Pertanika J. Sci. & Technol. 29 (3): 1875 - 1899 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

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

Review article

A Comprehensive Review of Automated Essay Scoring (AES) Research and Development

Chun Then Lim[1]*, Chih How Bong[1], Wee Sian Wong[1] and Nung Kion Lee[2]

[1]Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[2]Faculty of Cognitive Sciences & Human Development, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

ABSTRACT

Automated Essay Scoring (AES) is a service or software that can predictively grade essay based on a pre-trained computational model. It has gained a lot of research interest in educational institutions as it expedites the process and reduces the effort of human raters in grading the essays as close to humans' decisions. Despite the strong appeal, its implementation varies widely according to researchers' preferences. This critical review examines various AES development milestones specifically on different methodologies and attributes used in deriving essay scores. To generalize existing AES systems according to their constructs, we attempted to fit all of them into three frameworks which are content similarity, machine learning and hybrid. In addition, we presented and compared various common evaluation metrics in measuring the efficiency of AES and proposed Quadratic Weighted Kappa (QWK) as standard evaluation metric since it corrects the agreement purely by chance when estimate the degree of agreement between two raters. In conclusion, the paper proposes hybrid framework standard as the potential upcoming AES framework as it capable to aggregate both style and content to predict essay grades Thus, the main objective of this study is to discuss various critical issues pertaining to the current development of AES which yielded our recommendations on the future AES development.

Keywords: Attributes, automatic essay scoring, evaluation metrics, framework, human raters, recommendation

ARTICLE INFO

Article history:
Received: 20 January 2021
Accepted: 24 May 2021
Published: 31 July 2021

DOI: https://doi.org/10.47836/pjst.29.3.27

E-mail addresses:
limchunthen95@gmail.com (Chun Then Lim)
chbong@unimas.my (Chih How Bong)
weesian.wong@gmail.com (Wee Sian Wong)
nklee@unimas.my (Nung Kion Lee)
* Corresponding author

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

INTRODUCTION

The introduction of automated grading essays is an innovative attempt to reduce the effort of examining essays and eliminate assessment biases and its discrepancies. Automated Essay Scoring (AES) appears as a standalone computer software or distributed services that evaluates and scores a written prose (Shermis & Burstein, 2003). The objective of AES is to overcome the time, cost, and reliability issues in manual assessment of essays. It should be made clear that AES is not intended to fully replace the human assessors but to be employed as part of low-stakes classroom assessments to assist teachers' essay marking routine. On the other hand, it can be adopted in large-scale high-stakes assessments for the purpose of increasing reliability, where the AES serves as an additional rater for cross-examination.

AES aims at developing models that can grade essays automatically or with reduced involvement of human raters. It is a Natural Language Processing (NLP) based method and application of assessing educational works especially on writing tasks. AES systems may rely not only on grammars, but also on more complex features such as semantics, discourse, and pragmatics. Thus, a prominent approach to AES is to learn scoring models from previously graded samples, by modelling the scoring process of human raters. When given the same set of essays to evaluate and enough graded samples, AES systems tend to achieve high agreement levels with trained human raters. There are three common AES frameworks: Content Similarity Framework (CSF) which assigns grades or scores to new essays based on closer similarity of the reference essays' scores, Machine Learning Framework (MLF) which treats AES as classification or regression task and classifies the new essays into correspond grade category by using machine learning algorithms, and Hybrid Framework which combines the characteristics of both frameworks.

In a general AES process, the collected essays usually stored as text or Microsoft Word format. Hence, first step is to convert the group of collected essays into Microsoft Excel or json format which contains the content and grade for each essay. The essay content will then undergo pre-processing step such as tokenization, stop word removal, stemming and lemmatization to remove noise. After that, there will be a major difference in different frameworks. In cosine similarity framework, the pre-processed essay content will undergo word representation step which convert the essay into vector form and compared their similarity with gold standard's grade. For machine learning framework, pre-processed essay content will undergo feature selection, useful features will be extracted and act as input data for predict model. Then, suitable machine learning algorithms will be used to train the predict model to classify new essays into corresponding grades. The performance of AES system is then evaluated with measurement metric. The most common evaluation metric is using accuracy of to show the proportion of true results against the total number of predicted grades examined.

Project Essay Grade® (PEG) proposed by Ellis in 1966, is one of the earliest AES systems. The project determines the quality of the essay by focusing on essay writing style (Page, 2003; Rudner & Gagne, 2001). Subsequently, Intelligent Essay Assessor™ (IEA) by Pearson (2010) introduced an AES which could consider the essay's semantic context. In addition, e-rater®, from Educational Testing Service (ETS) is another revolutionary grading tool that uses computational methods to make sense of human natural language by the means of tagging, chunking, and other labels, based on a collection of learner's actual language uses. To date, My Assess, from IntelliMetric® model, is probably the pioneer essay scoring tools solely based on Artificial Intelligence (AI), which with is modelled by 45 computers on human intelligence (Shermis & Burstein, 2003).

Over the course of 40 years, we have started noticing that the trend of AES development and many commercial applications are pretty much emerged in the Western continents, especially in the United States. However, in recent years, there were many literatures on AES reported in Asia: Malaysia, Thailand, Philippine, and Indonesian covering mainly the English language as well as other languages.

To understand how AES has grown to its current state to meet the need in different regions and to anticipate its future development, a detailed survey on AES is essential. Thus, the purpose of this paper is to determine the recent progress of AES and generalize frameworks accordingly. Moreover, this paper lays a spectrum of the development frameworks for the reader by discussing the findings presented in recent research papers.

BACKGROUND

In this section, we provide an overview of AES systems that have achieved great success in commercialization and attracted greater publicity. These systems are mainly proprietary software developed in Western countries. For each product, we will present its vendor/ developer, primary focus, essay feature, scoring mechanism, and number of training samples required.

Project Essay Grader® (PEG)

Ellis Page's Project Essay Grader (PEG) is considered as the first AES (Page, 1966). It focuses on evaluating essays based on its writing style by using *trin*s and *proxes.* PEG assumes that there exist intrinsic qualities in a person's writing style known as *trins,* which can be measured or correlated with observable components denoted as *proxes* (Rudner & Gagne, 2001). For example, the fluency of an essay (*trin*) can be correlated with the amount of vocabulary (*proxe*). With training set of 100 to 400, PEG facilitates statistical regression analysis to estimate essay scores. To date, PEG has developed more than 500 *trins* to be used to score essays (Measurement Incorporated, 2020).

Intelligent Essay Assessor™ (IEA)

Intelligent Essay Assessor (IEA) is introduced by Pearson Knowledge Technologies (PKT) to assess the quality of essay contents (Foltz et al., 1999; Pearson, 2010)). IEA scores essay by using the Latent Semantic Analysis (LSA), which is a computational distribution model to assess the semantics similarity of texts (Landauer et al., 1998). LSA is operated on domain-specific corpus and the essays are represented through the multidimensional semantic space of the meaning of their contained words and the similarity is derived by comparing with other essay semantic representation (Foltz et al., 1999). IEA differed from other AESs on the aspects that the scores derived from LSA are aligned closely to human graders (Landauer et al., 2020), compared to the scores which are derived by correlation of essay features. In addition, IEA uses NLP techniques to extract essay attributes such as sophistication of lexical uses, grammatical, mechanical, stylistic, and organizational aspects of essays (Zupanc & Bosnic, 2015). Comparing with other ASE, IEA requires only a relatively small number of 100 pre-scored training essays sample for scoring a prompt-specific essay (Dikli, 2006).

IntelliMetric®

Vantage Learning proposed IntelliMetric as a proprietary AES to score essays operationally since 1998 (Vantage Learning, 2020). IntelliMetric is regarded as the very first AES system leveraging on Artificial Intelligence (AI) and Machine Learning (ML) to simulate the scoring process (Dikli, 2006; Hussein et al., 2019). In producing an essay score, IntellMetric uses more than 400 features (including semantics, syntactics, and discourses), which can be categorized into five groups of IntelliMetric Feature Model: focus and unity (coherence), organization, development and elaboration, sentence structure, mechanics, and conventions in its scoring process. IntelliMetric claims itself of having multiple automated scoring systems at work, each using a different mathematical model (e.g., Linear Analysis, Bayesian and LSA) for essay scoring (Vantage Learning, 2005; Vantage Learning, 2020). Such multiple scoring engine within IntelliMetric emulates the equivalent of a panel of multiple judges for achieving a more accurate final score as compared with a single scoring engine in others. Another distinctive feature of IntelliMetric is its ability in scoring essays in other languages besides English (Elliot, 2003). However, one of the downsides of IntelliMetric is it requires a minimum of at least 300 scored essays to be operated (Zupanc & Bosnic, 2015).

E-rater®

E-rater is developed and used by the Educational Testing Service (ETS) since 1999 (Attali & Burstein, 2006). It relies on patented NLP techniques to extract linguistic features for evaluating the style and content of an essay. E-rater 2.0 makes used of of syntactic,

discourse and topical-analysis module to analyze essay features (Dikli, 2006). The features are grammatical errors, word usage errors, mechanics error, style, organization segments and vocabulary content (Shermis et al, 2010). To date, E-rater version extends the essay scoring features into two areas:

(i) writing quality: grammar, usage, mechanics, style, organization, development, word choice, average word length, proper prepositions, and collocation usage

(ii) content or use of prompt-specific vocabulary (Ramineni & Williamson, 2018).

E-rater uses regressing modelling to assign a final score to an essay. In addition, a collection of approximately 250 training essay samples is required for the regression model (Zupanc & Bosnic, 2015).

The comparison of these well-known AES is shown in Table 1.

Table 1
Summary of well-known AES

AES System	Vendor/ Developer	Main Focus	Essay Scoring Mechanism	Essay-Scoring Features	Training Samples Required
PEG	Measurement Incorporated	Style	Statistical	*Trins & Proxes*	100 - 400
IEA	Pearson Knowledge Technologies	Content	LSA	• Content Style • Mechanics	100
IntelliMetric	Vantage Learning	Style & Content	AI - cognitive processing, computational linguistics, and classification	• Focus & Unity (Coherence) • Organization • Development & Elaboration • Sentence Structure • Mechanics & Conventions	300
e-rater		Style & Content	Regression Analysis	• Grammatical Errors • Word Usage Errors • Mechanics Errors • Style • Organizational Segment • Vocabulary Contents	250

PAST LITERATURE

In this section, we summarize the development of AES in previous studies and categorized their findings based on the type of attribute, methodology, prediction model and findings (Table 2).

The attribute refers to the aspects evaluated by the proposed models which include style, content, and hybrid. Style attributes focus on linguistic features such as spelling mistakes, essay length and stop word count. Content attributes works to verify the correctness of

Table 2
Summary of AES

Type of Attribute	Methodology	Prediction Model	Measure & Finding	Reference
Style	Natural language processing (Linguistic features)	Nonparametric Weighted Feature Extraction, Stepwise Regression and Discriminant Analysis	Accuracy (51.3%)	(Pai et al., 2017)
		Linear Regression	Close to human rater	(Ramalingam et al., 2018)
		Random Forest	Quadratic Weighted Kappa (0.8014)	(Chen & He, 2013)
		Bayesian Linear Ridge Regression (BLRR)	Quadratic Weighted Kappa (0.784)	(Phandi et al., 2015)
Hybrid	Natural language processing (Rhetoric, Organisation, Content)	Rule-based Expert System	Correlation with rater (0.57)	(Ishioka & Kameda, 2006)
	Vector Space Models (VSM)	Support Vector Regression (SVR)	Average correlation (0.6107)	(Peng et al., 2010)
	Natural language processing (Rhetoric, Organisation, Content and Length)	Rule-based Expert System	Pearson's correlation coefficient (0.562)	(Imaki & Ishihara, 2013)
	Latent Semantic Analysis, number of words, number of spelling mistakes, and word distance.	Linear Regression	Correlation with rater (0.78) Accuracy (96.72%)	(Alghamdi, et al., 2014)
	Latent semantic features	Support Vector Machine for Ranking	Pearson Corelation (0.7248)	(Jin & He, 2015)
	Latent Semantic Analysis, Rhetorical Structure Theory (RST) and hand-crafted features	Rule-based Expert System	Accuracy (78.33%)	(Al-Jouie & Azmi, 2017)
	Latent Semantic Analysis and Feature extraction	Linear Regression	Accuracy (47.16%)	(Contreras et al., 2018)
Content	Latent Semantic Analysis	Artificial Neural Network (ANN)	Mean of error (0.44)	(Loraksa & Peachavanish, 2007)
		Cosine similarity	Accuracy for small class (69.80 % – 94.64 %) Accuracy for medium class (77.18 % - 98.42 %)	(Ratna et al., 2007)

Table 2 (*continue*)

Type of Attribute	Methodology	Prediction Model	Measure & Finding	Reference
			Accuracy (83.3%)	(Amalia et al., 2019)
		Learning Vector Quantization	Accuracy (96.3%)	(Ratna et al., 2018)
	Topic classified by SVM and assessed by LSA	Frobenius norm	Average accuracy for Japanese (89.175%);	(Ratna et al., 2019a)
			Accuracy for Bahasa Indonesia (72.01%)	(Ratna et al., 2019b)
	GLSA	Cosine similarity	Precision, Recall and F1 scores (0.98)	(Islam & Hoque, 2013)
	Latent Semantic Analysis, Disco2, Damera-levenshtein and N-gram	Similarity degree	Correlation with rater (0.82)	(Shehab et al., 2018)
	Latent Semantic Analysis and Winnowing algorithm	Cosine similarity	Accuracy for LSA (87.78%); Accuracy for Winnowing (86.72%)	(Ratna et al., 2019c)
	Latent Semantic Analysis with multi-level keywords	Compared document vector	Human raters' agreement (86%)	(Ratna et al., 2015)
	Modified LSA and syntactic features	Cosine similarity	RMSE (0.268)	(Omar & Mezher, 2016)
	Enhanced Latent Semantic Analysis	Cosine similarity	Gap with rater (0.242)	(Sendra et al., 2016)
	Latent Semantic Analysis, Probabilistic Latent Semantic Analysis	Cosine similarity	Spearman correlation (0.78)	(Kakkonen et al., 2005)
	Arabic WordNet (AWN)	Cosine similarity	Pearson Corelation (98%)	(Awaida et al., 2019)
	Concept Indexing	Cosine similarity	Exact Agreement Accuracy (0.452)	(Ong et al., 2011)
	Contextualized Latent Semantic Indexing	Support Vector Machine (SVM)	Rating Agreement (89.67)	(Xu et al., 2017)
	Statistic (One-hot encoding)	Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bi-directional Long Short-Term Memory (BiLSTM)	Quadratic Weighted Kappa for LSTM + CNN (0.761)	(Taghipour & Ng, 2016)
	Statistic (Word embedding)	Convolutional Neural Network (CNN)	Average kappa value (0.734)	(Dong & Zhang, 2016)

Table 2 (*continue*)

Type of Attribute	Methodology	Prediction Model	Measure & Finding	Reference
		CNN and Ordinal Regression (OR)	Accuracy (82.6%)	(Chen & Zhou, 2019)
		Siamese Bidirectional Long Short-Term Memory Architecture (SBLSTMA)	Average Quadratic Weighted Kappa (0.801)	(Liang et al., 2018)
	Statistic (Training word vector)	2-Layer Neural Networks	Quadratic Weighted Kappa (0.9448)	(Nguyen & Dery, 2016)
	Hierarchical Recurrent Neural Network	CNN, LSTM, BiLSTM	Average Quadratic Weighted Kappa (0.773)	(Chen & Li, 2018)
	Natural language processing (Unigram Language Model)	Machine Learning Classifier	Mean accuracy (51.5%)	(Wong & Bong, 2019)
	Natural language processing (Word features, syntactic features, and dependency relation features)	Logistic Regression and k-Nearest Neighbors	Correlation with the rater (0.92)	(Cheon et al., 2015)

content meaning and similarity between an essay with the graded essays. Hybrid attribute facilitating both style and content.

Methodology refers to the methods used to identify the features from the essays. The prediction model records the techniques or algorithms used to predict the score or grade of the essay.

From Table 2, we can see that different researchers have developed their own methodology and used different prediction models. This has proliferated the development of AES because each methodology and prediction model does not seem to be universally accessible to other researchers. Hence, a standard framework of AES needs to be proposed so that all researchers can use, modify, and enhance it in the future. Moreover, the evaluation metric of AES also needs to be standardized so that the performance of AES system can be compared with one and another. Lastly, most of the articles did not described in details of their dataset used in research and therefore other researchers cannot reproduce the same result as they stated in their article.

THE GENERAL FRAMEWORKS OF AES

Despite the increasing number of literature reporting novel approaches for AES implementation, we can summarize them into three major general frameworks: content similarity, machine learning and hybrid.

Content Similarity Framework

The idea of content similarity framework (CSF) is to assign grades or scores to new essays based on closer similarity of the reference essays' scores. The framework requires a gold standard: a collection of human graded reference essays, covering all spectrum of grades or scores on the respective topics. The workflow of content similarity framework is illustrated in Figure 1.

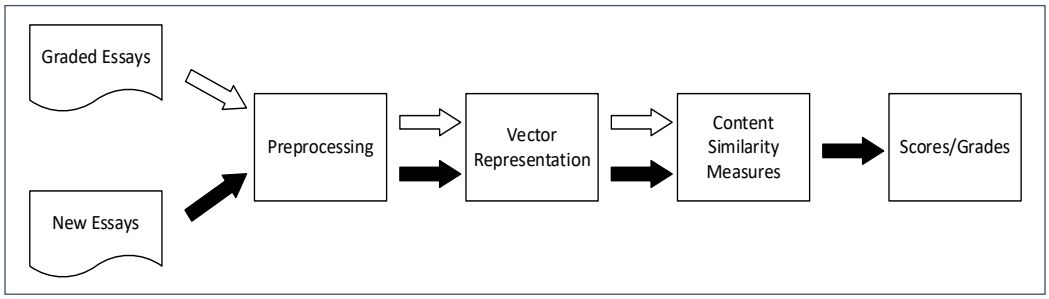

Figure 1. The workflow of a content similarity framework

In this framework, selected essays firstly undergo pre-processing step, which includes tokenization, stop words removal, stemming and lemmatization to reduce the noise in essays. The subject of similarity can be based on (a) syntactic or, (b) semantics indicators, or a combination of both.

(a) Syntactic Indicators. Syntactic indicators refer to the essays' surface features such as part-of-speech, stemmed words, word connectors, and word count. Whereas the semantics indicators refer to the meaning of word, phrase, sentence, and text. It is commonly regarded that the semantics indicators are used to justify the whole or partial essays' semantics similarity (Islam & Hoque, 2013; Omar & Mezher, 2016; Sendra et al., 2016; Landauer et al., 2000; Ghosh & Fatima, 2008). The common syntactic indicators used in AES are spelling checking, stemming, lemmatization, word segmentation (Loraksa & Peachavanish, 2007), n-gram (Islam & Hoque, 2013; Chen & Zhou, 2019; Xu et al., 2017), and normalization (Taghipour & Ng, 2016; Ratna et al., 2019a). These were the essays' surface features, and they are found to be useful in grading essays (Ong et al., 2011). In addition, there were works reported to facilitate external knowledge bases such as WordNet (Omar & Mezher, 2016; Shehab et al., 2018) and ontology (Contreras et al., 2018) to improve grading efficacy.

In addition, the Japanese Scoring System (JESS) demonstrates an example of facilitating syntactic features on essay grades based on three syntactic categories: rhetoric, organization, and contents (Ishioka & Kameda, 2006). These categories are quantified by readability, percentage of long, different words, passive sentences, orderly presentation

idea, and topical vocabularies. The essay score is then derived based on the deduction mechanism of the essay's perfect score. However, the uses of syntax and style alone are not enough to determine the merits of the essay. Thus, JESS used syntactic indicators and semantic indicators in its content analysis.

(b) Semantics Indicators. In recent years, AES solely based on syntactic indicators are getting scarce, as many developments discovered that semantics indicators render more accurate grades or scores. In natural language modelling, a semantic space aims to create representations of the natural language that can represent the context. The most basic semantic space can be tracked back to Vector Space Model (VSM), which was used to derive content similarity based on the co-occurrence word in essays (Al-Jouie & Azmi, 2017).

One of the most popular syntactic-blind semantics indicators is Latent Semantic Analysis (LSA), which excels at deriving content analysis (Landauer et al., 1998). LSA is a distributional model used to derive meaning from a text. LSA was deemed "… a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text". With LSA, essays are represented as a term-document matrix, which in turn is approximately reduced using singular value decomposition (SVD). The dimension reduction process in LSA is to induce the probable similarity of every word to every other if they are ever occurred in in other essays at a common context. Experiment results showed that the addition of LSA over syntactic features improves the scoring performance of AES (Omar & Mezher, 2016). Many contents similarity-based AES used LSA or any of its variations in deriving grades or scores (Awaida et al., 2019; Amalia et al., 2019; Alghamdi, et al., 2014; Contreras et al., 2018; Ong et al., 2011; Shehab et al., 2018).

On the other hand, Generalized Latent Semantic Analysis (GLSA) is a variant of LSA which considers word sequence and has been reported to be capable of improving the efficacy of AESs (Islam & Hoque, 2013; Sendra et al., 2016). Almost all modern development of AESs reported are composed of both syntactic and semantics features.

The study of essay grading using CSF took two inputs: key answers and student answers (Amalia et al., 2019). Both essay inputs are pre-processed through noise removal, case conversion, tokenization, stopwords removal negation, conversion, stemming, synonym conversion, which are then represented using a term-document matrix where each row corresponded to the term and each column corresponded to the document. Each cell in the matrix represents the occurrence of the term to the documents. A zero value indicate the absence of the term in the documents.

The most common similarity computation is derived through LSA, which is a 2-step process: Singular Vector Decomposition (SVD) and cosine similarity measure. SVD is responsible to decompose the term-document matrix as $D=U\Sigma VT$. The k largest singular

values (which is dimensionality) is used to approximate D as $D \approx U_k \Sigma_k V^T_k$. Its purpose is to the discover "latent" concepts in the matrix. SVD is first applied to the key answers, where each of the student answer is then go through the same pre-processing processes and match it to the most similar key answers using cosine similarity measure to determine the scores. The workflow of essay grading using CSF is illustrated in Figure 2 (Amalia et al., 2019).

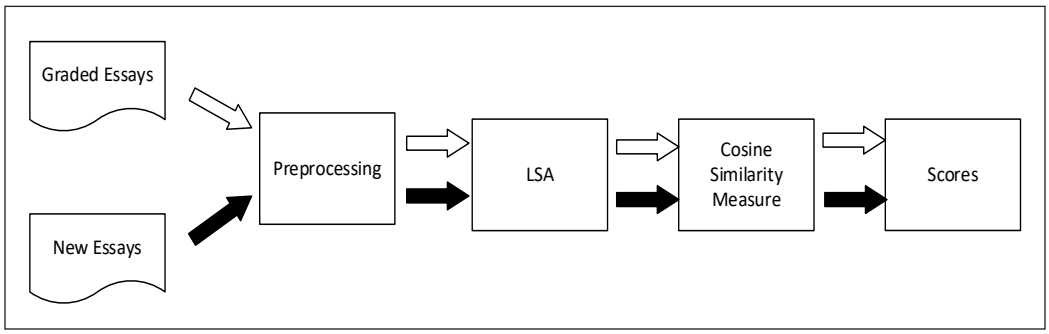

Figure 2. A case of essay grading using CSF (Amalia et al., 2019)

Machine Learning Framework

As shown in Figure 3, in Machine Learning Framework (MLF), essay grading is treated as a multiclass classification problem in which each grade is represented as a class. Modelling requires computational functions to generalize all essays into multi classes. Since AES has been seen as the document classification problem, the machine learning algorithms used are mainly from the categories of regression and classification. The workflow of machine learning framework is illustrated in Figure 3.

Pre-processing is the first process which prepares the data and removes noises. Similar to CSF, all essays will undergo tokenization, stop word removal, stemming and lemmatization processes. Next, the essays are processed to retain significant features in the Feature Selection process. Typical features in MLF were words (Cheon et al., 2015), syntactic and dependency features. Feature selection is an important step in many machine

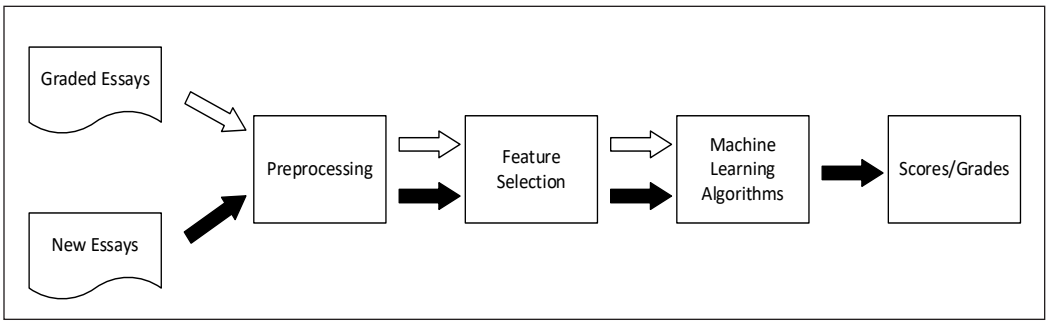

Figure 3. The workflow of a machine learning framework

learning tasks with the purpose to identify a significant feature subspace in reducing redundant features and reducing complex computational space, yielding the optimal essay representation. In this context, feature selection reduced the number of words to prevent the curse of dimensionality that can eventually degrade the accuracy of classification. In general, there are two dimensionality reduction techniques: feature elimination or feature selection. After dimensionality reduction, the selected features will act as inputs to train the machine learning model such as Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Naive Bayes and Artificial Neural Network (ANN).

Like CSF, a machine learning framework requires a gold standard, however, the graded essays are compulsorily to be processed and transformed into a computational model to be used for prediction of grades and scores on new essays. This is one of the significant differences between the CSF and MLF.

The study reported in Taghipour and Ng (2016) adopted MLF to derive essay scores. Both key and student answers are pre-processed with tokenization, case conversion and normalize the essay score in the range of [0,1]. Feature selection is performed through Enhanced AI scoring engine (EASE) is used to derive length-based representation, POS, word overlap with the key answers, and bag of n-gram. The features are then fed into machine learning algorithms such as support vector regression (SVR), Bayesian linear ridge regression (BLRR) and a variant of neural networks to derived student answer scores, resulted marginal increment against the baseline. The workflow of machine learning framework reported in Taghipour and Ng (2016) is illustrated in Figure 4.

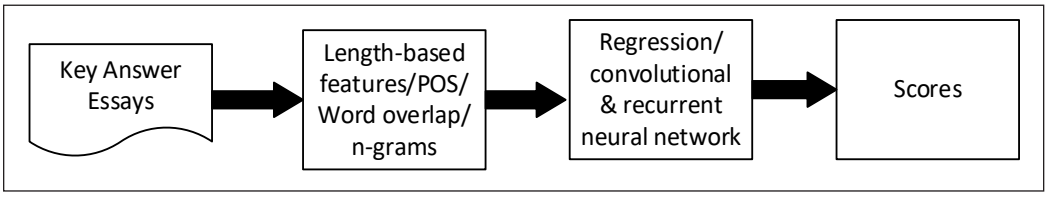

Figure 4. A MLF reported in (Taghipour & Ng, 2016)

Looking at the recent trends in AES, the machine learning framework is gaining popularity in recent years due to the efficacy of SVM (Ratna et al., 2019b; Ratna, et al., 2019a; Xu et al., 2017; Awaida et al., 2019; Chen & Li, 2018) and the ability to represent text context with word embedding (Liang et al., 2018; Taghipour & Ng, 2016) propelled by Artificial Neural Network (ANN) (Loraksa & Peachavanish, 2007; Taghipour & Ng, 2016; Dong & Zhang, 2016; Liang et al., 2018).

Hybrid Framework

The emerging hybrid framework (HF) combines both the goodness of content similarity and machine learning framework in which it capable of aggregating both style and content

to derive essay grades or scores. Different from the general MLF, where machine learning algorithms are directly used to derive the grades or score, the machine learning algorithms in the framework is used to generalize syntactic features (indices, topics, and domain specific keywords) where CSF is used to retrieve the closest key answer score in the semantic space. The process pipeline of hybrid framework is similar with the ML framework except both machine learning algorithms and content similarity measure are used to derive essay scores or grades. Figure 5 illustrates a general workflow of the hybrid framework.

The emerging of recent studies incorporating the hybrid framework have been seen the combining the use of linear regression on selected features, derived from an ontology and using LSA to measure content similarity (Contreras et al., 2018), generalizing essay scores through vectorization with artificial neural network and LSA (Loraksa & Peachavanish, 2007) and twostep process classification with SVM and LSA Framework (Ratna et al., 2019a; Ratna, et al., 2019b)

The studies reported in Ratna et al. (2019c) used two-step grading process as the HF: used Support Vector Machine (SVM) to classify essay's topic and LSA to compute the similarity between student and key answers. A pre-trained SVM model on key answers' topics is meant to rule out unrelated topic essays and to route it to the related key answers

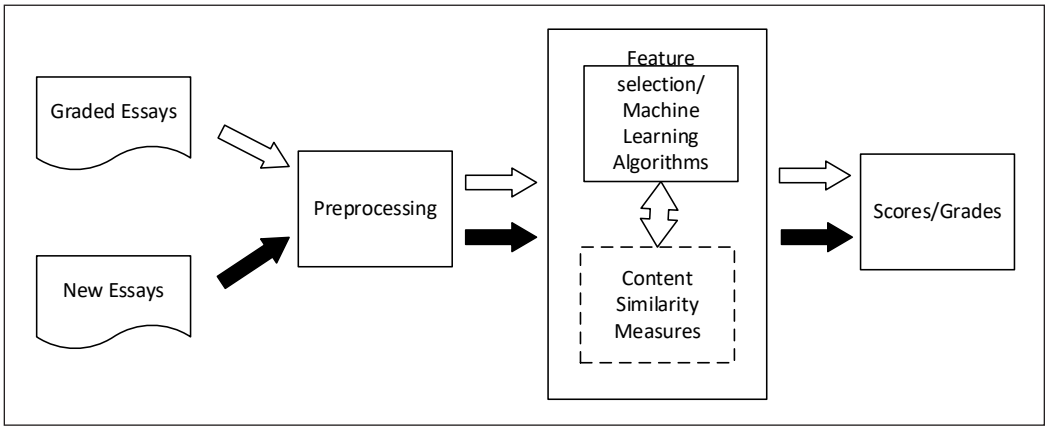

Figure 5. The workflow of a hybrid framework

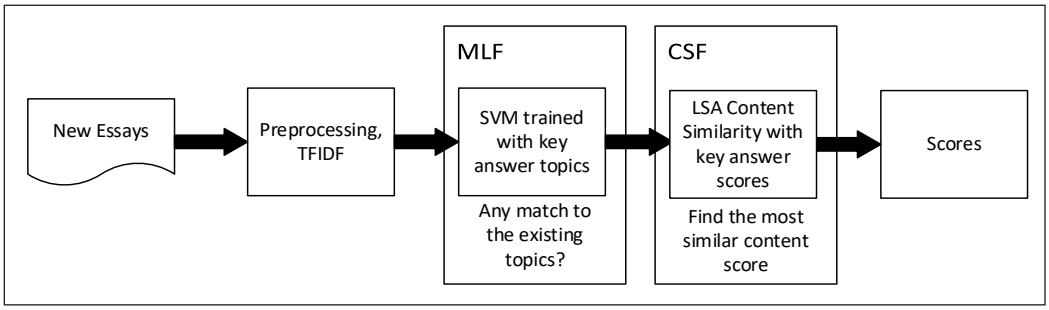

Figure 6. A Hybrid framework reported in (Ratna et al., 2019c)

essay. The LSA is intended to build a semantic space for key answer essays, where each of the student essay's score in the range of [0,100] is obtained by finding the most similar key answer score, using Frobenius norm similarity measure. The studies report substantial accuracy (>95%) as compared to human raters. The workflow of hybrid framework reported in Ratna et al. (2019c) is illustrated in Figure 6.

AES EVALUATION METHOD

In this section, we categorized the automated essay scoring evaluation methods from past literature into five categories. All evaluation methods require the human rater to provide a score or a grade for each essay to act as a reference. Then, the evaluation method will compare the human rater's score with the AES model predicted score to examine the accuracy of the proposed model.

Overall Accuracy

This evaluation method compares the average score manually marked by the human rater and scores automatically generated by the proposed model to examine the accuracy. The predicted result is classified into three classes, which are Exact Agreement Accuracy (EAA), Adjacent Agreement Accuracy (AAA) and Overall Accuracy (OA). Exact Agreement Accuracy is the number of essays with human score equal to AES score over the total numbers of essays, as denoted in Equation 1.

$$EAA = \frac{num\_human\_equal\_aes}{total\ number\ of\ test\ essays} \quad [1]$$

The Adjacent Agreement Accuracy is defined as the ratio of the number of test essays with a human score equal to machine score ± 1 over the total number of test essays and is given by the Equation 2.

$$AAA = \frac{num\_human\_equal\_aes\ \pm\ 1}{Total\ number\ of\ test\ essays} \quad [2]$$

Overall Accuracy is the sum of EAA and AAA and is defined as the ratio of the number of test essays with a human score equal to AES score ± 1 over the total number of test essays (Equation 3).

$$OA\ =\ EAA\ +\ AAA \quad [3]$$

This evaluation method is simple and able to evaluate the accuracy of the scoring model intuitively. However, the result can be inaccurate when the dataset contains an imbalance ratio of data.

Root Mean Square Error

Root Mean Square Error (RMSE) is used to examine the similarity between the human rater's score set and system predicted score set. RMSE is calculated by Euclidean distance hence it can use to evaluate real values such as essay score. Therefore, the accuracy of predict model depends on the similarity between the human rater's score and system predicted score based on Euclidean distance (Omar & Mezher, 2016). RMSE can be calculated as Equation 4.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(HS_i - SS_i)^2}{n}} \quad [4]$$

where HS_i is human rater's score, SS_i is the system predicted score and n is number of essays. The smaller value of RMSE indicates that the predicted score is more similar to human rater's score and therefore achieve a better scoring.

Mean of Error and Standard Deviation of Errors

The mean error and standard deviation of errors are used to examine the difference between the human rater's score, and the system predicted score. This evaluation method requires a state-of-art system as a baseline to examine the performance of proposed system. The mean of error between human rater's score and baseline system predicted score is calculated and compare with the mean of error between human rater's score and proposed system predicted score. The smaller value of mean of error and standard deviation of error represent the system has better performance in essay scoring. The mean of error and standard deviation of error is calculated by using Equation 5 and 6:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad [5]$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \grave{} x)^2}{n}} \quad [6]$$

$\overline{x}$ is the arithmetic mean from all errors,
x_i is an absolute value of an error between human score and machine score computed by $x_i = \mid HumanScore_i - MachineScore_i \mid, i = 1 \ldots n$
n is the number of data set
SD is a standard deviation of error

Pearson's correlation coefficient, Cohen's Kappa Coefficient and Quadratic Weighted Kappa (QWK)

Pearson's correlation coefficient, Cohen's Kappa coefficient and QWK scores are used to examine the degree of agreement between the human rater's score, and the system predicted score. The output of these evaluate methods will between 0 and 1 when 1 means that the system predicted score is complete agreement towards human rater's score and 0 means that is random agreement between system predict score and human rater's score. In interpreting any kappa value, K can be considered as poor when lower than 0.4, fair to good when K between 0.4 and 0.75, and excellent when K is greater than 0.75. To interpret Pearson's correlation, r can be considered as very small when r lower than 0.2, small when r between 0.2 and 0.4, medium when r between 0.4 and 0.6, large when r between 0.6 and 0.8, and very large when r greater than 0.8 (Cheon et al., 2015). Theses evaluation methods are suitable for evaluating the multiclass classification model since it calculates a confusion matrix between the predicted and actual values. Usually, these methods require a state-of-art scoring model as a baseline to examine the performance of the proposed model.

Mean of Accuracy

The scoring model is evaluated by obtaining the mean of accuracy of each test date. In achieving this, the difference between human rater's score and system predicted score will be derived. The accuracy for each test data will be calculated by using Equation 7.

$$Accuracy = 100\% - \frac{|\text{Human score} - \text{System score}|}{100} * 100 \qquad [7]$$

It should be noted that |Human score – System score| is an absolute value. By using this formula, the smaller the difference between human rater's score and system predicted score, the higher the accuracy of the scoring model. This evaluation method is suitable for evaluate the actual score of essays but not essay grade due to the gap between essay grade is usually small and result in overrated scoring model accuracy.

ISSUES

There Is No One-Size-Fits-All Solution

Most of the existing essay scoring systems reported thus far performed well in grading pure English essays or essays written in pure European language. However, the system graded a 10-15% lower score on essays containing Asian local content (Ghosh & Fatima, 2008). This is due to the influence of local languages and English written by non-native English speakers. For example, Asian students tend to obtain lower scores in TOEFL exams. Besides, Wong and Bong (2019) claim that the direct adoption of contemporary automated

essay scoring system in the Asian context may not be practical as it may lead to the issue of assessment reliability and validity. Reliability means how well an assessment tool can produce stable and consistent results even in different time and place, whereas validity means how good an assessment tool measures what it is supposed to measure.

Existing Automated Essay Scoring Systems Predict Essay Grade Ineffectively under a Prompt-Independent Setting

Essay prompt refers to an essay title or topic which the writer requires to treat as the main content of the essay. From the literature, majority of the essay scoring mechanism rely on rated essays as the gold standard and the performance of the system is highly dependent on these training data (Jin et al., 2018). However, this approach of training is hard to perform especially when the rated essays for a target prompt are difficult to obtain or even inaccessible due to legal, copyright or privacy issues which lead to inefficacy. In addition, essays for different prompts may differ a lot in the uses of vocabulary, structure, and grammatical characteristics. Hence, these models can hardly be generalized and fail to grade them accurately for non-target prompts. This situation is because prompt-dependent models are designed to learn the features from prompt-specific essays.

Effective AES Systems Requires Expert Tuning

A study by Alikaniotis et al. (2016) stated that the predictive features of the automated essay scoring system need to be manually crafted by human experts to achieve satisfactory performance and the process will consume a lot of time and work. The lack of human experts results in a decrease in AES performance and an increase in the AES development time.

The Relationship between Essay's Features and Grade is not Linear

Most of the existing automated essay scoring systems assume linear relationship between the features of the essay and the essay grade. However, the study by Fazal et al. (2011) stated that there exists a non-linear relationship between the feature vector and essay grade. For example, most of the existing automated essay scoring systems treat the length of essay as an important feature and use it to indicate the quality of essay which means the longer the essay, the higher the essay grade. However, in some specific cases, an essay will be assigned with a lower grade even though it is very long because of the irrelevant content in the essay. Moreover, automated essay scoring system will delete unnecessary and non-meaningful words to filter out keywords to evaluate during processing the essay. This makes automated essay scoring system easy to be fooled by students because it is unable to distinguish between good writing and baloney. In order words, the act of using and repeating some key words from the prompt, fill up lots of space may result in a higher

grade from the automated essay scoring system (Greene, 2018). It is also reported in Davis (2014) that essays can obtain high scores even with gibberish which makes no sense to human readers.

AES are Sensitive to Noise

Fazal et al. (2011) also reported that the noise from the essays will have a negative effect on the performance and efficiency of automated essay scoring system. These noises include punctuation errors, syntax-based errors, morphological, context-based spelling errors, and misspellings. Hence, it is important to rule out the noise before the modeling process because it will lower the performance.

Essay Feedback Given by Automated Essay Scoring System is Unable to Increase Student's Grade Significantly

Darus et al. (2003) had conducted a study in Malaysia to investigate the improvement of students after revising their essays based on the feedbacks given by automated essay scoring system. Based on the results, the revisions made did not significantly increase the score of revised essays for most of the students. Furthermore, students who participate in the relevant research found that the feedback given by automated essay scoring system is useful and informative to a certain extent although their score remained the same. This suggests that feedbacks given in AES are not sufficient in pointing out areas that could potentially increase the students' scores.

AES do not Consider Context and Rating Criteria

To develop a context-aware AES, the essay context information should be reinforced to build an AES which can distinguish poor, ordinary, and excellent essays. Besides, semantic features should be added to assist the AES in context grading process. On the other side, some existing AES only focus on the essay content and did not consider rating criteria behind the essay during grading process (Liang et al., 2018). It is because a few rating criteria are difficult to integrate in AES and therefore lower the accuracy of the system.

DISCUSSION

Standardizing AES Framework

The review of previous studies on AES has shown that various frameworks have been used and this has proliferated the development of AES because each framework does not seem to be universally accessible to other researchers. After reviewed all frameworks used in AES system, we suggest hybrid framework as standard framework for AES as shown in Figure 5 because it is a hybrid framework combines both the goodness of content similarity

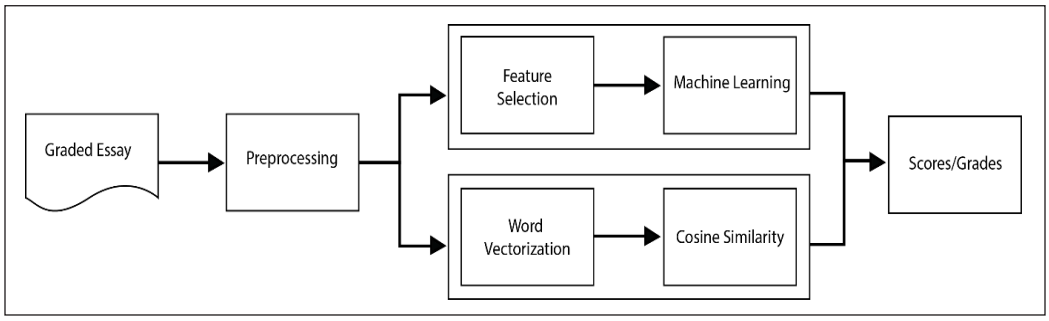

Figure 7. The workflow of proposed standard AES framework

and machine learning framework in which capable to aggregate both style and content to predict essay grades or scores. Besides, this framework enables researchers to apply different predictive models and NLP techniques on it and examine their performance. The workflow of proposed standard AES framework is illustrated in Figure 7.

For Malaysia University English Test (MUET), the marking scheme of an essay is the aggregation of task fulfilment, language, and organisation (Malaysian Examinations Council, 2014). Task fulfilment refers to the ability to understand topic and developing ideas which can be benchmarked with semantic indicators. Whereas the language and organisation refer to the number of grammar error, number of spelling error, use of appropriate vocabulary and coherence of content, are regarded as syntactic indicators. Hence, the machine learning part of our proposed framework is aimed to evaluate language and organisation using where the semantic indicators can be evaluated with content similarity measures.

In this framework, the essays undergo pre-processing such as word tokenization, predicting Parts of Speech, lemmatization and stop words removal to reduce the noises in the essay. Next, the features of the essay such as essay length, word count and misspelled word count may need to be extracted to ensure these features will not result in biased decision after vectorization of essay content. Then, the pre-processed essay undergoes word vectorization step to vectorize the essay content and feature selection to extract syntactic features from essay. Finally, the features and vectors are both treated as the inputs and used to train the predictive model using machine learning and content similarity measures.

Recommended Evaluation Metric

One of the issues with existing literatures is each researcher used different evaluation metrics to examine the performance of AES. This causes discrepancy when comparing the results with another works.

Accuracy has been seen as the most common evaluation metrics used in evaluating the performance of AES by calculating the percentage of matched pairs with human raters.

However, the result may be skew and biased toward the majority class when dealing with imbalanced datasets (Tanha et al., 2020). In real life, dataset normally contains imbalance proportion of grades since the distribution of the essays grade are normal.

Hence, this paper suggests that using Quadratic Weighted Kappa (QWK) score as a standard evaluation metric for AES. QWK is belong to kappa-like family and kappa coefficient has been proven that it can provide valuable information on the reliability of ordinal scale data (Sim & Wright, 2005). Moreover, Wong and Bong (2019) had stated that Kappa value is a better measurement than simple percent agreement calculation in AES because it corrects the percent-agreement for the case of agreement that would be expected purely by chance when estimating the degree of agreement between two raters.

CONCLUSION

Automated Essay Scoring (AES) is a software or service which grade essays with high human rater agreement. The objective of AES is to reduce the cost of time and effort on grading essays. In the current stage, the presence of AES cannot replace human raters in essay grading tasks, but it can assist human raters as a second rater. Most of the AES are developed in Western countries and there is still no commercial AES developed in the Asian region. According to past literature, most studies focused on evaluating essays based on its content by using Latent Semantic Analysis (LSA) technique. Moreover, most of the reported implementations treat AES as a supervised document classification task. Hence, this paper proposed three types of supervised general frameworks (content similarity, machine learning and hybrid) based on the past literature and a new framework which evaluates essays based on content and linguistic features. This review has also shown that different AES research used different evaluation methods to examine the proposed model performance, and this causes difficulty in doing a proper comparative study. Hence, this paper proposes Quadratic weighted kappa (QWK) as a standard method to evaluate AES performance and this can help to standardize the development of AES.

ACKNOWLEDGEMENTS

This study is supported by Kementerian Pendidikan Tinggi Malaysia Prototype Research Grant Scheme PRGS/1/2019/ICT01/UNIMAS/02/1 and Exploratory Research Grant Scheme ERGS/ICT07(01)/1018/2013(15).

REFERENCES

Alghamdi, M., Alkanhal, M., Al-Badrashiny, M., Al-Qabbany, A., Areshey, A., & Alharbi, A. (2014). A hybrid automatic scoring system for Arabic essays. *AI Communications, 27*(2), 103-111. https://doi.org/10.3233/aic-130586

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 715-725). Association for Computational Linguistics. https://doi.org/10.18653/v1/p16-1068

Al-Jouie, M., & Azmi, A. (2017). Automated evaluation of school children essays in Arabic. *Procedia Computer Science, 117*, 19-22. https://doi.org/10.1016/j.procs.2017.10.089

Amalia, A., Gunawan, D., Fithri, Y., & Aulia, I. (2019). Automated Bahasa Indonesia essay evaluation with latent semantic analysis. *Journal of Physics: Conference Series, 1235,* Article 012100. https://doi.org/10.1088/1742-6596/1235/1/012100

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-29.

Awaida, S. A., Shargabi, B. A., & Rousan, T. A. (2019). Automated Arabic essays grading system based on F-score and Arabic wordnet. *Jordanian Journal of Computers and Information Technology (JJCIT), 5*(3), 170-180. https://doi.org/10.5455/jjcit.71-1559909066

Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1741-1752). Association for Computational Linguistics.

Chen, M., & Li, X. (2018). Relevance-based automated essay scoring via hierarchical recurrent model. In *2018 International Conference on Asian Language Processing (IALP)* (pp. 378-383). IEEE Conference Publication. https://doi.org/10.1109/ialp.2018.8629256

Chen, Z., & Zhou, Y. (2019). Research on automatic essay scoring of composition based on CNN and OR. In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 13-18). IEEE Conference Publication. https://doi.org/10.1109/icaibd.2019.8837007

Cheon, M., Seo, H. W., Kim, J. H., Noh, E. H., Sung, K. H., & Lim, E. (2015). An automated scoring tool for Korean supply-type items based on semi-supervised learning. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 59-63). Association for Computational Linguistics and Asian Federation of Natural Language Processing. https://doi.org/10.18653/v1/w15-4409

Contreras, J. O., Hilles, S., & Abubakar, Z. B. (2018). Automated essay scoring with ontology based on text mining and nltk tools. In *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)* (pp. 1-6). IEEE Conference Publication. https://doi.org/10.1109/icscee.2018.8538399

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. https://doi.org/10.11113/jt.v39.472

Davis, B. (2014). *Essay grading computer mistakes gibberish for genius*. Retrieved August 28, 2020, from http://www.realclear.com/tech/2014/04/29/essay_grading_computer_mistakes_gibberish_for_genius_6784.html

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment (JTLA), 5*(1), 1-35.

Dong, F., & Zhang, Y. (2016). Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1072-1077). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1115

Elliot, S. M. (2003). IntelliMetric: From here to validity. In M. D., Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Routledge

Fazal, A., Dillon, T., & Chang, E. (2011). Noise reduction in essay datasets for automated essay grading. In *OTM Confederated International Conferences On the Move to Meaningful Internet Systems* (pp. 484-493). Springer. https://doi.org/10.1007/978-3-642-25126-9_60

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. I*nteractive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), 939-944.

Ghosh, S., & Fatima, S. S. (2008). Design of an automated essay grading (AEG) system in Indian context. In *TENCON 2008-2008 IEEE Region 10 Conference* (pp. 1-6). IEEE Conference Publication. https://doi.org/10.1109/tencon.2008.4766677

Greene, P. (2018). *Automated essay scoring remains an empty dream*. Retrieved September 8, 2020, from Forbes: https://www.forbes.com/sites/petergreene/2018/07/02/automated-essay-scoring-remains-an-empty-dream/#4474e4f74b91

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, Article e208 https://doi.org/10.7287/peerj.preprints.27715v1

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

Ishioka, T., & Kameda, M. (2006). Automated Japanese essay scoring system based on articles written by experts. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 233-240). Association for Computational Linguistics. https://doi.org/10.3115/1220175.1220205

Islam, M. M., & Hoque, A. L. (2013). Automated Bangla essay scoring system: ABESS. In *2013 International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 1-5). IEEE Conference Publication. https://doi.org/10.1109/iciev.2013.6572694

Jin, C., & He, B. (2015). Utilizing latent semantic word representations for automated essay scoring. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)* (pp. 1101-1108). IEEE Conference Publication. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.202

Jin, C., He, B., Hui, K., & Sun, L. (2018). TDNN: a two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 1088-1097). Association for Computational Linguistics. https://doi.org/10.18653/v1/p18-1100

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005). Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 29-36). Association for Computational Linguistics.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems, 15*, 27-31.

Liang, G., On, B. W., Jeong, D., Kim, H. C., & Choi, G. S. (2018). Automated essay scoring: A siamese bidirectional LSTM neural network architecture. *Symmetry, 10*(12), Article 682. https://doi.org/10.3390/sym10120682

Loraksa, C., & Peachavanish, R. (2007). Automatic Thai-language essay scoring using neural network and latent semantic analysis. In *First Asia International Conference on Modelling & Simulation (AMS'07)* (pp. 400-402). IEEE Conference Publication. https://doi.org/10.1109/ams.2007.19

Malaysian Examinations Council. (2014). *Malaysian university English test (MUET): Regulations, test specifications, test format and sample questions*. Retrieved March 15, 2021, from https://www.mpm.edu.my/images/dokumen/calon-peperiksaan/muet/regulation/Regulations_Test_Specifications_Test_Format_and_Sample_Questions.pdf

Measurement Incorporated. (2020). *Automated essay scoring*. Retrieved June 4, 2020, from https://www.measurementinc.com/products-services/automated-essay-scoring

Nguyen, H., & Dery, L. (2016). Neural networks for automated essay grading. *CS224d Stanford Reports*, 1-11.

Omar, N., & Mezher, R. (2016). A hybrid method of syntactic feature and latent semantic analysis for automatic Arabic essay scoring. *Journal of Applied Sciences, 16*(5), 209-215. https://doi.org/10.3923/jas.2016.209.215

Ong, D. A., Razon, A. R., Guevara, R. C., & Prospero C. Naval, J. (2011, November 24-25). Empirical comparison of concept indexing and latent semantic indexing on the content analysis of Filipino essays. In *Proceedings of the 8th National Natural Language Processing Research Symposium* (pp. 40-45). De La Salle University, Manila.

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

Page, E. (2003). Project essay grade: PEG. In M. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Lawrence Erlbaum Associates Publishers.

Pai, K. C., Lu, Y., & Kuo, B. C. (2017). Developing Chinese automated essay scoring model to assess college students' essay quality. In *Proceedings of the 10th International Conference on Educational Data Mining* (pp. 430-432).

Pearson. (2010). *Intelligent essay assessor (IEA)™ fact sheet*. Pearson Education. Retrieved June 4, 2020, from https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf

Peng, X., Ke, D., Chen, Z., & Xu, B. (2010). Automated Chinese essay scoring using vector space models. In *2010 4th International Universal Communication Symposium* (pp. 149-153). IEEE Conference Publication. https://doi.org/10.1109/IUCS.2010.5666229

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural*

Language Processing (pp. 431-439). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1049

Ramalingam, V. V., Pandian, A., Chetry, P., & Nigam, H. (2018). Automated essay grading using machine learning algorithm. In *Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012030). IOP Publishing. https://doi.org/10.1088/1742-6596/1000/1/012030

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series, 2018*(1), 1-31. https://doi.org/10.1002/ets2.12192

Ratna, A. A. P., Budiardjo, B., & Hartanto, D. (2007). SIMPLE: System automatic essay assessment for Indonesian language subject examination. *Makara Journal of Technology, 11*(1), 5-11.

Ratna, A. A. P., Purnamasari, P. D., & Adhi, B. A. (2015). SIMPLE-O, the Essay grading system for Indonesian Language using LSA method with multi-level keywords. In *The Asian Conference on Society, Education & Technology 2015* (pp. 155-164). The International Academic Forum.

Ratna, A. A. P., Arbani, A. A., Ibrahim, I., Ekadiyanto, F. A., Bangun, K. J., & Purnamasari, P. D. (2018). Automatic essay grading system based on latent semantic analysis with learning vector quantization and word similarity enhancement. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality* (pp. 120-126). Association for Computing Machinery. https://doi.org/10.1145/3293663.3293684

Ratna, A. A. P., Kaltsum, A., Santiar, L., Khairunissa, H., Ibrahim, I., & Purnamasari, P. D. (2019a). Term frequency-inverse document frequency answer categorization with support vector machine on automatic short essay grading system with latent semantic analysis for japanese language. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)* (pp. 293-298). IEEE Conference Publication. https://doi.org/10.1109/ICECOS47637.2019.8984530

Ratna, A. A. P., Khairunissa, H., Kaltsum, A., Ibrahim, I., & Purnamasari, P. D. (2019b). Automatic essay grading for Bahasa Indonesia with support vector machine and latent semantic analysis. In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)* (pp. 363-367). IEEE Conference Publication. https://doi.org/10.1109/ICECOS47637.2019.8984528

Ratna, A. A. P., Santiar, L., Ibrahim, I., Purnamasari, P. D., Luhurkinanti, D. L., & Larasati, A. (2019c). Latent semantic analysis and winnowing algorithm based automatic Japanese short essay answer grading system comparative performance. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)* (pp. 1-7). IEEE Conference Publication. https://doi.org/10.1109/ICAwST.2019.8923226

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation, 7*, Article 26.

Sendra, M., Sutrisno, R., Harianata, J., Suhartono, D., & Asmani, A. B. (2016). Enhanced latent semantic analysis by considering mistyped words in automated essay scoring. In *2016 International Conference on Informatics and Computing (ICIC)* (pp. 304-308). IEEE Conference Publication. https://doi.org/10.1109/IAC.2016.7905734

Shehab, A., Faroun, M., & Rashad, M. (2018). An automatic Arabic essay grading system based on text similarity Algorithms. *International Journal of Advanced Computer Science and Applications, 9*(3), 263-268. https://doi.org/10.14569/IJACSA.2018.090337

Shermis, M. D., & Burstein, J. (2003). Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates Publishers.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education, 4*(1), 20-26. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268.

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882-1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/d16-1193

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data, 7*(1), 1-47. https://doi.org/10.1186/s40537-020-00349-y

Vantage Learning. (2005). *How IntelliMetric™ works*. Retrieved June 4, 2020, from http://www.vantagelearning.com/docs/intellimetric/IM_How_IntelliMetric_Works.pdf

Vantage Learning. (2020). *IntelliMetric®: Frequently asked questions*. Retrieved June 4, 2020, from http://www.vantagelearning.com/products/intellimetric/faqs/#LongUsed

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

Xu, Y., Ke, D., & Su, K. (2017). Contextualized latent semantic indexing: A new approach to automated Chinese essay scoring. *Journal of Intelligent Systems, 26*(2), 263-285. https://doi.org/10.1515/jisys-2015-0048

Zupanc, K., & Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica, 4*(39), 383-396.