Pertanika J. Sci. & Technol. 29 (2): 919 - 941 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Assessing Malaysian University English Test (MUET) Essay on Language and Semantic Features Using Intelligent Essay Grader (IEG)

Wee Sian Wong* and Chih How Bong

Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

ABSTRACT

Automated Essay Scoring (AES) refers to the Artificial Intelligence (AI) application with the "intelligence" in assessing and scoring essays. There are several well-known commercial AES adopted by western countries, as well as many research works conducted in investigating automated essay scoring. However, most of the products and research works are not related to the Malaysian English test context. The AES products tend to score essays based on the scoring rubrics of a particular English text context (e.g., TOEFL, GMAT) by employing their proprietary scoring algorithm that is not accessible by the users. In Malaysia, the research and development of AES are scarce. This paper intends to formulate a Malaysia-based AES, namely Intelligent Essay Grader (IEG), for the Malaysian English test environment by using our collection of two Malaysian University English Test (MUET) essay dataset. We proposed the essay scoring rubric based on its language and semantic features. We analyzed the correlation of the proposed language and semantic features with the essay grade using the Pearson Correlation Coefficient. Furthermore, we constructed an essay scoring model to predict the essay grades. In our result, we found that the language featured such as vocabulary count and advanced part of speech were highly correlated with the essay grades, and the language features showed a greater influence on essay grades than the semantic features. From our prediction model, we observed that the model yielded better accuracy results based on the selected high-correlated essay features, followed by the language features.

Keywords: Artificial intelligence, automated essay scoring, intelligent system in education, machine learning, MUET, natural language processing

ARTICLE INFO

Article history:
Received: 30 November 2020
Accepted: 10 February 2021
Published: 30 April 2021

DOI: https://doi.org/10.47836/pjst.29.2.12

E-mail addresses:
weesian.wong@gmail.com (Wee Sian Wong)
chbong@unimas.my (Chih How Bong)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Automated Essay Scoring (AES) is a measurement technology in which computers evaluate written work (Shermis & Burstein, 2003). The purpose of the AES is to automate the essay scoring process, and thus reducing labor-intensive marking activities, overcoming time, cost, and yet assuring high assessment reliability. From the perspective of the Computer Science discipline, AES is a Natural Language Processing (NLP) application – a subfield of Artificial Intelligence (AI) focusing on the interactions between computers and human (natural) languages. The main task of AES is to *"intelligently"* classify the given essays into the discrete classes corresponding to the grades defined in the assessment; in another context, it is a document classification problem. The classification of essays is achievable based on the rationale that essays with similar grades shall possess similar characteristics in term of their style and content, which can be measured and quantified by essay characteristics such as words use, syntax structure, and content similarity which in turn can be generalized by employing computer algorithms and statistical methods.

Since the inception of the first AES, named Project Essay Grader (PEG) (Page, 1966), there were numerous commercial AES applications developed as the means of essay assessment in the western world. The typical examples of AES were Project Essay Grader (PEG) (Measurement Incorporated, n.d.), Intelligent Essay Assessor (IEA) (Pearson Education, 2010), IntelliMetric (Vantage Learning, n.d.), and e-rater (Educational Testing Service, n.d.). They employed several methods to assess different aspects of an essay. PEG mainly utilizes the observable essay features denoted as "*proxes"* to evaluate the writing style; and IEA focuses on assessing the essay content by Latent Semantic Analysis (Landauer et al., 1998). E-rater relies on NLP techniques to evaluate essays; while IntelliMetric claimed itself as the first AES that fully leverages various AI techniques in essay scoring. Both the e-rater and IntelliMetric focus on assessing the content and writing style of the essay.

Besides the commercial AES application, there are many research works carried out in investigating automated essay scoring. A large quantity of AES research works involves feature engineering to score a particular dimension of an essay. For example, Crossley and McNamara (2016) used the Tool for the Automatic Analysis of Cohesion (TAACO) to investigate the relationship of text cohesion and essay quality. Zupanc and Bosnić (2017) implemented AES by evaluating the semantic aspect of essays. Cozma et al. (2018) used word embedding based on a pre-trained large corpus to represent n-gram features for scoring essays' lexical attributes. Other aspects of essay scoring can be found in works such as grammatical and mechanic errors detection (Crossley et al., 2019a), prompt adherence modeling (Persing & Ng, 2014), essay stance classification (Persing & Ng, 2016), argumentative structure modeling (Nguyen & Litman, 2018).

In the Malaysian context, there are very few active research works of AES, especially from the discipline of Computer Science. For all we know, there is one AES-related research works focusing on grammar checking of tenses in essays by using a heuristic approach. This work can be found in the literature by Maasum et al. (2012), and Omar et al. (2009). Besides the tenses error detection, other aspects of the essays such as content, language, and organization. are ignored in their work. Apart from this work, Wong and Bong (2019) carried out a feasibility study of adopting AES in the Malaysian English test context based on its reliability and validity requirement. They assessed a well-known AES called LightSide (LightSide, 2019) for determining its feasibility in scoring the Malaysian University English Test (MUET) essays. On the other hand, all other AES works in Malaysia are mainly related to adopting AES as the pedagogical or instructional tool, for studying the AES impact on students' writing skills. Such works include the impact study of Criterion - the instructional application of e-rater (Darus et al., 2003), MY Assess - the instructional interface of Intellimetric (Govindasamy et al., 2013), PaperRate.com (Manap et al., 2019), and Automated Essay Scorer with Feedback – AESF (Ng et al., 2019).

To the best of our knowledge, all the well-established AES, such as PEG, IEA, e-rater, and Intellimetric were emerged and grown in the western world, especially in the United State. They are all western software products, tied with the native English speaking environment and specific test setting. Such AES by their nature may not be suitable when applied to the Malaysian context, and probably of producing an invalid score. The inapplicability of adopting those AES in the Malaysian educational setting can be described by the characteristic below:

(i) The Association of the AES Scoring Rubrics to a particular Test Setting (Wong & Bong, 2019)
The assessment criteria of the AES tend to be associated with the specific scoring rubrics of a particular English assessment context. For example, IntelliMetric's scoring rubric is formulated mostly based upon GMAT; while the e-rater's rubric is designed based on TOEFL.

(ii) The Proprietary Nature of the AES (Wong & Bong, 2019)
Most, if not all of the AES is commercial software, incorporating the proprietary scoring algorithm, which is not accessible and understandable by the users.

(iii) The lack of dataset standardization
All the well-known AES are developed by educational assessment institutions, using their essay dataset which is not available for public use. The Automated Student Assessment Prize (ASAP) corpus (Kaggle, 2012) is the only free available essay dataset that is widely used by most AES research works. However, this ASAP corpus incorporates different assessment criteria and does not contain any

paragraph information, which makes them inappropriate to be used for formulating AES based on the Malaysian educational context.

To address the issues, i.e. the absence of specific AES tailored for the Malaysian English educational setting, this research work is carried out to formulate a Malaysia-based AES, namely Intelligent Essay Grader (IEG). We characterized this IEG as the AES that is custom-made for Malaysian educational assessment and distinguish itself from other commercial proprietary AES software in the following aspects:

(i) The IEG assesses essays based on its unique set of scoring rubrics associated with the specific Malaysian English test context.
(ii) The IEG's scoring rubrics shall be transparent, understandable, and accessible by the end-users.
(iii) The IEG employs a scoring function that is formulated by using the local Malaysian English test dataset. This local context dataset is essential for demonstrating the validity of our developed AES, i.e. it assesses what it claims to assess.

MATERIALS AND METHODS

Malaysian University English Test (MUET)

For the provision of the Malaysian English test context, we have chosen the Malaysian University English Test (MUET) as our use case study. MUET is the English test designed to measure the English language proficiency of pre-university students for entry into tertiary education (Malaysian Examination Council, 2014). It tests all the four language skills of Listening, Speaking, Reading, and Writing; and assess candidates' level of proficiency based upon an aggregated score range of zero to 300, which correlates with a banding system ranging from Band 1 to Band 6 (Malaysian Examination Council, 2014). The MUET writing paper comprises two writing tasks: transferring information from a non-linear source to a linear text and a piece of extended writing which may cover the rhetorical style of analytical, descriptive, persuasive, and argumentative. This study is carried out particularly to tackle the essay assessment of the second task in the MUET writing paper, i.e. the extended writing based on a given topic.

Scoring Rubrics of MUET Essay

In the MUET writing paper, students are assessed on their ability to write various types of text covering a range of rhetorical styles. As specified in the document of *"MUET Regulations, Test Specifications, Test Format and Sample Questions"* (Malaysian Examination Council, 2014), the scoring rubric of the MUET essay covers the aspect of accuracy, appropriacy, coherence and cohesion, use of language functions, and task fulfilment. Table 1 shows the specification of each scoring rubrics based on the document.

Table 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 suggestion 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

Source: Malaysia Examination Council (2014)

Mapping of Essay Dimension with MUET Scoring Rubrics

In addressing such wide-scope and sometimes ambiguous MUET essays scoring rubrics as specified in Table 1, this work tries to formulate a scoring scheme for MUET essays mainly from two aspects, namely the language and semantic dimensions. As illustrated by Figure 1, we hypothesized that our proposed language and semantic dimensions should be correlated with the scoring rubrics of the MUET essays in Table 1. We justified our hypothesis as below:

(i) It is common knowledge that the language dimension, such as spelling error, grammatical error, lexical richness, and parts of speech shall reflect the rubric of accuracy (e.g. using correct spelling and grammar), and appropriacy (e.g. using varied vocabulary and sentences) as specified in the MUET Test Specification.

(ii) The semantic dimension is highly related to the coherence and cohesion rubrics in the MUET Test Specification. The close relationship of semantic dimensions

with coherence and cohesion is described in works by Foltz (2007), Janda, et al. (2019), Zupanc and Bosnić (2014).

(iii) The semantic dimension which is defined as relating to the meaning of words, sentences, and texts can be associated with the use of language functions rubric (e.g. defining, describing and explaining, presenting argument), and the task fulfilment rubric which is particularized as content and topic treatment in the MUET Test Specification.

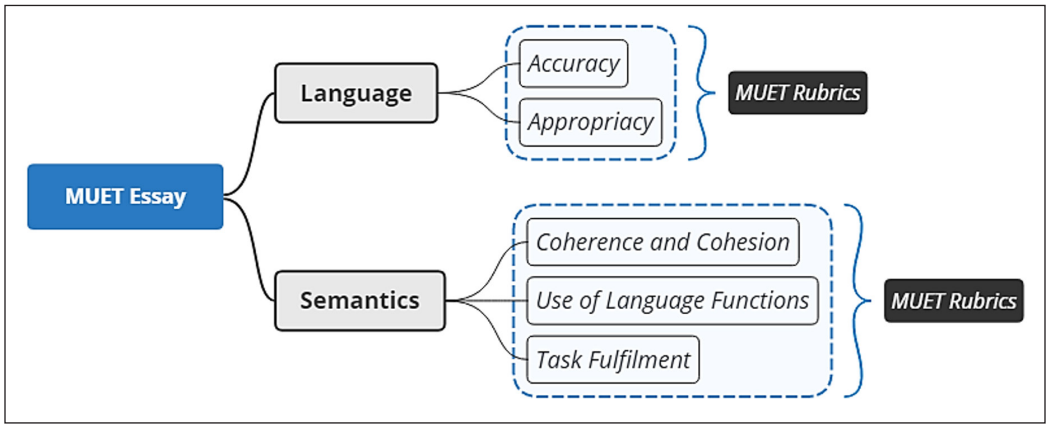


Figure 1. Mapping of IEG dimension with MUET scoring rubrics

Essay Features

As depicted by Figure 1, we formulated our IEG through the perspective of the language and the semantic dimension of the essay. We then represented the essay dimensions using three groups of features: language, local semantics, and global semantics. Each feature quantifies a particular attribute of the dimension, which overall devises a fine-grained and quantifiable approach in grading the essays.

Language. In our IEG model, the language dimension represents the language features and surface features of an essay. The language features cover the linguistic attributes of writing such as lexical richness, parts of speech, spelling error, and grammatical error; while the surface features are related to the attributes of length and *"count"* of the essay such as word count, sentence count, paragraph count, and average words per sentence. Table 2 specifies the 14 language features used in IEG.

Local Semantics. We categorized the essay semantic dimension into local and global semantics. Derived from the works of Crossley et al. (2019b), Crossley and McNamara (2016), we defined the local semantic dimension as the semantic similarity at the span of

Table 2
Language features of IEG

Features	Acronym	Description
Word Count	word_count	Number of Words
Sentence Count	sent_count	Number of Sentences
Paragraph Count	para_ccount	Number of Paragraphs
Average Words Per Sentence	word_sent	Word Count divided by Sentence Count
Average Words Per Paragraph	word_para	Word Count divided by Paragraph Count
Average Sentences Per Paragraph	sent_para	Sentence Count divided by Paragraph Count
Vocabulary Count	vocab_count	Count of Distinct Word
Lexical Richness	lexical_rich	Vocabulary Count divided by Word Count
Spelling Error	spell_err	Number of Spelling Error
Spelling Error Rate	spell_err_rate	Spelling Error divided by Word Count
Grammatical Error	grammar_err	Number of Grammatical Error
Grammatical Error Rate	grammar_err_rate	Grammatical Error divided by Word Count
Advanced Parts of Speech	adv_pos	Accumulated count of Adjective, Adverb, Past Participle, Present Participle Parts of Speech
Advanced Parts of Speech Rate	adv_pos_rate	Advanced Parts of Speech / Word Count

sentence-level (e.g. noun synonym overlap between adjacent sentences). We used eight features in representing the local semantics of an essay, which included the co-occurrence of WordNet Synonym (Miller, 1995) between sentences, the sentences similarity computed by LSA (Landauer et al., 1998), LDA (Blei et al., 2003) and Word2Vec (Mikolov et al., 2013), with the span of one and two-sentence chunk. Table 3 shows the eight local semantic features used in IEG.

Global Semantics. Based on the works of Crossley et al. (2019b), Crossley and McNamara (2016), we defined the global semantic dimension as semantic similarity at the paragraph level (such as the noun overlap between subsequent paragraphs). Similar to local semantics, we adopted eight features in representing the global semantics of an essay; including the co-occurrence of WordNet Synonym (Miller, 1995) between paragraph, the paragraph similarity computed by LSA (Landauer et al., 1998), LDA (Blei et al., 2003) and Word2Vec (Mikolov et al., 2013), with the span of one and two-paragraph chunk. Table 4 shows the eight global semantic features used in IEG.

Dataset

As one of the emphases in our study is to use the Malaysian English test dataset, it is essential that we trained and tested our IEG based on the local context dataset, instead of any public available essay dataset. In the experiment, we used the MUET dataset collected from several schools in Kuching, Sarawak. All the essays collected were written by Form Sixth

Table 3
Local semantic features of IEG

Features	Acronym	Description
Synonym Overlap (Sentence, Noun)	syn_overlap_sent_noun	Average sentence to sentence overlap of noun synonyms based on WordNet
Synonym Overlap (Sentence, Verb)	syn_overlap_sent_verb	Average sentence to sentence overlap of verb synonyms based on WordNet
LSA Cosine Similarity (Adjacent Sentences)	lsa_1_all_sent	Average Latent Semantic Analysis cosine similarity between all adjacent sentences
LSA Cosine Similarity (Two Adjacent Sentences)	lsa_2_all_sent	Average Latent Semantic Analysis cosine similarity between all adjacent sentences (with a two-sentence span)
LDA Divergence (Adjacent Sentences)	lda_1_all_sent	Average Latent Dirichlet Allocation divergence score between all adjacent sentences
LDA Divergence (Two Adjacent Sentences)	lda_2_all_sent	Average Latent Dirichlet Allocation divergence score between all adjacent sentences (with a two-sentence span)
Word2Vec Similarity (Adjacent Sentences)	word2vec_1_all_sent	Average word2vec similarity score between all adjacent sentences
Word2Vec Similarity (Two Adjacent Sentences)	word2vec_2_all_sent	Average word2vec similarity score between all adjacent sentences (with a two-sentence span)

Table 4
Global semantic features of IEG

Features	Acronym	Description
Synonym Overlap (Paragraph, Noun)	syn_overlap_para_noun	Average paragraph to paragraph overlap of noun synonyms based on WordNet
Synonym Overlap (Paragraph, Verb)	syn_overlap_para_verb	Average paragraph to paragraph overlap of verb synonyms based on WordNet
LSA Cosine Similarity (Adjacent Paragraphs)	lsa_1_all_para	Average Latent Semantic Analysis cosine similarity between all adjacent paragraphs
LSA Cosine Similarity (Two Adjacent Paragraphs)	lsa_2_all_para	Average Latent Semantic Analysis cosine similarity between all adjacent paragraphs (with a two-paragraph span)
LDA Divergence (Adjacent Paragraphs)	lda_1_all_para	Average Latent Dirichlet Allocation divergence score between all adjacent paragraphs
LDA Divergence (Two Adjacent Paragraphs)	lda_2_all_para	Average Latent Dirichlet Allocation divergence score between all adjacent paragraphs (with a two-paragraph span)
Word2Vec Similarity (Adjacent Paragraphs)	word2vec_1_all_para	Average word2vec similarity score between all adjacent paragraphs
Word2Vec Similarity (Two Adjacent Paragraphs)	word2vec_2_all_para	Average word2vec similarity score between all adjacent paragraphs (with a two-paragraph span)

students, which were then graded by qualified graders. In addition, to correctly represent the real-word MUET context, we used the real essay prompts between July and November 2014. The datasets consist of two essay prompts labeled as Dataset-1 and Dataset-2, with 259 essay samples in Dataset-1 and 200 essay samples in Dataset-2. The essay samples

Table 5
Grade distribution of MUET essay dataset

Topic	Dataset-1	Dataset-2
	The imbalance between the number of boys and girls pursuing university education creates social problems. To what extent is this statement true? Discuss.	Playing computer games is beneficial for everyone. Discuss.
Grade 1	20	0
Grade 2	73	17
Grade 3	130	103
Grade 4	15	67
Grade 5	11	12
Grade 6	10	1
Total Sample	259	200

are normally distributed with the highest frequency in the Grade-3 essay. Table 5 shows the corresponding grade distribution of the essay samples in Dataset-1 and Dataset-2.

Experiment Setting

There are two ultimate goals in this study:

(i) To identify, analyze, and rank each essay feature that infers the essay grade.
(ii) To attempt a machine learning framework to predict the essay grade based on the different feature groups.

To achieve the goals, we attempted to answer two specific questions of our automated essay scoring construct for MUET:

(i) What are the essay features that are influencing the essay grades?
(ii) What are the optimal essay features to be formulated in IEG for scoring MUET essays?

The answers to the questions are detailed in the following section, namely:

(i) The Correlation of IEG Essay Features with MUET Essay Grades
(ii) The Prediction of MUET Essay Grades using IEG

Figure 2 illustrates the overall process flow of the experiment carried out in this IEG work.

RESULTS AND DISCUSSIONS

The Correlation of Intelligent Essay Grader (IEG) Essay Features with MUET Essay Grades

To answer the question regarding the essay features which influence the MUET essay grades, we examined the correlation of each feature value with the essay grade. We hypothesize that the significant features will have a higher magnitude of correlation. By

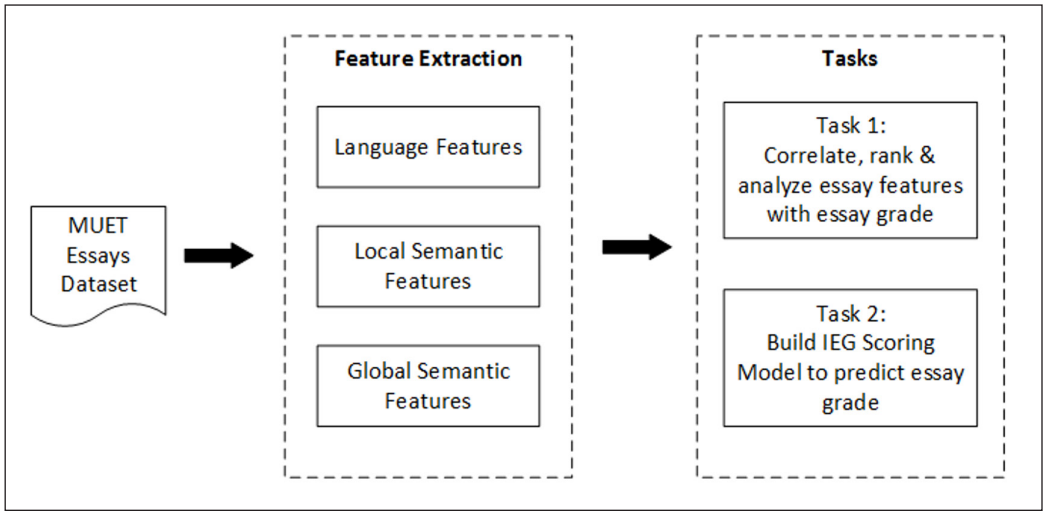


Figure 2. IEG process flow

using the Pearson Correlation Coefficient (Benesty et al., 2009), we investigated and discussed the relationship between the three essay dimensions of interest and the essay grades in the following aspects:

(i) Essay features ranking by Pearson Correlation Coefficient
(ii) Comparison of language dimension and semantic dimension effect on essay grades
(iii) Comparison of local semantic dimension and global semantic dimension effect on essay grades

Evaluation Metric – Pearson Correlation Coefficient, *r*. We used the Pearson Correlation Coefficient, *r* (Benesty et al., 2009) as the metric to evaluate the correlation of essay features and essay grades. Pearson Correlation Coefficient measures the effect size *r*, i.e. the strength of the linear relationship between two variables. The coefficient value ranges from -1 to 1, with a value greater than 0 indicates a positive correlation, while a value less than 0 indicates a negative correlation The larger the absolute value of the coefficient, the stronger the relationship between the variables. As represented by Equation 1 in the following, we quantified the strength of association between the essay features and the essay grades with this Pearson Correlation Coefficient. We used the Pearson Correlation Coefficient in this work as all the essay features (independent variables x) are continuous variable, whereby the essay grades (dependent variable y) is a continuous interval variable.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad [1]$$

where

x_i = value of a particular feature for the ith essay

y_i = grade of the ith essay

$\bar{x}$ = mean value of a particular essay feature

$\bar{y}$ = mean value of essay grade

Essay Features Ranking by Pearson Correlation Coefficient. The sorted Pearson Correlation Coefficient, *r-value,* between each essay feature and essay grade is attached in Appendix A (Table A1). To further analyze and interpret the relationship of the essay features and the essay grades, we plotted the data into a line chart to illustrate the effect size of each essay feature against the essay grade. As shown in Figure 3, we can notice that the language features such as vocabulary count, advanced parts of speech, word count, and paragraph count are highly correlated to the essay grades. On the other hand, language features such as grammatical error and spelling error rate demonstrated a strong negative correlation to the essay grades.

From the plot, we found out that:

(i) Essay features such as vocabulary count, advanced part of speech, word count, average words per paragraph, sentence count, and average sentences per paragraph are strongly positive-correlated with the essay grades. The higher of these features value indicates the higher essay grades. We also computed the average score for

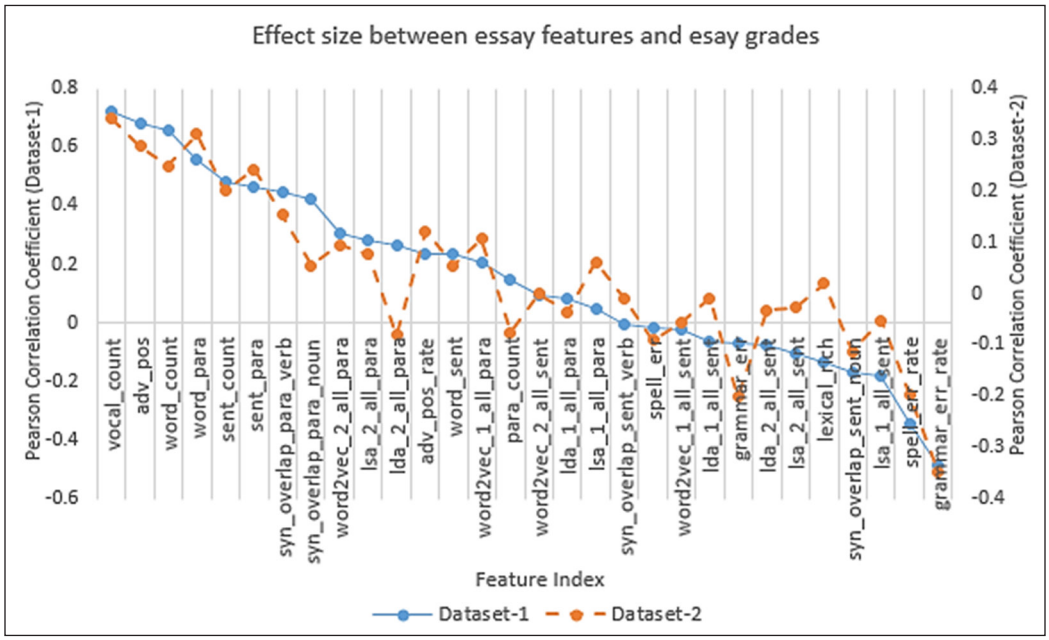


Figure 3. Effect size between essay features and essay grades

Table 6
Average score of highly correlation features in (a) Dataset-1, and (b) Dataset-2

(a)						
Essay Grade (y)	Essay Features (x)					
	vocal_count	adv_pos	word_count	word_para	sent_count	sent_para
1	79.7	31.2	190.5	41.7	12.3	12.3
2	132.8	58.0	351.1	66.5	21.0	21.0
3	157.8	70.3	430.4	73.2	25.4	25.4
4	190.9	86.3	509.1	88.8	28.3	28.3
5	270.6	128.7	646.5	112.2	33.6	33.6
6	293.3	142.6	752.8	142.1	35.3	35.3
(b)						
Essay Grade (y)	Essay Features (x)					
	vocal_count	adv_pos	word_count	word_para	sent_count	sent_para
1	N/A	N/A	N/A	N/A	N/A	N/A
2	183.2	85.6	488.8	83.1	25.1	4.3
3	192.2	86.7	480.8	86.5	26.3	4.7
4	209.7	95.4	530.6	95.4	28.6	5.2
5	249.4	124.7	626.6	110.3	31.8	5.6
6	235.0	88.0	508.0	127.0	27.0	6.8

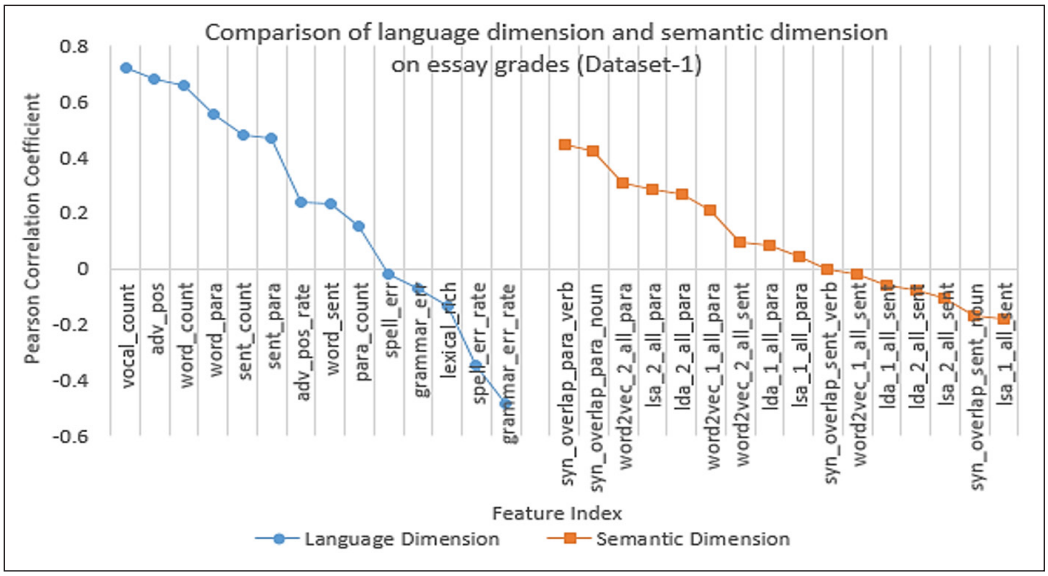
each feature as the following according to the essay grades. As shown in Table 6, we can notice that each average score shows an uptrend from the lower to higher essay grades.

(ii) Essay features such as grammatical error rate and spelling error rate are strongly negative-correlated with the essay grades. The stronger these features value; the lower the essay grades. The error rate (grammatical error rate and spelling error rate), which shows a higher *r-value*, compared with the error count (grammatical error count and spelling error count), indicates the use of error rate instead of its absolute value is more effective in predicting essay grade.

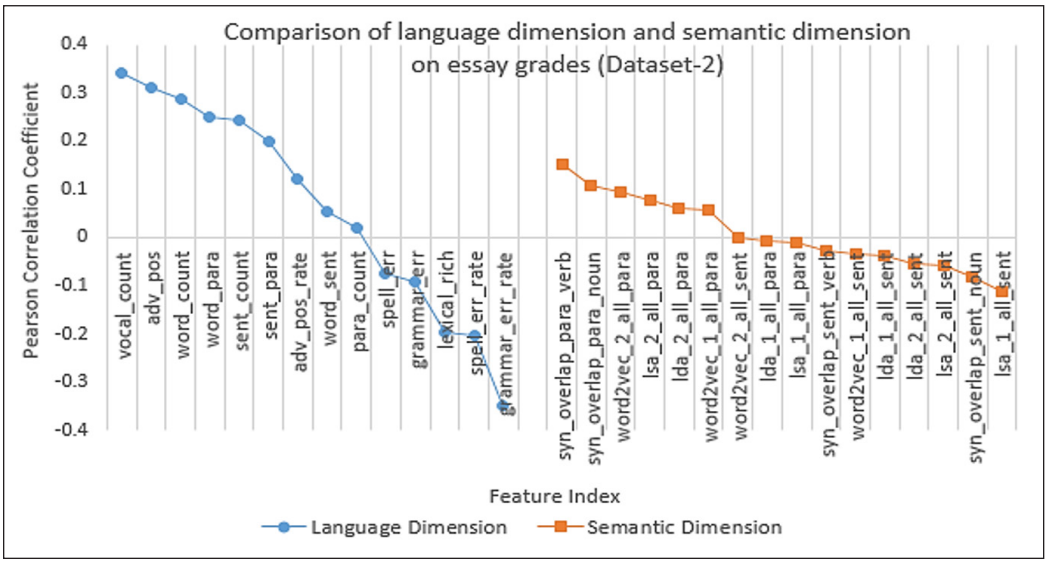
(iii) Holistically, the *r-value* of both Dataset-1 and Dataset-2 shows a consistent trend where the ranking of the sorted essay features are similar and close to each other for both the dataset. This observation justifies that our result of ranking the influential essay features are consistent and reliable.

Comparison of Language Dimension and Semantic Dimension Effect on Essay Grades. Figure 4 shows the comparison of language dimension and semantic dimension effect on essay grades in Dataset-1 and Dataset-2. From the plot, we observed that the language

dimension plot exhibits a greater gradient value with a steeper slope, compared with the semantic dimension plot. This greater gradient value denotes the language dimension covers a wider scope of effect size in both the positive and negative correlation, and thus indicate a greater influence on essay grades than the semantic dimension. This result of the higher impact of language features than semantic features in determining essay grades has been reported in works done by Crossley and McNamara (2011) and McNamara et al. (2010).



(a)

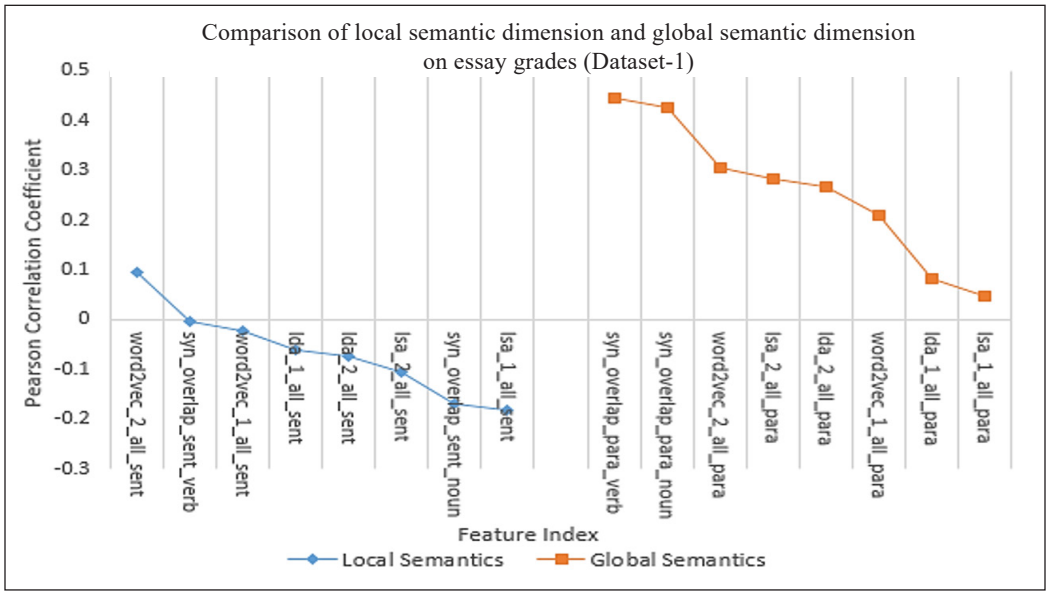


(b)

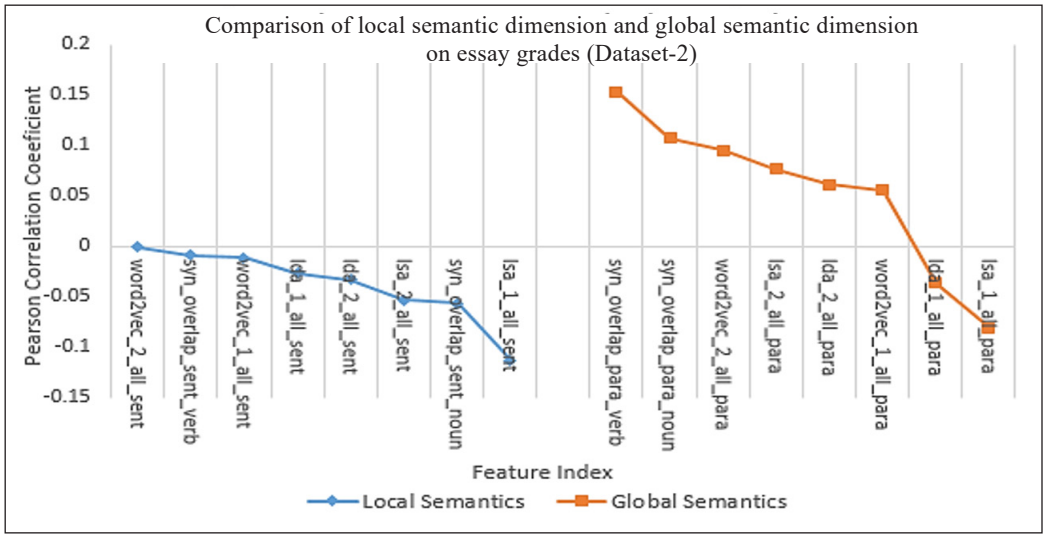
Figure 4. Comparison of language dimension and semantic dimension on essay grades in (a) Dataset-1, and (b) Dataset-2

Comparison of Local Semantic Dimension and Global Semantic Dimension Effect on Essay Grades. Figure 5 shows the comparison of local semantic dimension and global semantic dimension effect on essay grades in Dataset-1 and Dataset-2. From the plot, we observed that:

(i) The global semantic dimension plot exhibits a greater gradient value with a steeper slope, compared with the local semantic dimension plot. This greater gradient



(a)



(b)

Figure 5. Comparison of local semantic dimension and global semantic dimension on essay grades in (a) Dataset-1, and (b) Dataset-2

value denotes the global semantic dimension covers a wider scope of effect and thus indicates a greater impact on essay grades than the local semantic dimension. This little to no effect of local semantic features on essay grade is supported by the studies carried out by Crossley and McNamara (2011) and McNamara et al. (2010).

(ii) Most of the global semantic features show the *r-value* greater than 0. This finding denotes the global semantic dimension (corresponding to the paragraph scope) is positively correlated with the essay grades. The higher semantic similarity between paragraphs tends to produce higher essay grades. The work done by Crossley and McNamara (2016) exhibits the same observation of positive-correlation between global semantic features and essay quality.

(iii) Most of the local semantic features show the *r-value* less than 0. This finding denotes the local semantic dimension (associated with the sentence scope) is negatively correlated with the essay grades. The higher semantic overlap between sentences tends to produce lower essay grades. The same finding of the negative-correlation between local semantic features and essay quality has been reported in the work done by Crossley and McNamara (2016).

The Prediction of MUET Essay Grades using Intelligent Essay Grader (IEG)

To investigate the optimal essay features in scoring MUET essays, we attempted a machine learning framework to predict the essay grade based on the different feature groups defined in our method. We treated the problem of essay grading as a multiclass classification task where each grade was represented as a class.

Intelligent Essay Grader (IEG) Scoring Model. Figure 6 illustrates the framework of our IEG Scoring Model. The gold standard of our experiment was the two MUET datasets which we collected. Each essay in the collection is preprocessed by removing all punctuations and stop-words, changing all words to lower case, and lemmatizing the words. We then computed and extracted the essay features from each essay. The extracted essay features were then fed into various machine learning classifiers for training the essay scoring model.

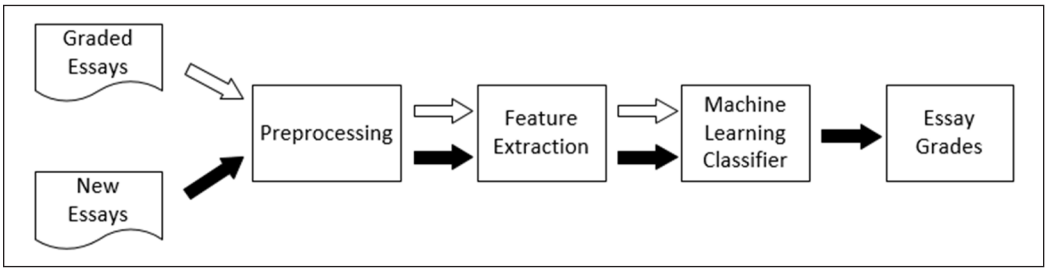


Figure 6. IEG scoring model

In the task, we employed four well-known machine learning classifiers, namely Logistic Regression (Cramer, 2002), Neural Network (Rumelhart et al., 1985), Random Forest (Breiman, 2001), and Support Vector Machine (Cortes & Vapnik, 1995), for building the IEG scoring model. Finally, we used the Leave-One-Out Cross Validation (Leave-One-Out Cross-Validation, 2011) to evaluate each of the machine learning classifiers in predicting the unseen essay grade.

To further investigate the classification accuracy based on different essay features, we built the IEG scoring model by using different features group as specified below:

(i) Language and Semantic Dimensions
All essay features as specified in Tables 2, 3, and 4.

(ii) Language Dimension Only
The language features group as specified in Table 2.

(iii) Semantic Dimension Only
The semantic features group as specified in Tables 3, and 4.

(iv) Top-7 Correlated Features
The Top-7 essay features with the highest absolute correlation value as listed in Appendix A (Table A1) *(vocal_count, adv_pos, word_count, word_para, grammar_err_rate, sent_count, sent_para).*

(v) Top-6 Correlated Features
The Top-6 essay features with the highest absolute correlation value as listed in Appendix A (Table A1) *(vocal_count, adv_pos, word_count, word_para, grammar_err_rate, sent_count).*

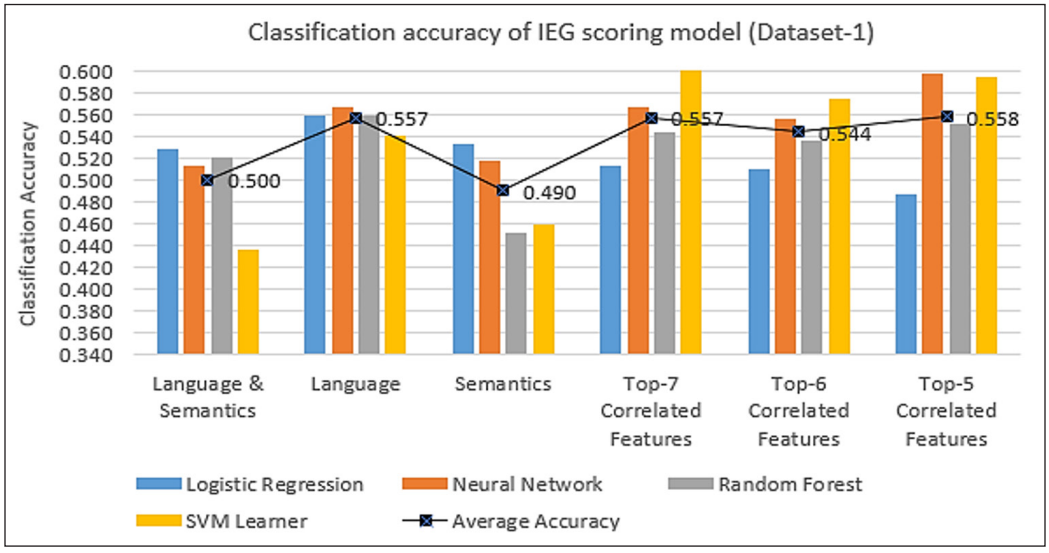
(vi) Top-5 Correlated Features
The Top-5 essay features with the highest absolute correlation value as listed in Appendix A (Table A1) *(vocal_count, adv_pos, word_count, word_para, grammar_err_rate).*

Evaluation Metric – Classification Accuracy, CA. Classification accuracy, *CA* (Accuracy, 2017) measures the ratio of numbers of the correct predictions to the total numbers of input samples. We evaluated our formulated IEG scoring model based on this classification accuracy as expressed in Equation 2.

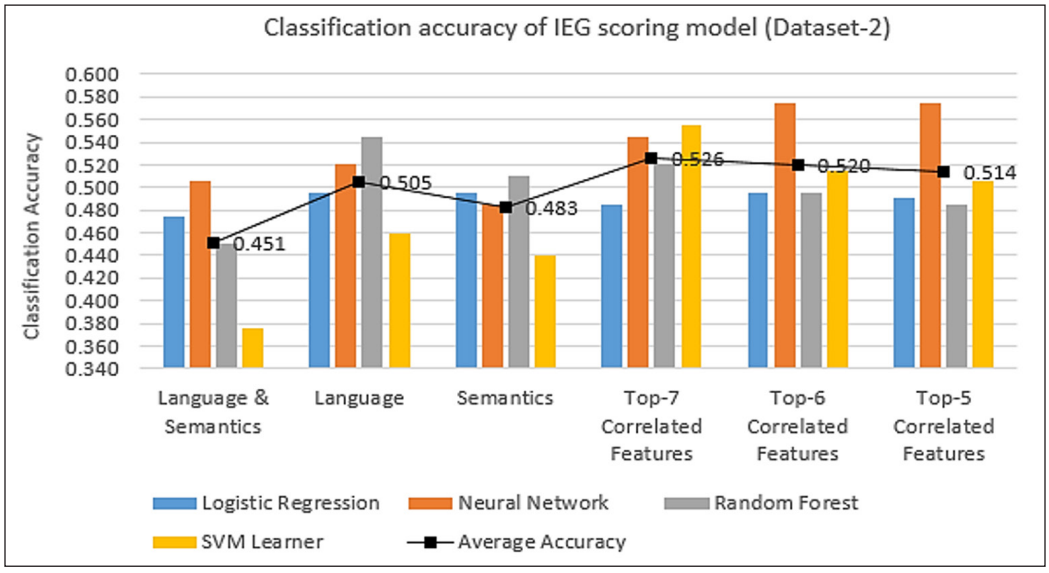
$$CA = \frac{Numbers\ of\ correct\ prediction}{Total\ numbers\ of\ essays\ to\ be\ predicted} \quad [2]$$

Results. The Classification Accuracy, *CA* of each essay feature group by employing various machine learning classifiers on Dataset-1 and Dataset-2 is attached in Appendix

B (Table B1). To facilitate effective visualization of the various CA values obtained by different essay feature groups, we plotted two combo charts (Figure 7); to show the CA values obtained in Dataset-1 and Dataset-2. In Figure 7, the CA values of each feature group with different machine learning classifiers are shown by the column chart; while the mean CA-values of a particular feature group averaged by the corresponding machine classifiers are shown in the line chart.



(a)



(b)

Figure 7. Classification accuracy of IEG scoring model in (a) Dataset-1, and (b) Dataset-2

Based on the average accuracy value in the plot, we observed the following results:

(i) The IEG scoring model based on the selected high correlated essay features (either Top-7, Top-6, or Top-5 highly correlated features) shows better CA results, compared with the prediction based on the language and semantic dimension alone or the combination of both. This result is consistent with the essay feature ranking by effect size shown in Figure 3.

(ii) The IEG scoring model based on the language dimension performed better than the scoring model based on the semantic dimension or the combination of both language and semantic dimensions. This result is consistent with the comparison of language dimension and semantic dimension effect size observed in Figure 4.

(iii) Among the different feature groups, the semantic dimension and the combination of both semantic and language dimensions produce lower CA results. The semantic dimension yields the lowest CA value in Dataset-1; while the combination of both semantic and language dimensions produces the lowest CA value in Dataset-2.

CONCLUSION

This paper discusses our work in formulating an Automated Essay Scoring, namely Intelligent Essay Grader (IEG), based on the English assessment context in Malaysia. In our work, we employed a total of 30 essay features based on language and semantic dimensions in assessing the MUET essays. We studied the relationship of each essay feature with essay grade and built an essay scoring function in predicting the essay grade. Based on our experiment, we summarized our findings and contribution as below:

(i) Instead of relying upon the publicly available corpus which does not reflect the essay writing in Malaysia, we collected and used the dataset of two essay prompts from MUET past year papers. This real-world dataset is essential in ensuring the validity of the constructed IEG.

(ii) Compared with other features, the essay language features such as vocabulary count, advanced part of speech, word count, average words per paragraph are significantly positive-correlated with essay grades; while grammatical error and spelling error are significantly negative-correlated with essay grades.

(iii) The essay language features show a higher effect on essay grades, compared with essay semantic features.

(vi) The global semantic features of the essay indicate a greater impact on essay grades, compared with the local semantic features.

(v) The global semantic features of the essay are mostly positively-correlated with the essay grades; while the local semantic features are mostly negatively-correlated with the essay grades.

(vi) Our formulated IEG scoring function produces the classification accuracy results, which is consistent with our findings observed in the correlation of the essay features and the essay grades. The scoring function yields better accuracy results based on the selected high-correlated essay features, following by the language features.

Based on the findings, we perceived the essential need to improve the classification accuracy of the IEG scoring model. In our future work. we are planning to:

(i) Refine the language and semantic features by removing the insignificant and redundant feature indexes to reduce the multicollinearity problem.
(ii) Identify other feasible high-impact language and semantic features on essay grades.
(iii) Incorporate other aspects of essay features such as essay content, organization, and argument strength for providing a comprehensive scope of IEG essay scoring.

ACKNOWLEDGEMENT

This study is supported and made possible by the RACE/B(6)/1098/2014(06) and ERGS/ICT07(01)/1018/2013(15) by the Malaysia Ministry of Higher Education. The authors would like to extend their greatest gratitude to the Malaysia Ministry of Higher Education for granting the funding to conduct this research.

REFERENCES

Accuracy. (2017). Accuracy. In C. Sammut & G. I. Webb (Eds.) *Encyclopedia of machine learning and data mining* (pp. 1-48). Springer. https://doi.org/10.1007/978-1-4899-7687-1_3

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. https://doi.org/10.1007/BF00994018

Cozma, M., Butnaru, A. M., & Ionescu, R. T. (2018). Automated essay scoring with string kernels and word embeddings. *Computation and Language, 2018*, 1-7.

Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute Working Paper No. 2002-119/4*. https://doi.org/10.2139/ssrn.360300

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning, 21*(2-3), 170-191.

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research, 7*(3), 351-370.

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. *Behavioral Research Methods, 51*(1), 14-27. https://doi.org/10.3758/s13428-018-1142-4

Darus, S., Stapa, S. H., & Hussin, S. (2003). Experimenting a computer-based essay marking system at Universiti Kebangsaan Malaysia. *Jurnal Teknologi, 39*(E), 1-18.

Educational Testing Service. (n.d.). *About the e-rater® scoring engine.* Retrieved October 30, 2020, from https://www.ets.org/erater/about

Foltz, P. W. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 167-184). Lawrence Erlbaum Associates.

Govindasamy, P. N., Tan, B. H., & Yong, M. F. (2013). Lower six students' preferred mode of feedback for essay revision. *Malaysian Journal of ELT Research, 9*(2), 82-104.

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

Kaggle (2012). *The Hewlett foundation: Automated essay scoring.* Retrieved October 30, 2020, from https://www.kaggle.com/c/ASAP-AES

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284. https://doi.org/10.1080/01638539809545028

Leave-One-Out Cross-Validation. (2011). Leave-One-Out Cross-Validation. In C. Sammut, & G. I. Webb (Eds.) *Encyclopedia of machine learning.* Springer. https://doi.org/10.1007/978-0-387-30164-8_469

LightSide. (2019). *LightSide researcher's workbench.* Retrieved January 11, 2021, from http://ankara.lti.cs.cmu.edu/side

Malaysian Examination Council. (2014). *Malaysian University English Test (MUET) - regulations, test specifications, test format and sample questions.* Retrieved October 30, 2020, from https://www.mpm.edu.my/images/dokumen/calon-peperiksaan/muet/regulation/Regulations_Test_Specifications_Test_Format_and_Sample_Questions.pdf

Manap, M. R., Ramli, N. F., & Kassim, A. A. M. (2019). Web 2.0 automated essay scoring application and human ESL essay assessment: A comparison study. *European Journal of English Language Teaching,* 5(1), 146-161. https://doi.org/ 10.5281/zenodo.3461784

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

Measurement Incorporated. (n.d.). *Automated Essay Scoring - Project Essay Grade (PEG®).* Retrieved October 31, 2020, from https://www.measurementinc.com/products-services/automated-essay-scoring

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111-3119.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Ng, S. Y., Bong, C. H., Hong, K. S., & Lee, N. K. (2019). Developing an automated essay scorer with feedback (AESF) for Malaysian University English Test (MUET): A design-based research approach. *Pertanika Journal of Social Science & Humanities,* 27(3), 1451-1468.

Nguyen, H., & Litman, D. (2018). Argument mining for improving the automated scoring of persuasive essays. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 5892-5899.

Omar, N., Razali, N. A. M., & 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.), *Lecture notes in computer science, Vol 5589: Rough Sets and Knowledge Technology* (pp. 475-482). Springer. https://doi.org/10.1007/978-3-642-02962-2_60

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan, 47*(5), 238-243.

Pearson Education. (2010). *Intelligent Essay Assessor (IEA)™ Fact Sheet* [Fact sheet]. Retrieved October 31, 2020, from https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf

Persing, I., & Ng, V. (2014). Modeling prompt adherence in student essays. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1*, 1534-1543.

Persing, I., & Ng, V. (2016). Modeling stance in student essays. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1*, 2174-2184.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Shermis, M. D., & Burstein, J. (2003). Introduction. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. xiii-xvi). Lawrence Erlbaum Associates.

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- 442.

Vantage Learning, (n.d.). *Intellimetric®.* Retrieved October 31, 2020, from http://www.intellimetric.com/direct

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), 69-78. https://doi.org/10.11113/ijic.v9n1.220

Zupanc, K., & Bosnic, Z. (2014). Automated essay evaluation augmented with semantic coherence measures. In R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, & X. Wu (Eds.), *2014 IEEE International Conference on Data Mining* (pp. 1133-1138). IEEE Conference Publication. https://doi.org/10.1109/ICDM.2014.21

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

APPENDIX A

Ranking of IEG essay features and MUET essay grades by Pearson Correlation Coefficient

Table A1
Ranking of IEG essay features and MUET essay grades by Pearson Correlation Coefficient

Feature Index	Essay Dimension	Pearson Correlation Coefficient, r		Feature Ranking	
		Dataset-1	Dataset-2	Dataset-1	Dataset-2
vocal_count	Language	0.724	0.342	1	1
adv_pos	Language	0.679	0.288	2	3
word_count	Language	0.658	0.249	3	4
word_para	Language	0.556	0.312	4	2
sent_count	Language	0.480	0.201	5	6
sent_para	Language	0.467	0.243	6	5
syn_overlap_para_verb	Global Semantic	0.445	0.153	7	7
syn_overlap_para_noun	Global Semantic	0.425	0.056	8	13
word2vec_2_all_para	Global Semantic	0.306	0.095	9	10
lsa_2_all_para	Global Semantic	0.283	0.077	10	11
lda_2_all_para	Global Semantic	0.268	-0.080	11	25
adv_pos_rate	Language	0.238	0.121	12	8
word_sent	Language	0.233	0.054	13	14
word2vec_1_all_para	Global Semantic	0.209	0.107	14	9
para_count	Language	0.150	-0.076	15	24
word2vec_2_all_sent	Local Semantic	0.096	0.000	16	16
lda_1_all_para	Global Semantic	0.082	-0.036	17	21
lsa_1_all_para	Global Semantic	0.047	0.062	18	12
syn_overlap_sent_verb	Local Semantic	-0.002	-0.011	19	18
spell_err	Language	-0.018	-0.091	20	26
word2vec_1_all_sent	Local Semantic	-0.021	-0.056	21	23
lda_1_all_sent	Local Semantic	-0.061	-0.008	22	17
grammar_err	Language	-0.069	-0.202	23	29
lda_2_all_sent	Local Semantic	-0.074	-0.033	24	20
lsa_2_all_sent	Local Semantic	-0.106	-0.027	25	19
lexical_rich	Language	-0.135	0.020	26	15
syn_overlap_sent_noun	Local Semantic	-0.167	-0.113	27	27
lsa_1_all_sent	Local Semantic	-0.183	-0.053	28	22
spell_err_rate	Language	-0.345	-0.197	29	28
grammar_err_rate	Language	-0.484	-0.349	30	30

APPENDIX B

Classification accuracy of IEG scoring model

Table B1
Classification accuracy of IEG scoring model in (a) Dataset-1, and (b) Dataset-2

(a)

	Language and Semantics	Language	Semantics	Top-7 Correlated Features	Top-6 Correlated Features	Top-5 Correlated Features
Logistic Regression	0.529	0.560	0.533	0.514	0.510	0.486
Neural Network	0.514	0.568	0.517	0.568	0.556	0.598
Random Forest	0.521	0.560	0.452	0.544	0.537	0.552
SVM Learner	0.436	0.541	0.459	0.602	0.575	0.595
Average Accuracy	0.500	0.557	0.490	0.557	0.544	0.558

(b)

	Language & Semantics	Language	Semantics	Top-7 Correlated Features	Top-6 Correlated Features	Top-5 Correlated Features
Logistic Regression	0.475	0.495	0.495	0.485	0.495	0.490
Neural Network	0.505	0.520	0.485	0.545	0.575	0.575
Random Forest	0.450	0.545	0.510	0.520	0.495	0.485
SVM Learner	0.375	0.460	0.440	0.555	0.515	0.505
Average Accuracy	0.451	0.505	0.483	0.526	0.520	0.514