



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)

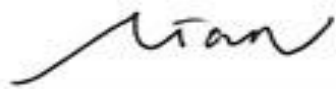
Faculty of Computer Science and Information Technology

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
<i>ABSTRAK</i>	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.