Contemporary Educational Technology, 2023, **15**(4), ep475
ISSN: 1309-517X (Online)

OPEN ACCESS

**Research Article**

# Exploring effective methods for automated essay scoring of non-native speakers

**Kornwipa Poonpon** [1]
0000-0002-8349-2834

**Paiboon Manorom** [1]
0009-0002-2165-1120

**Wirapong Chansanam** [1*]
0000-0001-5546-8485

[1] Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen, THAILAND
[*] Corresponding author: wirach@kku.ac.th

**Citation:** Poonpon, K., Manorom, P., & Chansanam, W. (2023). Exploring effective methods for automated essay scoring of non-native speakers. *Contemporary Educational Technology, 15*(4), ep475. https://doi.org/10.30935/cedtech/13740

**ARTICLE INFO**

Received: 19 May 2023

Accepted: 15 Sep 2023

**ABSTRACT**

Automated essay scoring (AES) has become a valuable tool in educational settings, providing efficient and objective evaluations of student essays. However, the majority of AES systems have primarily focused on native English speakers, leaving a critical gap in the evaluation of non-native speakers' writing skills. This research addresses this gap by exploring the effectiveness of automated essay-scoring methods specifically designed for non-native speakers. The study acknowledges the unique challenges posed by variations in language proficiency, cultural differences, and linguistic complexities when assessing non-native speakers' writing abilities. This work focuses on the automated student assessment prize and Khon Kaen University academic English language test dataset and presents an approach that leverages variants of the long short-term memory network model to learn features and compare results with the Kappa coefficient. The findings demonstrate that the proposed framework and approach, which involve joint learning of different essay representations, yield significant benefits and achieve results comparable to state-of-the-art deep learning models. These results suggest that the novel text representation proposed in this paper holds promise as a new and effective choice for assessing the writing tasks of non-native speakers. The result of this study can apply to advance educational assessment practices and promote equitable opportunities for language learners worldwide by enhancing the evaluation process for non-native speakers

**Keywords:** automated essay scoring, non-native speakers, machine learning, long short-term memory network, Thailand

## INTRODUCTION

In today's interconnected world, the ability to communicate effectively in English has become increasingly important. English proficiency assessments play a crucial role in determining language competency, particularly for non-native speakers. Traditionally, the evaluation of essays written by these individuals has been a labor-intensive and time-consuming process, requiring human graders with expertise in language assessment. However, advancements in technology have paved the way for automated essay scoring (AES) systems, offering a promising alternative that combines the power of artificial intelligence (AI) and natural language processing (NLP).

The rapid advancement of technology and the development of NLP techniques have given rise to various automated applications across various fields. Among these applications, AES systems have emerged as an essential tool in recent decades. AES is a computer-based assessment system employing appropriate features to score or grade student responses automatically. The origins of AES research can be traced back to 1966

with the introduction of the project essay grader (PEG) by Ajay et al. (1973). To assign grades to essays, PEG evaluates various writing characteristics, including grammar, diction, and construction. In 1999, Foltz et al. introduced the intelligent essay assessor, which employed latent semantic analysis to evaluate the content and generate an overall score. Shermis et al. (2001) released a modified version of PEG that primarily focused on grammar checking and showed a correlation between human evaluators and the system. Subsequently, Powers et al. (2002) proposed the E-rater, Rudner et al. (2006) developed Intellimetric, and Rudner and Liang (2002) created the Bayesian essay test scoring system.

AES systems usually utilized NLP techniques to assess style and content to assign scores to essays. During the 1990s, as mentioned above, most essay scoring systems followed traditional approaches such as pattern matching and statistical-based methods. However, in the past decade, there has been a shift towards regression-based and NLP techniques in essay grading systems. AES systems developed in 2014 and later, e.g., Dong et al. (2017) and Rupp et al. (2019), incorporated deep learning techniques to induce syntactic and semantic features, resulting in improved performance compared to earlier systems. One of the critical components of AES systems is content scoring, which is pivotal in calculating marks for a given essay answer.

Two primary approaches to content scoring include similarity-based and instance-based methods. Student answers are compared with model answers in similarity-based methods, and scores are assigned based on similarity. Conversely, the instance-based approach involves learning a model that identifies the features associated with different score levels without needing reference answers (Steimel & Riordan, 2020). Unlike similarity-based approaches, instance-based scoring does not rely on external knowledge (Driessens & Džeroski, 2005; Horbach & Zesch, 2019; Shaker & Hüllermeie, 2012). Studies done in the past that used instance-based methodologies include those by Beseiso and Alzahrani (2020), Chen and Zhou (2019), Cozma et al. (2018), Li et al. (2018), Steimel and Riordan (2020), and Tashu and Horváth (2020).

Several studies have highlighted the benefits of AES systems in language assessment. The systems have revolutionized the field of language assessment by providing efficient and objective evaluations (Evanini et al., 2015; Haberman, 2011; Rupp et al., 2019). They can alleviate the manual correction burden on scorers or instructors and reduce subjectivity in the grading process (e.g., Attali & Burstein, 2006; Lahitani et al., 2016; Munir et al., 2016; Phoophuangpairoj & Pipattarasakul, 2022; Zupanc & Bosnic, 2020). They can also reduce grading time, increase scalability, provide feedback, and minimize bias associated with human grading (Foltz et al., 2020). While AES has shown remarkable success in assessing essays written by native speakers, the evaluation of essays composed by non-native speakers presents unique challenges.

The language proficiency of non-native speakers encompasses a wide spectrum, varying from those with limited knowledge of English to individuals who are highly proficient but may still exhibit subtle errors or non-conventional language use (Weigle, 2002; Wolfe, 2016). These variations make it crucial to develop and refine automated scoring systems specifically tailored to the needs of non-native speakers. Such systems should not only consider linguistic accuracy but also consider factors such as language complexity, syntactic patterns, and cultural nuances (e.g., Ramesh & Sanampudi, 2022; Zechner et al., 2017).

The reviewed previous studies have suggested challenging aspects and gaps for future research in the field of using AES systems for assessing non-native speaker essays. These challenges include

(1) developing AES system to handle better challenges for non-native speaker essay evaluation,

(2) determining biases in AES system, especially for those that could affect non-native speakers,

(3) creating AES system, which provides a more holistic evaluation, including creativity, critical thinking, and coherent argumentation,

(4) integrating different approaches to content scoring in AES system to improve accuracy and robustness,

(5) developing AES system, which provides detailed feedback to non-native speakers to help them to improve their language proficiency,

(6) creating AES system for other languages apart from English, and

(7) investigating the ethical implications of using AES, especially for non-native speakers.

The areas for further research identified in the essay writing assessment would help to advance and refine AES system to be more effective and beneficial for non-native speakers, especially in terms of dealing with the diversity of non-native speaker essays and ensuring fairness and objectivity.

This research, therefore, addresses the gaps in AES for non-native speakers. In particular, the study aims to develop a robust and effective AES system that accommodates the unique linguistic characteristics exhibited by non-native speakers, by leveraging state-of-the-art techniques in NLP and machine learning. It also proposes AES system development framework for assessing non-native speakers' writing abilities. These innovative approaches will not only enhance the accuracy and reliability of writing assessment, but also contribute to ongoing efforts in advancing language assessment practices and facilitating fair and comprehensive evaluation of non-native speakers' writing proficiency.

## METHODOLOGY

### A Corpus of English Essays

This study compiled a corpus of written essays produced by non-native speakers of English. The corpus included two sets of written data: the automated student assessment prize (ASAP) dataset (Kaggle, 2012) and the Khon Kaen University Academic English Language Test (KKU-AELT) essays. The Hewlett Foundation made this dataset available through the Kaggle competition in 2012. The data set included a total of three sub-corpora on an ASAP ("https://www.kaggle.com/c/asap-sas/") essays and short answers. It has nearly 12,976 essays, with up to 3,000 essays provided for each prompt. These prompts are designed to test $7^{th}$ to $10^{th}$-grade students, and scores are given in the range of [0-3] and [0-60]. The essays range from an average length of 150 to 550 words per response. Each essay is scored by two raters, and if there is a significant disagreement, a third rater evaluates the essay, and their grade becomes the final score (Beseiso & Alzahrani, 2020; Yang et al., 2020). The final score is either calculated from the first two raters' scores or is the score given by the third rater. However, it is important to note that there are limitations to these corpora. Firstly, there is a different score range for other prompts. Secondly, the evaluation of essays in these corpora relies on statistical features such as named entity extraction and lexical features of words.

The second set of data was KKU-AELT written corpus with an official permission from the Center for English Language Excellence, Khon Kaen University. This corpus encompasses a substantial collection of 4,817 individual responses from ten thought-provoking prompts, which provide a diverse and comprehensive dataset for analysis. The essays ranges from an average length of 100 to 500 words per response. Notably, each response was scored by two trained human raters. The interrater reliability of KKU-AELT responses is ensured at least 0.80 (Srisawat & Poonpon, 2023). KKU-AELT corpus committed to ensure a rich and multifaceted representation of non-native speakers' linguistic proficiency. The variety of prompts stimulates various responses would enable the researchers to delve into the intricacies of language use, cognitive expression, and linguistic evolution within a specific context. This corpus's significant size would reinforce the statistical significance results and allow researchers to draw more accurate and dependable conclusions. The dual evaluation process by human graders enhances the credibility of the results obtained from this corpus and introduces an element of cross-validation, reducing potential biases and inconsistencies that might arise from a single-grader evaluation. This approach not only attests to the meticulousness of the research methodology but also substantiates the trustworthiness of the generated data.

In the study, ASAP and KKU-AELT data sets were combined into one. ASAP dataset was primarily utilized as an established benchmark dataset, as widely used for comparing results across various AES systems. Finally, the dataset contains 17,793 essays (12,976 ASAP and 4,817 KKU-AELT essays). The dataset encompasses various essay lengths, from 100 to 550 words. Integrating KKU-AELT corpus into ASAP dataset would offer rich source data that reflects the complexities of non-native speakers. This would yield benefits for the ongoing development and integration of AES systems, which can introduce a degree of objectivity and consistency to the evaluation process.

### Automated Essay Scoring System Development Process

AES system is a method of assigning scores to essays using machine learning algorithms. The process of KKU-AELT AES system development involved several stages, as follows.

### *Stage 1: Data preparation*

All essays in the corpus of English essays were formatted and cleaned in an electronic form and converted into a readable format for KKU-AELT AES system. At this stage, spelling errors in the essays were eliminated and pre-processing tasks such as document tokenization, stop word removal, and stemming were performed.

### *Stage 2: Compute similarity techniques in KKU-AELT automated essay scoring system*

This study employed instance-based approach in similarity techniques to generate scores. In the instance-based techniques, instances are usually learned through a supervised classification algorithm and stored in memory. When new instances need to be evaluated, a set of similar instances is retrieved from memory to predict the target value for the new instances. In this study, the lexical properties of correct answers were learned by models and stored in memory. To predict the score of new answers, the model classified them based on the learned answers. The instance-based techniques require a substantial dataset for training and testing to achieve high accuracy and, in supervised machine learning algorithms, the data must be labeled. At this point, it is challenging that the labeled data is task-dependent and cannot be universally applied to all AES systems and that considerable instructor involvement has to be used to create labeled data specific to a particular AES system. This can be time-consuming (Ghosh & Fatima, 2010; Kulkarni et al., 2014). Therefore, in this study, transfer learning was used to reduce the amount of data required to train a model and decrease training time. Transfer learning involves applying knowledge from one domain to another to enhance learning performance in the target domain (Zhuang et al., 2020). By leveraging pre-existing, pre-trained models, transfer learning allowed the construction of models with relatively limited training data, which is particularly advantageous for AES systems that necessitate labeled datasets (Roy et al., 2016).

### *Stage 3: Machine learning algorithms*

At this stage, a technique of machine learning algorithms called neural network (NN) plays a crucial role in AES system. NN is an AI technique that enables computers to analyze data in a manner reminiscent of the human brain. It falls under the category of deep learning, a machine learning approach, employing interconnected nodes or neurons arranged in layers, similar to the structure of the human brain. NN can be utilized to comprehend essay representations, enabling the application of a scoring function to calculate essay scores based on the learned features (Yang et al., 2020). Among several algorithms, e.g., recurrent neural network (RNN), convolutional neural network, long short-term memory (LSTM) was employed in developing AES system of the present study.

LSTM is well-known for its proficiency in effectively acquiring knowledge from sequential data (Hochreiter & Schmidhuber, 1997). LSTMs is an RNN, which excels in processing sequential data like essay content and is used in AES for non-native speakers with various benefits, e.g., understanding context and relationships in sequences, accommodating varying essay lengths, and maintaining context awareness. LSTMs automatically learn linguistic nuances to assist in identifying patterns and complexities in non-native essays by constructing syntax models and semantics models to distinguish genuine errors from unconventional but valid ones in terms of language use. To leverage the pre-trained models, LSTMs grasp language structure and offer constructive feedback to enhance assessment quality. In conclusion, LSTMs present a promising approach for AES of non-native speakers to manage the diverse essay intricacies and effectively capture context and patterns (Huang et al., 2018; Mesgar & Strube, 2018).

Various NLP libraries were also used to thoroughly examine feature extraction, as the effective AES systems heavily rely on a well-defined set of features and appropriate models. One such library is the natural language toolkit (NLTK), which is commonly utilized to extract statistical features such as part-of-speech, word count, sentence count, and more. However, relying solely on NLTK may lead to overlooking important semantic features in the essays. To capture these semantic features, popular libraries such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) were also used to extract the semantic content from the essays. Additionally, some systems directly trained the model using word embeddings to predict the essay scores, bypassing the need for explicit feature extraction. GloVe has demonstrated superior performance to Word2Vec in word analogy and similarity tasks (Zhao et al., 2017). When scoring essays, various factors were considered, including length, coherence, topical information, semantic relationships, dependencies, syntactic

constructions, grammatical relations, and the complexity of sentences within the essay. Deep neural networks (DNNs) and RNNs can automatically learn features, encode the necessary information for essay evaluation, and capture complex patterns in the data through non-linear neural layers (Alikaniotis et al., 2016; Taghipour & Ng, 2016). It is worth noting that the performance of an NN model is often directly related to the number of training samples used, particularly when seeking generalized results.

The following description illustrated how LSTM algorithm was used in the present study to develop a model capable of automatically scoring essays. The primary objective of the study was to devise a scoring methodology that exhibited accuracy and precision and minimized potential disparities. To accomplish this goal, employing many-to-one LSTM algorithm was essential. Many-to-one LSTM algorithm is a type of LSTM architecture that allows the model to take multiple inputs and provide a single output. This architecture is ideal for the essay scoring task since it enables the model to process each sentence in the essay as a separate input and generate a single output score for the entire essay. Using this algorithm can ensure that the scoring methodology is consistent and unbiased across different essays and test-takers. Additionally, a 2-layer LSTM model was defined. A 2-layer LSTM, a promising approach for AES of non-native speakers, was used to capture intricate language patterns, such as grammatical errors and unconventional sentence structures, manage the hierarchical features, such as the overall organization and coherence of ideas in an essay, maintain context awareness, which is essential for understanding non-native speaker essays that may use different contextual cues and syntactic structures, learn more intricate features from the data, which is beneficial for extracting subtle linguistic patterns and nuances, generalize better to unseen essays, which is vital for assessing non-native speaker essays that vary significantly in style, content, and proficiency levels, mitigate overfitting, which is vital for dealing with potentially limited training data for non-native speaker essays. These consisted of the inherent complexity of AES for non-native speakers. In short, a 2-layer LSTM can better understand non-native speaker essays and make more accurate assessments of their quality.

Then, a large dataset of essays and their corresponding scores were utilized to train our model. The data were pre-processed by tokenizing the text and converting it into a numerical format suitable for LSTM model. Then the data were split into training and testing sets and trained the model on the training data. Once the model was trained, its performance on the testing data was evaluated and fine-tuned as necessary. Overall, the approach in the present study combined the power of deep learning with the effectiveness of LSTM and many-to-one architecture to develop a reliable and accurate essay scoring system.

### *Stage 4: Evaluation of KKU-AELT automated essay scoring system*

The evaluation stage aims to assess the performance of the developed system or model. It is important to be noted that there are alternative statistical techniques, which can be used to assess the performance. For instance, quadrated weighted kappa (QWK) measures the agreement between human and system evaluation scores, yielding values between zero and one. Mean absolute error represents the difference between human-rated and system-generated scores, while mean square error calculates the average squared difference between the human-rated and system-generated scores, resulting in only positive values. The Pearson's correlation coefficient determines the correlation between two variables, offering three possible values (0, 1, -1). A value of "0" indicates no relationship between the human-rated and system scores, "1" signifies a positive correlation, and "-1" represents a negative correlation. A kappa score of one signifies perfect agreement. One such technique is the Fleiss kappa score, which is appropriate when evaluating the agreement between more than two raters. In addition, other statistical techniques, such as Pearson's and intraclass correlation coefficients, can be used to assess inter-rater agreement in different contexts. Cohen's kappa score (K) is a widely used statistical technique for determining inter-rater agreement or the agreement between two or more evaluators who rate the same set of objects or subjects. It measures the extent to which the evaluators' ratings agree beyond what would be expected by chance alone. In this study, K was employed as a reliable statistical approach, beneficial for evaluating subjective tasks such as essay grading to evaluate our scoring model performance.

The following sections provide a comprehensive overview of K and its properties. Illustrative examples are given to demonstrate the interpretation of the kappa score and its limitations. By understanding K, the researchers can better assess performance of the essay scoring model and identify areas for improvement.

1. To evaluate the level of agreement between two raters who classify n items into C categories. For this purpose, we use K, which is a statistical measure of inter-rater reliability. K is defined, as follows:

$$K \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where $p_o$ represents the relative observed agreement between the assessors, which can be equated to precision. On the other hand, $p_e$ refers to the hypothetical probability of chance agreement. It is determined by using observed data to calculate the probability of random observers assigning items to each category.

After the written-response scores are graded, the accuracy of the scoring model can be measured using different performance measures. Shermis and Hamner (2013) evaluated the performance of nine AES systems in ASAP competition using two distributional difference measures and four agreement measures. Two agreement measures used are described here: exact-agreement percentage and kappa. Exact-agreement percentage measures the exact matching between human and computer scores as a percentage. K is a measure of agreement that takes into consideration chance agreement. K value serves as a measure of agreement, where 1.0 represents a perfect agreement, and 0.0 indicates agreement equivalent to chance. Viera and Garrett (2005) have proposed guidelines for interpreting K values. A K value less than zero signifies agreement lower than chance, while K values ranging from 0.01 to 0.20 indicate slight agreement. K values between 0.21 and 0.40 suggest fair agreement, while K values between 0.41 and 0.60 represent moderate agreement. K values indicate substantial agreement between 0.61 and 0.80, and K values between 0.81 and 0.99 indicate almost perfect agreement. Therefore, a K value of 0.81 or higher is the reference standard for perfect agreement. The validation dataset can be used to evaluate the performance of AES scoring model, where the computer-scored essay scores are compared with human scores using measures of agreement. A kappa score of zero indicates that there is no agreement between raters beyond what would be expected by chance, while a kappa score of one indicates perfect agreement. F1-score is a metric commonly used in machine learning to assess the performance of a classification model. The value of 0.017 indicates the harmonic mean of precision and recall for the given model and dataset.

2. When both raters demonstrate complete agreement, K attains a value of one. Conversely, if there is no agreement beyond what would be expected by chance (as determined by the hypothetical probability $p_e$), the value of $k$ diminishes to zero. This signifies a negative statistical probability, suggesting either a lack of substantial agreement between the two assessors or an agreement level that is even inferior to what would be anticipated by random chance.

As part of our NLP essay analysis, further processing of the sentences was undertaken based on their linguistic competence. This involves the utilization of tokenization techniques, where we employ NLTK library in conjunction with regular expressions. Subsequently, the compositions were transformed into vectors by employing Word2Vec algorithm and leveraging the pre-trained GloVe model. To construct our NN architecture, LSTM algorithm was employed, and our feature vectors were fed into this framework. In order to evaluate and compare their performance, the flexibility to create multiple NN architectures exists. The ultimate goal is to identify NN architecture that yields the most favorable outcomes. In this study, the primary objective is to train the model and save it in h5 file format, which is specific to Keras technology. It is common practice to store models as pickle files (.pkl), while Tensorflow models can be saved as protobuf files (.pb).

### *Stage 5: Developing KKU-AELT automated essay scoring system & its application*

To streamline the development of a model that aligns with an automated essay-scoring system, code examples were furnished at the outset of each Python script on the web page of our system hosted on Google Colab (https://colab.research.google.com/drive/1YYCdy5sO8U_-I4vyZQzcAuNAwFhTdt6X?usp=sharing). This stage gains credibility through insights from hyperparameter tuning and ablation studies, which validate the chosen model configurations and provide a deeper understanding of their impact on AES process, as follows. Hyperparameter tuning process involves systematically adjusting various parameters that govern the model's behavior. To explore different hyperparameter settings, the study gains insights into how the model's performance is affected in the context of AES for non-native speakers. Tuning hyperparameters such as learning rate, batch size, LSTM layer size, and dropout rates could offer significant benefits. First, hyperparameter tuning process can identify the learning rate leading to faster training convergence. Finding an optimal learning rate for non-native speaker essays, where linguistic complexity varies, would help to

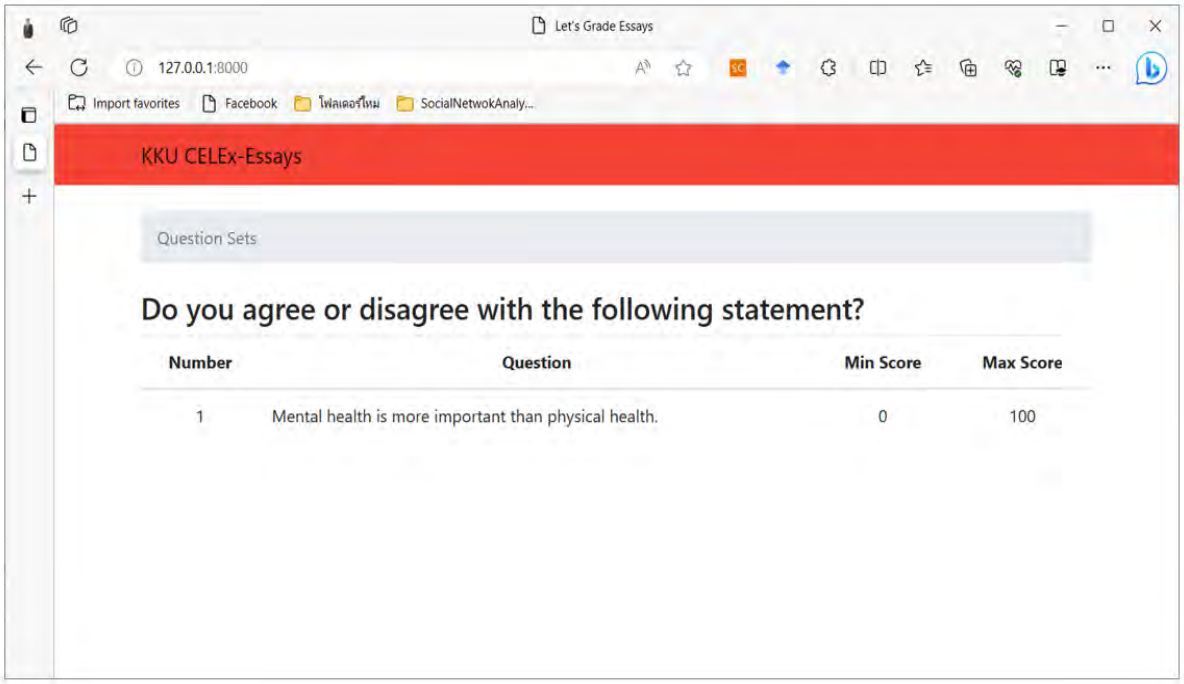


**Figure 1.** Home screen of KKU-AELT essay system (Source: Authors, using Microsoft Edge)

balance model updates and better adapt to different essay styles. Second, tuning the size of LSTM layers impacts the model's ability to capture intricate language patterns. The ablation study comparing different layer sizes can reveal the optimal configuration that effectively handles minor errors and unconventional language use. Third, dropout is a regularization technique that prevents overfitting when the Hyperparameter tuning can determine the dropout rate that minimizes overfitting while maximizing the model's generalization to diverse non-native essays. Next, ablation studies systematically remove certain components or layers from the model to understand their contributions. In the context of AES for non-native speakers, ablation studies could shed light on effects of hierarchical layers by removing one of LSTM layers in 2-layer LSTM architecture. With this, researchers can gauge how much hierarchical modelling contributes to the model's performance.

This would help to validate the choice of a deeper architecture for capturing nuanced language features. Regarding the contextual and non-contextual learning, ablating LSTM layer responsible for context learning can assess its impact on maintaining context awareness. This informs the model's ability to understand the progression of ideas in non-native essays. Moreover, temporarily removing dropout layers can quantify their contribution in mitigating overfitting, demonstrating their importance in adapting the model to the variations in non-native essays. Therefore, incorporating insights from hyperparameter tuning and ablation studies enhances the study's credibility by demonstrating a systematic approach to configuring the model. It substantiates the authors' rationale for selecting specific hyperparameters and architectural choices. Additionally, these insights highlight the adaptability of the chosen model to the diverse linguistic characteristics and challenges posed by non-native speaker essays, making the study's findings more robust and applicable. Upon loading the essay test, the grading procedure is initiated by activating the "Submit" button. Subsequently, the system retrieves the pre-existing models and employs them to conduct analysis and processing, thereby effectively demonstrating the scoring process.

To develop an AES system for the present study, two essential files were utilized for processing: Firstly, the "Python modelling by LSTM Model.ipynb" file was employed to train and save the resulting model. Secondly, the "/mysite/grader/views.py" file was used to receive the content, display the results on the web page, and generate essay scores using the previously created and saved model (from file 1). The "/mysite/grader/views.py" file played a critical role in the system as it functions as a content receiver from web pages. Its primary function was to retrieve essay writing information and utilize the recorded model to generate scores for the essays.

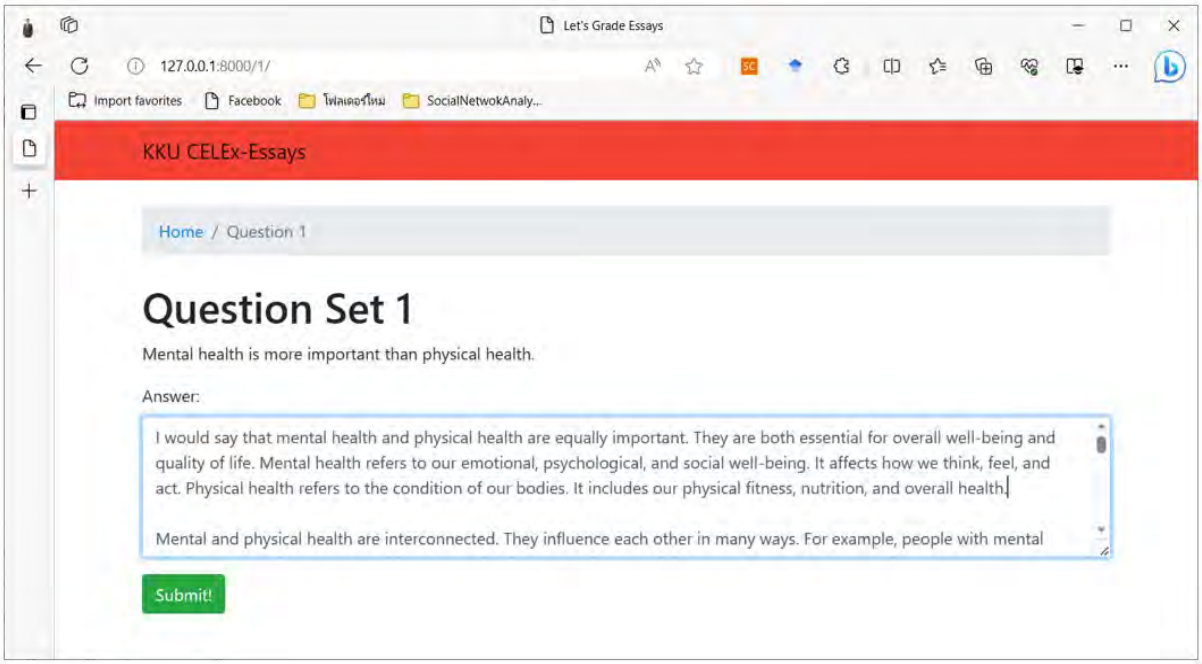


**Figure 2.** Input answer screen of KKU-AELT essay system (Source: Authors, using Microsoft Edge)

Django libraries and Utils were employed to easily call the index function whenever a user interacted with our web application, which in turn loaded the index.html file and its associated files. In addition, the essay function was utilized to display the essay.html file, which offered a complete list of available essay options for the user or test taker to choose from, as shown in **Figure 1**.

In the process of analyzing essays for automated scoring, KKU-AELT AES system utilized the question function to handle the content submitted by the user through a POST request. This function was responsible for storing the content in their respective fields and processing it further. A series of codes were employed to convert the content into testdataVectors, which was done by using the documented Word2Vec model and the previously defined getAvgFeatureVec function. This function calculated the average feature vectors that are necessary for the essay-scoring process.

Once the feature vectors were generated, the next step was to load Final_lstm.h5 file, which contains the model weights. This allowed us to apply the relevant weights and biases to our NN model to generate a prediction. The prediction process was executed, based on the predefined criteria or scoring conditions, which were carefully defined to ensure an accurate and reliable result.

AES system developed in the present study ensured that all essays submitted by users were processed with the utmost care and attention to detail. The graphic user interface is shown in **Figure 2**. The rigorous approach to essay scoring used in the study incorporated multiple stages of processing, including linguistic competency analysis and deep learning algorithms. By leveraging the power of NLP and machine learning, we were able to provide accurate and reliable scores for essays of all types, allowing students and educators to gauge their performance and improve their skills.

The system for AES has been implemented successfully, and it can return the predicted score to the Question.html file. The score is displayed prominently on HTML page, guaranteeing its visibility to the user (**Figure 3**). The implementation of this feature is crucial to ensure that the user can access the essay score easily and without delay. This user-friendly approach is significant in enhancing the overall experience of the user and contributes to the credibility and reliability of AES system. The design of Question.html file and the user interface was created with a user-centered approach, emphasizing accessibility and ease of use. The importance of a user-friendly interface cannot be overstated as it helps to create a positive user experience.

To summarize AES system methodology, the selection of LSTM networks and Django framework was justified by their distinct advantages in tackling the research complexities. LSTM network was adapted to handling sequential data like essays, capturing context, dependencies, and nuanced language patterns to

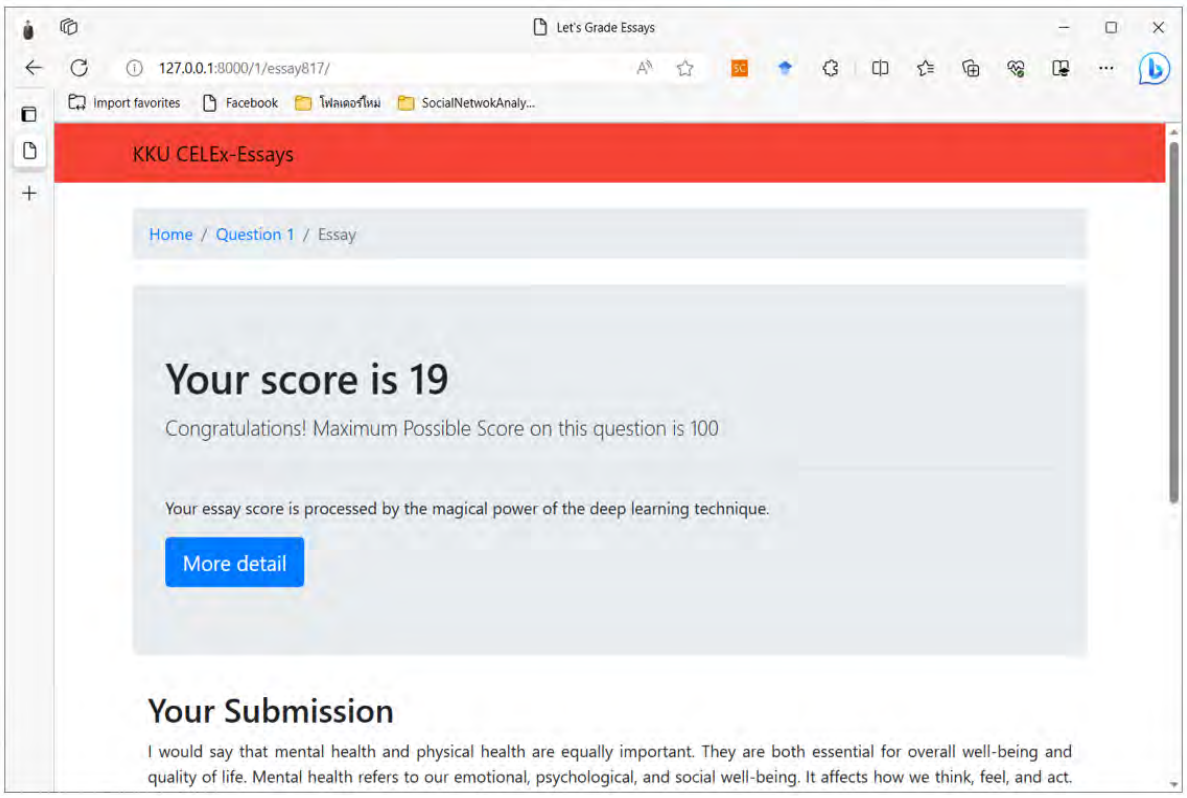


**Figure 3.** Score report screen of KKU-AELT essay system (Source: Authors, using Microsoft Edge)

facilitate accurate scoring by automatically extracting relevant features. Django framework was opted for suitability in web application development, ensuring user-friendly access to AES system since its modular design organizes components, enhancing input, scoring, and result visualization clarity together with the framework's database integration, simplifies data management, while its scalability tools ensure system reliability over time. Also, Django supports user interaction, which is vital for feedback and system refinement. In conclusion, the study benefits from LSTM networks' capabilities in sequence modelling and feature extraction, addressing nuances in non-native essays. Django's adoption amplifies user accessibility through a user-friendly web application, effectively integrating AES system for educators, students, and researchers. This synergy offers a robust solution for the intricate challenges of evaluating non-native speaker essays.

## RESULTS

The results show that KKU-AELT AES system can be developed using machine learning algorithms and its evaluation is reported below. Moreover, the study reveals a framework, drawn from AES development process and procedures, for developing an AES system for non-native speakers of English.

### Developed KKU-AELT Automated Essay Scoring System & Its Evaluation

The study found that KKU-AELT AES system can be developed using machine learning algorithms in a method of assigning scores to essays by mapping input text features to an output score. Supervised training methods use pre-scored samples from human raters to create a mapping function, while unsupervised methods create the function without pre-scored data. KKU-AELT AES system process involved data preparation, feature extraction, mapping features to output scores, and score classification using algorithms like sequential minimal optimization for support vector machines (SVMs). k-fold cross-validation was used to evaluate and improve the scoring model's accuracy. The final step was score classification and error analysis, where the model was used to grade essays. Prompt-specific models were used for KKU-AELT AES system. Developing an effective KKU-AELT AES system requires combining expertise in machine learning, NLP, and essay evaluation. After developing KKU-AELT AES system, the system was evaluated to see the effectiveness of the system. This study found that the average K after a 5-fold cross-validation was 0.364.

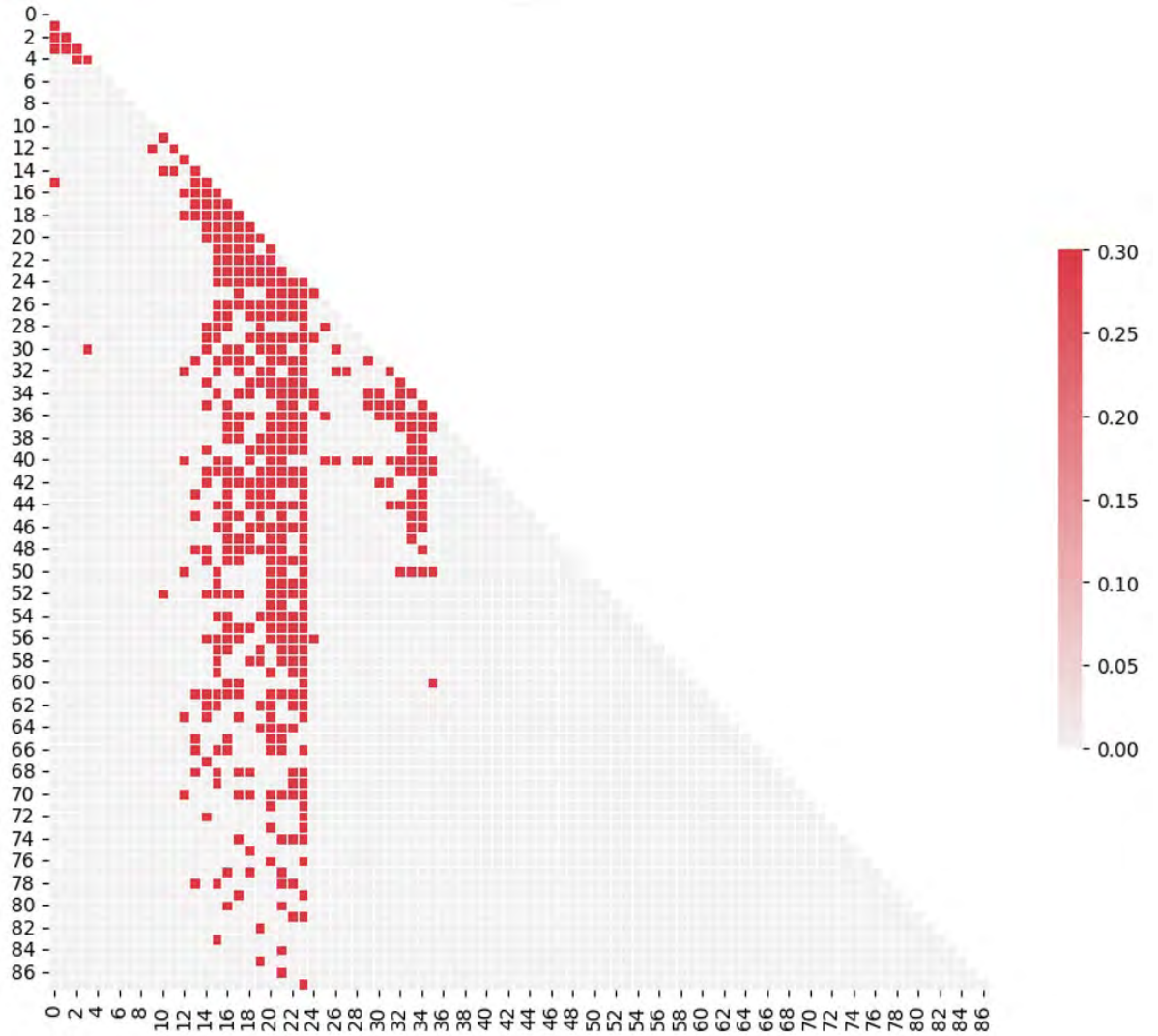


**Figure 4.** Heat-map of confusion matrix (Source: Authors, using Python in Colab)

In the initial utilization of ASAP dataset exclusively, a remarkably high K of 0.960 was observed, signifying excellent agreement. Nonetheless, for the purpose of this study, which aimed to develop an AES system tailored to non-native speakers, KKU-AELT essay dataset were incorporated. As a result, the final K obtained from this amalgamation was 0.364. This indicates that there is a moderate level of agreement between raters, beyond what would be expected by chance. It is important to interpret kappa scores with caution as kappa scores can be affected by a number of factors, such as the number of raters, the complexity of the task, and the variability of the data. However, the fact that this is the highest kappa score found in this dataset is a positive sign. It suggests that the raters are becoming more consistent in their ratings, and that the agreement between them can also be improved.

Another measure for AES system evaluation is a confusion matrix. A confusion matrix is a table that is used to evaluate the performance of a classification model. It is a way of summarizing the accuracy of a model by showing the number of instances that were correctly classified and the number of instances that were incorrectly classified. The confusion matrix for AES system prediction is illustrated in **Figure 4**.

The confusion matrix is usually used to identify areas, where the model is performing well and areas where the model is performing poorly. According to **Figure 4**, if the model has a high accuracy but a low precision, the model is likely to be predicting a lot of false positives. This can be a problem if the false positives are costly or disruptive. The confusion matrix can also be used to identify the types of errors that the model is making. For example, if the model is predicting a lot of false negatives, then the model is likely to be missing a lot of actual positive instances. This can be a problem if the missed instances are important or valuable. Overall,

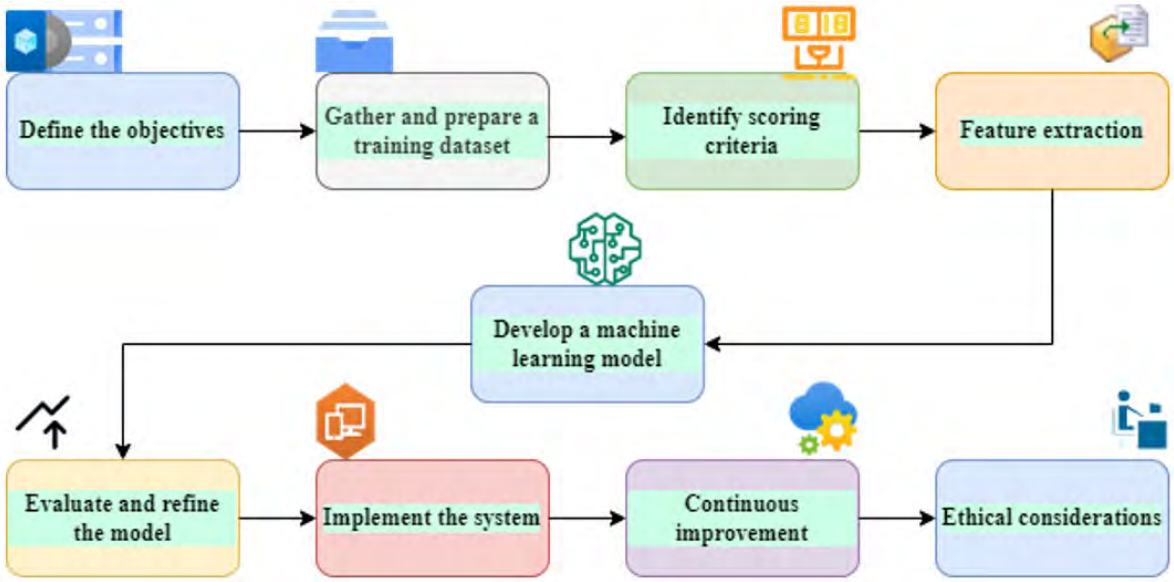


**Figure 5.** KKU-AELT AES system development framework (Source: Authors, using https://app.diagrams.net)

the confusion matrix is a useful tool for evaluating the performance of a classification model. It can be used to identify areas, where the model is performing well and areas, where the model is performing poorly. This information can be used to improve the model or to make better decisions about how to use the model.

F1-score is a measure of accuracy that considers both precision and recall. Precision is the percentage of predicted positive results that are actually positive. Recall is the percentage of actual positive results that are predicted positively. Based on the results, the obtained F1-score of 0.021508824524596713 reflected a very low level of accuracy for the model, which is important to note that the F1-score is a measure that balances both precision and recall, considering the trade-off between false positives and false negatives. In this context, the model's F1-score of 0.0215 implies that the model struggles to find a balance between correctly identifying positive instances and minimizing incorrect classifications. Meanwhile, F1-score does not directly translate to a percentage of correct predictions, it indicates that the model's overall performance is quite limited in accurately identifying positive outcomes. Especially, the model has correctly predicted only around 2.15% of the positive results, highlighting the need for substantial enhancements in predictive capabilities. There are a few possible reasons why the model is not very accurate. One possibility is that the data is not very well-labeled. Another possibility is that the model is not very complex. It may be possible to improve the accuracy of the model by using a more complex model or by using better data.

## Proposed KKU-AELT Automated Essay Scoring System Development Framework

The study proposes a framework for developing KKU-AELT AES system, which is a complex task, as shown in **Figure 5**.

The following are the general steps described in developing KKU-AELT AES system the framework.

1. **Define the objectives:** Clearly outline the goals and objectives of AES system. Consider factors such as accuracy, reliability, efficiency, and scalability.
2. **Gather and prepare a training dataset:** Collect a large dataset of essays that are manually scored by human experts. Ensure that the dataset covers a wide range of topics and writing styles. Preprocess the data by cleaning it, removing irrelevant information, and standardizing the formatting.
3. **Identify scoring criteria:** Determine the scoring criteria that the system will use to evaluate essays. This can include aspects like grammar, vocabulary, coherence, organization, and argumentation. Create a rubric or scoring guide that specifies the different levels or dimensions for each criterion.
4. **Feature extraction:** Extract relevant features from essays to represent their content and quality. This can involve linguistic features like sentence length, word frequency, grammar patterns, and semantic similarity. Also, consider using advanced techniques like NLP to analyze the essays in more depth.

5. **Develop a machine learning model:** Select an appropriate machine learning algorithm for your AES system. Common approaches include regression models, SVMs, random forests, LSTM, or NNs. Train the model using the prepared dataset, using the manually scored essays as the ground truth.
6. **Evaluate and refine the model:** Assess the performance of AES system by using a separate validation dataset. Measure metrics such as accuracy, precision, recall, and F1-score. Identify areas, where the model may be lacking and refine it accordingly. Consider using techniques like cross-validation to ensure robustness.
7. **Implement the system:** Once AES system has a trained and validated model, build an interface or API that allows users to submit essays for scoring. Ensure that the system is user-friendly and provides clear instructions. Test the system thoroughly to identify and fix any bugs or performance issues.
8. **Continuous improvement:** Monitor the performance of AES system over time and gather user feedback. Periodically update the model and scoring criteria based on new data and insights. This iterative process will help improve the accuracy and reliability of the system.
9. **Ethical considerations:** Pay attention to ethical considerations, such as bias detection and mitigation, ensuring fairness across different demographic groups, and maintaining privacy and data security.

## DISCUSSION

This study employed an NN approach utilizing learning vector quantization to train essays scored by humans. Following training, the network is capable of assigning scores to ungraded essays. Initially, essay processing steps were conducted to eliminate spelling errors and perform pre-processing tasks such as document tokenization, stop word removal, and stemming. Subsequently, the pre-processed essay was fed into NN. The model then provided feedback on the essay's relevance to the topic. The obtained Kappa coefficient score is 0.364. Although not according to Williamson et al. (2012), a desirable value for QWK in automated scores for high-stakes tests should be at least 0.70. However, the performance of AES systems using the methods mentioned above is commendable. QWK is common for AES assessment, but limitations identified suggest additional metrics are needed for accurate evaluation (Doewes et al., 2023). AES via NNs enhances accuracy and efficiency. Uncertainty exists about which word embedding techniques offer superior AES accuracy. This study assesses fine-tuned modern embeddings' impact on an LSTM AES model. Pretrained GloVe outperformed Word2Vec, elevating accuracy. Findings aid AES research in optimal word representations and fine-tuning strategies (Firoozi et al., 2022). NN models have shown promise but lack rating criteria integration. A symmetrical model, Siamese bidirectional long short-term memory architecture (SBLSTMA), pairs essays with rating criteria samples. Applied to ASAP dataset, SBLSTMA proves superior to prior NN methods (Liang et al., 2018). Furthermore, deep learning networks employing LSTM and pre-trained GloVe word embeddings were also utilized. This approach extracted various features, such as sentence count, word count per sentence, the number of out-of-vocabulary words in the sentence, language model score, and text perplexity. The network predicted the goodness scores for each essay, with higher scores indicating a higher rank and vice versa (Nguyen & Dery, 2016; Mathias & Bhattacharyya, 2018a, 2018b, 2020). AES seeks efficient grading. DNNs replace feature engineering. A hybrid approach, integrating DNN and RNNs, faces issues. A novel method combining handcrafted essay-level features with DNN-AES models yields substantial accuracy improvement (Uto et al., 2022).

According to the results, AES has several advantages in KKU-AELT essay writing assessments. First, it can improve scoring consistency. KKU-AELT AES system yield scores that consistently agree with those of human raters at a level as high, if not higher, as the level of agreement among human raters themselves. This agrees with the previous studies by Lahitani et al. (2016) and Zupanc and Bosnic (2020) showing that AES systems yield scores that are consistent and aligned with human graders' assessments. Second, it can reduce time required for scoring and reporting. Manually scoring constructed-response tasks can be time-consuming and expensive. KKU-AELT AES system can offer a potential solution using computer technology to score written responses quickly and accurately. Similarly, several empirical investigations have underscored the time-saving potential of AES systems. For example, the research by Hussein et al. (2019), Phoophuangpairoj and Pipattarasakul (2022), and Ramnarain-Seetohul et al. (2022) has shown that automated systems significantly reduce the time required for scoring and reporting, compared to manual grading methods. This efficiency is

vital for streamlining assessment processes. Third, KKU-AELT AES system can minimize scoring costs associated with scoring constructed-response tasks and number of human raters. This finding is in line with some empirical studies (Lahitani et al., 2016; Steimel & Riordan, 2020) in that the implementation of automated systems can curtail the expenses associated with human raters, making the assessment process more economical. Moreover, KKU-AELT AES system can provide students with immediate feedback on constructed-response tasks. The system can provide immediate feedback on their written responses, allowing them to identify areas for improvement and adjust their study strategies accordingly. Likewise, a number of previous studies (e.g., He et al., 2022; Ramnarain-Seetohul et al., 2022) have established that AES systems effectively provide students with immediate feedback on their written responses. This feedback mechanism assists learners in recognizing areas for improvement, guiding their study strategies, and enhancing their writing skills (Phoophuangpairoj & Pipattarasakul, 2022).

In sum, the advantages of AES system as observed in KKU-AELT essay writing assessment are empirically supported. These include improving consistency with human raters, reducing scoring time, minimizing scoring costs, and providing students with immediate feedbacks. The empirical evidence from various studies emphasizes the potential of KKU-AELT AES system as a valuable solution for addressing challenges related to writing skill assessment in graduate students.

### Limitations & Recommendations

While KKU-AELT AES system offers efficient and consistent assessments, it is important to acknowledge its limitations. First, AES system has limited ability to assess certain aspects of writing, such as grammar and syntax, but struggles to evaluate more complex aspects of writing, such as creativity or critical thinking, nuanced arguments, or originality. Recognizing the subjective nature of these aspects, automated systems should be utilized as complementary tools, allowing human graders to assess these dimensions. Second, in terms of context-specific understanding, challenges arise when dealing with context-specific language use. AES system might misinterpret figurative language or idiomatic expressions. Ensuring algorithms that can decipher context accurately is crucial for a more reliable scoring process. The final limitation lies in the limited human touch. In this study, the automated system lacked the human touch, making them unable to fully comprehend nuances in writing. Collaboration between automated systems and human evaluation can be further investigated to provide a balanced and comprehensive assessment.

To cope with the limitations, the following recommendations are offered. First, integrating human evaluation alongside automated systems should be promoted to enhance the assessment process and address the limitations of automated scoring. Human graders can evaluate creativity, originality, and nuanced arguments that automated systems might struggle with, providing a more comprehensive evaluation. By combining automated scoring with human evaluation, addressing challenges in context-specific language, and enhancing feedback mechanisms, educators can harness the benefits of technology while maintaining a comprehensive and accurate assessment of English writing skills. Second, to improve the understanding of context-specific or domain-specific language use, refining the algorithms to recognize specialized vocabulary and context can enhance the accuracy of automated scoring, making the system more versatile. Finally, enhancing the feedback mechanism to provide scores and specific suggestions for improvement can make the system more valuable for learners. Incorporating detailed feedback on grammar, coherence, and structure can enhance the skill.

## CONCLUSIONS

AES is a process of evaluating and scoring written English language proficiency using computer algorithms. Traditionally, assessing English language proficiency has been a time-consuming and labor-intensive task, requiring human graders to read and evaluate essays or written responses. However, with advances in NLP and machine learning, it has become possible to develop automated systems that can assess and score English language proficiency more efficiently. KKU-AELT AES system worked by analyzing various linguistic features of given pieces of writing. These features included grammar, vocabulary usage, sentence structure, coherence, and overall fluency. The algorithms behind these systems were trained on large datasets of pre-scored essays or writing samples, allowing them to learn patterns and correlations between the features and

the corresponding scores assigned by human graders. To create an automated English scoring system, a large corpus of essays or writing samples (i.e., ASAP and KKU-AELT) was collected and annotated with human-assigned scores. This dataset was then used to train a machine learning model, i.e., an NN, to predict the scores based on the linguistic features extracted from the texts. Once the model was trained, it was used to automatically score new essays or written responses. The system took the input text, analyzed its linguistic features, and produced a predicted score based on the learned patterns from the training data. The predicted score was used to assess the writer's English language proficiency, evaluate their writing skills, and provide feedback. KKU-AELT AES system has been found to offer several advantages. It is efficient, as it can process and score a large number of essays or written responses in a short amount of time. It is consistent, as it can apply the same criteria and algorithms to all submissions, reducing potential bias or subjectivity. Additionally, it can provide immediate feedback to students, allowing them to improve their writing skills and monitor their progress. However, it is important to note that KKU-AELT AES system are not perfect and have limitations. While the system can assess certain aspects of writing, it may struggle with evaluating creativity, critical thinking, originality, or nuanced arguments. It may also face challenges with understanding context-specific or domain-specific language use. Therefore, it is crucial to use AES system as a supportive tool alongside human evaluation, rather than a complete replacement.

In conclusion, KKU-AELT AES system is a process that uses machine learning algorithms to evaluate the quality of written English of non-native English speakers. It analyzes various linguistic features such as grammar, vocabulary, coherence, and organization to assign a score to a piece of text. This technology can be used in language proficiency testing, educational assessments, and language learning platforms. It provides a quick and objective evaluation of English writing, allowing raters or professors to efficiently assess and provide feedback on a large number of student essays. KKU-AELT AES system does not comprehend written text, but instead uses related variables to grade written responses. The system demonstrates optimal performance when handling tasks characterized by distinct input (language features) and tangible output (rater scores). A machine learning algorithm establishes the relationship between input and output, leveraging empirical evidence from a training dataset. Consequently, a dependable collection of prompt-specific essays accompanied by human scores is essential for the computer to grade written responses accurately. The accuracy of the grading process can be evaluated through agreement measures.

**Author contributions: KP:** conceived & designed analysis, collected data, contributed data & analysis tools, performed analysis, & wrote paper; **PM:** conceived & designed analysis, contributed data & analysis tools, & wrote paper; **WC:** conceived & designed analysis, contributed data & analysis tools, & wrote paper. All authors approved the final version of the article.

**Funding:** This article was supported by the Fundamental Fund of Khon Kaen University. The research on "a comparative study of English as a foreign language learners' academic writing ability evaluated by learning algorithm and human rater judgement" by Khon Kaen University has received funding support from the National Science, Research and Innovation Fund ("NSRF").

**Acknowledgements:** The authors would like to thank the Center for English Language Excellence, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand, for supporting the essay dataset for this study.

**Ethics declaration:** The authors declared that since the study used pre-existing data that is openly accessible and did not necessitate approval from an ethics committee, ethical review and approval were waived.

**Declaration of interest:** Authors declare no competing interest.

**Data availability:** Data generated or analyzed during this study are available from the authors on request. The code supporting this study is available at: https://colab.research.google.com/drive/1YYCdy5sO8U_-I4vyZQzcAuNAwFhTdt6X?usp=sharing

## REFERENCES

Ajay, H. B. (1973). Strategies for content analysis of essays by computer. *University of Connecticut*. https://search.proquest.com/openview/739b97ecbfd94af0356f4da011575ef8/1?pq-origsite=gscholar&cbl=18750&diss=y

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *arXiv, 1606*, 04289. https://doi.org/10.18653/v1/P16-1068

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2.0. *Journal of Technology, Learning, and Assessment, 4*(3), i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications, 11*(10), 204-210. https://doi.org/10.14569/IJACSA.2020.0111027

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

Cozma, M., Butnaru, A. M., & Lonescu, R. T. (2018). Automated essay scoring with string kernels and word embeddings. *arXiv, 1804*, 07954. https://doi.org/10.18653/v1/P18-2080

Doewes, A., Kurdhi, N., & Saxena, A. (2023). Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining* (pp. 103-113). International Educational Data Mining Society.

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning* (pp. 153-162). https://doi.org/10.18653/v1/K17-1017

Driessens, K., & Džeroski, S. (2005). Combining model-based and instance-based learning for first order regression. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 193-200). https://doi.org/10.1145/1102351.1102376

Evanini, K., Hauck, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the TOEFL junior® comprehensive writing and speaking test. *ETS Research Report Series, 1*, 1-11. https://doi.org/10.1002/ets2.12052

Firoozi, T., Bulut, O., Epp, C. D., Naeimabadi, A., & Barbosa, D. (2022). The effect of fine-tuned word embedding techniques on the accuracy of automated essay scoring systems using neural networks. *Journal of Applied Testing Technology, 23*, 21-29.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2), 939-944. https://doi.org/10.1201/9781351264808-1

Foltz, P. W., Yan, D., & Rupp, A. A. (2020). The past, present, and future of automated scoring for complex tasks. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 1-11). CRC Press.

Ghosh, S., & Fatima, S. (2010). Design of an automated essay grading (AEG) system in Indian context. *International Journal of Computer Applications*, *1*(11), 60-65. https://doi.org/10.5120/237-391

Graves, A., & Graves, A. (2012). *Supervised sequence labelling*. Springer. https://doi.org/10.1007/978-3-642-24797-2_2

Haberman, S. J. (2011). Use of e-rater® in scoring of the TOEFL iBT® writing test. *ETS Research Report Series,* 2, 1-13. https://doi.org/10.1002/j.2333-8504.2011.tb02261.x

He, Y., Jiang, F., Chu, X., & Li, P. (2022). Automated Chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3007-3016).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Horbach, A., & Zesch, T. (2019). The influence of variance in learner answers on automatic content scoring. *Frontiers in Education, 4*, 28. https://doi.org/10.3389/feduc.2019.00028

Huang, X., Sun, J., & Sun, J. (2018). A car-following model considering asymmetric driving behavior based on long short-term memory neural networks. *Transportation Research Part C: Emerging Technologies, 95*, 346-362. https://doi.org/10.1016/j.trc.2018.07.022

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208

Kaggle. (2012). *The Hewlett Foundation: Automated essay scoring*. https://www.kaggle.com/c/asap-aes/overview/evaluation

Kulkarni, C., Socher, R., Bernstein, M. S., & Klemmer, S. R. (2014). Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the 1st ACM Conference on Learning @ Scale Conference* (pp. 99-108). ACM. https://doi.org/10.1145/2556325.2566238

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016, April). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *Proceedings of the 4th International Conference on Cyber and IT Service Management* (pp. 1-6). IEEE. https://doi.org/10.1109/CITSM.2016.7577578

Li, X., Chen, M., Nie, J., Liu, Z., Feng, Z., & Cai, Y. (2018). Coherence-based automated essay scoring using self-attention. In M. Sun, T. Liu, X. Wang, Z. Liu, & Y. Liu (Eds.), *Chinese computational linguistics and natural language processing based on naturally annotated big data* (pp. 386-397). Springer. https://doi.org/10.1007/978-3-030-01716-3_32

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), 682. https://doi.org/10.3390/sym10120682

Mathias, S., & Bhattacharyya, P. (2018). ASAP++: Enriching ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Mathias, S., & Bhattacharyya, P. (2018). Thank "goodness"! A way to measure style in student essays. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 35-41). https://doi.org/10.18653/v1/W18-3705

Mathias, S., & Bhattacharyya, P. (2020). Can neural networks automatically score essay traits? In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 85-91). https://doi.org/10.18653/v1/2020.bea-1.8

Mesgar, M., & Strube, M. (2018). A neural local coherence model for text quality assessment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 4328-4339). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1464

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv, 1310*, 4546. https://doi.org/10.48550/arXiv.1310.4546

Munir, H., Wnuk, K., & Runeson, P. (2016). Open innovation in software engineering: A systematic mapping study. *Empirical Software Engineering, 21*, 684-723. https://doi.org/10.1007/s10664-015-9380-x

Nguyen, H., & Dery, L. (2016). *Neural networks for automated essay grading*. https://cs224d.stanford.edu/reports/huyenn.pdf

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543). https://doi.org/10.3115/v1/D14-1162

Phoophuangpairoj, R., & Pipattarasakul, P. (2022). Preliminary indicators of EFL essay writing for teachers' feedback using automatic text analysis. *International Journal of Educational Methodology, 8*(1), 55-68. https://doi.org/10.12973/ijem.8.1.55

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*(2), 103-134. https://doi.org/10.1016/S0747-5632(01)00052-8

Ramesh, D., & Sanampudi, S.K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*, 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Ramnarain-Seetohul, V., Bassoo, V., & Rosunally, Y. (2022). Similarity measures in automated essay scoring systems: A ten-year review. *Education and Information Technologies, 27*(4), 5573-5604. https://doi.org/10.1007/s10639-021-10838-z

Roy, S., Dandapat, S., Nagesh, A., & Narahari, Y. (2016). *Wisdom of students: A consistent automatic short answer grading technique*. NLP Association of India.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *Journal of Technology, Learning and Assessment, 4*(4).

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1*(2). https://ejournals.bc.edu/index.php/jtla/article/view/1668

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany. *ETS Research Report Series, 1*, 1-23. https://doi.org/10.1002/ets2.12249

Shaker, A., & Hüllermeier, E. (2012). IBLStreams: A system for instance-based classification and regression on data streams. *Evolving Systems, 3*, 235-249. https://doi.org/10.1007/s12530-012-9059-0

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 335-368). Routledge. https://doi.org/10.4324/9780203122761-20

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the world wide web. *Assessment & Evaluation in Higher Education, 26*(3), 247-259. https://doi.org/10.1080/02602930120052404

Srisawat, C., & Poonpon, K. (2023). Revision of an academic English writing rubric for a graduate school admission test. *PASAA, 65*, 234-262.

Steimel, K., & Riordan, B. (2020). Towards instance-based content scoring with pre-trained transformer models. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

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). https://doi.org/10.18653/v1/D16-1193

Tashu, T. M., & Horváth, T. (2020). Smart score-short answer scoring made easy using Sem-LSH. In *Proceedings of the 14th International Conference on Semantic Computing* (pp. 145-149). IEEE. https://doi.org/10.1109/ICSC.2020.00028

Uto, M., Xie, Y., & Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6077-6088). https://doi.org/10.18653/v1/2020.coling-main.535

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360-363.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.https://doi.org/10.1017/CBO9780511732997

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement Issues and Practices, 31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1-10. https://doi.org/10.1016/j.asw.2015.06.002

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1560-1569). https://doi.org/10.18653/v1/2020.findings-emnlp.141

Yang, Z., Yu, Y., You, C., Steinhardt, J., & Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. In *Proceedings of the International Conference on Machine Learning* (pp. 10767-10777). PMLR.

Zechner, K., Yoon, S., Bhat, S., Leong, C. W. (2017). Comparative evaluation of automated scoring of syntactic competence of non-native speakers. *Computers in Human Behavior, 76*, 672-682. https://doi.org/10.1016/j.chb.2017.01.060

Zhao, Z., Liu, T., Li, S., Li, B., & Du, X. (2017). Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 244-253). https://doi.org/10.18653/v1/D17-1023

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555

Zupanc, K., & Bosnić, Z. (2020). Improvement of automated essay grading by grouping similar graders. *Fundamenta Informaticae* [*Fundamentals of Informatics*]*, 172*(3), 239-259. https://doi.org/10.3233/FI-2020-1904