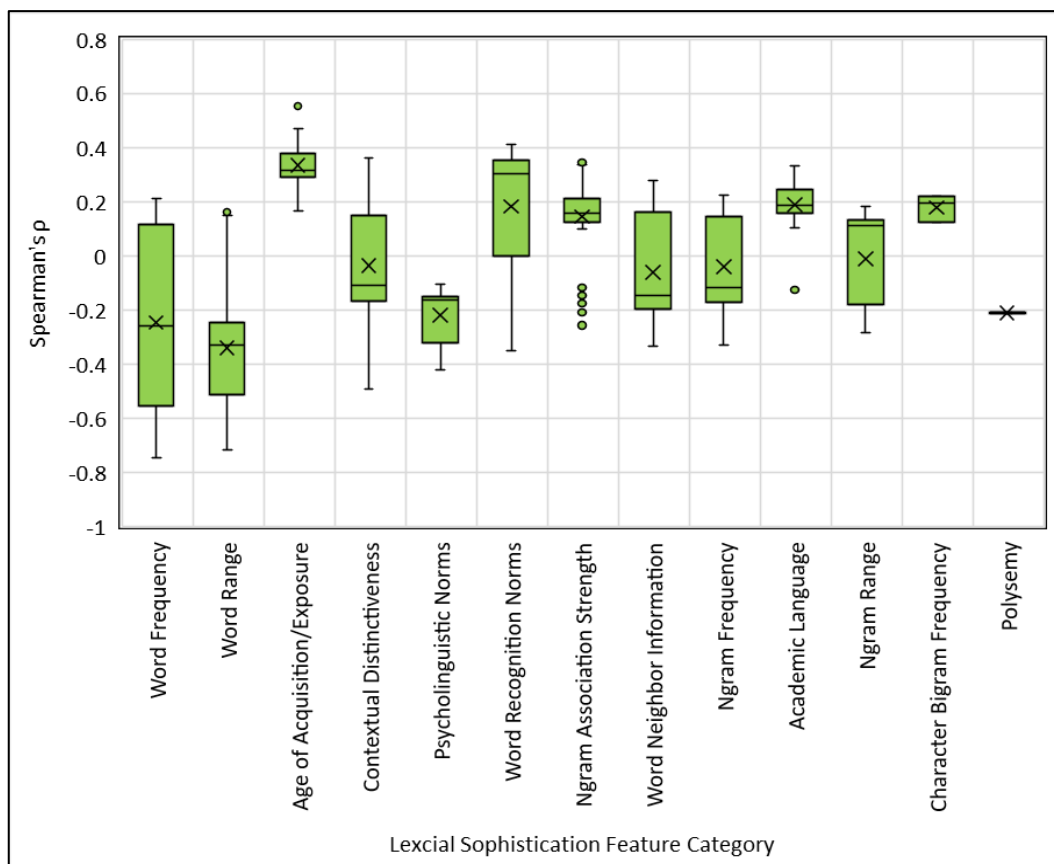


Figure 5.9: Correlation between Lexical Sophistication Feature Set and Essay Scores

The results outlined in points (ii) and (iii) above, which discuss the relationship between the lexical sophistication indices (word frequency, word range, and age of acquisition/exposure) and essay scores, are consistent with the findings presented in the research by Kyle et al. (2018). In their investigation of lexical sophistication indices in text written by both L1 and L2 English learners, Kyle et al. (2018) found that the word frequency and word range indices showed a negative correlation with the lexical proficiency rubric score. Conversely, the age of acquisition/exposure indices exhibited a positive correlation with the rubric score.

Besides the correlation analysis of lexical sophistication feature categories, another boxplot is plotted in Figure 5.10 to investigate the usage of word form (raw word or lemma) and word uniqueness (word type or word token) in predicting the essay score. The underlying theory is that a better scheme shall demonstrate a higher correlation magnitude with essay scores. As shown in Figure 5.10, the lemma type (unique words in lemma form) is a better essay score predictor, followed by the raw type (unique words in raw form).

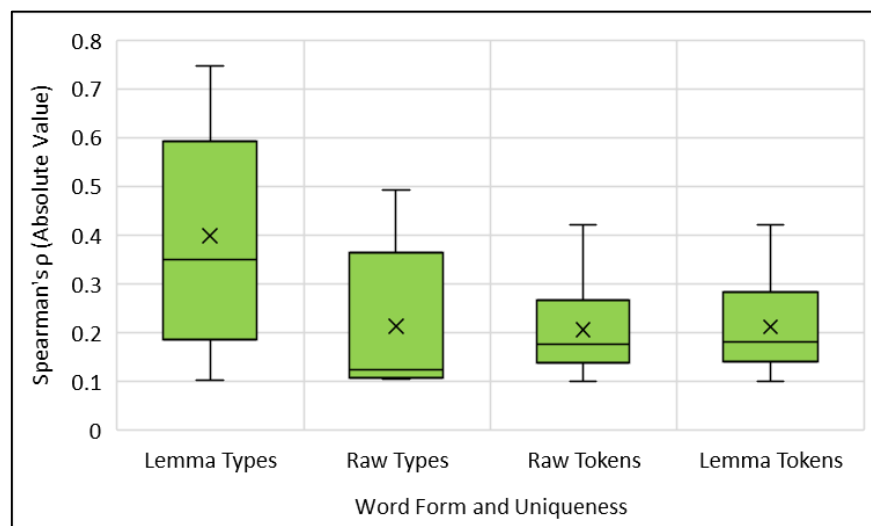


Figure 5.10: Correlation between Applied Word Form and Uniqueness with Essay Scores

5.2.7 Syntactic Complexity

Figure 5.11 shows the distribution of correlation coefficients between the syntactic complexity feature set and essay scores. The boxplot indicates:

- i. The syntactic complexity feature set shows a low to negligible correlation with essay scores.
- ii. Among the four syntactic complexity categories examined, they all show a weak correlation with essay scores, except for the syntactic sophistication category, which exhibits a relatively higher correlation, particularly when focusing on specific indices. Some indices of syntactic sophistication, particularly those associated with Verb-Argument Constructions (VACs), demonstrate moderate correlation values around the range of -0.3 to -0.4.

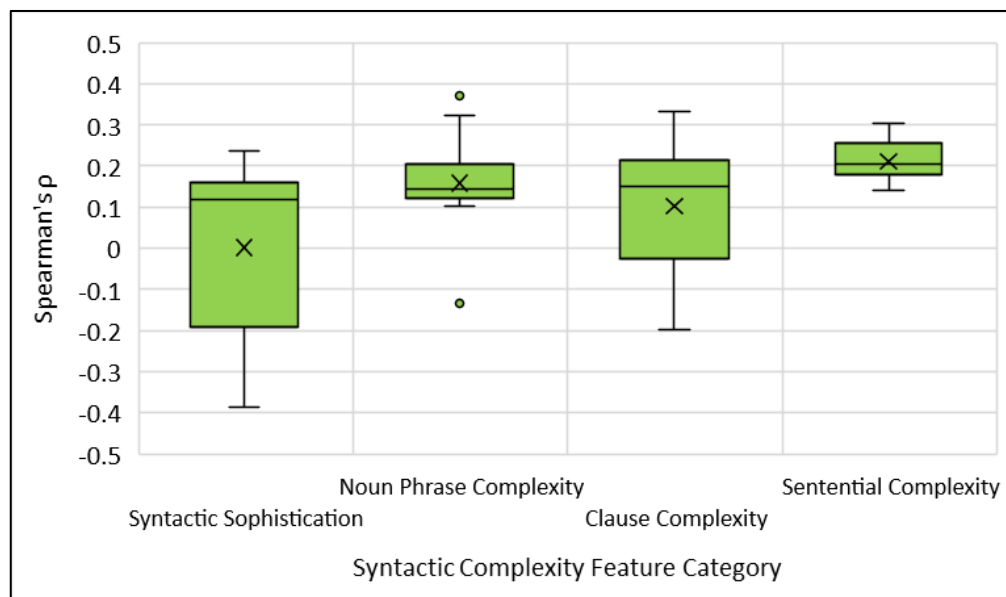


Figure 5.11: Correlation between Syntactic Complexity Feature Set and Essay Scores

5.2.8 Readability

Figure 5.12 shows the distribution of correlation coefficients between the readability feature set and essay scores. Besides the boxplot, another column chart is plotted in Figure 5.13 to illustrate the correlation value between the fine-grained readability feature indices and the essay scores. The plots suggest:

- i. The readability feature set, particularly the CARES and CAREC feature indices shows a high positive correlation with essay scores.
- ii. Among the three readability feature categories studied, the new crowdsourced readability, represented by CARES and CAREC demonstrates the highest degree of correlation with the essay scores. The CARES and CAREC readability formulas encompass multiple linguistic aspects, such as length-based features, text cohesion, lexical diversity, and lexical sophistication, enabling a comprehensive evaluation of writing. As a result, they have the potential to be leveraged as a strong and reliable predictor of essay scores. This finding is one of the novel results of this research.
- iii. The classic and Coh-Metrix readability show a moderate correlation with essay scores. The respective formulas use only a narrow scope of the length-based features in computing their readability scores and are incapable of providing a holistic linguistic assessment of essay quality.

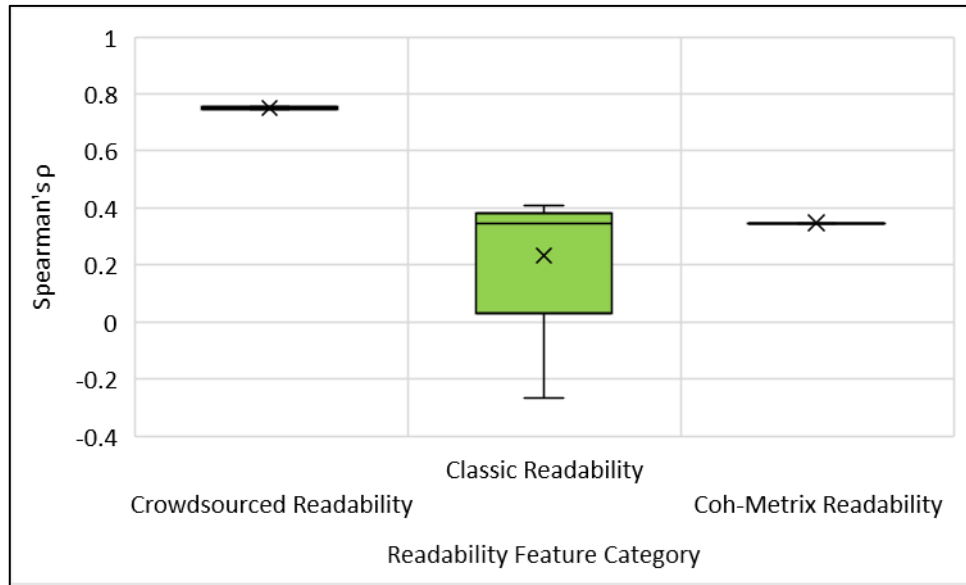


Figure 5.12: Correlation between Readability Feature Set and Essay Scores

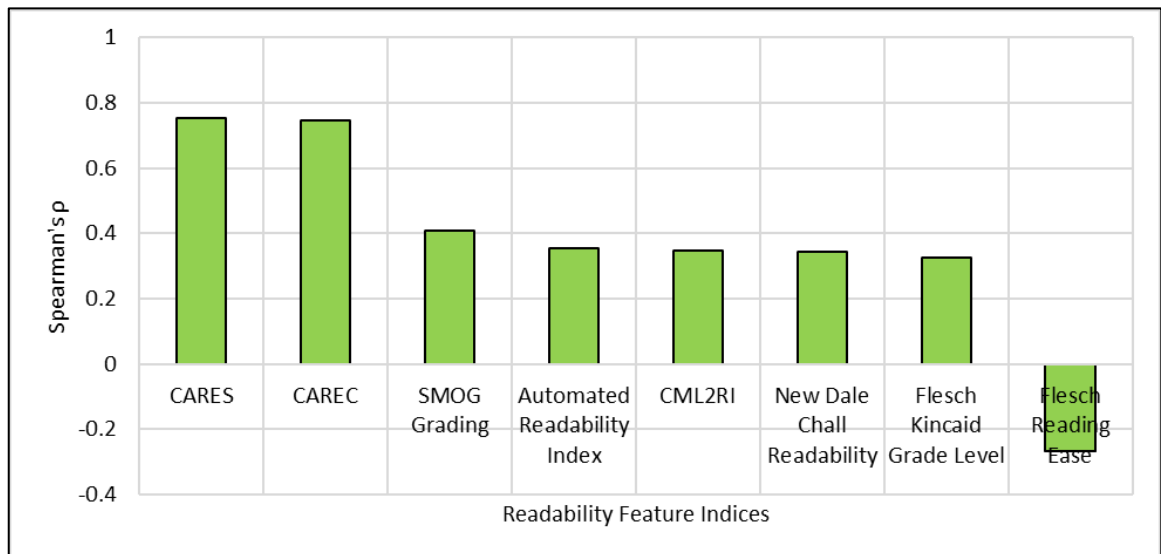


Figure 5.13: Correlation between Readability Feature Indices and Essay Scores

5.3 Comprehensive Linguistic Feature Set Evaluation

Figure 5.14 depicts the distribution of Spearman's ρ value based on the comprehensive linguistic feature sets investigated in this research. The boxplot demonstrates the main results of this research as described below.

- i. The readability, lexical diversity, and surface feature set strongly correlate with essay scores. Some of the relevant feature indices show a correlation value of around 0.7.
- ii. Text cohesion and semantic feature sets are considered moderately correlated with essay scores. Some of the feature indices show a correlation range of around 0.5 to 0.7.
- iii. Linguistic error and syntactic complexity feature sets render a low to negligible effect on essay scores.
- iv. Lexical sophistication overall shows a low to negligible correlation with essay scores except in specific feature indices such as word frequency and word range, which renders a high negative correlation with essay scores.

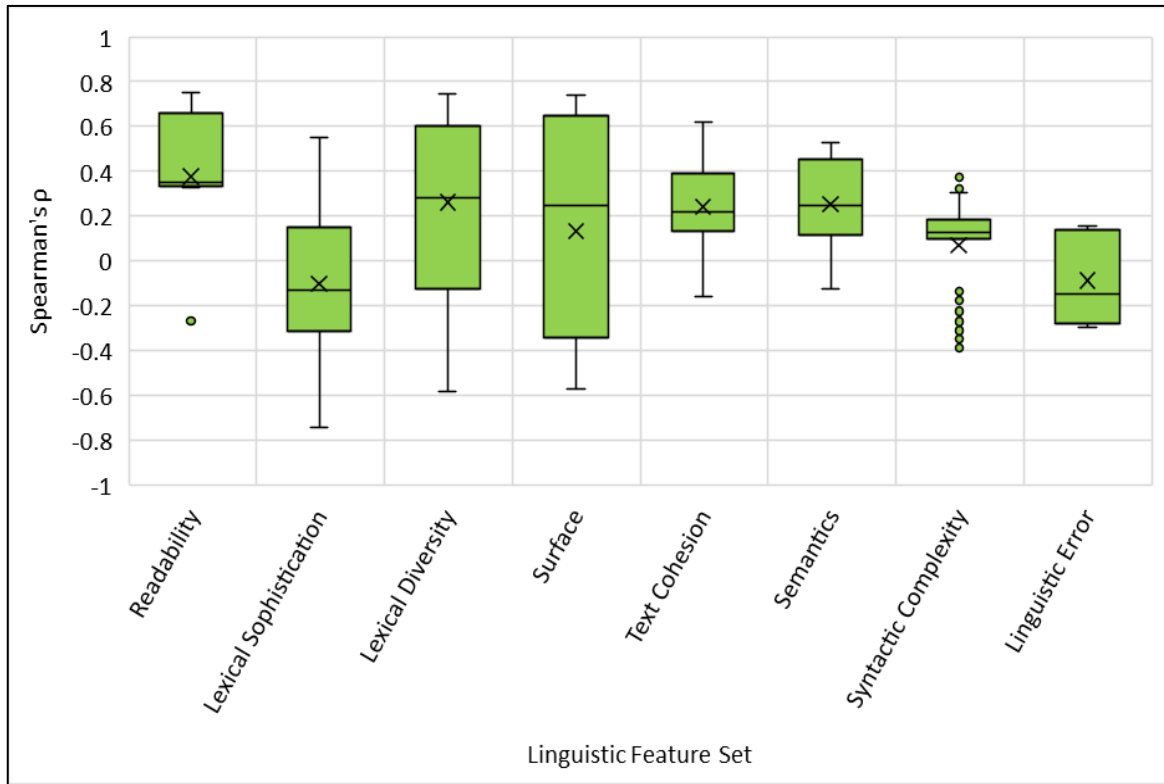


Figure 5.14: Correlation between Comprehensive Linguistic Feature Set and Essay Scores

5.4 Comprehensive Linguistic Feature Category Evaluation

Figure 5.15 depicts the dispersion of Spearman's ρ value based on the comprehensive linguistic feature categories investigated in this research. The boxplot renders the complete “linguistic category spectrum” sorted by its correlation value. This innovative visualization model can be adopted for other linguistics feature analysis tasks in future work. The plot demonstrates the main results of this research as described below.

- i. The crowdsourced readability feature categories (represented by CARES and CAREC indices) exhibit a strong correlation with essay scores, demonstrating positive correlation values exceeding 0.7. This finding is novel and of utmost

importance which is worth further investigation as there is a lack of research in leveraging readability indices for essay scoring.

- ii. The lexical raw count, length-based features, and advanced parts of speech also strongly correlate essay scores, exhibiting positive correlation values exceeding 0.7 in some of their feature indices. The identification of advanced parts of speech as highly correlated feature represents a novel and unique finding in this research.
- iii. A specific case of word frequency and word range feature indices derived from COCA and TOEFL11 corpus show a high negative correlation with essay scores.
- iv. The moderate-correlation feature categories (with a correlation magnitude range within 0.5 to 0.7) include global cohesion, root TTR, simple TTR, short-paragraph, lexical density, age of acquisition/exposure, and global semantics.

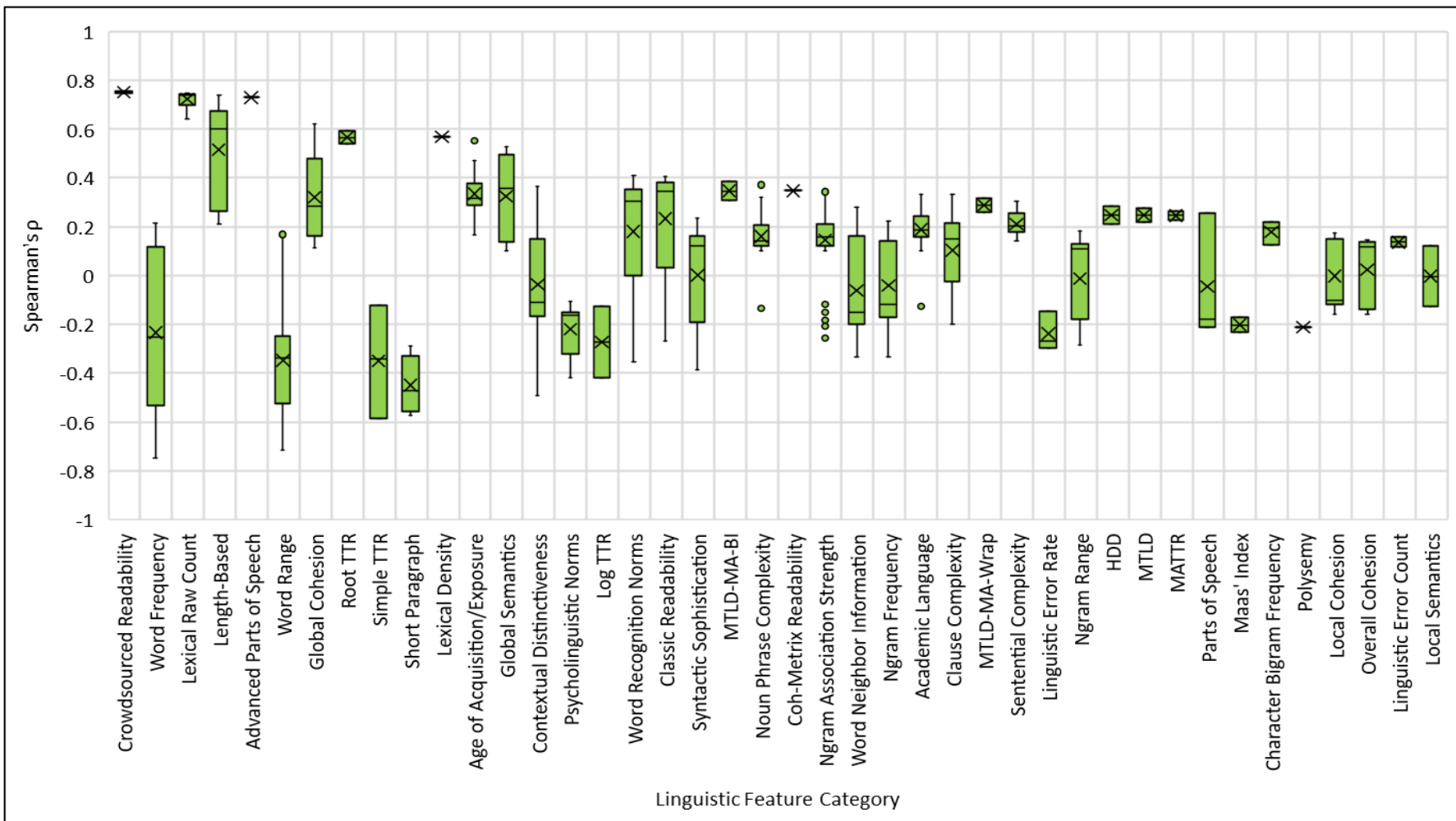


Figure 5.15: Correlation between Comprehensive Linguistic Feature Category and Essay Scores

5.5 Feature Selection

Table 5.2 shows the number of linguistic feature indices used in every feature selection scheme. Based on the 1,709 linguistic feature indices extracted from the MUET dataset, features that were not statistically significant ($p\text{-value} > 0.05$) were discarded. The remaining 1,176 features were further filtered by removing the small effect features that did not reach an absolute correlation threshold of Spearman's $\rho \geq 0.1$. The remaining 1001 features were then selected based on different feature selection schemes (either Correlation Rank or MRMR) to produce the final number of features for the automated essay scoring classifier. The complete listing and description of the 1,709 linguistic feature indices in this research are available for reference in “1_Features_Specification.xlsx”, as specified in Appendix B, with a hyperlink to the directory containing the file.

Table 5.2: Number of Linguistic Feature Indices Used by Different Feature Selection Schemes

Feature Selection Scheme	Number of Feature Indices
No Feature Selection	1,709
No Feature Selection ($p\text{-value} \leq 0.05$)	1,176
No Feature Selection ($p\text{-value} \leq 0.05$ AND $ \text{Spearman's } \rho \geq 0.1$)	1,001
COR-03	75
COR-05	36
COR-07	14
MRMR-10	10
MRMR-25	25
MRMR-50	50
MRMR-100	100

Table 5.3 and Table 5.4 comprise the list of the selected feature indices based on the COR-07 and MRMR-10 feature selection schemes. The COR-07 scheme selected fourteen high-correlated feature indices from the surface, lexical diversity, lexical sophistication, and readability feature set. The MRMR-10 scheme selected ten optimal feature indices based on their underlying mechanism, covering text cohesion, lexical diversity, lexical sophistication, syntactic complexity, and readability feature set. Due to the lengthy feature indices list, the other two listings of the selected features based on the COR-05, MRMR-25, and baseline schemes are attached in Appendices C, D, and E. The linguistic feature indices selected by the remaining feature selection schemes in this research (COR-03, MRMR-50, and MRMR-100) are available for reference in “4_Features_Selected.xlsx”, as specified in Appendix B, with a hyperlink to the directory containing the file.

Table 5.3: Linguistic Feature Indices Selected by COR-07 Feature Selection

Feature Set	Feature Category	Feature Indices
Surface	Length-Based	Word Count
Surface	Advanced Parts of Speech	Advanced Parts of Speech
Lexical Diversity	Lexical Raw Count	Number of Content Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Content Word Types
Lexical Diversity	Lexical Raw Count	Number of Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Word Types
Lexical Diversity	Lexical Raw Count	Number of Function Word Tokens
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Magazine)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Newspaper)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (TOEFL11 All Words)

Table 5.3 continued

Feature Set	Feature Category	Feature Indices
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Academic)
Lexical Sophistication	Word Range	Mean Lemma Type Range Score (TOEFL11 All Words)
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Speed (CARES)
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Comprehension (CAREC)

Table 5.4: Linguistic Feature Indices Selected by MRMR-10 Feature Selection

Feature Set	Feature Category	Feature Indices
Text Cohesion	Global Cohesion	Number of paragraphs with any function lemma overlaps with the next paragraph
Lexical Diversity	Lexical Raw Count	Number of Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Word Types
Lexical Diversity	Lexical Raw Count	Number of Content Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Content Word Types
Lexical Diversity	Lexical Raw Count	Number of Function Word Tokens
Lexical Sophistication	N-gram Association Strength	Trigram Bigram to Unigram Association Strength (COCA Academic)
Syntactic Complexity	Syntactic Sophistication	Average Delta p Score Construction (Cue) - Verb (Outcome) (Types only) - COCA Magazine
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Comprehension (CAREC)
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Speed (CARES)

5.6 Performance of Automated Essay Scoring Using Linguistic Features

This section discusses the performance of the proposed Automated Essay Evaluator (AEE) in essay scoring. During the experiment, the optimal linguistic features were identified by the Correlation Rank and MRMR Feature Selection Scheme. Subsequently, these selected features were evaluated with six different machine learning classifiers, namely Naive Bayes, Generalized Linear Model, Random Forest, Gradient Boosted Tree, Multilayer Perceptron, and Support Vector Machine to score essays. The AEE performances were evaluated using QWK and accuracy as the primary evaluation metrics. The other evaluation metrics such as precision, recall, and F1 score are provided when necessary. The discussion of the AEE performance is organised into the following five sub-sections, where sub-sections 5.6.1, 5.6.2, and 5.6.3 are the main results of this research.

5.6.1 Effectiveness of Linguistic Features in Predicting Essay Score

This section examines the effectiveness of linguistic features in essay scoring. It accomplishes this by comparing the classification performance of the selected linguistic features by the Correlation Rank and MRMR Feature Selection schemes. Table 5.5 shows the mean value of each evaluation metric for all the proposed schemes. These mean values were obtained by averaging the performance of each feature selection scheme across the six machine learning classifiers.

Table 5.5: Performance of Baseline, Correlation Rank, and MRMR Feature Selection Schemes

Scheme	QWK	Accuracy	Precision	Recall	F1 Score
Baseline	0.34	0.54	0.47	0.54	0.48
COR-03	0.57	0.62	0.59	0.62	0.59
COR-05	0.63	0.64	0.63	0.64	0.63

Table 5.5 continued

Scheme	QWK	Accuracy	Precision	Recall	F1 Score
COR-07	0.64	0.65	0.64	0.65	0.64
MRMR-10	0.63	0.66	0.64	0.66	0.64
MRMR-25	0.64	0.66	0.64	0.66	0.64
MRMR-50	0.62	0.64	0.63	0.64	0.63
MRMR-100	0.60	0.62	0.62	0.63	0.61

To compare and visualize the results, several column charts are plotted in Figure 5.16, Figure 5.17, and Figure 5.18 to present the QWK, accuracy, precision, recall, and F1 score of each feature selection scheme. As exhibited in the column charts:

- i. Overall, the proposed Correlation Rank and MRMR Feature Selection Schemes (except COR-03 and baseline scheme) achieved an approximate score of 0.60+ for all the evaluation metrics. Compared to the AES vendors' performance (as shown in Table 2.9), the proposed schemes attain a similar level of accuracy value with most AES software (around 0.60+ in AES vendors) but a slightly lower score in the QWK value (around 0.70+ in AES vendors).
- ii. The proposed Correlation Rank and MRMR schemes outperformed the baseline scheme by a substantial lead. The baseline scheme obtained only 0.34 and 0.54 in QWK and accuracy values. The results denote that the Correlation Rank and MRMR schemes contribute to improving the AEE performance through the selection of optimal features.
- iii. Among all the proposed schemes, the COR-07, MRMR-10, and MRMR-25 schemes exhibit a slightly better result of QWK and accuracy values ranging from 0.63 to 0.66 (as depicted in Figure 5.16). These feature selection schemes require only 14, 10, and

25 features respectively from the 1,709 linguistic features studied. The selection of a smaller number of features offers the advantages of simplicity and reduced computational cost.

- iv. The Correlation Rank and MRMR Feature Selection Scheme demonstrate similar performance results. These outcomes are owing to their consistent selection of similar features. As a result, their prediction performance is nearly the same because they leverage similar features to score essays. Consider a comparison between COR-07, MRMR-10, and MRMR-25 Feature Selection Scheme, Table 5.6 presents the linguistic feature indices selected using these schemes. It shows a significant overlap of the same feature indices being selected (denoted by yellow highlighting for two overlaps and light red for three overlaps). Description of the feature indices in Table 5.6 can be referred to “1_Features Specification.xls” specified in Appendix B.

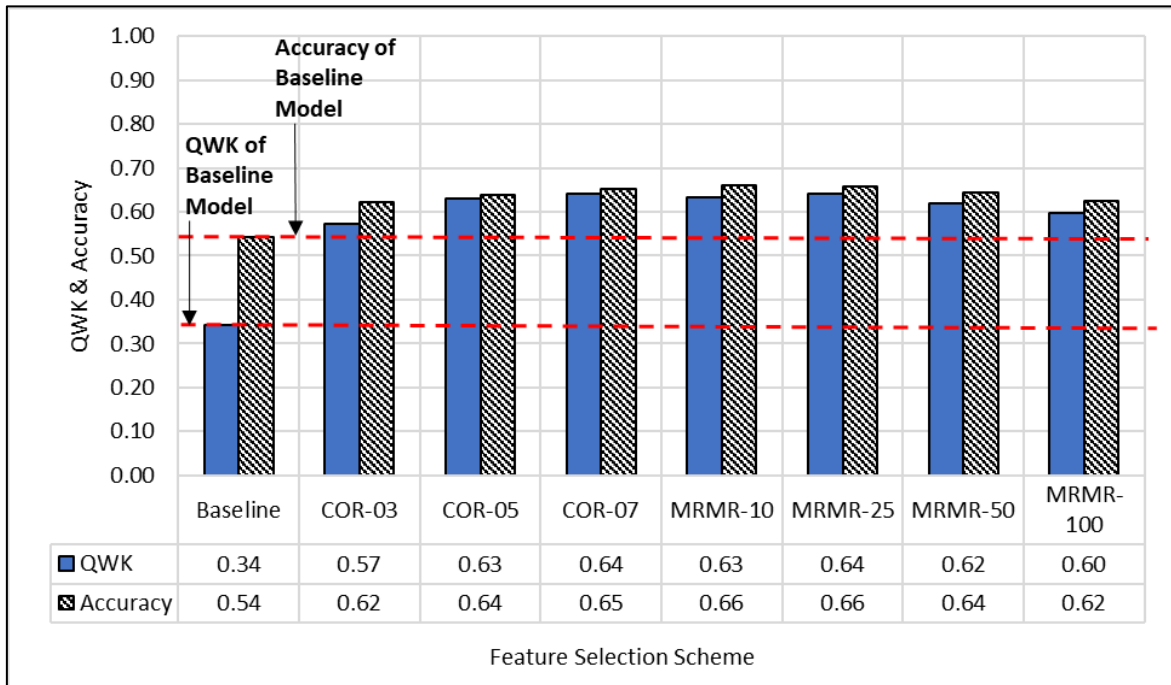


Figure 5.16: QWK and Accuracy of Baseline, Correlation Rank and MRMR Feature Selection

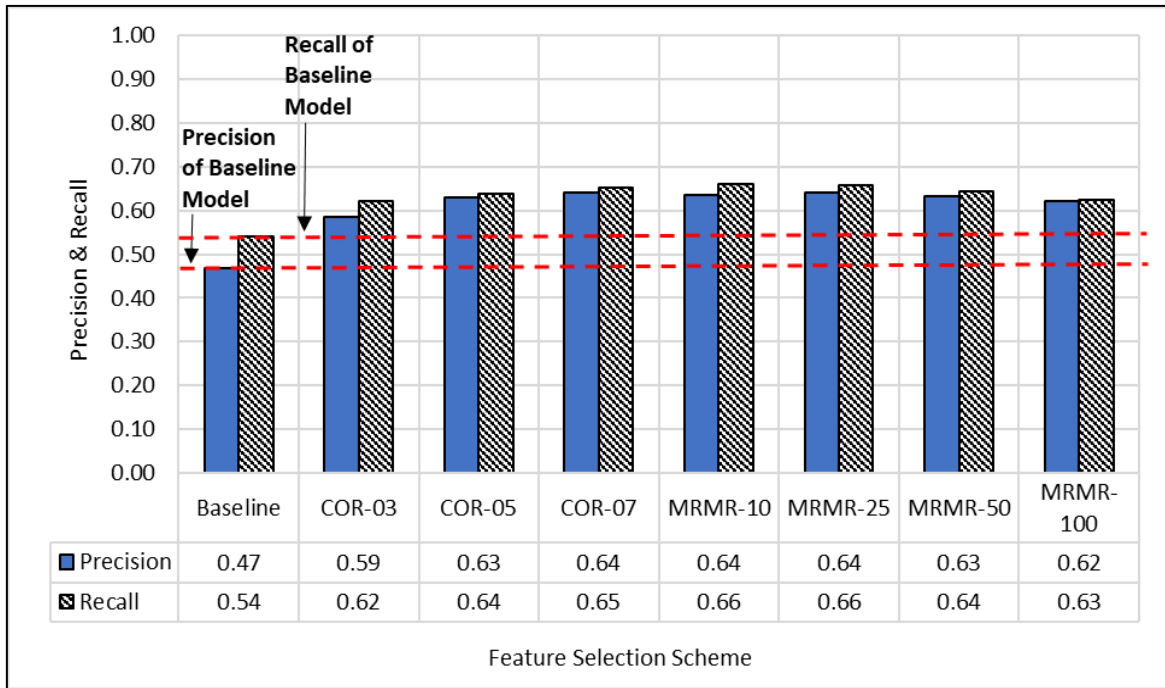


Figure 5.17: Precision and Recall of Baseline, Correlation Rank and MRMR Feature Selection

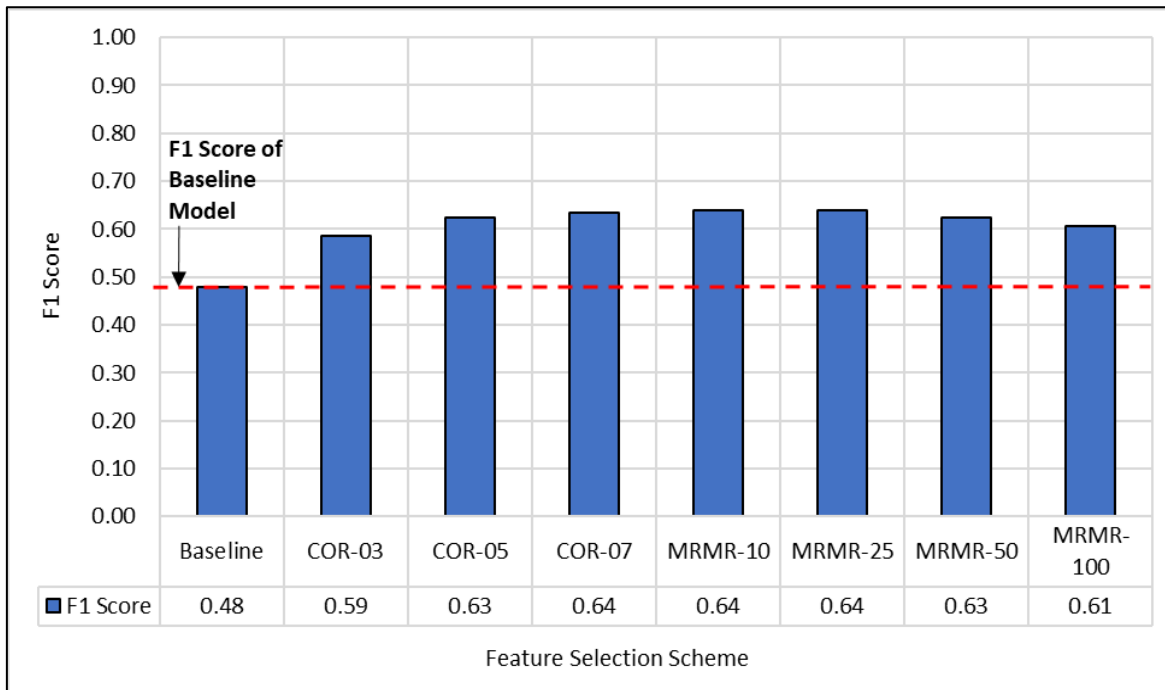


Figure 5.18: F1 Score of Baseline, Correlation Rank, and MRMR Feature Selection

Table 5.6: Similar Features Indices Selected by COR-07, MRMR-10, and MRMR-25 Feature Selection

COR-07	MRMR-10	MRMR-25
numberwords (2)	adjacent_overlap_binary_fw_para (2)	Short1
AdvancePOSCount (2)	basic_ntokens (3)	numberwords (2)
basic_ncontent_tokens (3)	basic_ntypes (3)	AdvancePOSCount (2)
basic_ncontent_types (3)	basic_ncontent_tokens (3)	adjacent_overlap_binary_fw_para (2)
basic_ntokens (3)	basic_ncontent_types (3)	adjacent_overlap_2_fw_para
basic_ntypes (3)	basic_nfunction_tokens (3)	basic_ntokens (3)
basic_nfunction_tokens (3)	COCA_academic_tri_2_MI (2)	basic_ntypes (3)
COCA_magazine_lemma_frequency_AW_TP (2)	mag_av_delta_p_const_cue_type (2)	basic_ncontent_tokens (3)
COCA_newspaper_lemma_frequency_AW_TP (2)	CAREC (3)	basic_ncontent_types (3)
NNS_Lemma_Freq_Types_WC_AW (2)	CARES (3)	basic_nfunction_tokens (3)
COCA_academic_lemma_frequency_AW_TP (2)		basic_nfunction_types
NNS_Lemma_Range_Types_WC_AW		simple_ttr_fw
CARES (3)		hdd42_fw
CAREC (3)		BNC_Written_Trigram_Freq_Normed_Log
		COCA_academic_tri_2_MI (2)
		COCA_academic_lemma_frequency_AW_TP (2)
		COCA_magazine_lemma_frequency_AW_TP (2)
		COCA_newspaper_lemma_frequency_AW_TP (2)
		COCA_spoken_lemma_frequency_AW_TP
		COCA_spoken_lemma_frequency_Log_AW_TP
		NNS_Lemma_Freq_Types_High_AW
		NNS_Lemma_Freq_Types_WC_AW (2)
		mag_av_delta_p_const_cue_type (2)
		CAREC
		CARES

 Features selected by two feature selection schemes

 Features selected by three feature selection schemes

5.6.2 Optimal Linguistic Features Selected by Feature Selection Schemes

Apart from the performance metrics described in Section 5.6.1, this section discusses the linguistic features indices selected by the Correlation Rank and MRMR Feature Selection schemes. The selected features are employed by the proposed AEE models to predict essay grades, yielding results reported in Section 5.6.1. The complete listing of the selected linguistic features can be referred to Table 5.3 (COR-07), Table 5.4 (MRMR-10), Appendix C (COR-05), Appendix D (MRMR-25), and Appendix E (Baseline Scheme). Due to the excessive length of the feature list, the features list based on each feature selection scheme, including the lengthy COR-03, MRMR-50, and MRMR-100, can be accessed via the hyperlink provided in Appendix B.

Table 5.7 shows the most frequent features selected by the COR-05, COR-07, MRMR-10, MRMR-25, and MRMR-50, which are the five highest-ranking effective feature selection schemes (as depicted in Figure 5.16). The AEE can adopt these consistently selected features as the optimal feature set for essay scoring. Table 5.7 indicates that:

- i. The lexical diversity features (number of word tokens, word types, etc.), and the readability features (CAREC, CARES) are consistently selected across all five feature selection schemes.
- ii. Subsequently, the lexical sophistication features (word frequency from COCA and TOEFL11 corpus) constitute another group of features that are consistently selected across four feature selection schemes.
- iii. Furthermore, the surface features (word count, Advanced POS), specific text cohesion features (paragraph count with function lemma overlap with next paragraph), lexical diversity features (function word type, function word TTR),

lexical sophistication features (n-gram association strength, word range based on TOEFL11 corpus), and syntactic complexity features (syntactic sophistication measure) are the other frequently selected features, each with a selection frequency of 3.

- iv. Among the selected features, the readability feature indices of CARES and CAREC and the surface feature of Advanced POS are novel features identified in this research for essay scoring, which is rarely found in existing AES literature.

Table 5.7: Linguistic Feature Indices Frequently Selected by the COR-05, COR-07, MRMR-10, MRMR-25, and MRMR-50 Feature Selection

Feature Set	Feature Category	Feature Description	Frequency of Selection
Lexical Diversity	Lexical Raw Count	Number of word tokens	5
Lexical Diversity	Lexical Raw Count	Number of word types	5
Lexical Diversity	Lexical Raw Count	Number of content word tokens	5
Lexical Diversity	Lexical Raw Count	Number of content word types	5
Lexical Diversity	Lexical Raw Count	Number of function word tokens	5
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Comprehension	5
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Speed	5
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Academic)	4
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Magazine)	4
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Newspaper)	4

Table 5.7 continued

Feature Set	Feature Category	Feature Description	Frequency of Selection
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (TOEFL11 All Words)	4
Surface Features	Length-based	Word Count	3
Surface Features	Advanced Parts of Speech	Advanced Parts of Speech	3
Text Cohesion	Global Cohesion	Number of paragraphs with any function lemma overlaps with the next paragraph	3
Lexical Diversity	Lexical Raw Count	Number of function word types	3
Lexical Diversity	Simple TTR	TTR of function words	3
Lexical Sophistication	N-gram Association Strength	Trigram Bigram to Unigram Association Strength (COCA Academic)	3
Lexical Sophistication	Words Range	Mean Lemma Type Range Score (TOEFL11 All Words)	3
Syntactic Complexity	Syntactic Sophistication	Average Delta p Score Construction (Cue) - Verb (Outcome) (Types only) - COCA Magazine	3

5.6.3 Comparison of AEE Performance with LightSide AES

This section evaluates the performance of the proposed AEE with the benchmark vendor model, LightSide AES (LightSide, 2019). The evaluation includes a comparison of the QWK value and classification accuracy achieved by the Correlation Rank and MRMR Feature Selection Schemes with those of the Lightside_Unigram and Lightside_Composite Schemes (specified in Section 4.5.2). Figure 5.19 presents the relevant QWK and accuracy values plotted in a column chart. The chart renders the following observation:

- i. The Correlation Rank and MRMR schemes proposed in this research demonstrate a substantial performance advantage over the LightSide AES. The LightSide_Unigram and LightSide_Composite features show notably low QWK scores of 0.39 and 0.33. Similarly, their prediction accuracy is low with values of 0.58 and 0.57, respectively.
- ii. The performance of LightSide AES to score the MUET essay dataset in this experiment was notably lower compared to the results reported by Shermis (2014). As shown in Table 2.9, Shermis (2014) utilized the LightSide AES to score the Automated Student Assessment Prize (ASAP) Essay Dataset achieving QWK and accuracy of 0.74 and 0.62, respectively. The observed differences in performance could be attributed to the ASAP dataset, which comprises essays written in native US English. LightSide AES may be associated with the native US English assessment context, and not construct-relevance in scoring Malaysian MUET essays. Furthermore, the number of essays used to train the LightSide model in Shermis' paper consists of 12,978 essays, which is significantly larger than the sample size used in this experiment (724 essays). Employing substantially larger training samples tends to develop a more effective and reliable AES classifier.

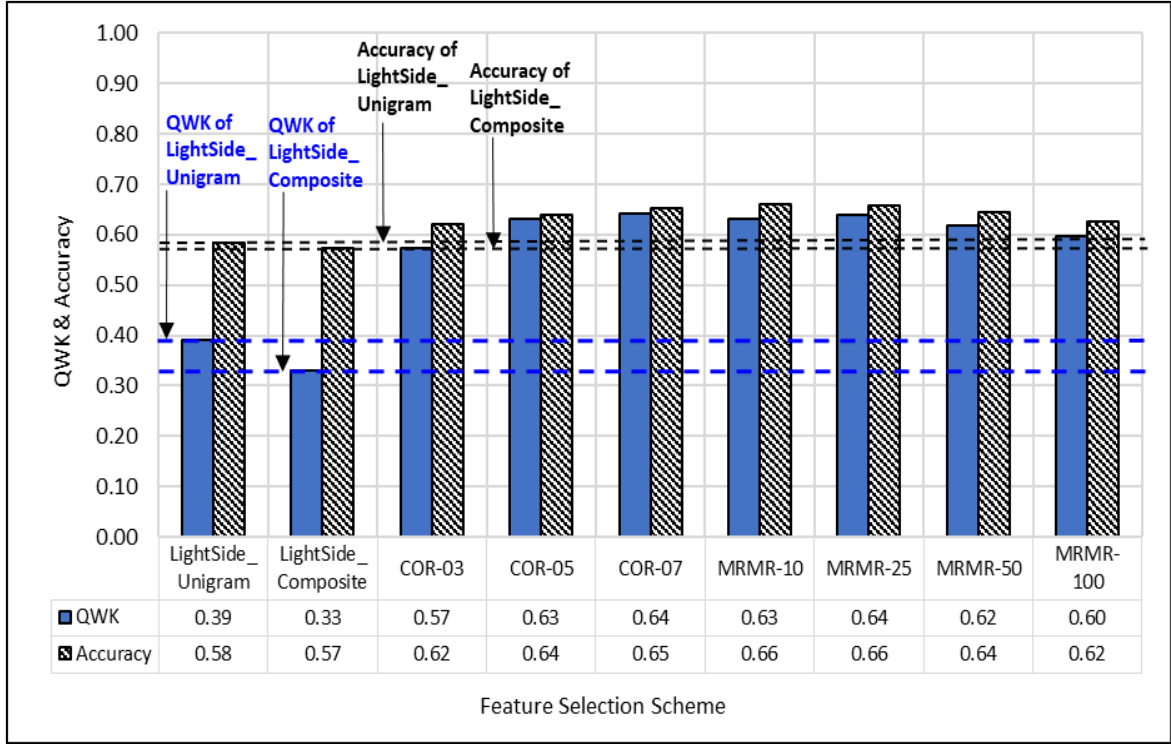


Figure 5.19: Comparison of QWK and Accuracy of Correlation Rank and MRMR Feature Selection with LightSide AES Features

5.6.4 Comparison of AEE Models with the Highest and Lowest Performance

This section analyses the performance of the proposed Correlation Rank and MRMR Feature Selection Scheme with different machine learning classifiers. In this research, eight feature selection schemes (COR-03, COR-05, COR-07, MRMR-10, MRMR-25, MRMR-50, MRMR-100, and baseline scheme) were employed to run on six machine learning algorithms (Naive Bayes, Naive Bayes, Generalized Linear Model, Random Forest, Gradient Boosted Tree, Multilayer Perceptron, and Support Vector Machine). The combination results in a total of 48 AEE models. The performance result of these 48 AEE models with their QWK, accuracy, precision, recall, and F1 score is provided in Appendix F.

To analyse the performance of the 48 AEE models, the top ten AEE models with the highest QWK and accuracy are plotted as clustered column charts in Figure 5.20. The chart reveals the following findings:

- i. The top ten AEE models predominantly consist of the following feature selection schemes: MRMR-25 (appearing three times), MRMR-50 (appearing two times), COR-07 (appearing two times), and MRMR-50 (appearing two times).
- ii. The most effective machine learning models for essay scoring in this experiment are Generalized Linear Model (appearing four times) followed by the Gradient Boost Tree (appearing three times).

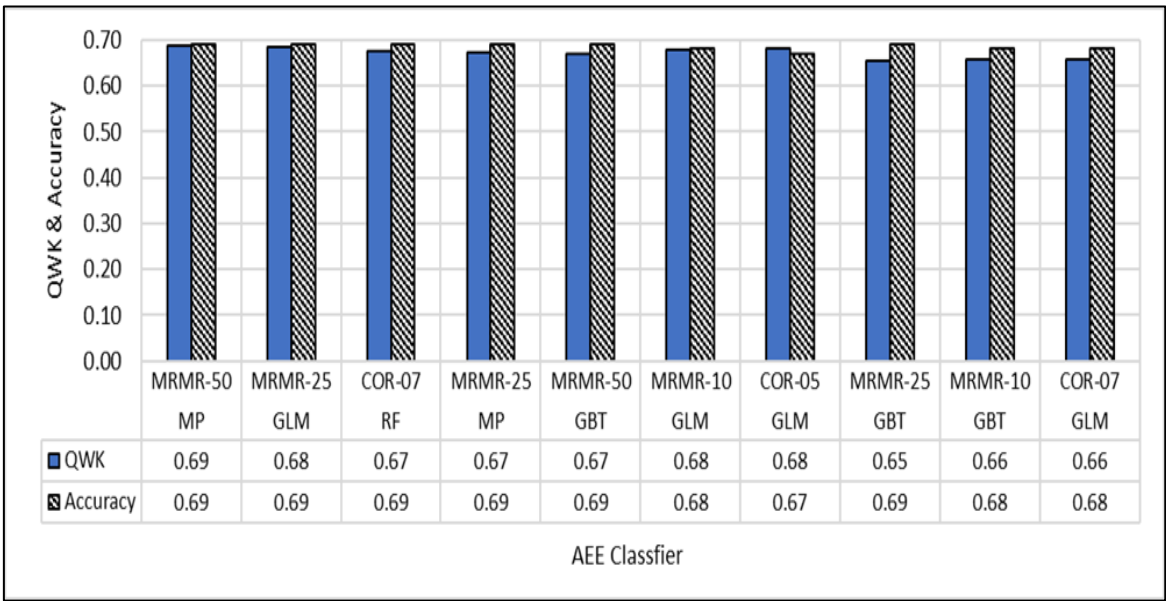


Figure 5.20: AEE Models with High QWK and Accuracy

On the contrary, Figure 5.21 illustrates the ten AEE models exhibiting the lowest QWK and accuracy value. The chart provides the following insights:

- i. Most low-ranked AEE models are associated with the baseline scheme (appearing six times in the list). The baseline scheme consists of 50 randomly selected linguistic features.
- ii. By excluding the baseline features, no machine learning model performs significantly poorly in scoring essays in this experiment. Nonetheless, the Random Forest and Support Vector Machine algorithms exhibit low QWK scores under the COR-03 and MRMR-100 schemes.

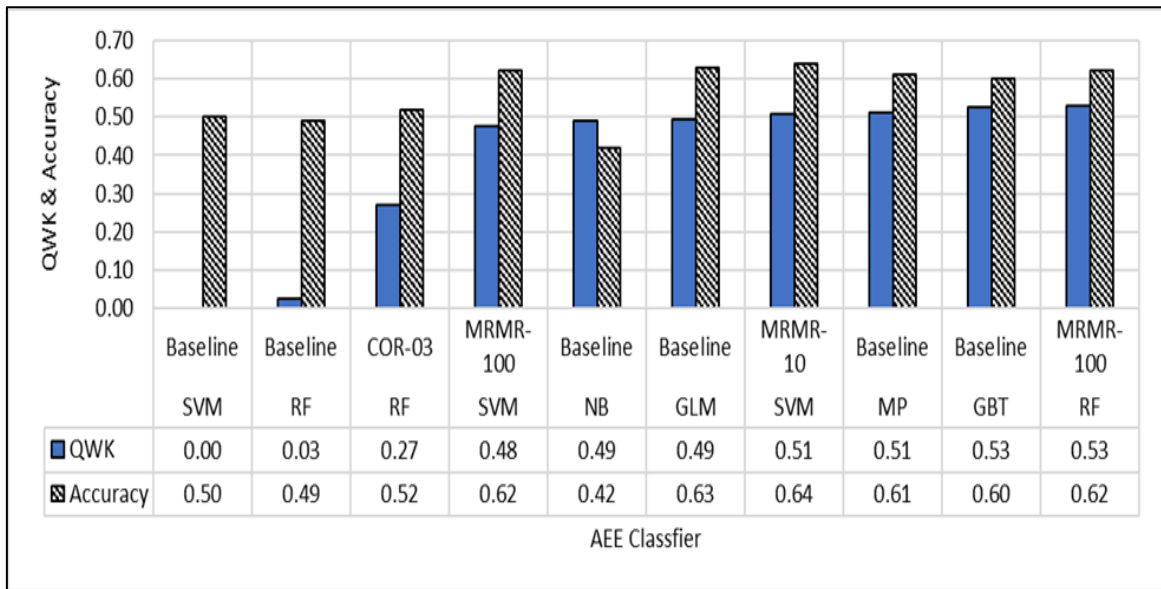


Figure 5.21: AEE Models with Low QWK and Accuracy

We noticed an unexpected result in Figure 5.21, where the Support Vector Machine and Random Forest demonstrate inferior performance under specific feature selection schemes. In many machine learning research, Support Vector Machines and Random Forest are typically expected to outperform simpler classifiers like Naive Bayes and Generalized Linear Models. The author attempts to explain the results as follows:

- i. The Support Vector Machine classifier exhibits poor performance with features selected by MRMR-100 (QWK = 0.48) and MRMR-10 schemes (QWK = 0.51). Table 5.8 and Table 5.9 show the confusion matrix of the Support Vector Machine using the MRMR-100 and MRMR-10 features. The tables underscore a notable bias in the SVM classifier towards the majority classes of Band-3 and Band-4 (highlighted by yellow-coloured cells), resulting in misclassifications of all other classes (highlighted by orange-coloured cells). This bias issue is common when Support Vector Machines are applied to imbalanced datasets. The issue is well-documented in the literature by Tao et al. (2019) and Piao et al. (2023). To mitigate this problem, two potential approaches can be considered: data-level resampling such as SMOTE (detailed in Section 6.5.2), or algorithm-level enhancement. The algorithm-based solution involves integrating the SVM into boosting algorithms, such as Adaptive Boosting (AdaBoost), to enhance classifier performance (Tao et al., 2019; Piao et al., 2023). Boosting algorithms assign greater weights to misclassified instances in each iteration, increasing the likelihood of correct classification in subsequent iterations.

Table 5.8: Confusion Matrix - Support Vector Machine with MRMR-100 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	0	0	4	0	0	0	0.00
True 2	0	0	40	1	0	0	0.00
True 3	0	0	287	73	0	0	0.80
True 4	0	0	92	165	0	0	0.64
True 5	0	0	5	51	0	0	0.00
True 6	0	0	0	6	0	0	0.00
Precision	0	0.00	0.67	0.56	0.00	0.00	
QWK: 0.48 Accuracy: 0.62 Adjacent Accuracy: 0.98							

Table 5.9: Confusion Matrix - Support Vector Machine with MRMR-10 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	0	0	4	0	0	0	0.00
True 2	0	0	40	1	0	0	0.00
True 3	0	0	301	59	0	0	0.84
True 4	0	0	96	161	0	0	0.63
True 5	0	0	2	54	0	0	0.00
True 6	0	0	0	6	0	0	0.00
Precision	0.00	0.00	0.68	0.57	0.00	0.00	0
QWK: 0.51 Accuracy: 0.64 Adjacent Accuracy: 0.98							

- ii. The Random Forest classifier performed poorly in features selected by COR-03 schemes (QWK=0.27, accuracy = 0.52). Table 5.10 shows the confusion matrix of the Random Forest classifier employing COR-03 features. It shows that the Random Forest classifier is unable to detect a particular class of Band-4 but misclassifies them mostly as Band-3 (highlighted by orange-coloured cells), which leads to inferior performance results. The author suspects that the results may be due to irrelevant features and multicollinearity between the extracted features introduced by COR-03 feature selection. COR-03 has the highest number of features (75 features), compared with COR-05 (36 features), and COR-07 (14 features). Furthermore, it does not consider multicollinearity as performed by MRMR feature selection. These highly correlated features may reduce the diversity of trees, causing difficulties in detecting certain classes and thus resulting in poor performance. Random Forest's performance improved with a smaller number but more correlated features, such as COR-07 scheme (QWK=0.67, accuracy = 0.69), and MRMR-10 (scheme (QWK=0.62, accuracy = 0.66). Table 5.11 and Table 5.12 show the confusion matrix of Random Forest, employing the COR-07 and MRMR-10 feature selection scheme. Compared

with COR-03 features, the Random Forest classifier exhibits its ability to identify Band-4 essays using the COR-07 and MRMR-10 features (highlighted by yellow-coloured cells).

Table 5.10: Confusion Matrix - Random Forest with COR-03 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	3	1	0	0	0	0	0.75
True 2	0	14	27	0	0	0	0.34
True 3	0	5	355	0	0	0	0.99
True 4	0	1	253	0	3	0	0.00
True 5	0	0	48	0	5	3	0.09
True 6	0	0	2	0	3	1	0.17
Precision	1.00	0.67	0.52	0.00	0.45	0.25	
QWK: 0.27 Accuracy: 0.52 Adjacent Accuracy: 0.93							

Table 5.11: Confusion Matrix - Random Forest with COR-07 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	3	1	0	0	0	0	0.75
True 2	1	15	25	0	0	0	0.37
True 3	0	6	268	84	2	0	0.74
True 4	0	0	55	197	5	0	0.77
True 5	0	0	1	42	11	2	0.20
True 6	0	0	0	2	2	2	0.33
Precision	0.75	0.68	0.77	0.61	0.55	0.50	
QWK: 0.67 Accuracy: 0.69 Adjacent Accuracy: 0.99							

Table 5.12: Confusion Matrix – Random Forest with MRMR-100 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	3	1	0	0	0	0	0.75
True 2	1	16	24	0	0	0	0.39
True 3	0	5	313	40	2	0	0.87
True 4	0	0	114	138	5	0	0.54
True 5	0	0	6	40	6	4	0.11
True 6	0	0	0	2	3	1	0.17
Precision	0.75	0.73	0.68	0.63	0.38	0.20	
QWK: 0.62 Accuracy: 0.66 Adjacent Accuracy: 0.99							

5.6.5 Evaluation of the Scoring of Different Essay Grades

This section analyses the scoring of different essay grades by using confusion matrices. As specified in Table 3.2, the distribution of grades in the MUET essay dataset is imbalanced, with Band-1 (0.5%), Band-2 (5.7%), Band-3 (49.7%), Band-4 (35.5%), Band-5 (7.7%), and Band-6 (0.8%). The confusion matrix results provide a detailed breakdown of the AEE's classification performance for each essay grade.

In this research, 48 AEE models were created, resulting in 48 confusion matrices. For the sake of simplicity and ease of discussion, the Generalized Linear Model was selected for the analysis of these confusion matrices in this section. Furthermore, the selection was made considering the Generalized Linear Model's consistent effectiveness in scoring essays across different feature selection schemes. Table 5.13 to Table 5.20 presents eight confusion matrices corresponding to the eight feature selection schemes applied to the Generalized Linear Model classifier. These confusion matrices show the number of essays accurately classified for each grade, with the relevant precision and recall values. As shown in the eight confusion matrices:

- i. The classifier (excludes the baseline model) demonstrates accurate essay grade prediction in most cases (indicated by the grey-coloured diagonal cells). However, there are instances where the classifier may misclassify with one-class discrepancy (indicated by the yellow-coloured cells) and occasionally two-class discrepancy misclassification (indicated by orange-coloured cells).
- ii. In general, the proposed correlation rank and MRMR feature selection schemes produce satisfactory classification results, with QWK scores ranging from 0.59 to 0.68, accuracies between 0.66 and 0.69, and adjacent accuracies from 0.99 to 1.

Table 5.13: Confusion Matrix - Generalized Linear Model with COR-03 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	0	2	2	0	0	0	0.00
True 2	0	6	35	0	0	0	0.15
True 3	0	0	308	52	0	0	0.86
True 4	0	0	74	182	1	0	0.71
True 5	0	0	2	50	4	0	0.07
True 6	0	0	0	3	3	0	0.00
Precision	0.00	0.75	0.73	0.63	0.50	0.00	
QWK: 0.62 Accuracy: 0.69 Adjacent Accuracy: 0.99							

Table 5.14: Confusion Matrix - Generalized Linear Model with COR-05 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	2	2	0	0	0	0	0.50
True 2	1	13	27	0	0	0	0.32
True 3	0	11	288	61	0	0	0.80
True 4	0	0	77	164	14	2	0.64
True 5	0	0	2	33	18	3	0.32
True 6	0	0	0	1	4	1	0.17
Precision	0.67	0.50	0.73	0.63	0.50	0.17	
QWK: 0.68 Accuracy: 0.67 Adjacent Accuracy: 0.99							

Table 5.15: Confusion Matrix - Generalized Linear Model with COR-07 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	2	2	0	0	0	0	0.50
True 2	1	10	30	0	0	0	0.24
True 3	0	3	297	60	0	0	0.82
True 4	0	0	84	167	5	1	0.65
True 5	0	0	2	38	15	1	0.27
True 6	0	0	0	2	3	1	0.17
Precision	0.67	0.67	0.72	0.63	0.65	0.33	
QWK: 0.66 Accuracy: 0.68 Adjacent Accuracy: 0.99							

Table 5.16: Confusion Matrix - Generalized Linear Model with MRMR-10 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	1	3	0	0	0	0	0.25
True 2	1	12	28	0	0	0	0.29
True 3	0	5	305	50	0	0	0.85
True 4	0	0	88	158	11	0	0.61
True 5	0	0	2	36	15	3	0.27
True 6	0	0	0	2	1	3	0.50
Precision	0.50	0.60	0.72	0.64	0.56	0.50	
QWK: 0.68 Accuracy: 0.68 Adjacent Accuracy: 0.99							

Table 5.17: Confusion Matrix - Generalized Linear Model with MRMR-25 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	1	3	0	0	0	0	0.25
True 2	2	12	27	0	0	0	0.29
True 3	0	5	296	59	0	0	0.82
True 4	0	0	76	172	9	0	0.67
True 5	0	0	2	38	14	2	0.25
True 6	0	0	0	1	3	2	0.33
Precision	0.33	0.60	0.74	0.64	0.54	0.50	
QWK: 0.68 Accuracy: 0.69 Adjacent Accuracy: 0.99							

Table 5.18: Confusion Matrix - Generalized Linear Model with MRMR-50 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	0	2	2	0	0	0	0.00
True 2	0	2	39	0	0	0	0.05
True 3	0	1	311	48	0	0	0.86
True 4	0	0	93	163	1	0	0.63
True 5	0	0	2	51	3	0	0.05
True 6	0	0	0	2	4	0	0.00
Precision	0.00	0.40	0.70	0.62	0.38	0.00	
QWK: 0.59 Accuracy: 0.66 Adjacent Accuracy: 0.99							

Table 5.19: Confusion Matrix - Generalized Linear Model with MRMR-100 Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	0	2	2	0	0	0	0.00
True 2	0	4	37	0	0	0	0.10
True 3	0	0	311	49	0	0	0.86
True 4	0	0	85	171	1	0	0.67
True 5	0	0	2	50	4	0	0.07
True 6	0	0	0	2	4	0	0.00
Precision	0.00	0.67	0.71	0.63	0.44	0.00	
QWK: 0.61 Accuracy: 0.68 Adjacent Accuracy: 0.99							

Table 5.20: Confusion Matrix - Generalized Linear Model with Baseline Features

	Pred. 1	Pred. 2	Pred. 3	Pred. 4	Pred. 5	Pred. 6	Recall
True 1	0	2	2	0	0	0	0.00
True 2	0	3	35	3	0	0	0.07
True 3	0	2	298	58	2	0	0.83
True 4	0	0	101	156	0	0	0.61
True 5	0	0	4	52	0	0	0.00
True 6	0	0	0	6	0	0	0.00
Precision	0.00	0.43	0.68	0.57	0.00	0.00	
QWK: 0.49 Accuracy: 0.63 Adjacent Accuracy: 0.98							

To comprehend the vast data contained in the confusion matrices, Table 5.21 is devised to present a summary of the mean class precision and recall value for each essay grade with the number of essay samples in the essay grade. These mean values were computed by averaging the class precision and recall across all the proposed feature selection schemes (excluding the baseline feature selection) in the Generalized Linear Model.

Table 5.21: Class Precision and Recall of Each Essay Grade

Grade	Number of Essays	Percentage	Precision	Recall
Band-1	4	0.6	0.31	0.21
Band-2	41	5.7	0.60	0.21
Band-3	360	49.7	0.72	0.84
Band-4	257	35.5	0.63	0.65
Band-5	56	7.7	0.51	0.19
Band-6	6	0.8	0.21	0.17

In addition, a column chart is plotted in Figure 5.22 to visualize the results. As depicted in the column chart:

- i. The highest precision values are observed in the majority classes (Band-3 and Band-4), followed by Band-2 and Band-5. Conversely, the under-represented classes (Band-1 and Band-6) show the lowest precision values.
- ii. Similarly, the majority classes (Band-3 and Band-4) exhibit the highest recall values, while other classes (Band-1, Band-2, Band-5, and Band-6) show lower recall values.
- iii. The obtained results are attributed to the highly imbalanced distribution of the MUET dataset used in this research, which contains only a small number of poorly written Band-1 essays (4 out of 724 essays), and well-written Band-6 essays (6 out of 724

essays). The classifier lacks sufficient representative examples from Band-1 and Band-6 essays to effectively learn their specific patterns. This led to lower class precision and recall values for these under-represented classes.

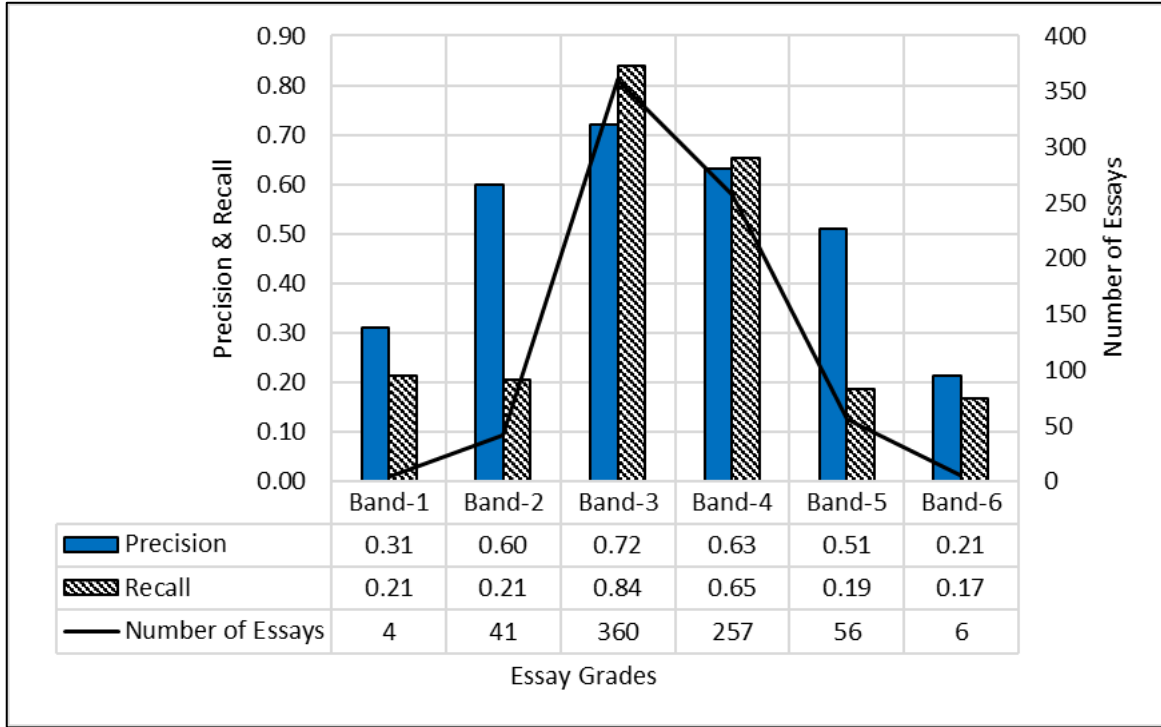


Figure 5.22: Class Precision and Recall of Each Essay Grade

Based on the above results, the formulated AEE faced challenges in the imbalanced MUET essay dataset, leading to a classification bias towards the majority classes. Addressing this imbalance typically involves two methods: an algorithm-driven or data-driven approach (Leevy et al., 2018; Kumar et al., 2021). The algorithm-driven approach maintains the dataset unchanged but adapts the machine learning algorithm to focus on learning tasks related to under-represented classes. This approach often involves ensemble learning techniques like bagging and boosting. Bagging reduces variability in classification by generating multiple training sets from the original dataset, creating a classifier for each

set, and then combining the models to achieve the final classification. Boosting also utilizes multiple training sets but adjusts the weights of each classifier iteratively based on misclassifications to produce the final classification result. The data-driven approach focuses on balancing class distribution through under-sampling and over-sampling of data. The under-sampling approach removes instances from the majority classes, while the over-sampling approach adds instances to the under-represented classes. Both the under-sampling and over-sampling approaches can be performed randomly or via specific algorithms. An example of a widely used algorithmic over-sampling technique is SMOTE, which is further described in Section 6.5.2.

5.7 Chapter Summary

Overall, this research found that readability, lexical diversity, surface feature, and lexical sophistication (limited to its specific indices of word frequency and word range derived from COCA and TOEFL11 corpus) are strong predictors of MUET essay scores. Text cohesion and semantics are moderately correlated with essay scores, while linguistic error and syntactic complexity are weak predictors of essay scores. Furthermore, this research found that readability feature indices (CARES and CAREC) which are rarely used in scoring essays show highly positive correlation coefficients with essay scores and seem to be good predictors of essay scores. The linguistic features selected by the Correlation Rank and MRMR Feature Selection Scheme in this research surpassed the baseline scheme, which involved the selection of 50 random features. Among all the proposed schemes, the COR-07, MRMR-10, and MRMR-25 schemes exhibit a slightly better result with QWK and accuracy values ranging from 0.63 to 0.66. Furthermore, the AEE formulated in this research outperformed the LigthSide AES vendor in scoring MUET essays. This linguistic-based

essay scoring AEE can serve as the basis for developing a complete full-fledged local Malaysian AES system, with the incorporation of essay content features.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

This is the final chapter of the thesis - discussing the overall findings, contributions, implications, and future works derived from this research. Section 6.1 provides a brief summary of this research. Section 6.2 summarizes the findings of this research by answering the research objectives and questions outlined in Chapter 1. Section 6.3 presents the contributions of this research to the existing knowledge of automated essay scoring, while Section 6.4 summarizes the implications of this research to different academic disciplines. Section 6.5 briefly discusses the limitations and potential future works from this research. Lastly, this whole thesis is closed by concluding remarks in Section 6.6.

6.1 Summary of the Research

This research is carried out to provide a systematic and detailed investigation of linguistic features in assessing essay quality. This thesis investigates multiple and comprehensive linguistic feature sets, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability in influencing essay scores. The investigated linguistic features involve a broad scope of features, which should be able to cover most, if not of all the linguistic features of an essay.

To conduct the investigation, related literature has been reviewed to research and identify possible linguistic features affecting writing quality. The identified features have been organized into logical groups of feature sets, feature categories, and feature indices to facilitate systematic analysis. For the provision of the Malaysian English test environment,

this research selected the MUET test with its essay dataset as the use case. An experiment framework has been designed to conduct the analysis and scoring of essays. The experiment involves the extraction of linguistic features from MUET essays, the correlation analysis of linguistic features and essay scores, the feature selection of significant features, and the formulation of multiple Automated Essay Evaluator (AEE) models to score essays. Finally, this research discussed the results of the linguistic feature evaluation and automated essay scoring via boxplot diagrams and other charts.

6.2 Summary and Discussion of Research Findings

This section discusses the research findings based on the research objectives outlined in Chapter 1. A summary of these findings can be found below.

6.2.1 Evaluation of Comprehensive Linguistic Features in Essay Scoring

Research Question (1): What are the potential comprehensive linguistic features that can be employed to score essays?

Based on Spearman's and Kendall's rank correlation coefficient, eight linguistic feature sets were analysed in the aspect of their correlation with MUET essay scores. The results are summarized below.

i. Surface Features

The surface feature set, particularly the length-based features and advanced parts of speech renders a high positive correlation with essay scores. The high correlation of advanced parts of speech with essay scores is one of the novel findings in this research.

ii. Linguistic Errors

The linguistic error feature set either in the aspect of spelling or grammatical errors shows a very small or negligible correlation with essay scores.

iii. Text Cohesion

The text cohesion feature set, particularly the global cohesion shows a moderate positive correlation with essay scores. In contrast, the local cohesion renders a weak negative correlation with essay scores. Both findings align with the result of research in writing assessment.

iv. Semantics

The semantic feature set, particularly the global semantics shows a moderate positive correlation with essay scores. In contrast, the local semantics render a weak negative correlation with essay scores. Both findings align with the result of research in writing assessment.

v. Lexical Diversity

The lexical diversity feature set generally demonstrates a significant correlation with essay scores. In particular, the lexical raw count renders a high positive correlation with essay scores, followed by the root TTR which shows a moderate positive correlation.

vi. Lexical Sophistication

Most of the feature category in the lexical sophistication feature set shows a low to negligible correlation with essay scores. Exception found in the word frequency and word range feature indices derived from the COCA and TOEFL11 corpus, which show a high negative correlation with essay scores.

vii. Syntactic Complexity

The syntactic complexity feature set either in the scope of syntactic sophistication, noun phrase complexity, clause complexity, or sentential complexity shows a low to negligible correlation with essay scores.

viii. Readability

The readability feature set, particularly the CARES and CAREC feature indices exhibits a high positive correlation with essay scores. This novel finding is of utmost importance and indicates potential for further investigation as there is a lack of research in leveraging readability indices for essay scoring.

6.2.2 Feature Selection of Optimal Linguistic Features in Essay Scoring

Research Question (2): Among the comprehensive linguistic features identified in RO (1), what are the optimized linguistic features that can be selected to score essays?

The optimal linguistic features are determined by their correlation strength with essay score and the most frequently selected features across the proposed Correlation Rank and Minimum Redundancy Maximum Relevance (MRMR) Feature Selection Schemes. The proposed AES can adopt these highly correlated and consistently selected features to assess an essay from its linguistic perspective. Table 6.1 presents the optimal linguistic features identified in this research. Among the selected features, the readability feature indices of CARES and CAREC and the surface feature of Advanced POS are novel features identified in this research for essay scoring, which is rarely found in existing AES literature.

Table 6.1: Optimal Linguistic Features Identified in this Research

Linguistic Feature Set	Linguistic Features
Lexical Diversity	Word Tokens, Word Types, TTR
Readability	CAREC, CARES
Surface Features	Word Count, Advanced POS
Lexical Sophistication	Word frequency (COCA and TOEFL11 corpus), Word Range (TOEFL11 corpus), N-gram Association
Text Cohesion	Lemma overlaps between paragraphs
Syntactic Complexity	Syntactic sophistication measure

6.2.3 Performance of the Selected Linguistic Features in Essay Scoring

Research Question (3): What is the performance of the selected linguistic features in predicting essay scores when applied to multiple AEE models?

This research evaluates the performance of the selected linguistic features by using them to score MUET essays. The selected linguistic features based on Correlation Rank and MRMR Feature Selection are applied to six different machine learning classifiers. Their outcomes are then compared to a baseline model and benchmarked against a LightSide AES vendor.

i. Baseline Model

The selected linguistic features in this research significantly outperformed the baseline model, which consists of 50 randomly selected features.

ii. LigthSide AES

The selected linguistic features demonstrated a notable performance advantage over the LigthSide AES upon scoring the same MUET essay dataset in this research.

iii. Overall Performance

In general, the selected linguistic feature achieved an approximate score of 0.60+ for both QWK and accuracy across most AEE models. When compared to the findings in the state-of-the-art AES system literature, the selected linguistic features in this research achieve a comparable level of accuracy to most AES software (typically around 0.60+ in AES vendors) but exhibit a slightly lower score in terms of QWK value when compared to AES vendors (around 0.70+ in AES vendors).

6.2.4 Mapping of MUET Scoring Rubrics with Linguistic Features

Research Question (4): How to map the MUET essay rubrics with linguistic feature sets for explaining the AEE's scoring in this research?

Another objective of this research is to establish an explanation of the proposed linguistic feature set contributing to the scoring rubrics used by MUET essays. It aims to avoid the AES black box nature, as stated in the problem statements, by explaining the formulated AEE linguistic features scoring with the MUET essay rubrics.

The author attempts to perform this analysis by theoretically mapping the MUET essay rubric specifications to the proposed linguistic feature sets. To establish the mapping between these two domains, the author presents Table 6.2 to map the MUET rubric specifications with the essay feature groups and linguistic feature sets formulated in the AEE. The "MUET Rubrics" and "MUET Rubric Specifications" columns in Table 6.2 are directly sourced from MUET essay rubrics (specified in Table 3.1), whereas the "Feature Group" and "AEE Linguistic Feature Set" columns are provided by the author. Furthermore, the highlighted light-yellow rows in the table indicate the MUET essay rubrics that can probably be scored by the proposed linguistic feature set in this research.

- i. According to the "accuracy" rubric specification, it primarily focuses on assessing linguistic errors while also evaluating syntactic complexity through the criterion of "using correct sentence structure". Therefore, this MUET rubric can probably be assessed by the linguistic errors and syntactic complexity feature set proposed in this research.
- ii. The "appropriacy" rubric can be assessed using the lexical diversity and sophistication features ("varied vocabulary", "language appropriate", and "conventions appropriate" in the specification), and syntactic complexity ("varied sentences" in the specification). Furthermore, the readability feature can contribute to scoring this rubric since readability, particularly the CARES and CAREC feature indices encompass multiple aspects of linguistic attributes, such as lexical diversity and sophistication, syntactic complexity, and length-based features.
- iii. As implied by its name, the "coherence and cohesion" rubric can be evaluated using the text cohesion feature. Furthermore, this rubric can be scored by semantic features. In this research, the semantic feature approximates the essay coherence and cohesion using the WordNet synonyms, LSA, LDA, and Word2vec similarity between adjacent sentences and paragraph segments.
- iv. The "use of language functions" rubric specifies using language for comparing and contrasting, giving reasons, presenting arguments, drawing conclusions, etc. These specifications go beyond the scope of linguistic-based scoring and can only be evaluated using content features. The content features reviewed in Section 2.4, such as prompt relevance, thesis clarity, organization & development, and argumentation are well-suit to assess this rubric.

- v. The "task fulfilment" rubric specifies how well an essay fulfils the assigned tasks by presenting relevant ideas, providing adequate content, and treating the topic. The assessment is closely related to content-based assessment. However, this rubric, especially the "providing adequate content" specification, might be approximated and evaluated using surface or length-based features. A particular length of the essay is essential to provide adequate content for ideas presentation and topic treatment. On the other hand, the readability feature set which encompasses various linguistic attributes, including the length-based features, can also be used to assess this rubric.

Table 6.2: Detailed Mapping of MUET Rubric Specifications with AEE Linguistic Feature Set

MUET Rubrics	MUET Rubric Specifications	Feature Group	AEE Linguistic Feature Set
Accuracy	Using correct spelling and mechanics	Linguistic	Linguistic Errors
	Using correct grammar		Linguistic Errors
	Using correct sentence structures		Syntactic Complexity
Appropriacy	Using varied vocabulary & expression	Linguistic	Lexical Diversity, Lexical Sophistication, Readability
	Using clear varied sentences		Syntactic Complexity
	Using language appropriate for intended purpose and audience		Lexical Diversity, Lexical Sophistication, Readability
	Observing conventions appropriate to a specific situation or text type		Lexical Diversity, Lexical Sophistication, Readability
Coherence and Cohesion	Develop and organising ideas	Linguistic & Content	Text Cohesion, Semantics, Readability
	Using appropriate markers and linking devices	Linguistic	Text Cohesion, Semantics, Readability

Table 6.2 continued

MUET Rubrics	MUET Rubric Specifications	Feature Group	AEE Linguistic Feature Set
	Using anaphora appropriately together with cohesive devices	Linguistic	Text Cohesion, Semantics, Readability
Use of Language Functions	Defining, describing, explaining	Content	-
	Comparing and contrasting		
	Classifying		
	Giving reasons		
	Giving opinions		
	Expressing relationships		
	Make suggestions and recommendations		
	Expressing agreement and disagreement		
	Persuading		
	Interpreting information from non-linear texts		
	Drawing conclusion		
	Stating & justifying points of view		
	Presenting an argument		
Task Fulfilment	Presenting relevant ideas	Content	-
	Providing adequate content	Content & Surface	Surface Features, Readability
	Show a mature treatment of topic	Content	-

The author presents Figure 6.1 as a summary of the relationship between the linguistic feature set used in this research and the MUET essay scoring rubrics. Compared with the detailed description in Table 6.2, Figure 6.1 provides a simplified presentation,

displaying only the mapping between MUET essay rubrics and the corresponding linguistic feature set of the AEE. As illustrated in Figure 6.1, the linguistic-based scoring proposed in this research can potentially address three out of five MUET essay rubrics, i.e., the accuracy, appropriacy, and coherence & cohesion (depicted by green rectangles in Figure 6.1). However, one of the MUET essay rubrics, namely the use of language functions, lies outside the scope of linguistic-based scoring (depicted by red rectangles in Figure 6.1). Another task fulfilment rubric can be partially addressed by the surface feature and readability feature (represented as orange rectangle with dotted connector in Figure 6.1). Besides, a significant observation is that readability can be used to assess multiple MUET essay rubrics, including appropriacy, coherence/cohesion, and task fulfilment. This finding may explain why the readability feature set is considered one of the optimal features to score MUET essays, as shown by the result of this research.

In conclusion, this section has explained the proposed linguistic feature set contributing to the MUET essay scoring rubrics. The explanation assumes that certain MUET rubrics can be evaluated by the linguistic features in this AEE, while other rubrics lie beyond the scope of linguistic-based assessment. As a result, the AEE's prediction capability is limited and achieves a "moderate" level of around 65% QWK and accuracy. However, the explanation may have its limitations since it relies upon the description of the MUET rubric specification and lacks empirical results to support its claim. The relevant experiment cannot proceed because the given marks in the MUET dataset are holistic scores. The dataset does not provide analytic scoring for individual rubrics. With the availability of analytic scores, future research can be conducted to investigate the use of each linguistic feature set or combinations of them to score individual MUET rubrics.

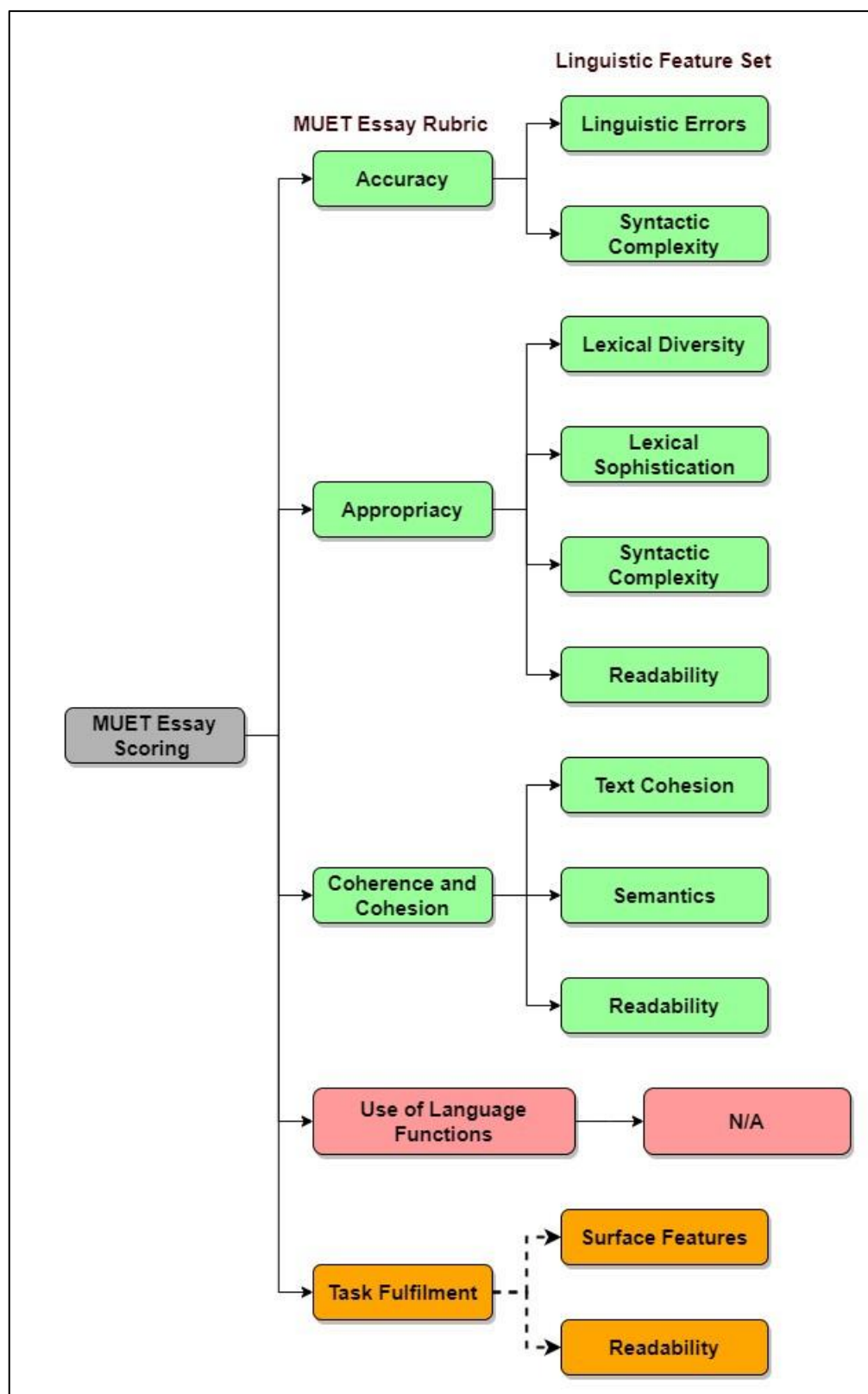


Figure 6.1: High Level Mapping of MUET Rubrics with AEE Linguistic Feature Set

6.3 Contributions of the Research

The contributions of this research are two-fold. It contributed to the existing knowledge of automated essay scoring, both in the body of knowledge and practical context described below.

6.3.1 Contribution to Body of Knowledge

- i. This research proposed multiple and comprehensive linguistic feature sets, namely the surface features, linguistic errors, text cohesion, semantics, lexical diversity, lexical sophistication, syntactic complexity, and readability in scoring essays. These linguistic feature sets were formulated based on the reviews of the literature regarding the effect of linguistic factors on writing quality. This methodology of linguistic-based assessment is the first of its kind, adopting most, if not all the existing linguistic features in scoring essay quality. The proposed linguistic features are both inclusive and extensive, i.e., inclusive in the context that they include most, if not all the feasible linguistic dimension, and extensive in which they cover a broad extent of linguistic features.
- ii. This research organized the comprehensive and diversified linguistic features into a proper specification or taxonomy of feature sets, feature categories, and feature indices. A specification of eight feature sets, 43 feature categories, and 1,709 fine-grained feature indices were detailed in this research (Section 3.3 and Table 3.14). This proposed linguistic feature specification can be leveraged to facilitate the systematic investigation, analysis, and scoring of linguistic features in future research work.

6.3.2 Contribution to Practice

This research contributed to the practical context by using the proposed linguistic features in analysing and scoring MUET essays. The contributions are described below.

- i. Several novel results were observed via the correlation analysis of linguistic features and MUET essay scores. These remarkable results include the findings of CARES and CAREC readability feature, word frequency and word range based on COCA and TOEFL11 corpus, and advanced parts of speech as a strong predictor of essay score. The complete analysis results were detailed in Sections 5.2, 5.3, and 5.4, with the overall summary in Figure 5.14 and Figure 5.15.
- ii. This research attempted to formulate an AES classifier, namely the Automated Essay Evaluator (AEE), to predict MUET essay scores based on the selected important linguistic features. The AEE demonstrates its effectiveness in scoring essays with QWK and classification accuracy around 0.60++, which is closely comparable with state-of-the-art AES systems. It outperformed the LightSide AES in scoring the MUET essays. To further explain the linguistic features in scoring essays, this research includes a theoretical mapping to associate the five MUET essay scoring rubrics with the proposed linguistic feature sets.
- iii. Last but not least, this research presented an innovative method of linguistic feature sets and categories visualization via boxplots of correlation coefficient distribution between the feature indices and essay scores. These plots which render the linguistic feature spectrum sorted by its correlation value are shown in Figure 5.14 and Figure 5.15. This visualization method can be automated through computer programming to provide an instant yet complete visualization of the linguistics feature analysis of an essay.

6.4 Implications of Linguistic-based Automated Essay Analysis and Scoring

This research contributed to the area of automated essay scoring and extended the knowledge boundary of linguistic features analysis, particularly in MUET essays. Based on the research contribution, this thesis implies benefits to several academic disciplines listed below.

6.4.1 Implication for Computer Science

This automated essay scoring work, with the detailed analysis of essays from their comprehensive linguistic features, is the first of its kind. The specification of linguistic feature sets, feature categories, and fine-grained feature indices in this research could be used by computer science researchers, particularly from the discipline of natural language processing and computational linguistics. No research has compiled such great numbers of linguistic features and combined them into meaningful groupings. They can leverage these linguistics features to conduct further research in automated essay scoring or writing assessment. For example, the researchers can incorporate other essay feature dimensions (e.g., argumentation, organization) to provide a full-scale AES for MUET essay scoring or formulate advanced machine learning models to improve classification accuracy. Besides, the collection of the MUET essay dataset in this research is valuable for future research, perceiving the lack of essay datasets in the local Malaysian context. This local context corpus is essential for demonstrating the validity of any future-developed AES for assessing the Malaysian English test.

6.4.2 Implication for Writing Assessment

This research attempted a novel approach to associate the MUET official essay scoring rubric with the proposed linguistic feature sets. Each feature set reflects a particular

dimension of linguistic attributes and can be used to provide a fine-grained and measurable approach to scoring essays. For example, the linguistic errors feature set can be associated with the MUET rubric of accuracy, and the specific feature index of grammatical error in the feature set can be used to assess the essay score for this rubric. These quantifiable feature indices are essential to provide a reliable and valid construct of essay scoring in high-stakes assessments. This thesis can serve as a reference document for educational testing organizations, particularly the Malaysian Examinations Council (MEC; Malaysian Examinations Council, n.d.), in analysing and reviewing the MUET scoring rubrics, or documenting the MUET Regulations and Test Specifications.

6.4.3 Implication for Writing Instruction

The findings of this research also have potential implications for writing instruction. The essay analysis and scoring process can be further automated via computer programming to device educational software for helping language teachers in teaching English. Leveraging contemporary information technology, such as web applications, web services, data analytics, and visualization, the result of the comprehensive linguistic analysis of essays can be used as the underlying database to construct automated essay scoring with feedback application. Besides producing a holistic final essay score, this writing instruction application can provide constructive and instant feedback to students in every linguistic dimension. Immediacy in feedback and recursive cycles of writing practice through the application shall enhance and motivate students in essay writing and relieve teachers' burden in teaching English.

6.5 Limitations and Future Works

While there are many findings derived from this research, there are, however, some limitations to acknowledge. These limitations imply the open challenges and potential for future works. This section briefly discusses the direction of future works based on the three limitations of the current research.

6.5.1 The Constraint of Linguistic Feature Scoring

This is the first and foremost constraint to be addressed. The formulated AES only scores essays exclusively based on linguistic features. It did not consider other features. Recognizing that writing quality is influenced by various factors beyond linguistics, the AES shall be enhanced to incorporate other essay features for realizing a full-fledged AES.

One of the significant aspects to consider in future works is the incorporation of content features. The content-based essay assessment can be approached in the following manner:

- i. **Assessment of specific aspects of content features**

These future works focus on specific aspects of content, such as prompt relevance, thesis clarity, organization and development, and augmentation. Future research on this content features can be extended from previous research such as thesis clarity (Persing & Ng, 2013), prompt relevance (Persing & Ng, 2014), and argumentation (Ke, 2019), which were reviewed in Section 2.4.

- ii. **Assessment of essays with domain knowledge**

These future works involve the evaluation of domain and subject-specific knowledge in an essay. Implementing these AES systems often involves the integration of the subject-specific database, ontologies, and domain model. Semantic analysis

techniques such as LSA, LDA, and word embedding approaches such as Word2vec and Global Vectors for Word Representation (GloVe; Pennington et al., 2014) can be utilized to analyse the essay's content based on the semantic understanding and contextual information they provided. An example of previous research in this area is the work of Zupanc and Bosnić (2017) on Semantic Essay Grader for Essays (SAGE), which is discussed in Section 2.1.9.

6.5.2 The Limitation of Essay Datasets

The current research is limited to the MUET essay dataset used as the gold standard for essay scoring. The dataset comprises only one essay prompt, which may constrain the generalization derived for the analysis and scoring of essays. The training data must be sufficiently large for the prediction model to be accurate. Furthermore, the distribution of the essay grade is highly imbalanced which causes ineffectiveness in learning and scoring essays, particularly those belonging to under-represented classes.

In future works, the following strategies can be applied to address the issues posed by the limited essay datasets:

- i. Expanding the collection of the MUET essay dataset.

This is the most intuitive and straightforward strategy. It could involve collaboration with other related schools or institutions to collect a larger and more diverse set of MUET essay samples. Another innovative method is to create an online platform to facilitate students to submit their essays for various purposes (e.g., feedback provision, competition). The platform can be leveraged as a repository to collect a large number of essays.

ii. Cross-prompt scoring / Domain adaptation

Cross-prompt scoring or domain adaptation involves the specific techniques of adapting scoring models that were initially trained on one essay prompt (the source or non-targeted prompt) to effectively score essays on a different essay prompt (the target prompt). This approach demonstrates the solution to address the insufficient number of essays written for the targeted prompt. In future works, it may be possible to build upon the pioneering domain adaptation work of Phandi et al. (2015), who employed Bayesian linear ridge regression to adapt their AES system for scoring the target prompt using training data from the source prompt. Another relevant reference is the research conducted by Ridley et al. (2020, 2021) who used prompt-agnostic essay scorers, which do not require access to the targeted prompt during model training.

iii. Transfer learning

Transfer learning involves employing knowledge obtained from one task or domain to improve performance on another related task or domain. In the context of AES, it involves pre-training a language model on a large corpus of text and then fine-tuning it on the small target essay dataset. Similar to cross-prompt scoring, this approach can also help mitigate the insufficient essay dataset issue. Future work can be developed based on well-known large language models such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018) and Generative Pre-trained Transformers (GPT; Radford et al., 2018). BERT, developed by Google, is a bidirectional model designed for contextual language understanding. The model is trained in a bidirectional manner (both left and right context for each word), enabling the encoding of deeper contextual information for better language

understanding. On the other hand, GPT, developed by OpenAI, is a unidirectional model consisting of billions of parameters specially designed for language generation tasks.

iv. Data resampling

The issue of class imbalance occurs in many machine learning problems, as in the case of automated essay scoring. It can create challenges in the training and scoring essays as the classifier may struggle to learn meaningful patterns from under-represented classes and be biased towards majority classes. Various data sampling can be applied to address the issue and thus enhance the classifier's performance. To increase under-represented class samples in the essay dataset, over-sampling techniques can be employed to create duplicate or synthetic instances of the under-represented class. One of the popular over-sampling techniques is the Synthetic Minority Over-sampling Technique (SMOTE; Fernández et al., 2018). It generates synthetic under-represented class instances by interpolating between the original instance to its k-nearest neighbours in the feature space. Research on resampling techniques for automated essay scoring can be found in the literature by Haendchen et al. (2019).

6.5.3 The Constraint of the Machine Learning Models

This research primarily focuses on incorporating optimized linguistic features for essay scoring, with less emphasis on its machine learning model. The research used the basic configurations of the proposed six machine learning modes without further model hyperparameter fine-tuning to obtain optimal results. Furthermore, the research did not attempt deep neural network approaches, which have attracted much attention recently due to their promising results.

The following approaches can be incorporated in future works to address these issues and enhance AES performance:

i. Deep neural network automated essay scoring (DNN-AES)

Many recent AES systems relied on deep neural network approaches (Ke & Ng, 2019). These DNN-AES models demonstrate their benefits by eliminating the need for manual handcrafted feature engineering and potentially achieving accurate classification results. Instead of relying on linguistic feature engineering as used in this research, future research may explore using neural approaches to automate feature extraction of MUET essay features. These deep neural network approaches could include CNN, RNN, LSTM, attention mechanisms, and transformer models (Vaswani et al., 2017) to extract hidden features in essays. Research related to DNN-AES can be found in the works of Taghipour and Ng (2016) on neural-based AES; Alikaniotis et al. (2016) on SSWE AES; Dong et al. (2017) on Attention-based AES; and Tay et al. (2018) on SKIPFLOW, which were reviewed in Section 2.1.8.

However, these DNN approaches may encounter challenges due to the limited number of MUET essay samples. As specified in Table 2.2, the DNN-AES systems relied on a vast number of ASAP essays, approximately 7,787 essays, to train their AES models. In contrast, this research utilized 724 essays, marking a tenfold decrease in sample size. Before incorporating this DNN approach, it is essential to expand the MUET dataset, as explained in Section 6.5.2. This may involve various strategies such as increasing the manual collection of MUET essays or employing techniques like cross-prompt scoring with other essay datasets to enlarge the training corpus. Transfer learning models can also be explored to utilize knowledge from

other related domains to score essays, such as leveraging pre-trained language models trained on large text corpora.

ii. Fine-tuning of machine learning hyperparameters

Future work can be conducted to fine-tune hyperparameters for enhancing the performance of the proposed six machine learning classifiers (Naïve Bayes, Generalized Linear Model, Random Forest, Gradient Boosted Tree, Multilayer Perceptron, and Support Vector Machine) applied in this research. This fine-tuning process typically involves iterative experimentations to select optimal hyperparameters using grid search, random search, and others. The key hyperparameters to configure differ based on the respective machine learning models. For example, hyperparameters for Random Forest include the number of trees, criterion for splitting, maximum depth, etc., while Multilayer Perceptron requires training cycles, learning rate, momentum, etc. To improve the performance of the AEE in this research, future work can be incorporated to fine-tune the hyperparameters of the proposed six machine learning classifiers as specified in Table 4.6.

6.6 Concluding Remarks

Most large-scale high-stakes assessments require written responses to be included as a fundamental component of the assessment task. As a result, AES has become an important tool to support the assessments of 21st-century skills. Although AES systems have been around for decades, there still isn't a standard set of features found to be the most predictive of essay scores. The generalizability of many features also has not been well documented because most AES systems are developed only for commercial proprietary purposes.

This research provides a detailed investigation of the analysis and scoring of essays from their comprehensive linguistic features. Several findings on the relationship between linguistic features and essay scores have been reported in this research. The formulated AEE has demonstrated its validity in scoring essays based on deep linguistic features, instead of shallow surface features. However, this research is only a small step towards the realization of the full-fledged Malaysian-based AES, which requires the incorporation of other essay features. Furthermore, it is hoped that the proposed comprehensive linguistic features in this research can be applied in various related investigations, extending beyond essay scoring alone.

REFERENCES

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.
- Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, 14(3), 455–459.
- Alelyani, S., Tang, J., & Liu, H. (2018). Feature selection for clustering: A review. In C. C. Aggarwal & C. K. Reddy (Eds.), *Data Clustering: Algorithms and Applications* (pp. 29–60). Chapman and Hall/CRC.
- Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *ArXiv Preprint ArXiv:1606.04289*.
- Allen, L., Crossley, S., Kyle, K., & McNamara, D. S. (2014). The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. In J. C. Stamper, Z. A. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 300–303). International Educational Data Mining Society (IEDMS).
- Amorim, E., Cançado, M., & Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 229–237.

- Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater® V.2.0 (Report No. RR-04-45)*. Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2004.tb01972.x>
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2.0. *The Journal of Technology, Learning, and Assessment*, 4(3), 1–31.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2), 283–316.
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459.
- Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1), 63–88.
- Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 142–173). Routledge.
- BETSY. (n.d.). *Bayesian Essay Test Scoring sYstem*. Retrieved July 14, 2023, from <https://edres.org/betsy/>
- Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2), i–15.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

- BNC Consortium. (2007). *British National Corpus, XML edition*. Literary and Linguistic Data Service. <http://hdl.handle.net/20.500.14106/2554>
- Bouchlaghem, Y., Akhiat, Y., & Amjad, S. (2022). Feature selection: A review and comparative study. *E3S Web of Conferences*, 351, 01046.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Briscoe, T., Medlock, B., & Andersen, Ø. (2010). *Automated assessment of ESOL free text examinations*. University of Cambridge, Computer Laboratory. <https://doi.org/10.48456/tr-790>
- Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers*, 16(6), 502–532.
- Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911.
- Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42–65.
- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3), 27–36.

- Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, 70–79.
- Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.
- Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28.
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 1–27.
- Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater ® 's performance on TOEFL ® essays. *ETS Research Report Series*, 2004(1), i–38.
- Chotlos, J. W. (1944). A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2), 75.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213.
- Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th International Conference on Machine Learning*, 160–167.
- Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4), 497–505.

- Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.
- Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, 45(2), 355–362.
- Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415–443.
- Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011a). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1), 84–101.
- Crossley, S. A., Bradfield, F., & Bustamante, A. (2019a). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research*, 11(2), 251–270.
- Crossley, S. A., Kyle, K., & Dascalu, M. (2019b). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1), 14–27. <https://doi.org/10.3758/s13428-018-1142-4>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237. <https://doi.org/10.3758/s13428-015-0651-7>
- Crossley, S. A., & McNamara, D. S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy & McNamara (2007). *Language Teaching*, 41(3), 409–429.

- Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66–79. <https://doi.org/10.1016/j.jslw.2014.09.006>
- Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Grantee Submission*, 7(3), 351–370.
- Crossley, S. A., Skalicky, S., & Dascalu, M. (2019c). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3–4), 541–561. <https://doi.org/10.1111/1467-9817.12283>
- Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011b). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28(3), 282–311. <https://doi.org/10.1177/0741088311410188>
- Cunningham-Nelson, S., Mukherjee, M., Goncher, A., & Boles, W. (2018). Text analysis in education: A review of selected software packages with an application for analysing students’ conceptual understanding. *Australasian Journal of Engineering Education*, 23(1), 25–39.
- Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, 27(1), 11–28.
- Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(1), 19–26.
- Dascalu, M., McNamara, D., Crossley, S., & Trausan-Matu, S. (2016). Age of exposure: A model of word learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2928–2934.

- Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4), 447–464. <https://doi.org/10.1093/lc/fqq018>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1), 1–36.
- Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(02), 185–205.
- Doewes, A., Kurdhi, N., & Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. *16th International Conference on Educational Data Mining, EDM 2023*, 103–113.
- 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*, 1072–1077.
- Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 153–162.
- Douglas, S. R. (2013). The Lexical Breadth of Undergraduate Novice Level Writing Competency. *The Canadian Journal of Applied Linguistics*, 16, 152–170.

- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). Wiley & Sons.
- ETS. (2023). *The e-rater scoring engine*. <https://www.ets.org/erater.html>
- Falakmasir, M. H., Ashley, K. D., Schunn, C. D., & Litman, D. J. (2014). Identifying thesis and conclusion statements in student essays to scaffold peer review. *International Conference on Intelligent Tutoring Systems*, 254–259.
- Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61, 863–905.
- Filho, A. H., Concatto, F., do Prado, H. A., & Fernela, E. (2021). Comparing feature engineering and deep learning methods for automated essay scoring of Brazilian national high school examination. *Proceedings of the 23rd International Conference on Enterprise Information Systems-Volume 1*, 575–583.
- Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? The effect of text Length on writing assessment. *Frontiers in Psychology*, 11, 562462. <https://doi.org/10.3389/fpsyg.2020.562462>
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221–233.
- Foltz, P. W. (2016). Advances in automated scoring of writing for performance assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of Research on Technology Tools for Real-World Skill Development* (Vol. 2, pp. 659–678). Information Science Reference, IGI Global. <https://doi.org/10.4018/978-1-4666-9441-5.ch025>

- Garner, J., Crossley, S., & Kyle, K. (2020). Beginning and intermediate L2 writer's use of n-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching*, 58(1), 51–74.
- Ge, S., & Chen, X. (2019). The application of deep learning in automated essay evaluation. *International Symposium on Emerging Technologies for Education*, 310–318.
- Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & de Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, 48(10), 950–962. <https://doi.org/10.1111/medu.12517>
- Goddard, C. (2011). *Semantic analysis: A practical introduction*. Oxford University Press.
- Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5), 403–420. <https://doi.org/10.1007/BF02163027>
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234.
- Graesser, A. C., McNamara, D. S., McNamara, M. M., & Cai, Z. Q. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202.
- Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), 123–145.
- Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique [Problems and methods of linguistic statistics]*. Springer Dordrecht.

- Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18(3), 218–238.
<https://doi.org/10.1016/j.asw.2013.05.002>
- Haendchen Filho, A., Concatto, F., Nau, J., do 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.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Routledge.
- Harika, Y., Latha, I. S., Sai, V. L., Krishna, P. S., & Suneetha, M. (2017). Automated essay grading using feature selection. *International Research Journal of Engineering and Technology (IRJET)*, 4(3), 2852–2855.
- Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 30(2), 87–93.
- Herdan, G. (1960). *Type-token mathematics. A textbook of mathematical linguistics*. The Hague: Mouton.
- Hoblos, J. (2020). Experimenting with latent semantic analysis and latent dirichlet allocation on automated essay grading. *2020 7th International Conference on Social Network Analysis, Management and Security, SNAMS 2020*, 1–7.
<https://doi.org/10.1109/SNAMS52053.2020.9336533>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

- Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. National Council of Teachers of English, Champaign, Illinois.
- Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems : A literature review. *PeerJ Computer Science*, 5(e208), 1–16. <https://doi.org/10.7717/peerj-cs.208>
- IBM. (n.d.). *SPSS statistics* | IBM. Retrieved July 14, 2023, from <https://www.ibm.com/products/spss-statistics>
- Ifenthaler, D. (2022). Automated essay scoring systems. In *Handbook of open, distance and digital education* (pp. 1–15). Springer.
- Jagaiah, T. (2017). *Analysis of syntactic complexity and its relationship to writing quality in argumentative essays* [Doctoral dissertation, University of Connecticut]. Digital Commons @ UConn. <https://opencommons.uconn.edu/dissertations/1571>
- Janda, H. K., Pawar, A., Du, S., & Mago, V. (2019). Syntactic, semantic and sentiment analysis: The joint effect on automated essay evaluation. *IEEE Access*, 7, 108486–108503. <https://doi.org/10.1109/ACCESS.2019.2933354>
- Johnson, W. (1944). Studies in language behavior: A program of research. *Psychological Monographs*, 56(2), 1–15.
- Jong, Y.-J., Kim, Y.-J., & Ri, O.-C. (2023). Review of feedback in automated essay scoring. *ArXiv Preprint ArXiv:2307.05553*.
- Juanggo, W. (2018). Investigating lexical diversity and lexical sophistication of productive vocabulary in the written discourse of Indonesian EFL learners. *Indonesian Journal of Applied Linguistics*, 8(1), 38–48. <https://doi.org/10.17509/ijal.v8i1.11462>

- Kaggle. (2012). *The Hewlett Foundation: Automated essay scoring* | Kaggle.
<https://www.kaggle.com/c/asap-aes/data>
- Kampookaew, P. (2020). An analysis of grammatical errors made by Thai EFL university students in an EAP writing class: Issues and recommendations. *REFlections*, 27(2), 246–273.
- Kaplan, R. M., Wolff, S. E., Burstein, J. C., Lu, C., Rock, D. A., & Kaplan, B. A. (1998). Scoring essays automatically using surface features. *ETS Research Report Series*, 1998(2), i–12. <https://doi.org/10.1002/j.2333-8504.1998.tb01788.x>
- Ke, Z. (2019). *Automated essay scoring: Argument persuasiveness* [Master's thesis, The University of Texas at Dallas]. UTD Theses and Dissertations.
<https://hdl.handle.net/10735.1/6987>
- Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 6300–6308. <https://doi.org/10.24963/ijcai.2019/879>
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2), 81–93.
- Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, 34(4), 1060–1073.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (Report No. RBR 8-75)*. Institute for Simulation and Training, University of Central Florida.
<https://stars.library.ucf.edu/istlibrary/56>

- Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing*, 16(3), 154–169. <https://doi.org/10.1016/j.asw.2011.01.001>
- Kobrin, J., & Shaw, E. J. (2007). Does quantity equal quality? The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology*, 8(1), 1–15.
- Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004). *Continuous multivariate distributions, Volume 1: Models and applications*. John Wiley & Sons.
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.
- Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of imbalanced data: Review of methods and applications. *IOP Conference Series: Materials Science and Engineering*, 1099(1), 12077.
- Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, 5, 572367.
- Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990.
- Kyle, K. (n.d.). *NLP tools for the social science*. Retrieved July 14, 2023, from <https://www.linguisticanalysisistools.org/>
- Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. SchoolWorks @ Georgia State University. <https://doi.org/10.57709/8501051>

- Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513–535.
- Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, 49(4), 757–786. <https://doi.org/10.1002/tesq.194>
- Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2), 333–349.
- Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>
- Lagakis, P., & Demetriadis, S. (2021). Automated essay scoring: A review of the field. 2021 *International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–6.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284. <https://doi.org/10.1080/01638539809545028>
- Latifi, S., & Gierl, M. (2021). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, 38(1), 62–85.
- Latifi, S. M. F. (2016). *Development and validation of an automated essay scoring framework by integrating deep features of English language* [Doctoral dissertation,

University of Alberta]. UAlberta Education and Research Archive.
<https://doi.org/https://doi.org/10.7939/R37S7J134>

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1), 1–30.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. *European Conference on Machine Learning*, 4–15.

Li, F., Xi, X., Cui, Z., Li, D., & Zeng, W. (2023). Automatic essay scoring method based on multi-scale features. *Applied Sciences*, 13(11), 1–18.

Li, X., & Liu, J. (2017). Automatic essay scoring based on Coh-Metrix feature selection for Chinese English learners. In TT. Wu, R. Gennari, YM. Huang, H. Xie, & Y. Cao (Eds.), *Emerging Technologies for Education. SETE 2016. Lecture Notes in Computer Science* (Vol. 10108, pp. 382–393). Springer, Cham. https://doi.org/10.1007/978-3-319-52836-6_40

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12), 6999–7019.
<https://doi.org/10.1109/TNNLS.2021.3084827>

LightSide. (2019). *LightSide researcher's workbench*. <http://ankara.lti.cs.cmu.edu/side/>

Louis, A., & Higgins, D. (2010). Off-topic essay detection using short prompt texts. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 92–95.

- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496.
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62.
- Maasum, T. N. R. T. M., Stapa, S. H., Omar, N., Aziz, M. J. A., & Darus, S. (2012). Development of an automated tool for detecting errors in tenses. *GEMA Online Journal of Language Studies*, 12(2), 427–441.
- MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, 32(6), 1553–1574. <https://doi.org/10.1007/s11145-018-9853-6>
- Madala, D. S. V., Gangal, A., Krishna, S., Goyal, A., & Sureka, A. (2018). *An empirical analysis of machine learning models for automated essay grading*. PeerJ Preprints. <https://doi.org/10.7287/peerj.preprints.3518v1>
- Malaysian Examination Council. (2014). *Malaysian University English Test (MUET) - regulations, test specifications, test format and sample questions*. www.mpm.edu.my
- Malaysian Examinations Council. (n.d.). *Official portal Malaysian Examinations Council*. Retrieved July 21, 2023, from <https://www.mpm.edu.my/en/>
- Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12* (pp. 512–518). The MIT Press.

- Mass, H.-D. (1972). Über den zusammenhang zwischen wortschatzumfang und länge eines textes [On the relationship between vocabulary and the length of a text]. *Zeitschrift Für Literaturwissenschaft Und Linguistik*, 2(8), 73.
- Mayfield, E., Adamson, D., & Rosé, C. P. (2014). *LightSide researcher's workbench user manual*. http://ankara.lti.cs.cmu.edu/side/LightSide_Researchers_Manual.pdf
- Mayfield, E., & Rosé, C. P. (2013). LightSide: Open source machine learning for text. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated essay evaluation: Current applications and new directions* (pp. 124–135). Routledge Academic Press.
- Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of Reading*, 12(8), 639–646.
- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488.
- McCarthy, P. M., & Jarvis, S. (2010). MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. <https://doi.org/10.3758/BRM.42.2.381>
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1), 57–86. <https://doi.org/10.1177/0741088309351547>
- McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45(2), 499–515. <https://doi.org/10.3758/s13428-012-0258-1>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

- McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1), 1–43.
- Measurement Incorporated. (n.d.). *Home | Measurement Incorporated*. Retrieved July 21, 2023, from <https://www.measurementinc.com/>
- Medsker, L., & Jain, L. C. (1999). *Recurrent neural networks: design and applications*. CRC press.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. *ArXiv Preprint ArXiv:1310.4546*, cs.CL.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013b). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*, cs.CL.
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41. <https://doi.org/10.1145/219717.219748>
- Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69–71.
- Naber, D. (2003). *A rule-based style and grammar checker* [Diploma thesis, Universität Bielefeld]. GRIN Computer Science. <https://www.grin.com/document/108379>
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370–384.
- Nemati, M., & Azizi, M. (2013). Readability index of essays as an alternative to the scoring procedure in L2 academic writing. *Archives of Advances in Biosciences*, 4, 2–10. <https://doi.org/10.22037/jps.v4i0.4011>

- Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, 1–14.
- Ng, S. Y. (2018). *Impact of automated essay scorer with feedback for essay writing* [Doctoral dissertation, Universiti Malaysia Sarawak]. UNIMAS Institutional Repository. <https://ir.unimas.my/id/eprint/26604/>
- Nigam, A. (2017). Exploring Automated Essay Scoring for Nonnative English Speakers. *ArXiv Preprint ArXiv:1706.03335*.
- NLTK. (n.d.). *NLTK: Natural Language Toolkit*. Retrieved July 21, 2023, from <https://www.nltk.org/>
- Omar, N., Razali, N. A. Mohd., & Darus, S. (2009). Automated grammar checking of tenses for ESL writing. In P. Wen, Y. Li, L. Polkowski, Y. Yao, S. Tsumoto, & G. Wang (Eds.), *Rough Sets and Knowledge Technology. Lecture Notes in Computer Science* (Vol. 5589, pp. 475–482). Springer Berlin Heidelberg.
- Ong, N., Litman, D., & Brusilovsky, A. (2014). Ontology-based argument mining and automatic essay scoring. In G. Nancy, A. Kevin, L. Diane, R. Chris, & W. Vern (Eds.), *Proceedings of the First Workshop on Argumentation Mining* (pp. 24–28). Association for Computational Linguistics. <https://doi.org/10.3115/v1/W14-2104>
- Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, 24(4), 492–518.

- Östling, R., Smolentzov, A., Hinnerich, B. T., & Höglin, E. (2013). Automated Essay Scoring for Swedish. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 42–47. <http://www.ling.su.se/aes>
- Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243.
- Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education*, 62(2), 127–142.
- Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121–145. <https://doi.org/10.1177/0267658317694221>
- Pearson Education. (2010). *Intelligent Essay Assessor*. <https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf>
- Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.
- Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language*, 37(3), 382–410.
- Peng, Y., Sun, J., Quan, J., Wang, Y., Lv, C., & Zhang, H. (2023). Predicting Chinese EFL learners' human-rated writing quality in argumentative writing through multidimensional computational indices of lexical complexity. *Assessing Writing*, 56, 100722. <https://doi.org/10.1016/j.asw.2023.100722>

- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
- Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, 21, 104–111. <https://doi.org/10.1016/j.asw.2014.05.001>
- Persing, I., Davis, A., & Ng, V. (2010). Modeling organization in student essays. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 229–239.
- Persing, I., & Ng, V. (2013). Modeling thesis clarity in student essays. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 260–269.
- Persing, I., & Ng, V. (2014). Modeling prompt adherence in student essays. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1534–1543.
- Persing, I., & Ng, V. (2015). Modeling argument strength in student essays. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 543–552.
- Pham, T. H., & Raahemi, B. (2023). Bio-inspired feature selection algorithms with their applications: A systematic literature review. *IEEE Access*, 11, 43733–43758.
- Phandi, P., Ming Chai, K. A., & Tou Ng, H. (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.
<http://www.kaggle.com/c/asap-aes>

Piao, C., Wang, N., Yuan, C., & others. (2023). Rebalance weights AdaBoost-SVM model for imbalanced data. *Computational Intelligence and Neuroscience*, 2023, 1–26.

Pikulski, J. J. (2002). *Readability*. Houghton Mifflin.

Powers, D. E. (2005). “*Wordiness*”: *A selective review of its Influence, and suggestions for investigating Its relevance in tests requiring extended written responses (Report No. RM-04-08)*. Educational Testing Service.
<https://www.ets.org/Media/Research/pdf/RM-04-08.pdf>

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527.
<https://doi.org/10.1007/s10462-021-10068-2>

Randy, A., & Angel, A. (2023). Automated measures of lexical sophistication: Predicting proficiency in an integrated academic writing task. *Journal of Language and Education*, 9(1), 12–25.

RapidMiner. (n.d.). *RapidMiner | Best data science & machine learning Platform*. Retrieved July 21, 2023, from <https://rapidminer.com/>

Read, J. A. S. (2000). *Assessing vocabulary*. Cambridge University Press.

- Richards, J. C., & Schmidt, R. (n.d.). *Longman dictionary of language teaching and applied linguistics*. Routledge. <https://doi.org/10.4324/9781315833835>
- Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. (2020). Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *ArXiv Preprint ArXiv:2008.01441*.
- Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. (2021). Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15), 13745–13753.
- Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53, 23–69.
- Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4), 1–22.
- Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2), 1–22.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (pp. 318–362). MIT Press.
- Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany. *ETS Research Report Series*, 2019(1), 1–23.

- Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, 27(3), 343–360. <https://doi.org/10.1177/0267658310395851>
- Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.
- Sanders, T., & Maat, H. P. (2006). Cohesion and coherence: Linguistic approaches. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics* (2nd ed, pp. 591–595). Elsevier.
- Senter, R. J., & Smith, E. A. (1967). *Automated readability index (Report No. AMRL-TR-66-220)*. Aerospace Medical Research Laboratories (U.S.).
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76.
- Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation* (pp. 23–37). Routledge.
- Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, 26(3), 247–259.
- Shin, E. (2018). *A Neural Network approach to automated essay scoring: A comparison with the method of integrating deep language features using Coh-Metrix* [Master's thesis, University of Alberta]. UAlberta Education and Research Archive. <https://doi.org/10.7939/R3V11W25D>

- Shin, J., & Epp, C. D. (2023). Understanding the effect of cohesion in academic writing clarity using education data science. In A. Peña-Ayala (Ed.), *Educational Data Science: Essentials, Approaches, and Tendencies* (pp. 193–218). Springer.
- Shiva Shankar, R., & Ravibabu, D. (2019). Digital report grading using NLP feature selection. *Soft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018*, 615–623.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512.
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 18(2), 161–169.
- Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM—a tutorial into long short-term memory recurrent neural networks. *ArXiv Preprint ArXiv:1909.09586*.
- 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*, 1882–1891. <https://doi.org/10.18653/v1/d16-1193>
- Tao, X., Li, Q., Guo, W., Ren, C., Li, C., Liu, R., & Zou, J. (2019). Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. *Information Sciences*, 487, 31–56.
- Tay, Y., Phan, M., Tuan, L. A., & Hui, S. C. (2018). Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 5948–5955. <https://doi.org/10.1609/aaai.v32i1.12045>

- The University of Memphis. (n.d.). *Coh-Metrix version 3.0 indices*. Retrieved July 21, 2023, from http://cohmetrix.memphis.edu/cohmetrixhome/documentation_indices.html
- Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Bureau of Publications, Teachers College, Columbia University.
- Tian, Y., Kim, M., Crossley, S., & Wan, Q. (2020). *Systemic functional approach to cohesion and its application to automated cohesion analysis*. EasyChair Preprint. <https://easychair.org/publications/preprint/S8wR>
- Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to Basics: How Measures of Lexical Diversity Can Help Discriminate between CEFR Levels. *Applied Linguistics*, 39(3), 302–327. <https://doi.org/10.1093/applin/amw009>
- Uto, M., Xie, Y., & Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. *Proceedings of the 28th International Conference on Computational Linguistics*, 6077–6088.
- Uzun, K. (2018). Home-grown automated essay scoring in the literature classroom: A solution for managing the crowd? *Contemporary Educational Technology*, 9, 423–436. <https://doi.org/10.30935/cet.471024>
- Vantage Labs. (2023). *IntelliMetric® | direct*. <https://www.intellimetric.com>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 5998–6088). NIPS Foundation.

- Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, 19(1), 3–26.
- Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24, 175–186.
- Wan, C. (2019). Feature selection paradigms. In *Hierarchical Feature Selection for Knowledge Discovery: Application of Data Mining to the Biology of Ageing* (pp. 17–23). Springer, Cham. https://doi.org/10.1007/978-3-319-97919-9_3
- Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: A simple visual method to interpret data. *Annals of Internal Medicine*, 110(11), 916–921.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality: Opportunities and challenges. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 679–704). IGI Global. <https://doi.org/10.4018/978-1-4666-9441-5.ch026>
- Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology*, 82, 123–140.
- 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), 71–80. <https://doi.org/10.11113/ijic.v9n1.220>

- Xiao, J., & Zhou, Z. (2020). Research progress of RNN language model. *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications*, 1285–1288. <https://doi.org/10.1109/ICAICA50127.2020.9182390>
- Xie, Q. (2019). Error analysis and diagnosis of ESL linguistic accuracy: Construct specification and empirical validation. *Assessing Writing*, 41, 47–62. <https://doi.org/10.1016/j.asw.2019.05.002>
- Xu, W., & Liu, M. (2016). Using Coh-Metrix to analyze Chinese ESL learners' writing. *International Journal of Learning, Teaching and Educational Research*, 15(5), 16–26.
- Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, 100505. <https://doi.org/10.1016/j.asw.2020.100505>
- Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 224–232. <https://www.languagetool.org>
- Zupanc, K. (2018). *Semantics-based automated essay evaluation* [Doctoral dissertation, University of Ljubljana]. CORE. <https://core.ac.uk/download/pdf/159486467.pdf>
- Zupanc, K., & Bosnić, Z. (2015). Advances in the field of automated essay evaluation. *Informatica*, 39(4), 383–395.
- Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, 120, 118–132. <https://doi.org/10.1016/j.knosys.2017.01.006>
- Zupanc, K., & Bosnić, Z. (2018). Increasing accuracy of automated essay grading by grouping similar graders. *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, 1–6.

Zupanc, K., Savić, M., Bosnić, Z., & Ivanović, M. (2017). Evaluating coherence of essays using sentence-similarity networks. *Proceedings of the 18th International Conference on Computer Systems and Technologies*, 65–72.

APPENDICES

Appendix A: Readability Formula

Readability Index	Formula	
	Variable (i)	Coefficient
Flesch Reading Ease	Constant	206.835
	Average number of words per sentence	- (i x 1.015)
	Average number of syllables per word	- (i x 84.6)
Flesch Kincaid Grade Level	Constant	-15.59
	Average number of words per sentence	i x 0.39
	Average number of syllables per word	i x 11.8
Automated Readability Index	Constant	-21.43
	Average number of words per sentence	i x 0.5
	Average number of characters per word	i x 4.71
SMOG Grading	Constant	3
	Square root of polysyllabic words per 30 sentences	i x 1
New Dale-Chall Readability	Average number of words per sentence	i x 0.0496
	Percentage of difficult words	i x 0.1579
Crowdsourced Algorithm of Reading Comprehension (CAREC)	Constant	1.811
	Average age of acquisition (Kuperman) for all content words	i x 0.022
	Average bigram range score (COCA) for all words	i x 0.746
	Average trigram proportion score (BNC-written) for all words	- (i x 0.742)
	Average imageability score (MRC) for all content words	- (i x 0.001)
	Average frequency score (Brown) for all words	i x 0.0000625
	Average type token ratio of lemma trigrams for all trigrams	- (i x 0.699)
	Proportion of lemma types that occur in the next paragraph for all paragraphs	- (i x 0.111)
	Number of temporal connectives divided by number of words in text	- (i x 2.067)
	Proportion of noun lemma types that occur in the next paragraph for all paragraphs	i x 0.035
	Number of content word lemma types	i x 0.002
	Positive adjective scores derived from 4 different corpora	- (i x 0.08)
	Average standard deviation of word length for all words	i x 0.047

Appendix A continued

Readability Index	Formula	
	Variable (i)	Coefficient
	Average character entropy for all characters	- (i x 0.395)
	Average age of acquisition (Kuperman) for all content words	i x 0.022
	Average bigram range score (COCA) for all words	i x 0.746
	Average trigram proportion score (BNC-written) for all words	- (i x 0.742)
	Average imageability score (MRC) for all content words	- (i x 0.001)
	Average frequency score (Brown) for all words	i x 0.0000625
	Average type token ratio of lemma trigrams for all trigrams	- (i x 0.699)
	Proportion of lemma types that occur in the next paragraph for all paragraphs	- (i x 0.111)
	Number of temporal connectives divided by number of words in text	- (i x 2.067)
	Proportion of noun lemma types that occur in the next paragraph for all paragraphs	i x 0.035
	Number of content word lemma types divided by number of content words	i x 0.2
	Positive adjective scores derived from 4 different corpora	- (i x 0.08)
	Average standard deviation of word length for all words	i x 0.047
	Average character entropy for all characters	- (i x 0.395)
Crowdsourced Algorithm of Reading Speed (CARES)	Constant	-0.862
	Average word naming response time for all words	i x 0.003
	Average concreteness score (MRC) for all words	- (i x 0.001)
	Average semantic distinctiveness scores for all words	- (i x 0.461)
	Number of content word lemmas	i x 0.004
	Number of function words	i x 0.002
	Complex nominals per T-unit	i x 0.011
	Number of dependents per direct object	i x 0.023
	Average number of sentences per paragraph	- (i x 0.015)
	Average number of characters per word	i x 0.062
Coh-Metrix L2 Readability Index (CML2RI)	Constant	-43.142
	Number of sentences in text	i x 0.642
	Average frequency score (SUBTLEXus) for all content words logged	i x 12.671
	Proportion of noun and pronoun lemma types that occur in the next two sentences for all sentences	i x 29.619

Appendix B: Data Files – Linguistic Features Specification, Features Value, Features Correlation and Features Selected

File Name (Spreadsheet)	Description
1_Features_Specification	Specification of the eight linguistic feature sets, 43 feature categories, and 1,709 linguistic feature indices employed in this research.
2_Features_Value	Features value extracted from the MUET Essay Dataset.
3_Features_Correlation	Spearman's ρ and Kendall's τ correlation value between the features indices and the MUET essay score/essay band.
4_Features_Selected	Selected features based on Correlation Rank and MRMR Feature Selection Schemes (COR-03, COR-05, COR-07, MRMR-10, MRMR-25, MRMR-75, MRMR-100, and Baseline).

Hyperlink to the online directory containing the files:

https://drive.google.com/drive/folders/1mcm8-SGM_OIDwKln3JUvxKc7HokuoIO5?usp=sharing

Appendix C: Linguistic Feature Indices Selected by COR-05 Feature Selection

Feature Set	Feature Category	Feature Indices
Surface	Length-Based	Word Count
Surface	Advanced Parts of Speech	Advanced Parts of Speech
Surface	Length-Based	Words per Paragraph
Surface	Length-Based	Sentence Count
Surface	Short Paragraph	Number of short paragraphs (≤ 4 sentences)
Surface	Length-Based	Sentences per Paragraph
Surface	Short Paragraph	Number of short Paragraphs (≤ 5 sentences)
Surface	Short Paragraph	Number of short Paragraphs (≤ 3 sentences)
Text Cohesion	Global Cohesion	Number of function lemma types occur at least once in the next two paragraphs
Text Cohesion	Global Cohesion	Number of function lemma types occur at least once in the next paragraph
Text Cohesion	Global Cohesion	Number of lemma types occur at least once in the next two paragraphs
Text Cohesion	Global Cohesion	Number of lemma types occur at least once in the next paragraph
Text Cohesion	Global Cohesion	Number of content lemma types occur at least once in the next two paragraphs
Text Cohesion	Global Cohesion	Number of noun and pronoun lemmas occur at least once in the next two paragraphs.
Semantics	Global Semantics	Average paragraph to paragraph overlaps of noun synonym
Lexical Diversity	Lexical Raw Count	Number of Content Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Content Word Types
Lexical Diversity	Lexical Raw Count	Number of Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Word Types
Lexical Diversity	Lexical Raw Count	Number of Function Word Tokens

Appendix C continued

Feature Set	Feature Category	Feature Indices
Lexical Diversity	Lexical Raw Count	Number of Function Word Types
Lexical Diversity	Root TTR	Root TTR of Content Words
Lexical Diversity	Simple TTR	TTR of Function Words
Lexical Diversity	Lexical Raw Count	Percentage of Word Types that are Content Words
Lexical Diversity	Root TTR	Root TTR of all Words
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Magazine)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Newspaper)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (TOEFL11 All Words)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Academic)
Lexical Sophistication	Word Range	Mean Lemma Type Range Score (TOEFL11 All Words)
Lexical Sophistication	Word Range	Mean Lemma Type Range Score (COCA Magazine)
Lexical Sophistication	Word Range	Mean Lemma Type Range Score (COCA Newspaper)
Lexical Sophistication	Word Range	Mean Lemma Type Range Score (COCA Academic)
Lexical Sophistication	Age of Acquisition/ Exposure	Mean Age of Acquisition Score (Kuperman)
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Speed (CARES)
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Comprehension (CAREC)

Appendix D: Linguistic Feature Indices Selected by MRMR-25 Feature Selection

Feature Set	Feature Category	Feature Indices
Surface	Short Paragraph	Number of short paragraphs (≤ 1 sentences)
Surface	Length-Based	Word Count
Surface	Advanced Parts of Speech	Advanced Parts of Speech
Text Cohesion	Global Cohesion	Number of paragraphs with any function lemma overlap with next paragraph
Text Cohesion	Global Cohesion	Number of function lemma types occur at least once in the next two paragraphs
Lexical Diversity	Lexical Raw Count	Number of Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Word Types
Lexical Diversity	Lexical Raw Count	Number of Content Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Content Word Types
Lexical Diversity	Lexical Raw Count	Number of Function Word Tokens
Lexical Diversity	Lexical Raw Count	Number of Function Word Types
Lexical Diversity	Simple TTR	TTR of Function Words
Lexical Diversity	HDD	HDD Scores of Function Words
Lexical Sophistication	N-gram Frequency	Trigram Frequency - Logarithm Transformed (BNC Written)
Lexical Sophistication	N-gram Association Strength	Trigram Bigram to Unigram Association Strength (COCA Academic)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Academic)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Magazine)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Newspaper)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (COCA Spoken)

Appendix D continued

Feature Set	Feature Category	Feature Indices
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score - Logarithm Transformed (COCA Spoken)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (TOEFL11 High Proficiency Words)
Lexical Sophistication	Word Frequency	Mean Lemma Type Frequency Score (TOEFL11 All Words)
Syntactic Complexity	Syntactic Sophistication	Average Delta p Score Construction (Cue) - Verb (Outcome) (Types only) - COCA Magazine
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Comprehension (CAREC)
Readability	Crowdsourced Readability	Crowdsourced Algorithm of Reading Speed (CARES)

Appendix E: Linguistic Feature Indices Selected by Baseline Model

Feature Set	Feature Category	Feature Indices
Surface Feature	Short Paragraph	Number of short paragraphs (< 1 sentence)
Text Cohesion	Global Cohesion	Adjacent paragraphs overlap adjective lemmas (paragraph normed)
Text Cohesion	Global Cohesion	Adjacent paragraph overlap adverb lemmas (paragraph normed)
Text Cohesion	Local Cohesion	Adjacent sentences overlap adverb lemmas (sentence normed)
Text Cohesion	Global Cohesion	Binary adjacent two-paragraph overlap all lemmas
Text Cohesion	Local Cohesion	Binary adjacent sentence overlap pronoun lemmas
Text Cohesion	Overall Cohesion	Negative logical connectives
Semantic	Local Semantics	Average sentence to sentence overlaps of verb synonyms
Semantic	Global Semantics	Average word2vec similarity score between all adjacent paragraphs.
Lexical Sophistication	Academic Language	Academic Word List Sublist 3
Lexical Sophistication	Ngram Frequency	BNC Spoken Bigram Frequency Logarithm
Lexical Sophistication	Word Frequency	Brown Frequency (All Words)
Lexical Sophistication	Ngram Frequency	COCA Academic Bigram Proportion 40k
Lexical Sophistication	Ngram Association Strength	COCA Academic Trigram Unigram to Bigram Association Strength (MI)
Lexical Sophistication	Ngram Frequency	COCA Fiction Bigram Proportion 20k
Lexical Sophistication	Ngram Frequency	COCA Fiction Bigram Proportion 60k
Lexical Sophistication	Word Frequency	COCA Fiction Frequency (Content Words)
Lexical Sophistication	Word Frequency	COCA Fiction Frequency (Content Words) Logarithm
Lexical Sophistication	Ngram Range	COCA Fiction Trigram Range Logarithm
Lexical Sophistication	Ngram Range	COCA Fiction Trigram Range

Appendix E continued

Feature Set	Feature Category	Feature Indices
Lexical Sophistication	Ngram Frequency	COCA Magazine Bigram Proportion 60k
Lexical Sophistication	Ngram Range	COCA Magazine Bigram Range Logarithm
Lexical Sophistication	Ngram Frequency	COCA News Bigram Frequency Logarithm
Lexical Sophistication	Ngram Association Strength	COCA News Trigram Unigram to Bigram Association Strength (MI)
Lexical Sophistication	Ngram Association Strength	COCA Spoken Bigram Association Strength (MI ²)
Lexical Sophistication	Ngram Range	COCA Spoken Trigram Range
Lexical Sophistication	Word Frequency	COCA Magazine Frequency (Content Words) Logarithm
Lexical Sophistication	Word Range	COCA Magazine Range (Function Words)
Lexical Sophistication	Ngram Frequency	COCA Magazine Trigram Proportion 10k
Lexical Sophistication	Ngram Frequency	COCA News Bigram Frequency Logarithm
Lexical Sophistication	Ngram Range	COCA Spoken Bigram Range Logarithm
Lexical Sophistication	Ngram Association Strength	COCA Spoken Trigram Bigram to Unigram Association Strength (AC)
Lexical Sophistication	Contextual Distinctiveness	Free Association Response Tokens (Content Words)
Lexical Sophistication	Word Neighbor Information	Phonological Neighbourhood Frequency (homophones excluded) (Function Words)
Lexical Sophistication	Contextual Distinctiveness	McDonald Co-occurrence Probability (Content Words)
Lexical Sophistication	Word Frequency	TOEFL11 High Proficiency Frequency Function Words
Lexical Sophistication	Word Frequency	TOEFL11 Low Proficiency Frequency Content Words (log transformed)
Lexical Sophistication	Word Frequency	TOEFL11 Frequency Content Words (log transformed)
Lexical Sophistication	Word Range	TOEFL11 High Proficiency Range Content Words (log transformed)
Lexical Sophistication	Word Frequency	TOEFL11 High Proficiency Frequency Function Words

Appendix E continued

Feature Set	Feature Category	Feature Indices
Lexical Sophistication	Word Range	SUBTLEXus Range (All Words)
Lexical Sophistication	Word Frequency	Thorndike-Lorge Frequency (All Words) Logarithm
Lexical Sophistication	Word Frequency	Thorndike-Lorge Frequency (Function Words)
Syntactic Complexity	Syntactic Sophistication	Average lemma construction combination frequency - academic
Syntactic Complexity	Syntactic Sophistication	Collostruction ratio - all
Syntactic Complexity	Noun Phrase Complexity	Dependents per nominal (no pronouns)
Syntactic Complexity	Sentential Complexity	ordinate phrases per T-unit
Syntactic Complexity	Clause Complexity	Direct objects per clause
Syntactic Complexity	Syntactic Sophistication	Average construction frequency (types only) - magazine
Syntactic Complexity	Clause Complexity	Clausal prepositional complements per clause

Appendix F: Performance of the 48 AEE Models

Feature Selection	ML Classifier	AEE Model	QWK	Accuracy	Precision	Recall	F1 Score
COR-03	NB	COR-03_NB	0.69	0.53	0.62	0.53	0.55
COR-03	GLM	COR-03_GLM	0.62	0.69	0.67	0.69	0.65
COR-03	RF	COR-03_RF	0.27	0.52	0.34	0.52	0.38
COR-03	GBT	COR-03_GBT	0.66	0.67	0.66	0.67	0.67
COR-03	MP	COR-03_MP	0.66	0.66	0.66	0.66	0.66
COR-03	SVM	COR-03_SVM	0.54	0.66	0.56	0.66	0.61
COR-05	NB	COR-05_NB	0.67	0.52	0.62	0.52	0.55
COR-05	GLM	COR-05_GLM	0.68	0.67	0.66	0.67	0.66
COR-05	RF	COR-05_RF	0.57	0.64	0.62	0.64	0.61
COR-05	GBT	COR-05_GBT	0.65	0.67	0.65	0.67	0.66
COR-05	MP	COR-05_MP	0.67	0.67	0.66	0.67	0.66
COR-05	SVM	COR-05_SVM	0.54	0.66	0.56	0.66	0.61
COR-07	NB	COR-07_NB	0.67	0.53	0.61	0.53	0.55
COR-07	GLM	COR-07_GLM	0.66	0.68	0.67	0.68	0.66
COR-07	RF	COR-07_RF	0.67	0.69	0.69	0.69	0.67
COR-07	GBT	COR-07_GBT	0.66	0.67	0.66	0.67	0.66
COR-07	MP	COR-07_MP	0.64	0.68	0.66	0.68	0.66
COR-07	SVM	COR-07_SVM	0.54	0.66	0.56	0.66	0.61
MRMR-10	NB	MRMR-10_NB	0.69	0.63	0.64	0.63	0.63
MRMR-10	GLM	MRMR-10_GLM	0.68	0.68	0.67	0.68	0.67
MRMR-10	RF	MRMR-10_RF	0.62	0.66	0.64	0.66	0.63
MRMR-10	GBT	MRMR-10_GBT	0.66	0.68	0.67	0.68	0.67
MRMR-10	MP	MRMR-10_MP	0.64	0.67	0.65	0.67	0.65

Appendix F continued

Feature Selection	ML Classifier	AEE Model	QWK	Accuracy	Precision	Recall	F1 Score
MRMR-10	SVM	MRMR-10_SVM	0.51	0.64	0.54	0.64	0.59
MRMR-25	NB	MRMR-25_NB	0.67	0.54	0.61	0.54	0.56
MRMR-25	GLM	MRMR-25_GLM	0.68	0.69	0.67	0.69	0.67
MRMR-25	RF	MRMR-25_RF	0.62	0.67	0.64	0.67	0.64
MRMR-25	GBT	MRMR-25_GBT	0.65	0.69	0.68	0.69	0.68
MRMR-25	MP	MRMR-25_MP	0.67	0.69	0.68	0.69	0.68
MRMR-25	SVM	MRMR-25_SVM	0.54	0.66	0.56	0.66	0.61
MRMR-50	NB	MRMR-50_NB	0.66	0.52	0.63	0.52	0.55
MRMR-50	GLM	MRMR-50_GLM	0.59	0.66	0.62	0.66	0.62
MRMR-50	RF	MRMR-50_RF	0.58	0.64	0.62	0.64	0.61
MRMR-50	GBT	MRMR-50_GBT	0.67	0.69	0.69	0.69	0.68
MRMR-50	MP	MRMR-50_MP	0.69	0.69	0.68	0.69	0.68
MRMR-50	SVM	MRMR-50_SVM	0.53	0.66	0.56	0.66	0.61
MRMR-100	NB	MRMR-100_NB	0.63	0.49	0.61	0.49	0.51
MRMR-100	GLM	MRMR-100_GLM	0.61	0.68	0.65	0.68	0.64
MRMR-100	RF	MRMR-100_RF	0.53	0.62	0.61	0.62	0.58

Appendix F continued

Feature Selection	ML Classifier	AEE Model	QWK	Accuracy	Precision	Recall	F1 Score
MRMR-100	GBT	MRMR-100_GBT	0.67	0.67	0.67	0.67	0.67
MRMR-100	MP	MRMR-100_MP	0.66	0.67	0.66	0.67	0.66
MRMR-100	SVM	MRMR-100_SVM	0.48	0.62	0.53	0.62	0.57
Baseline	NB	Baseline_NB	0.49	0.42	0.56	0.42	0.45
Baseline	GLM	Baseline_GLM	0.49	0.63	0.56	0.63	0.59
Baseline	RF	Baseline_RF	0.03	0.49	0.26	0.49	0.33
Baseline	GBT	Baseline_GBT	0.53	0.60	0.58	0.60	0.58
Baseline	MP	Baseline_MP	0.51	0.61	0.59	0.61	0.59
Baseline	SVM	Baseline_SVM	0.00	0.50	0.25	0.50	0.33

Appendix G: Publications

Wong, W. S., & Bong, C. H. (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).
<https://doi.org/10.47836/pjst.29.2.12>

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). <https://doi.org/10.11113/ijic.v9n1.220>