Semi-Automatic Short-Answer Grading Tools for Thai Language using Natural Language Processing

Chatchai, Wangwiwattana*
University of the Thai Chamber of Commerce, Thailand
chatchai_wan@utcc.ac.th

Yuwaree, Tongvivat
University of the Thai Chamber of Commerce, Thailand
yuwaree_ton@utcc.ac.th

ABSTRACT

The past decade has witnessed enormous advancement in online educational resources. One noteworthy advancement has been the development of automatic learning platforms. The introduction of this new technology has raised questions about its effectiveness in aiding educators to improve the engagement of students and evaluate their achievement of learning outcomes. While the use of open-ended questions to assess learners' outcomes is valuable, the workload demanded of educators can increase considerably when open-ended questions are used in large classes. We have experimented with a semi-automatic method to help grade short open-ended questions answered in Thai language. Our method employed Keyword Matching and unsupervised document grouping. Fixed types of questions were tested using different algorithms. Keyword Matching was found to be an effective method for a relatively fixed, yet open-ended set of answers. For non-fixed types of answers, Document Clustering proved suitable. In generating grading tools, we adopted three methods: Keyword Matching; Sentence Vector Similarity Ranking; and Document Clustering with TF-IDF and K-Means. The algorithms were found to be useful for online learning and grading specific content-based answers which, in turn, may be used as a guide in directing educators who wish to elicit information to provide feedback.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Natural Language Processing**;

KEYWORDS

Natural Language Processing, MOOC, grading tool, Machine Learning

ACM Reference Format:
Chatchai, Wangwiwattana and Yuwaree, Tongvivat. 2022. Semi-Automatic Short-Answer Grading Tools for Thai Language using Natural Language Processing. In *2022 5th International Conference on Education Technology Management (ICETM 2022), December 16–18, 2022, Lincoln, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3582580.3582600

1 INTRODUCTION

The cutting-edge technology for online educational tools has removed former learning boundaries, enabling greater flexibility for learners as to when, where and how they study. For example, the online learning platform MOOCs, along with Hyflex make this possible [1], [2], [3], [4]. Educators are encouraged to adopt technologies that facilitate both online and face-to-face learning settings [5]. In shifting towards an online mode of learning, both as an alternative and supplementary to onsite classrooms, evaluating learning outcomes and providing important feedback can be a challenge when dealing with many students. While feedback serves as part of the learning process, allowing more opportunities for students to reflect upon their learning and how they can plan for further achievements, a considerable amount of time is required. Hence, the role of automatic grading comes into play.

A number of studies have been conducted to help develop approaches for automatic grading, and they have been adopted increasingly by educational institutions [6], [7], [8], [9]. Algorithms for grading in the classroom might need to be customized to be applicable for content-based and specialized subjects. In adopting an approach for evaluation that is specialized and subject-specific, learners are given freedom to express their thoughts. It is here that open-ended types of answers come into play. Thus, evaluating open-ended answers can help teachers evaluate how well students learn, as well as assess their levels of criticality that they have developed in the learning process, following Bloom Taxonomy for evaluating learning outcomes [10].

Although the pedagogical values of activating critical thinking in adopting open-ended answers, grading can be time-consuming, especially for teachers having large classes. It is possible for teachers to automatically grade short answers using online quiz forms; however, students need input the same answers as the teacher has designated as correct. This is not feasible because students may answer correctly but shape their language in a different form. Further technical issues involve current algorithms used for experimenting with automatic grading. While studies have shown that some models (i.e., LSA, Transformers, BERT, GPT-3) work better than others for essay grading [11], [12], [13], [14], depending on the types of questions and answers, adopting deep learning models are not necessarily promising when compared with simpler methods using lexical overlap [15]. Some of the studies that apply deep learning algorithms also consider the pragmatic sense of language (e.g., levels of politeness) in essay scoring [12]. In this present study, which centers on grading short open-ended answers, the pragmatic level of language is not a primary concern. The expected answers are relatively fixed, yet open-ended, allowing a certain degree of creativity in the answers. We propose a method that could be applicable for this type of grading. In doing so, we adopt TF-IDF,

[*] chatchai_wan@utcc.ac.th

ICETM 2022, December 16–18, 2022, Lincoln, United Kingdom

ACM ISBN 978-1-4503-9801-5/22/12…$15.00
https://doi.org/10.1145/3582580.3582600

K-means, word embedding, Sentence Vector Similarity Ranking and Sentence Embedding. Applications of Keyword Matching with Regular Expression are useful for fixed types of answers [16], while unsupervised clustering with TF-IDF and sentence vector similarity ranking works well for relatively non-fixed types of answers. After we have touched upon some of the concepts outlined above, Section 2 will then provide some key features in automatic grading in the Thai language, data collecting, and data pre-processing that will frame the methods and discussions. Section 3 proposes the design of a semi-automatic grading system for varied types of open-ended questions. Section 4 then describes results from adopting the models for grading, analyzing texts in Thai language and applications for automatic grading. Finally, Section 5 wraps up the ideas in this study and brings them to a conclusion. The following sections outline the nature of Thai language and current practice of NLP for Thai language and current applications on automatic grading.

1.1 The Nature of Thai Language and Current Practice of NLP for the Thai Language

Natural Language Processing ("NLP") is a branch of linguistics, computer science, and artificial intelligent, which focuses attention on how computers and human language interact, as well as how to teach computers to process and analyze vast amounts of natural language data. The goal is to create a machine that can "understand" the contents of documents, including the subtleties of language used in different contexts [17]. The rapid development of NLP for the Thai language has expanded continuously in recent years. Despite these achievements, the effectiveness and accuracy in the application of current algorithms for text analyses is still limited due to some unique features of the Thai language. One main issue is related to word segmentation, where the supposed meanings are conveyed from the sentence level to the word level and then to the level of the lexical unit. Sentences are typically in the Subject-Verb-Object (S-V-O) order [18], which is comparably the same in English. Yet, there are other variations in sentences including the O-V-S order and S-O-V order [19]. The mixed patterns and permutation of sentences could hinder the accuracy of results from text analyses if algorithms are trained to understand only S-V-O patterns of sentences.

Space is another unique issue in Thai because it is a scriptio continua language. Space does not function as a delimiter between words, but typically signifies that a new sentence follows on instead of having a full stop to separate sentences. Therefore, understanding how lexical units are connected to form intended meanings and the differentiation of lexical units and their functions in sentences is essential in understanding meanings. The spacing issue at the word level is related to the words juxtaposed together. A word can contain (smallest) lexical units that constitute meanings, but without space. Lexical units within a word can be polysemous and postulate different meanings depending on the way in which the consonant or vowels within words are segmented. For example, ตากลม can be segmented as: (1) ตาก (verb) ลม (noun) meaning *to be exposed to the air*, and (2) ตา (noun) กลม (adjective) has a literal meaning as *rounded eyes*. Part of speech (POS) is thus important because segmentations are delineated by word class and the smallest lexical units. The last issue involves the sentence patterns. Thai language has the S-V-O sentence pattern with possibilities of a mixture of O-V-S and S-O-V patterns, which generates different types of meanings. This would be challenging for machines to understand, due to variations being linked to certain grammatical patterns. Additionally, POS is important for meaning differentiation. These issues would be attributable to the correct understanding of meanings and accurate frequency word count, which is fundamental for our model selection.

One of the challenges in choosing models for analyzing Thai texts is word tokenization. Some models, for example PyThai NLP's newmm (1993); TLTK, are dictionary-based, which could tend to ignore words that are not listed in the dictionaries in tokenization [19]. This is an issue in terms of word strings. For example, มากกกกก with repeated ก is an emphasized version of มาก meaning *a lot*. In this case, we removed the emphasis string in the tokenization process. We selected a learning-based model ("ULMFit"), for text classification [20]. Although the model is not Thai-language-specific, the versatile universal capability for the model appears effectively in relation to our datasets, alongside the nature of Thai language and current practice of NLP for Thai language.

1.2 Automatic Grading

Studies have been conducted on automatic grading over the past decades. Many researchers have successfully experimented with models for automatic grading which are useful tools for being implemented on non-fixed types of answers [7], [21]. While some of these models such as EBOW and SLTM are focused on semantic tasks such as reading comprehension, other context-based models would benefit fixed types of answers. In this case, contextual clues are essential factors in the choice of models for grading. Deep-learning methods using Transformers and Bert are highly effective for assisting essay grading, and the increased efficacy justifies the high cost of power consumption [13], [14]. A key advantage with these models is that they facilitate the differentiation of author writing styles and content, which is useful for grading long answers. For grading very short answers, algorithms that require low power demand and less cost would be sufficient for the purpose.

This study focuses on grading open-ended short answers. We evaluate three types of models that are applicable, which are Keyword Matching; Sentence Vector Similarity Ranking; and document clustering with Term frequency-inverse document frequency ("TF-IDF") and K-Means. TF-IDF is a statistical representation of texts based on the word overlapping in the whole corpus. This approach is commonly used to classify texts due to its efficacy and straightforwardness. In this work, we represented text documents using TF-IDF. Simple K-means clustering is a well-known unsupervised learning technique for document clustering [22]. The word2vec algorithm employs a neural network model to learn word connections from a vast corpus of texts. Once trained, the model can find synonymous words or suggest extra words for an incomplete text. Word2vec represents each different word with a specific list of numbers known as a vector. The vectors are carefully chosen so that a simple mathematical function (cosine similarity between vectors) reflects the level of semantic similarity between the words represented by those vectors. We average word vectors in each sentence to estimate the similarity between two sentences. We call them "sentence vectors." The similarity can be quantified as the cosine of the angle between two sentence vectors. The vector representations suggest semantic similarity by classifying and differentiating contexts in given sentences when compared to/against contexts in other sentences [23]. Keyword Matching by using simple regular expression is employed in automatic grading systems. Many of those are used in a fixed set of responses, such as programming classes where the answers have a clear structure [24]. Some learning management systems (LMS) have integrated regular expression within their system. Once similar answers are grouped together, they can be used for grading answers as a group receiving the same scores.

2 METHODOLOGY

2.1 Data Collection

Data were collected from students' answers to five open-ended questions. The questions were designed to evaluate students' understanding of the subject contents, and their ability to express their criticality. In keeping with criticality-based evaluations, we follow Bloom et al.'s (1956) criteria [10]. We gathered brief responses to the five questions from 79 students enrolled in a Games and Esports Business course at a university in Bangkok in 2021. Respondents were first-year Thai-national university students. The test took place on-site at the university, but students responded to open-ended questions by typing their answers in an online MS form. The test information is shown below. This is an open-book test. The free-response questions are as follows:

- Q1: What is the second valuable source of income for the esports industry?
- Q2: What is the 3rd highest viewing hour game in the data above?
- Q3: How many people are expected to be interested in esports in 2021?
- Q4: Based on this information, which game (LOL, Dota2, CS GO) is worth investing in and why?
- Q5: Do students think loot boxes in video games are gambling? Why?

The first three questions (Q1, Q2, and Q3) assess students' knowledge on the subject materials, based on Bloom's taxonomy on the 'Understanding' and 'Applying' levels of criticality. Q4 and Q5 place a greater emphasis on the 'Analyzing' and 'Evaluating' criticality levels of Bloom's taxonomy [10], thus students are supported to freely express their own views and criticality in answering these two questions. Scores range from 0-2. A score of 0 was given to empty or irrelevant answers; a score of 1 is given to a correct answer without any reason or justification provided; and a score of 2 is given to an answer with a reason or some explanation. 78 short answers were retrieved from students. Altogether 390 data points were collected in total.

2.2 Data Preprocessing

Each question was analyzed independently. All questions were processed with the same workflow for the data analysis. Three stages were adopted in data preprocessing. The first stage was to tokenize the answers, followed by cleaning the answers by removing tags, question marks and stop words. The second stage then was to transform all answers using the TF-IDF method and adopting simple K-means to cluster the cleaned answers. The answers were subsequently evaluated based on the number of clusters using Elbow methods. The third stage was exporting the results to Excel for teachers to grade the groups of answers and show a dashboard.

2.3 Unsupervised Clustering with K-means Clustering and Sentence Vectors Similarity Scores

Q4 and Q5 encourage students' creativity. Thus, there are several potential responses. We utilized unsupervised learning to group together all responses with a similar topic. The summary of this process is shown in Figure 1.

Due to the modest size of our dataset (just 79 students), the word dictionary generated by TF-IDF may be sparse and have too high dimensionality. Simple K-means could be afflicted by these characteristics, known as the curse of dimensionality. To overcome this issue, Latent Semantic Analysis or LSA was used to lower the dimension. Latent Semantic Analysis (LSA) is performed using Singular Value Decomposition (SVD) to decompose a matrix containing word representations. The dimensionality of the word representation matrix is reduced to 100 components while there is still no loss of information. The data in the text are used to determine an algorithm's optimal number of clusters without supervision. The Elbow Method and the Silhouette coefficient are used in order to detect clusters and assess the quality of clustering. The elbow approach is a heuristic for picking the ideal number of clusters, which is suggested in an elbow-shaped curve. The silhouette coefficient is a way to measure how well clustered data points are. A higher silhouette coefficient score suggests a higher degree of clustering quality. To assess the efficacy of clusters, we use coherence ratings. The coherence score indicates the consistency of grading in each group. A group having a value of 1.0 means that responses in the group received the same grade. Scores below 1.0 imply that teachers must analyze a variety of student grades.

$$coherence_c = max\left(\frac{r_e}{n_e}\right)$$

Where c represents the number of clusters, r_e is the teacher rating grade in each evaluation metric , and n_e represents the number of all members in cluster evaluation matric e. Intuitively, it is a maximum ratio of all teacher evaluation scores in each cluster. This technique is value independence. That means its coherence score is not affected by the value of attributes in data sets. The second method we propose is Sentence Vectors Similarity Ranking. The goal is to score students' responses in order of how closely they resemble teachers' responses. Although Q4 is a more open-ended question, there is a fixed set of answers and results that teachers might expect. Teachers may create an expected answer in the form of text, and they might want to rank the answers that are most similar to a teacher's answers. To address that, we apply a sentence vector similarity ranking. In doing so, teachers' and students' answers are converted into sentence vectors. With this method, we can measure how close the teacher's answer is to the students' answers by using cosine similarity. Then we rank those

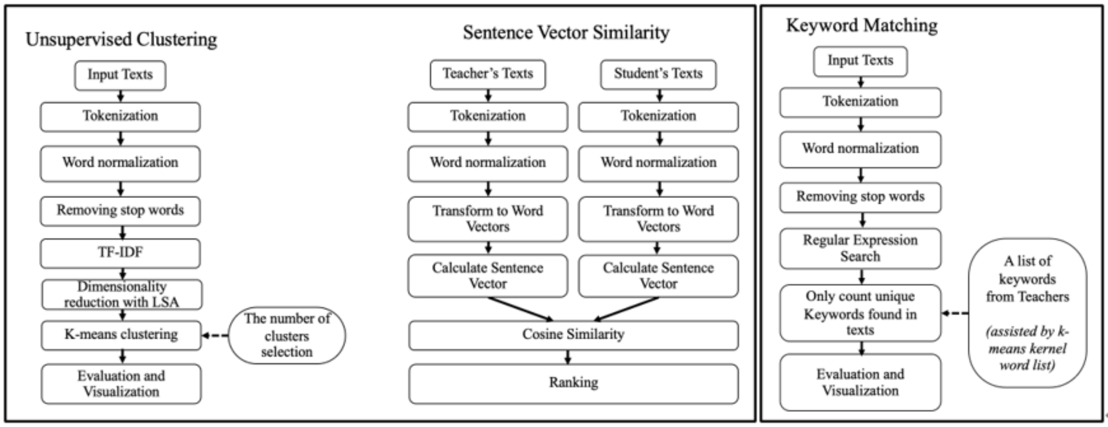
Unsupervised Clustering
Input Texts
Tokenization
Word normalization
Removing stop words
TF-IDF
Dimensionality reduction with LSA
K-means clustering
The number of clusters selection
Evaluation and Visualization
Sentence Vector Similarity
Teacher's Texts
Student's Texts
Tokenization
Tokenization
Word normalization
Word normalization
Transform to Word Vectors
Transform to Word Vectors
Calculate Sentence Vector
Calculate Sentence Vector
Cosine Similarity
Ranking
Keyword Matching
Input Texts
Tokenization
Word normalization
Removing stop words
Regular Expression Search
Only count unique Keywords found in texts
A list of keywords from Teachers
(assisted by k-means kernel word list)
Evaluation and Visualization

Figure 1: the summary of the unsupervised clustering process (left), sentence vector similarity process (center), the summary Keyword Matching algorithm (right)

similarity scores. The closest to 1.0 indicates the most similarity between two answers.

2.4 Keyword Matching

The concept of word matching is predicated based on the notion that queries have preset, fixed responses. Teachers have a set of expected responses. This strategy was used for evaluating answers from Q1-3. The summary of the algorithm is shown in Figure 1. The analytical process began with tokenizing and normalizing all student responses. The process of normalizing words is addressed in Section 2.2. Data preprocessing. Subsequently, we searched for unique terms in each student's response and counted the number of times each keyword appears. Although the answers to questions one through three are relatively established, they are nevertheless open-ended. In this regard, all possible keywords attributable to answers to the questions might be challenging for teachers to come up with. This problem is circumvented by extracting keywords from responses using 1) word clouds and 2) cluster kernels. We determined the frequency of each word in the corpus and generated a word cloud. The size of high-frequency words is larger than that of other words. Words and non-stop words that frequently occur suggest that they are noteworthy and potentially important based on students' views. In addition to word clouds, utilization of words in clusters of kernels was also adopted. TF-IDF and K-means clustering were utilized to extract these keywords. The approach can extract terms that typically appear in one student's response but not in the responses of other students. Combining these two techniques would enable instructors to select important keywords from the machine to help classify and score groups of answers.

3 RESULTS

We compared actual teacher evaluation scores for Questions 1-5 to examine them. Answers to questions 1, 2, and 3 are concise and specific. Although there is some variety in the answers, the result demonstrates that it can classify each item accurately (100 percent accuracy). However, keyword matching heavily depends on the teacher's input. A good keyword selection can have a significant impact on performance. Questions 4-5 are more open-ended allowing more freedom in the responses. The objective of these questions is to evaluate whether students can analyze the situation based on the knowledge they have learned. We use two methods that are unsupervised clustering with TF-IDF and K-means, and sentence vector similarity ranking with Q4-5 answers.

Table 1: Coherence scores of each cluster in question 4 and 5

Questions	Clusters	Coherence scores
Q4	Cluster 1	0.875
	Cluster 2	0.948
Q5	Cluster 1	0.846
	Cluster 2	1.000
	Cluster 3	1.000
	Cluster 4	0.833
	Cluster 5	0.941
	Cluster 6	1.000
	Cluster 7	0.909

Figure 2 demonstrates distinct variations in dispersion and silhouette coefficient values amongst different groups. This implies that the data in Q4 may be grouped by the K-means technique. In this study, we made use of the recommended cluster sizes of 2 for Q4 and 7 for Q5. Table 1 shows coherence scores of Q4 indicating that each cluster has similar student grades. Q5 also has similar results, clusters 2,3, and 6 reach 1.0 coherence.

The teacher's answer is "League of legends is probably the best investment option, because it has the highest *number of* viewers continuously *for a long time"*. The top three answers contain information similar to the teacher's answer; while the bottom three have considerably different answers from those of teachers shown in Table 2.

In Q5, the teacher's answer is "Loot box is not gambling because the percentage is clearly defined and players get the digital product every time; therefore, according to the law, it is not gambling". The top three answers contain information similar to the teacher's

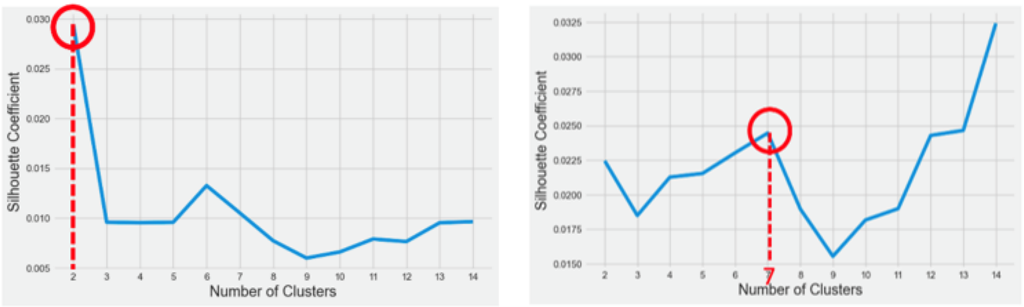

Figure 2: Silhouette coefficient of 2 - 14 clusters of Q4 (left) and Q5 (right)

Table 2: samples Q4 answers ordered by similarity scores

Example Answers	Similarity Scores
League of Legends because it has the most viewing hours. This means it has the highest number of views as well. Make investments in this game to get the most viewers.	0.890
LOL or League of Legends because of its high viewership, it is the most attractive investment and success.	0.884
LEAGUE OF LEGENDS because it generates the highest revenue.	0.868
. . .	
League of Legends I think so because there are a lot of people playing.	0.265
Roblox	0.
League of Legends	0.

Table 3: samples Q5 answers ordered by similarity scores

Example Answers	Similarity Scores
No. The loot box clearly describes the random rate of each reward. But if the rate of obtaining the item is not given, it may be considered unnecessarily exploitative.	0.838
Loot box is gambling, gambling according to the dictionary. The Royal Institute's edition states that "Playing for money or anything else through intelligence, dexterity, cunning, wit and skill, including luck." So, loot boxes are gambling because it's about taking random in-game money or money to get what we want.	0.806
It is not a gamble because it is another option to randomly receive the items in the game that some players want at a lower price. A full price item might be too expensive. But the probability amplitude for a chance of winning is low due to randomness, causing people to have to buy items at the full price instead.	0.805
. . .	
Gamble, nothing is certain, all bets.	0.367
-	0.

answer, while the bottom three students' answers are quite different shown in Table 3.

4 DISCUSSION

Each of the techniques we utilized in this study is appropriate for a different sort of situation. The answers for Q1-3 are quite fixed. The results show that for these tasks, a more straightforward approach, such as keyword matching with regular expressions, performs well. The teacher's pre-assigned keywords are crucial to this method's success. Teachers may be able to extract keywords useful for grading by using K-means, which could be applicable for most learning management systems. Thus, this semi-automatic technology should work with most classes in teaching that requires grading fixed open-ended answers.

Q4 has a predetermined set of solutions. The results show that TF-IDF and K-means can categorize answers, providing a 0.95 coherence score in this case. This makes it possible for teachers to assess each group of answers quickly. This technology could ease and expedite the grading process for short answers.

Due to the nature of Q5, which focuses on the expression of thoughts, students are given more freedom in their responses. The findings demonstrate that TF-IDF and K-means group content together reasonably well. Nonetheless, high coherence scores do not mean the system automatically predicts the correct score. It still requires teachers to grade each group. In addition, the second responses to question 5 have high similarity ratings, but their contents are different. (*Loot box is gambling, gambling according to the dictionary. The Royal Institute's edition states, "Playing for money or anything else through intelligence, dexterity, cunning, wit and skill. Including luck". So, loot boxes are gambling because it's about taking random in-game money or money to get what we want.*) The student believes loot boxes are gambling, but the teacher asserts that they are not. This may imply that the sentence vector similarity scores compare the similarity of word selections rather than contexts. Although we used a more modern learning technique in this study, creating a system that understands the entire context is still challenging, especially in Thai. The three methods presented in this paper exploit the fact that the answers in this assessment are not too varied. In this case, correct answers are expected. In this regard, creating an automatic grading system effectively for various types of answers is still a research gap for future.

5 CONCLUSION

This study is preliminary research aimed at facilitating instructors' use of semi-automatic tools. Techniques used in this study are typical for use in analyzing social media texts. In the applications for the pedagogical domain, the techniques work well for grading specialized and subject-specific short open-ended, yet relatively fixed answers. Since assessments and grading criteria vary among teachers and institutions, a fully automatic approach for grading is yet to be further developed. Nevertheless, a tool for grouping answers would augment the flexibility in evaluating short answers, enabling semi-automatic grading for many courses. The grading tools presented comprise three methods, beginning with evaluating the use of short-text clustering with TF-IDF and Simple K-means to student responses. Simple keyword matching presented can be effective to evaluate very short answers. Sentence vectors can be used to rank students' answers that are similar to pre-defined teacher's answers. Future research could explore other types of answers that determine the efficacy of employing these strategies to foresee more useful implications from automatic grading systems.

REFERENCES

[1] M. M. M. Abdelmalak and J. L. Parra, "Expanding Learning Opportunities for Graduate Students with HyFlex Course Design," *International Journal of Online Pedagogy and Course Design*, vol. 6, no. 4, pp. 19–37, Oct. 2016, doi: 10.4018/IJOPCD.2016100102.

[2] B. Beatty, "Hybrid courses with flexible participation: The hyflex course design," *Practical Applications and Experiences in K-20 Blended Learning Environments*, pp. 153–177, Dec. 2013, doi: 10.4018/978-1-4666-4912-5.CH011.

[3] M. Detyna, R. Sanchez-Pizani, V. Giampietro, E. J. Dommett, and K. Dyer, "Hybrid flexible (HyFlex) teaching and learning: climbing the mountain of implementation challenges for synchronous online and face-to-face seminars during a pandemic," *Learn Environ Res*, Apr. 2022, doi: 10.1007/s10984-022-09408-y.

[4] N. Thi, H. Minh, and T. T. Linh, "Designing Online English Grammar Exercises 10th Graders via Learning Management System Chamilo," *Journal of English Language Teaching and Applied Linguistics* , vol. 3, no. 5, pp. 55–63, May 2021, doi: 10.32996/JELTAL.2021.3.5.6.

[5] B. L. Moorhouse, Y. Li, and S. Walsh, "E-Classroom Interactional Competencies: Mediating and Assisting Language Learning During Synchronous Online Lessons:," https://doi.org/10.1177/0033688220985274, Feb. 2021, doi: 10.1177/0033688220985274.

[6] M. A. Hussein, H. Hassan, and M. Nassef, "Automated language essay scoring systems: A literature review," *PeerJ Comput Sci*, vol. 2019, no. 8, 2019, doi: 10.7717/PEERJ-CS.208.

[7] V. S. Kumar and D. Boulanger, "Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined?," *Int J Artif Intell Educ*, vol. 31, no. 3, pp. 538–584, Sep. 2021, doi: 10.1007/S40593-020-00211-5.

[8] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artif Intell Rev*, vol. 55, no. 3, p. 2495, Mar. 2022, doi: 10.1007/S10462-021-10068-2.

[9] V. Kumar and D. Boulanger, "Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value," *Front Educ (*Lausanne*)*, vol. 5, Oct. 2020, doi: 10.3389/FEDUC.2020.572367.

[10] B. S. Bloom, M. D. Engelhart, E. J. Furst, and D. R. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain.* 1956.

[11] J. Hoblos, "Experimenting with Latent Semantic Analysis and Latent Dirichlet Allocation on Automated Essay Grading," in *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Dec. 2020, pp. 1–7. doi: 10.1109/SNAMS52053.2020.9336533.

[12] S. Ludwig, C. Mayer, C. Hansen, K. Eilers, and S. Brandt, "Automated Essay Scoring Using Transformer Models," *Psych*, vol. 3, no. 4, pp. 897–915, Dec. 2021, doi: 10.3390/psych3040056.

[13] E. Mayfield and A. W. Black, "Should You Fine-Tune BERT for Automated Essay Scoring?," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 151–162. doi: 10.18653/v1/2020.bea-1.15.

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, doi: 10.48550/arxiv.1810.04805.

[15] L.-H. Chang, I. Rastas, S. Pyysalo, and F. Ginter, "Deep learning for sentence clustering in essay grading support," Apr. 23, 2021. https://arxiv.org/abs/2104.11556 (accessed Oct. 20, 2022).

[16] D. S. Morris, "Automatically grading Java programming assignments via reflection, inheritance, and regular expressions," *Proceedings - Frontiers in Education Conference*, vol. 1, 2002, doi: 10.1109/FIE.2002.1157985.

[17] J. Eisenstein, Introduction *to Natural Language Processing.* USA: Westchester Publishing Services, 2019.

[18] P. Singhapreecha, "Review of the book: A reference grammar of Thai," *Lingua*, vol. 117, no. 8, pp. 1497–1512, Aug. 2007, doi: 10.1016/J.LINGUA.2006.09.005.

[19] R. Arreerard, S. Mander, and S. Piao, "Survey on Thai NLP Language Resources and Tools," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Jun. 2022, vol. 4695, pp. 6495–6505.

[20] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 328–339, Jan. 2018, doi: 10.48550/arxiv.1801.06146.

[21] M. A. Hussein, H. Hassan, and M. Nassef, "Automated language essay scoring systems: A literature review," *PeerJ Comput Sci*, vol. 2019, no. 8, 2019, doi: 10.7717/PEERJ-CS.208.

[22] D. J. C. Mackay, *Information theory, inference, and learning algorithms Cambridge.* 2003.

[23] L. Logeswaran and H. Lee, "An efficient framework for learning sentence representations," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[24] O. Sychev, A. Anikin, and A. Prokudin, "Automatic grading and hinting in open-ended text questions," *Cogn Syst Res*, vol. 59, 2020, doi: 10.1016/j.cogsys.2019.09.025.