Thinking Skills Level Classification of Scientific Questions Using Bidirectional LSTM

Nanthawan Yaemsawat* and Nuanwan Soonthornphisaj*

Received: December 14, 2023
Revised: July 3, 2024
Accepted: July 19, 2024

* Corresponding Author: Nanthawan Yaemsawat, E-mail: nanthawan.ya@ku.th

DOI: 10.14416/j.it/2024.v2.003

Abstract

Science education with a suitable learning activity can help students enhance their thinking skills. Examination is one of the assessment tools to evaluate the student learning outcome in the domain of thinking skills. The Revised Bloom's Taxonomy, a well-known theory used to describe cognitive domains, divides thinking skills into two categories: basic and advanced thinking skills. Classifying questions according to their level of thinking abilities is an important task for teachers to design effective assessment tools. The objective of this study is to propose a model for classifying Thai language questions in science subjects. Initially, we used three algorithms: Bidirectional LSTM (BiLSTM), Naive Bayes (NB), and Support Vector Machine (SVM) for selecting Thai word tokenization algorithms. Then, we compare the model's performance using different feature sets. The combination of the question, training choice, and length of choice features with BiLSTM obtained an accuracy of 70%. Moreover, we employed part-of-speech (POS) tagging for feature selection. According to the findings, using nouns, verbs, adjectives, and adverbs enhances accuracy by 80.24%. This study shows the ability to use a model to categorize science questions to assist teachers in choosing questions that are appropriate to encourage higher-order thinking skills in students.

Keywords: Text Classification, Deep Learning, Learning Taxonomy, Science Education.

1. Introduction

Science education is aimed at developing important scientific skills in students, such as problem-solving, critical thinking, creative thinking, and decision-making. According to "The Future of Jobs Report 2023" from the World Economic Forum, analytical and creative thinking are the most significant skills for workers [1]. As a result, students must develop these skills, as they are necessary for the 21st century. The revised Bloom's taxonomy classifies learning outcomes into six levels of cognitive ability. These include remembering, understanding, applying, analyzing, evaluating, and creating [2], with each step representing a higher degree of cognitive growth. The questions at the top three levels are regarded as higher-order thinking abilities, whereas the questions at the bottom three levels are considered lower-order thinking skills [3]. Therefore, student assessments are essential to assess students' knowledge and skills. The teacher needs to design questions that are relevant to their learning outcome.

At present, many students in Thailand appear to lack 21st-century skills, especially higher-order thinking skills such as critical thinking [4]. As a result, teachers must create innovative methods of instruction for their students. Measurement from questions is a way of determining how learning activities help students improve their thinking skills. Therefore, teachers must create or select suitable questions from the question bank. The main issue is that teachers need to understand which questions require higher-order thinking abilities and which require lower-order thinking skills to assess and design effective tools. In addition, it was discovered that there are numerous institutions in Thailand, particularly in the countryside. Due to government budget constraints and limited teacher employment framework, teachers,

* Department of Computer Science, Faculty of Science, Kasetsart University Bangkok.

such as those with a degree in social sciences, must teach subjects in which they are not proficient, such as science. As a result, the design or selection of questions from the bank may be inappropriate. As a result of this research, we hope to help teachers understand the level of thinking of each question designed or chosen from the question bank so that they can choose questions that are appropriate for the instructional level objectives and encourage higher-order thinking skills.

Science questions can be classified automatically using machine learning (ML) techniques. According to previous studies, [5] implemented a K-Nearest Neighbors (KNN) and a Support Vector Machine (SVM) to classify questions into six levels based on the revised Bloom's taxonomy. A dataset containing 1,000 questions from an operating system course. They compared the performance of the two classifiers using the same dataset. The findings indicated that SVM outperformed KNN, with the highest F1-score of 92.3%. In addition, In the past few years, the classification of questions has become increasingly dependent on deep learning (DL). For example, [6] sought to improve the automatic revised Bloom's taxonomy CLOs and exam questions classification model by combining the proposed LSTM model with contextual domain embeddings. The dataset includes fields in computer science, electrical engineering, and business administration domains. They used different pre-trained embeddings to learn efficient word representations for their datasets and found that pre-trained embeddings, "namely, "Wiki Word Vectors", provided the highest accuracy (74% for the CLO dataset and 87% for exam questions). Difference classifiers from ML and DL were used for comparison, but the LSTM classifier performed the best.

Previous research has applied ML or DL to classify questions in English. Nevertheless, research that uses this method to classify science questions written in Thai is still scarce. To solve the problem, we propose a model for classifying science questions in Thai based on learning level, which is divided into 2 levels: basic thinking skills (lower-order thinking skills) and advanced thinking skills (higher-order thinking skills), using BiLSTM, NB, and SVM classifiers to compare performance. In this study, we describe approaches to effectively classifying science questions using ML and DL algorithms to develop an appropriate model that can assist teachers in creating suitable tools to measure and evaluate students, as well as improve students' higher-order thinking skills. As a result, utilizing this proposed model to classify science questions will significantly benefit Thai educators.

The remainder of this paper is as follows: Previous studies that classified questions using deep learning and machine learning are discussed in Section 2. Section 3 illustrates preprocessing and the processes of ML and DL to learn from the dataset, whereas Section 4 discusses the results of the experiment. Finally, Section 5 concludes this research paper.

2. Related Work

2.1 Bloom's Taxonomy

Benjamin Bloom was an educational psychologist who formulated a learning taxonomy that depicts the hierarchical behavioral characteristics of learners. He stated that every individual has three separate learning domains: cognitive, affective, and psychomotor [7].

In 2001, a team of psychologists under the direction of Anderson and Krathwohl—researchers, measurement, and evaluation experts—published A Taxonomy for Teaching, Learning, and Assessment [2]. Bloom's taxonomy is improved within the Cognitive Domain to facilitate the creation of learning objectives. After revision, this taxonomy gained widespread recognition in the academic community. Figure 1 shows six cognitive behavior levels the instructor desires students to possess. In addition, cognitive domain behavior has been divided into two levels based on the level of thinking skills: basic and advanced thinking skills. Note that the first three levels of thinking behavior: remembering, understanding, and applying are considered basic thinking skills. Analyzing, evaluating, and creating are considered advance thinking skills [3].

2.2 ML in Question Classification

Many studies use ML to classify computer programming

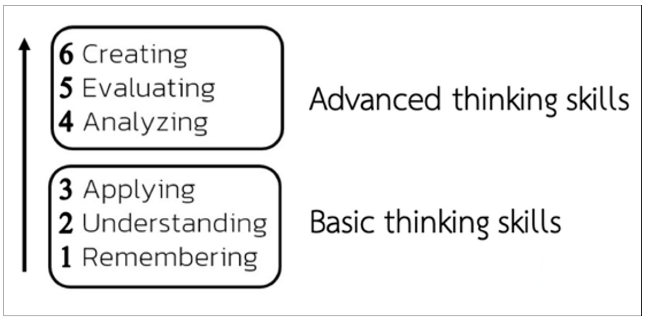


Figure 1. *The revised Bloom's Taxonomy.*

questions based on revised Bloom's taxonomy. For example, [8] aimed to classify programming questions utilizing KNN and NB classifiers as well as feature selection based on Chi-Square, Mutual Information, and Odds Ratio. They found that feature selection plays a key role in classification performance. The best F1 score result obtained from the KNN classifier is 90% using the Mutual Information feature selection, with an F1 measurement of 89%. [9] utilizes the revised Bloom's taxonomy to categorize online content for programming languages into three classes based on difficulty: beginner, intermediate, and advanced level, allowing learners to select content that best meets their requirements. The performance of the bag of words and POS tagging was compared using a random forest for classification. The results demonstrated that the use of the revised Bloom's taxonomy verbs and synonyms enhanced performance. The maximum accuracy obtained is 98%.

Some studies used ML to classify questions from different domains. [10] used ML to classify questions in a variety of areas, such as chemistry, biology, social marketing, math, law, and others, using the SVM, KNN, and Naïve Bayes (NB) classifiers. The dataset was collected from a variety of sources, including 600 questions. They observed that E-TFIDF, an upgraded version of the traditional TF-IDF feature, improves classifying questions better than traditional TF-IDF, while the SVM classifier excelled by up to 86%. Two years later, they compared feature extraction, including TF-IDF, TFPOS-IDF, and Word2Vec-TFPOSIDF. The question set was obtained from Yahya et al. (2012), and a few additional questions were gathered manually. The results revealed that Word2Vec-TFPOSIDF provided the best performance in this study, followed by TF-IDF and TFPOS-IDF [11].

Question classification based on revised Bloom's taxonomy has also been applied to languages other than English. [12] compared the performance of classification questions with two sets of questions translated from English to Chinese using Random Forests (RF), Logistic Regression, and XGBoost classifiers. POS tagging is used to generate keywords for training. The experimental results demonstrated that RF and selection keywords improve accuracy. The highest score was 86%. [13] categorized primary and high school questions. The dataset used would consist of mathematics and Indonesian language questions totaling 670. They trained both text and non-text questions to compare the results and use a pre-train model named IndoBert, which was learned from Indonesian text. The classifiers used were SVM and NB. The evaluation results show that the SVM classifier provided the highest value for mathematics, which is 82%, whereas the NB classifier provided the highest value for Indonesian language subjects at 63%. Both results are trained from text and questions.

Previous research has shown that ML can classify questions based on revised Bloom's taxonomy. In addition, feature selection plays an important role in improving classification performance. There is also research that uses questions in English and other languages from different domains to show that we can use ML to categorize questions in a variety of languages and subjects.

2.3 Deep Learning in Questions Classification

Word embedding is an essential approach for converting text to vectors before training deep learning algorithms. Many studies in the past have utilized various models to construct word embedding. [14] classified questions in English that were translated from Turkish using the Word2vec technique with a CBOW and skip-gram models. The dataset collects questions from numerous categories, such as animal, creative, food, and many more. They analyze models using CNN, LSTM, and a combination of CNN-LSTM and CNN-SVM classifiers to compare outcomes. The accuracy of the test data is quite high, with CNN and skip-gram, a type of Word2vec model, achieving 94%. [15] sought to identify an appropriate word embedding technique for five datasets to be pre-trained for contextual and non-contextual data. For non-contextual word embedding, Fast Text has an accuracy of 0.82, whereas for contextual RoBERTa, it has an accuracy of 0.85. Up to six techniques were compared. Once the appropriate word embedding technique was applied to the CNN classifier, the accuracy was found to be 86%, which was higher than previous studies in which the word embedding was in the dataset.

Some studies employ feature extraction for DL classification [16] employed the TF-TDF approach to extract features. They categorized 141 questions based on Bloom's taxonomy from a variety of domains and compared two classifiers, the Artificial Neural Network dataset (ANN) and SVM. The experimental findings showed that using TF-TDF to extract features enhanced classification performance, with the ANN classifier having the best performance of 85.2%. [17] applied the ETFPOS-IDF approach for examination question categorization, which is an improved version of the TFPOS-IDF proposed by [11] Three datasets were collected from a variety of fields, including computing, social science, business, and others. Three classifiers were used: SVM, ANN, and Random Forest. The results demonstrate that ETFPOS-IDF outperforms all previous studies' schemes in test question categorization, achieving 0.749 in accuracy and 0.746 in F1 score.

In addition, BERT (Bidirectional Encoder Representations from Transformers), a transformer-based AI deep learning technique, is also utilized in text categorization. [18] used BERT to classify computer education questions from the Canterbury Question Bank using Bloom's taxonomy. In the experiment, it was found that the dataset containing certain classes had an imbalance issue. As a consequence, the Application, Synthesis, and Evaluation classes must be eliminated to resolve class imbalance concerns. In Experiment 3, reducing the number of classes to be classified to the remaining three based on Bloom's taxonomy level produced the highest accuracy at 82.61%.

DL algorithms such as LSTM, CNN, and RNN are increasingly being utilized to categorize questions. Word2vec, ETFPOS-IDF, and other approaches have been developed and deployed. Despite the quantity of such research, the vast majority of data sets are in English or other non-Thai languages. To gain a suitable model, we must also investigate the work in Thai.

2.4 Thai Text Classification

ML and DL have been applied to classify Thai text in the education field based on revised Bloom's taxonomy, both in classifying questions and content written in blogs. [19] used multiple classifiers, such as NB, decision tree, SVM, and multilayer perceptron, to classify the revised Bloom's taxonomy-based questions in Thai. Datasets were collected from of several websites in Bloom's cognitive domain literature. They focus on feature selection. Cleaning data, word segmentation, part-of-speech tagging, and feature selection are all applied to each question. This experiment shows that verbs, adverb, adjectives, conjunctions, and question tags should be selected as features in Thai's exam classification. The highest accuracy of 71.2% with the SVM classifier. [20] desired to classify blogs according to information and communication technology (ICT) using the revised Bloom's taxonomy. Three classes are used to classify simple, moderate, and difficult blog content. Multiple classifiers, DL and ML, are utilized. The training data consists only of textual content. Utilize a dictionary and name entity recognition (NER) to tokenize words. The training text frequently contained English computer terminology. So, they preprocess in both languages. Deep neural networks (DNN) were found to have the highest classification performance of all classifiers in this experiment, with an accuracy of 87%.

In the educational domain, there are few Thai text classifications. We attempted to investigate more text processing in other domains, such as sentiment analysis. [21] classified Thai children's tales consisting of 1,115 sentences divided into three sentiments—negative, neutral, and positive—using combining BiLSTM and CNN classifiers with several combinations of the features, including POS tagging, semantic, and word embedding. The experiment result shows that the combination of POS tagging, word embedding, and semantic features provided the highest classification accuracy at 78.89%. [22] proposed a social media sentiment analysis model.

They collected a dataset from the Wongnai food platform website, which had 2000 positive and negative restaurant reviews. They employ the longest word pattern method for word segmentation and the Word2Vec method for word embedding. Gated Recurrent Unit (GRU), CNN, and LSTM classifiers were used to compare results. The results demonstrated that LSTM was more effective at classification than CNN and GRU, with an accuracy of 84%.

In recent years, there has been an increase in the amount of research that uses BiLSTM to categorize Thai text. [23] employed a BiLSTM classifier to analyze the text categorization of comments on Thai tourism-related YouTube channels. BiLSTM can narrow down the 43 remaining tourist categories to 33. The top four categories for places of interest are temples, food, coffee, and green tea, in that order. Performance of the proposed text classification at 85.78%. [24] used a BiLSTM classifier to detect clickbait. The Thai clickbait corpus consists of 30000 headlines collected from trendy websites shared on online social media and non-clickbait from newspapers, community blogs, and online magazines. DL models are used to compare outcomes. However, BiLSTM with word-level embedding performs the highest, with a 98% accuracy rate.

We found that the research on question classification in Thai using the revised Bloom's taxonomy is very limited. Therefore, we wish to develop a question classification model to be useful to Thai educators.

3. Proposed Method

This section describes the approach to preprocessing data, training data to build models, and evaluating experimental results. Figure 2 provides an overview of the modeling for the classification of Thai scientific questions based on learning level.

3.1 Dataset

This study, dataset contains the Thai questions in a science course. At the level of secondary education, grades 7-12, 1246 questions from the database of the Office of the Basic Education Commission (OBEC), Ministry of Education, have already been labeled for all six cognitive levels by educators. The majority of the questions are multiple-choice. The questions were separated into two categories. The first three levels of the revised Bloom's Taxonomy assessed basic thinking skills (lower-order thinking skills), while the top three levels assessed advanced thinking skills (higher-order thinking skills). Questions are distributed by their level labels. Table 1 demonstrates that the questions have been divided into two classes, so there is no class imbalance issue.

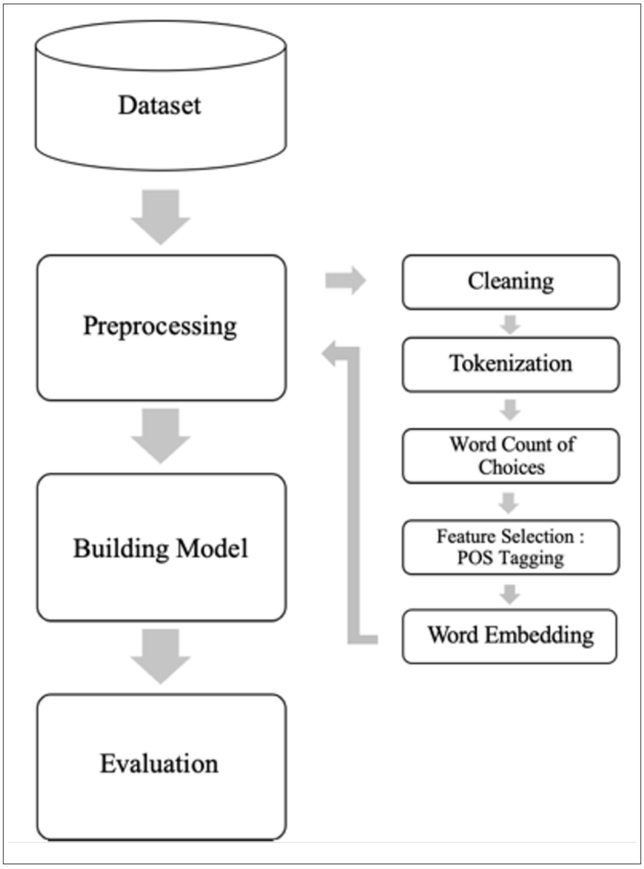


Figure 2. Overview of the proposed scientific question classification model.

3.2 Preprocessing

3.2.1 Cleaning and Tokenization

Based on the Natural Language Processing (NLP) tasks, we must preprocess the text data. Then, each question was cleaned by removing English letters and punctuation. We use the Newmm algorithm to tokenize words. The reasons are explained in the Experiment 1 section. We also compare the word sizes with and without removing stop words. In Table 2, it can be seen that if the stop word isn't removed, the average number of words per question is doubled, with 8 words per question for removing stop words and 14-15 words per question for non-removing stop words. However, based on the results of experiment 1, we have decided not to remove the stop words in this study because their performance is similar.

Table 1. *Class distribution of the dataset.*

Class	Cognitive Level	Question	Total
Basic Thinking Skills	Remembering	28	666
	Understanding	539	
	Applying	99	
Advanced Thinking Skills	Analyzing	573	580
	Evaluating	3	
	Creating	4	
Total			1,246

Table 2. *The size of words in the dataset.*

		Word Size			
Class	Question	Stop word removal	Avg.	Without stop word removal	Avg.
Basic Thinking Skills	666	5,405	8.07	9,666	14.51
Advanced Thinking Skills	580	4,990	8.60	9,100	15.69
Total	1,246	10,395	8.32	18,766	15.06

3.2.2 Word Embeddings

Word embedding, which represents a "word" as a "number", is one method to create machine-understandable features from words. The format of these numbers is a vector [25]. There are numerous word embedding models, including Word2vec, GloVe, FastText, and more. Each model employs a unique algorithm for vector generation.

Word embedding is an approach with which we represent documents and words before training. This study will utilize Thai2Vec from PythaiNLP, which has a function to generate vectors of words trained using the Word2vec family of techniques. The vocabulary was used to generate a 51,556-word vector from 300 dimensions [26], which will convert questions and multiple-choice answers into vectors for further training. In addition, to prevent the model from memorizing training data for overfitting, we set the vector length to 180 and scaled the dimension to the magnitude of the vector using the Principal Component Analysis (PCA) method [27].

3.3 Question Classification Algorithms

There are three main question classification approaches, including rule-based, machine-learning-based, and hybrid-based approaches [28]. We focus on machine learning approaches, including traditional methods and deep learning methods. In this paper, we experiment by using two standard algorithms of traditional methods, including Naïve Bayes and Support Vector Machines, and one approach of deep learning model, including Bidirectional Long Short-Term Memory.

3.3.1 Bidirectional Long-Short-Term Memory

BiLSTM is bidirectional long-short-term memory. The LSTM is a sort of recurrent neural network in which the previous step's RNN output is supplied as input to the next phase. It was invented by Hochreiter & Schmidhuber [29]. It solves the issue of long-term dependency on RNN, which cannot anticipate words stored in long-term memory. The LSTM can hold data for a long time and may be used for time series data processing, prediction, and classification. It has a chain structure made up of a neural network and multiple memory units known as cells. BiLSTM is a bidirectional LSTM, which means that the signal is propagated backward and forward. Figure 3 describes the BiLSTM layer's architecture, where X_i is the input token, Y_i is the output token, and A and A' are LSTM nodes. The combination of A and A' LSTM nodes is the final output of Y_i.

The BiLSTM architecture consists of two unidirectional LSTMs that process the sequence in both forward and backward directions. This architecture may be viewed as having two independent LSTM networks, one receiving the token sequence as it is and the other receiving it in reverse order. Both of these LSTM networks provide a probability vector as output, and the result is the sum of these probabilities [30]. It can be written as equation 1.

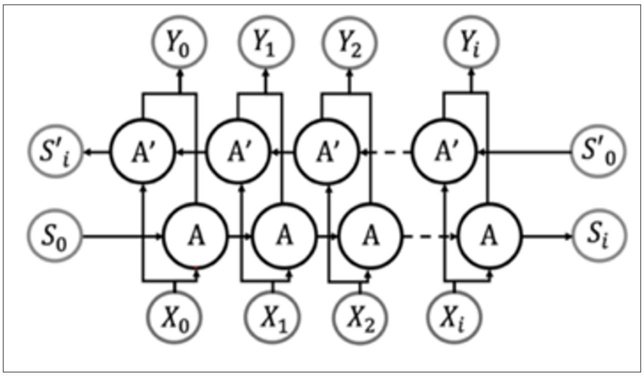


Figure 3. *BiLSTM layer architecture.*

$$p_t \quad = \quad p^f_t + p^b_t \qquad (1)$$

where p_t is the final probability vector of the network, p^f_t is the probability vector from the forward LSTM network, and p^b_t is the probability vector from the backward LSTM network.

3.3.2 Naïve Bayes

Naïve Bayes (NB) is a probabilistic classifier. This approach applied Bayes' theorem, which was developed by Thomas Bayes, an English mathematician [26]. This algorithm makes new assumptions that differ from Bayes' theorem. As a result, the word naive is utilized.

Bayes' theorem finds the probability of event A occurring when event B has already occurred. which can be found in the equation 2.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \qquad (2)$$

where P*(A)* is the probability of event A, and $P(B \mid A)$ is the probability of event B occurring when event A is known to have already occurred.

NB identifies a class by assuming that each feature's occurrence is independent, resulting in equation 3.

$$P(class|a_{1,}a_{2,}, \ldots, a_n) = \frac{P(class)P(a_1,a_2,\ldots,a_n|class)}{P(a_1,a_{2,},\ldots,a_n)} \qquad (3)$$

where a_i is the value of any feature that appears in the sample to be categorized, and P(a_1, a_2, ..., a_n) is the probability of occurrence of a feature with value a_1 followed by a feature with value a_2 until the occurrence of features with value a_n.

3.3.3 Support Vector Machines

Another classic technique used in this study is Support Vector Machines. SVM was first heard in COLT-92 in 1992, when Boser, Guyon, and Vapnik introduced it [31]. This method of supervised learning is utilized for classification and regression. It is a classifier that aims to divide the instants into the training data set by finding linear equations that separate the data in each class to achieve high-performance values (optimal hyperplane). From Figure 4, a hyperplane must have the longest distance and width determined from the hyperplane to a class instance (maximum margin).

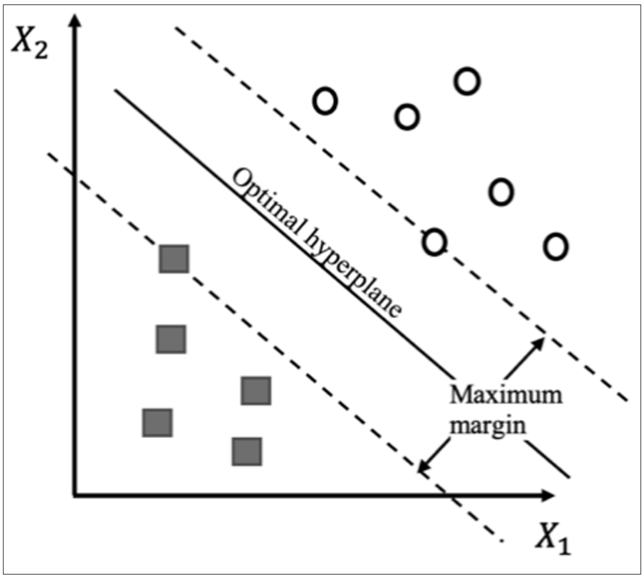


Figure 4. *SVM algorithm for learning on training data with two features.*

SVM attempts to find the support vector at the closest data points from each of the classes [31], which leads to finding the optimal hyperplane that can then be defined in Equation 4.

$$W.X + b = 0 \qquad (4)$$

where w represents the weight vector, x is the input feature vector, and b represents the bias. For all components of the training set, w and bwould meet inequalities 5 and 6, respectively:

$$w.x_i + b \geq 1 \, ; \, if \; Y_{i=} \; +1 \qquad (5)$$

$$w.x_i + b \leq 1 \, ; \, if \; Y_{i=} \; -1. \qquad (6)$$

where $Y_i \in \{-1,1\}$ In the case in Figture. 5, the support vectors are H_1 and H_2 When the support vector is identified, Equation 7 can be used to get the value that maximizes the margin.

$$Margin = \frac{1}{\|w\|^2} \qquad (7)$$

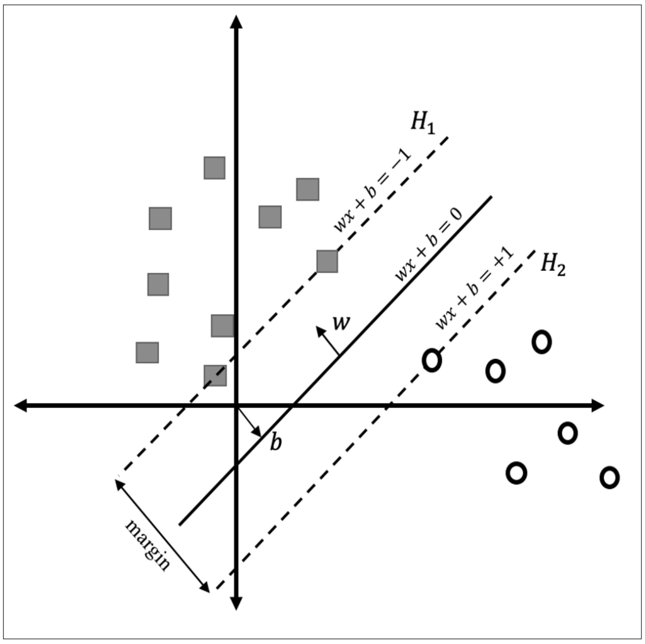
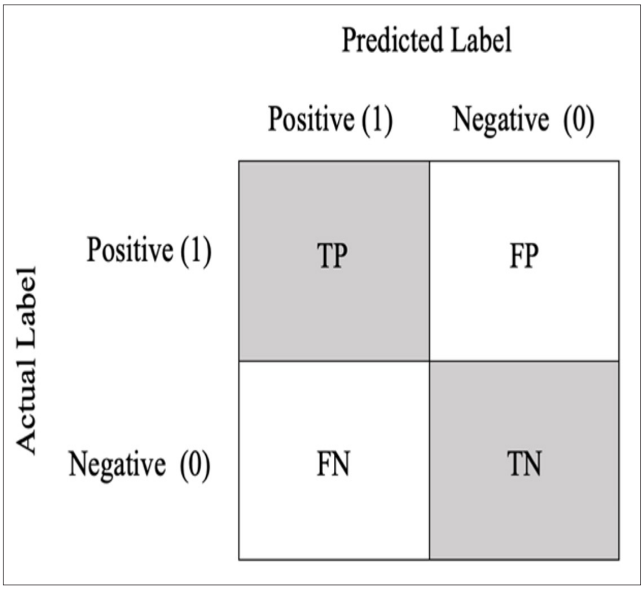


Figure 5. *Support vector.*

3.4 Model Evaluation

The model evaluation utilizes the K-Fold method (K = 10), which is more standardized than the Train-Test split because the model is constructed and evaluated 10 times, mitigating the issue of overfitting and producing accurate results. Precision, Recall, F1-Score, and Accuracy are the most commonly used evaluation metrics for text classification models to assess classification performance. True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are used to calculate these metrics. This combination of four numerals generates the confusion matrix depicted in Figure 6.



		Predicted Label	
		Positive (1)	Negative (0)
Actual Label	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 6. *Confusion matrix.*

Precision is the proportion of correctly predicted values for a particular class relative to the total number of predicted values for that class.

$$Precision \;=\; \frac{TP}{TP+FP} \qquad (8)$$

Recall is the ratio of all predicted values for a particular class to the actual values for that class.

$$Recall \;=\; \frac{TP}{TP+FN} \qquad (9)$$

F1-Score achieves a balance between precision and recall. The ratio of precision to recall is optimal.

$$F1 \text{ - } Score \;=\; \frac{2xPrecision \; x \; Recall}{Precision+Recall} \qquad (10)$$

Accuracy is a ratio of instances appropriately classified in relation to all values.

$$Accuracy \;=\; \frac{(TP+TN)}{TP+FP+TN+FN} \qquad (11)$$

3.5 Experimental Results

3.5.1 Experiment 1

We consider Thai word tokenization algorithms, Newmm [32], Longest Matching (LM) [33], and Deepcut [34], to look at their potential because Thai has no word spaces. The result of word splitting is significant because it affects the meaning of words. As a result, we must ensure that the appropriate algorithm is selected. Furthermore, we compare performance with and without stop words to determine whether stop words are required.

We employ both machine learning and deep learning algorithms (NB, SVM, and BiLSTM) to train data. Table 3 displays the results of the three algorithms. The highest accuracy with the BiLSTM algorithm is 55% for Newmm and Deepcut. However, we chose the Newmm approach for data splitting because it splits words faster than Deepcut. When comparing the results of removing or not removing stop words, it appears that not removing stop words has a bit higher results in BiLSTM and NB, thus we decided not to delete stop words in this study.

Table 3. *Performance of the NB, SVM, and BiLSTM models using different preprocessing techniques.*

Classifier	Types of dataset	Accuracy		
		Newmm	LM	Deepcut
BiLSTM	Stop word removal	0.53	0.53	0.54
	Without stop word removal	**0.55**	0.53	0.55
NB	Stop word removal	0.51	0.51	0.49
	Without stop word removal	0.52	0.51	0.51
SVM	Stop word removal	0.55	0.55	0.54
	Without stop word removal	0.53	0.54	0.54

3.5.2 Experiment 2

The objective of this experiment

1. To identify a feature set that can be used to train a model. We consider all three parts: 1. contexts in the question 2. multiple choices 3. length of choices.

2. To compare the performance of the model with different feature sets to improve the model's classification accuracy.

Part 1. Contexts in the question

In previous experiments, only questions were used for training. A question also contains other contexts, such as situations that are textual in addition to the question, images, and tables; however, we didn't train images and tables because we only want to learn text. So, context refers to situations only. We counted the frequency of contexts in each class, which is shown in Table 4, and found that context is not useful for categorizing classes because both classes have almost equal contexts. In the basic thinking skills class question, 64.40% of the contexts were found, which is similar to 57.71% in the advanced thinking skills class. Therefore, contexts will not be trained.

Table 4. *Percentage of contexts appearing in each class's questions.*

Class	Contexts	
	Yes	No
Basic Thinking Skills	64.40%	35.60%
Advanced Thinking Skills	57.71%	42.29%

Part 2. Multiple choices

Each question has multiple choices that may help in question classification because the words in the choices help make the question more complete and clearer about what learning level is needed. Table 5 shows that the question is classified as advanced thinking skills, but it only asks yes or no. However, the objective of the question is not only wants a yes or no answer but also the correct reasons, which are shown in the multiple choice. That is why the question is in the advanced thinking skills class. As a result, we will select multiple choices as one of the features to evaluate the model's performance.

Table 5. Example of a question that needs Advanced Thinking Skills.

Context	ภารโรงโรงเรียนแห่งหนึ่ง ตัดกิ่งไม้เพื่อให้ต้นไม้แตกกิ่ง โดยเมื่อตัดกิ่งไม้ที่มีใบแล้วนำกิ่งไม้วางไว้ในที่ร่มตอนกลางวัน ซึ่งกิ่งไม้นั้นยังสดอยู่ [A school janitor prunes the tree to make the tree branch. After cutting he places them in the shade during the day when the branch is still fresh.]
Question	จากข้อมูล ยังคงมีการสังเคราะห์แสงในใบไม้ หรือไม่ [From the given scenario, does the photosynthesis continue in the leaves?]
Multiple-choice	1. มี เพราะไมโทคอนเดรียเกิดปฏิกิริยาภายในเซลล์ [1. Yes, because there is reaction in mitochondria] 2. ไม่มี เพราะไซโทพลาสซึมไม่เกิดปฏิกิริยาภายในเซลล์ [2. No, because no reaction found in cytoplasm] 3. มี เพราะคลอโรพลาสต์ยังทำงานจึงเกิดกิจกรรมภายในเซลล์ [3. Yes, because chloroplasts are still active. Therefore, there is activity inside the cell.] 4. ไม่มี เพราะแวคิวโอลไม่ทำงานจึงไม่เกิดกิจกรรมภายในเซลล์ [4. No, because vacuoles do not work. So, no activity inside the cell]

Part 3. Length of choices

We also observed that the length of multiple choice can help in classifying classes. Table 6 compares the average word counts of choices and questions in the two classes. In terms of question length, the mean word count for both classes is similar. But in terms of the average word count of choices, the advanced thinking skills class has a higher average word count than the basic thinking skills class, at a difference of 16.83 words. Therefore, we will also train the word count of choices as another feature.

Table 6. Average word count of choices and questions.

Class	Average word count	
	Choice	Question
Basic Thinking Skills	20.74	14.23
Advanced Thinking Skills	37.57	15.47

Due to this experiment, we need to use the choices feature, but some of the questions in the dataset contain no choices. Consequently, it is necessary to eliminate those questions. Thus, the dataset was reduced by approximately 20%, leaving a total of 992 questions for training in Table 7.

Then, questions and multiple choices were preprocessed. After tokenizing the words in the choices, we receive the length of the choices in each question and then normalize those values, which is another feature that will be utilized for training.

Table 7. The class distribution of the dataset after eliminating questions that have no choices.

Class	Question
Basic Thinking Skills	486
Advanced Thinking Skills	506
Total	992

We use three classifiers for training: BiLSTM, NB, and SVM. We built three models to compare results using different feature sets. Model 1 employs questions with choices; Model 2 employs questions and the length of choices; and Model 3 combines questions, choices, and the length of the choices. The results indicated that all three models improved by approximately 20% from experiment 1. The BiLSTM classifier performed better than the other two classifiers, particularly in model 3. As shown in Table 8, the highest accuracy in this experiment was 0.70. This shows that multiple choices and the length of choices contribute to the model's ability to classify more accurately.

Table 8. *Performance of the model using different feature sets.*

Feature Set	Classifer	Accuracy	Precision	Recall	F1
Questions + Choices	BiLSTM	0.68	0.68	0.68	0.68
	NB	0.65	0.70	0.78	0.66
	SVM	0.59	0.75	0.57	0.49
Questions + Length of choices	BiLSTM	0.67	0.68	0.68	0.67
	NB	0.54	0.67	0.17	0.26
	SVM	0.65	0.77	0.52	0.53
Questions + Choices + Length of Choices	**BiLSTM**	**0.70**	**0.70**	**0.70**	**0.70**
	NB	0.61	0.76	0.41	0.39
	SVM	0.59	0.74	0.56	0.51

3.5.3 Experiment 3

We investigate the impact of POS tagging on feature selection to improve the model's performance in categorizing questions. The addition of multiple choice as a feature increased word size by 39% to 30,769 words. As a result, we wish to eliminate terms that do not affect question classification and select group words that help the classifier categorize questions more accurately.

POS tagging assigns a label for each word with related grammatical elements such as nouns, verbs, adjectives, or adverbs. Previous studies used POS tagging for selecting features to train the model. They found that it helps to improve accuracy [12], [19].

We choose the ORCHID POS tags set , in which the Thai text corpus is over 400,000 words and all word types are labeled, for tagging each word in the dataset. Then, the words in each tag are counted and ranked. The results show that the most POS tagging is NCMN, according to VACT and VSTA, as shown in Table 9.

We create five groups for the POS tag combinations to select features for model training. For example, Group 1 consists of words tagged POS: verb, noun, adposition, and adjective; Group 2 consists of words tagged POS: verb, noun, adposition, adjective, and determiner. Each group chooses a tag from the ORCHID POS tags, as shown in Table 10. The BiLSTM classifier is used for training with different POS tag groups to compare efficiency.

Table 9. *Top 10 Occurrences of POS tagging.*

Rank	Part-of-Speech tag	Abbreviation	Word (%)
1	Common noun	NCMN	62.84
2	Active verb	VACT	11.86
3	Stative verb	VSTA	3.93
4	Determiner, cardinal number expression	DCNM	2.59
5	Attributive verb	VATT	2.52
6	Cardinal number	NCNM	2.44
7	Adverb with normal form	ADVN	1.89
8	Unit classifier	CNIT	1.89
9	Preposition	RPRE	1.84
10	Subordinating conjunction	JSBR	1.41

Table 9. *Top 10 Occurrences of POS tagging.*

Part of the	ORCHID	Group				
Speech tag	POS tags	1	2	3	4	5
Verb	VACT, VSTA	✓	✓	✓	✓	✓
Noun	NCMN, CMTR, CNIT	✓	✓		✓	✓
Adposition	RPRE	✓	✓			✓
Adjective	VATT	✓	✓	✓	✓	✓
Determiner	DDAC, DDAN, DIBQ		✓			
Adverb	ADVN			✓	✓	✓
Subordinating Conjunction	JSBR			✓		

[1] https://pythainlp.github.io/docs/2.3/api/tag.html

The results of POS tag combinations show that group 4, which uses nouns, verbs, adjectives, and adverbs, had the highest performance. As illustrated in Figure 7, the highest was 80.24%. Type 1 has an accuracy of 79.84%, indicating that prepositions are not required as features. For groups 2 and 5, even though more POS tags were chosen, the classification remained lower than group 4. Group 3 demonstrates the importance of nouns as a learning feature because no nouns were chosen, resulting in a lower categorization efficiency than other groups.

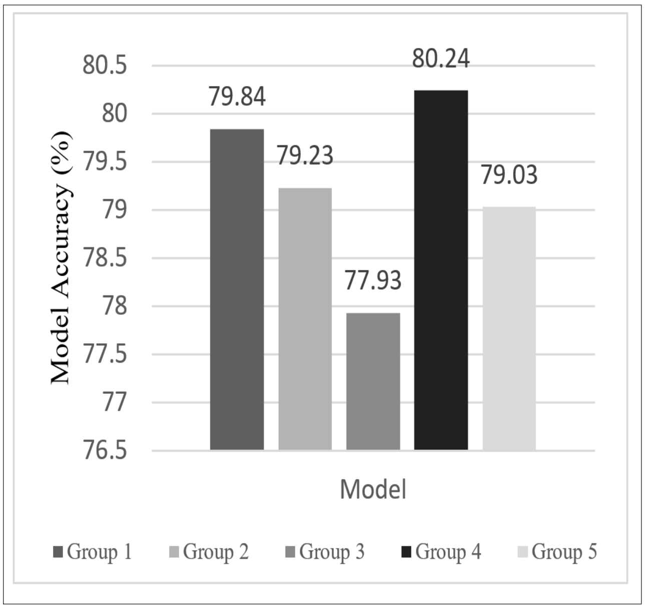


Figure 7. *Performance of feature selection from POS tag combinations in each group.*

As a result, the POS tags chosen as features will be listed in Table 11. The vocabulary has been reduced by 17% from the total number of terms, leaving 3,293 words out of 3,972.

Table 11. *The best group of POS tags as a feature for model training.*

Type	Part of the Speech tag
Nouns	NCMN
Verbs	VACT, VSTA
Adjectives	VATT
Adverbs	ADVN

Table 12 shows the experimental result when compared to the NB and SVM classifiers. The classification performance of the NB and SVM classifiers was similar to that of experiment 2, however, the accuracy of the BiLSTM classifier increased to 80%. This demonstrates that feature selection from POS tagging had an influence on the BiLSTM classifier and improved the model's performance.

Table 12. *The accuracy of classifiers following feature selection with POS tagging includes nouns, verbs, adjectives, and adverbs.*

Classifier	Accuracy	Precision	Recall	F1
Proposed BiLSTM model	**0.80**	**0.81**	**0.80**	**0.80**
NB	0.65	0.77	0.50	0.58
SVM	0.65	0.70	0.65	0.65

3.6 Experimental setting

Experiment 3 provides a suitable classification model for the questions using the BiLSTM classifier. Figture 8 depicts the model's architecture. We experiment several times to select a value for the hyperparameter until a suitable value is found, as shown in Table 13.

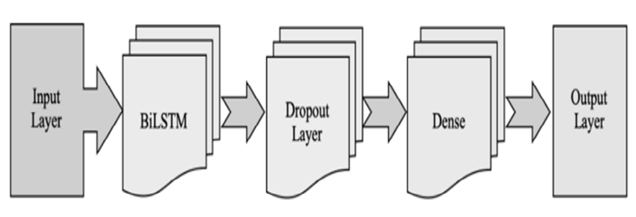


Figure 8. *Architecture of the proposed BiLSTM model.*

Table 13. *The hyperparameters tuned in this research.*

Hyperparameter	Optimal value
Embedding	180
The first Bidirectional-LSTM	10
The second Bidirectional-LSTM	10
epoch	50
Batch size	5
Optimizer	Adam
Activation in a dense layer	Softmax
Rate of the dropout layer	0.5
Learning rate	0.001
Dense output	2

4. Results and Discussion

In this paper, we aim to present a model for efficiently classifying Thai science questions. The results of the three experiments are shown in tables 3, 8, and 12, respectively. We improved the accuracy of the model by adding feature word selection to experiment 3, thereby optimizing the model. Using POS tagging to select features, Word2vec to convert words into vectors, and BiLSTM to classify resulted in the highest accuracy of 0.80, while precision, recall, and F1-scores were 0.81, 0.80, and 0.80, respectively, as shown in Table 12. Table 14 displays the values separated by class.

Table 14. *Class-by-class analysis of the proposed BiLSTM model.*

Class	Precision	Recall	F1-score
Advanced Thinking Skills	0.84	0.75	0.79
Basic Thinking Skills	0.77	0.85	0.81

The confusion matrix of the model presented in this study is depicted in Figture 9. To evaluate the K-fold assessment method, the model selects questions at random so that the class classification accuracy in each fold can be determined and then averaged across all folds (10 folds). The dataset contains 992 questions; there are two folds with 100 questions and eight folds with 99 questions for the test set. The model's classification findings revealed that in the basic thinking skills class, the model correctly classified 85%, while in the advanced thinking skills class, the model correctly classified 75% of the questions from all folds on average.

We test the model with an unseen test set of 36 questions: 27 in the Advanced Thinking Skills class and 9 in the Basic Thinking Skills class. These questions were provided by OBEC. However, the model has never been evaluated with this dataset before. When the model was employed to test question classification, Table 15 and the confusion matrix in Fig. 10 reveal that 30 questions were properly identified, with an accuracy of 0.83. The model correctly classified 24 questions from the Advanced Thinking Skills classification, resulting in 89%, and 6 questions from the Basic Thinking Skills classification, representing 67%.

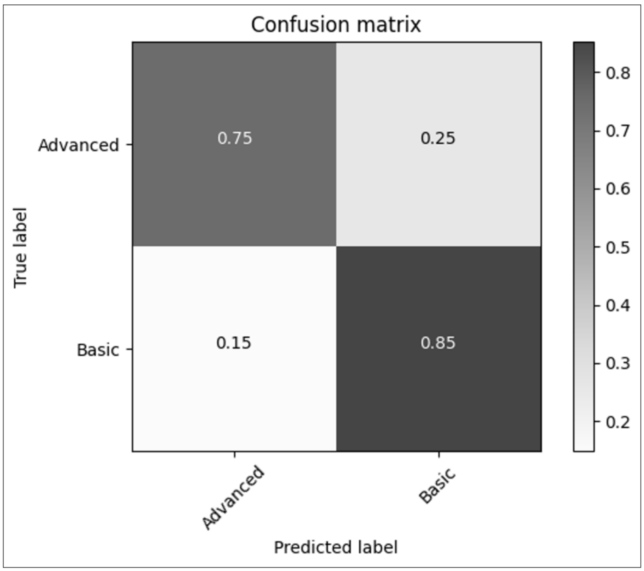


Figure 9. *Confusion matrix of the proposed model for data testing in the K-fold method.*

Table 15. *The proposed model's performance for unseen data testing.*

Class	Accuracy	Precision	Recall	F1-score
Advanced Thinking Skills	0.83	0.89	0.89	0.89
Basic Thinking Skills		0.67	0.67	0.67

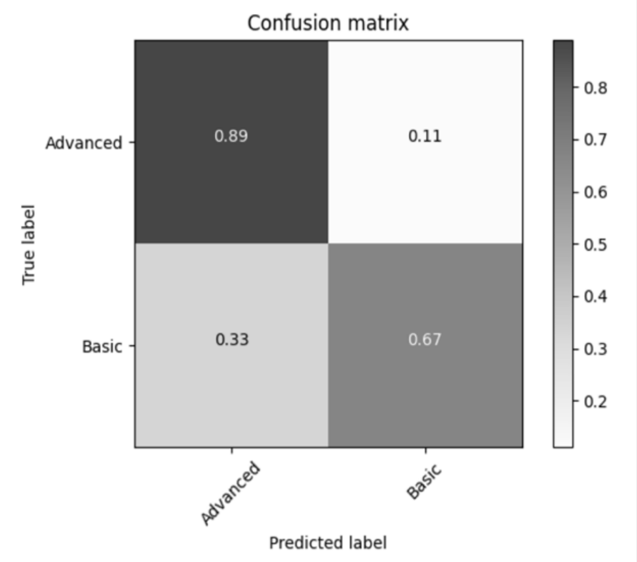


Figure 10. *Confusion matrix of the proposed model for unseen data testing.*

5. Conclusion

This research was conducted to develop an efficient model for classifying secondary school science questions into basic and advanced thinking skills based on revised Bloom's taxonomy. We achieved an accuracy of 80% for question categorization by improving the feature set and applying techniques through 3 experiments. The performance of the model was evaluated by an accuracy metric using 10-fold cross-validation. The results demonstrated that multiple choices influence the model's training. When the choices and the length of words in the choices were trained alongside the questions, the model learned more effectively, and the selection of words in the choice also affected the model's learning. In this study, the POS tagging method was used to tag POS terms such as nouns, verbs, adjectives, and adverbs, resulting in improved model performance. Previous research also used POS tagging for feature selection [9], [12], [19] showing that feature selection can increase classification performance [8], [9]. In addition, the BiLSTM algorithm was found to have better classification efficiency compared to Naïve Bayes and SVMs, resulting in an optimal model using BiLSTM to classify questions. This is consistent with previous work that used BiLSTM to classify Thai text [24], [27] showing the BiLSTM algorithm's high efficiency in classifying Thai text.

In the future, it may be possible to increase the size of the larger dataset so that the model can be classified as closely to reality and standardized as possible, as well as incorporate additional deep learning techniques for additional experiments, such as combining CNN with BiLSTM to produce a more

6. Rererences

[1] World Economic Forum, *Future of Jobs Report 2023.* Available Online at https://www.weforum.org/reports/the-future-of-jobs-report-2023, accessed on 30 September 2023.

[2] L.O. Wilson. "Anderson and Krathwohl–Bloom’s taxonomy revised." *Understanding the New Version of Bloom's Taxonomy*, 2016.

[3] J. Irvine. "A Comparison of Revised Bloom and Marzano's New Taxonomy of Learning." *Research in Higher Education Journal,* Vol. 33, 2017.

[4] K. Changwong, A. Sukkamart, and B. Sisan. "Critical thinking skill development: Analysis of a new learning management model for Thai high schools." *Journal of International Studies,* Vol. 11, No. 2, pp. 37-48, 2018.

[5] S. K. Patil and M. M. Shreyas. "A Comparative Study of Question Bank Classification based on Revised Bloom's Taxonomy using SVM and K-NN." *2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT),* Tumakuru, India, pp. 1-7, 2017.

[6] S. Shaikh, S. M. Daudpotta, and A. S. Imran. "Bloom’s Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings." *IEEE Access,* Vol. 9, pp. 117887-117909, 2021.

[7] M. Forehand. "Bloom's taxonomy." *Emerging perspectives on learning, teaching, and technology,* Vol. 41, No. 4, pp. 47-56, 2010.

[8] N. Ghalib and D. S. Hammad. "Classifying Exam Questions Based on Bloom's Taxonomy Using Machine Learning Approach." *Technologies for the Development of Information Systems (TRIS-2019),* pp. 260-269, 2020.

[9] J. Chandra and B. Thomas. "The Effect of Bloom’s Taxonomy on Random Forest Classifier for cognitive level identification of E-content." *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE),* pp. 1-6, 2020.

[10] M. Mohammed and N. Omar. "Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF." *International Journal on Advanced Science, Engineering and Information Technology,* Vol. 8, No. 4-2, pp. 1679-1685, 2018.

[11] M. Mohammed and N. Omar. "Question classification based on Bloom’s taxonomy cognitive domain using modified TF-IDF and word2vec." *PLOS ONE,* Vol. 15, No. 2, 2020.

[12] J. Huang, Z. Zhang, J. Qiu, L. Peng, D. Liu, P. Han, and K. Luo. "Automatic Classroom Question Classification Based on Bloom's Taxonomy." *Proceedings of the 13th International Conference on Education Technology and Computers,* pp. 33-39, 2022.

[13] Hasmawati, A. Romadhony, and R. Abdurohman. "Primary and High School Question Classification based on Bloom's Taxonomy." *2022 10th International Conference on Information and Communication Technology (ICoICT),* pp. 234-239, 2022.

[14] S. Yilmaz and S. Toklu. "A deep learning analysis on question classification task using Word2vec representations." *Neural Computing and Applications,* Vol. 32, No. 7, pp. 2909-2928, 2020.

[15] M. O. Gani, R. K. Ayyasamy, A. Sangodiah, and Y. T. Fui. "Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique." *Education and Information Technologies,* Vol. 28, pp. 15893-15914, 2023.

[16] M. Ifham, K. Banujan, B. T. G. S. Kumara, and P. M. A. K. Wijeratne. "Automatic Classification of Questions based on Bloom's Taxonomy using Artificial Neural Network." *2022 International Conference on Decision Aid Sciences and Applications (DASA),* pp. 311-315, 2022.

[17] M. O. Gani, R. K. Ayyasamy, S. M. Alhashmi, A. Sangodiah, and Y. T.Fui. "ETFPOS-IDF: A Novel Term Weighting Scheme for Examination Question Classification Based on Bloom's Taxonomy." *IEEE Access,* Vol. 10, pp. 132777-132785, 2022.

[18] J. Zhang, C. Wong, N. Giacaman, and A. L. Reilly. "Automated Classification of Computing Education Questions using Bloom's Taxonomy." *Proceedings of the 23rd Australasian Computing Education Conference,* pp. 58-65, 2021.

[19] K. Anekboon. "Feature Selection for Bloom's Question Classification in Thai Language." *Proceedings of the 2018 Computing Conference,* Vol. 1, pp. 152-162, 2019.

[20] C. Chootong and J. Charoensuk. "Cognitive level classification on information communication technology skills for blog." *International Journal of Electrical and Computer Engineering,* Vol. 12, No. 6, pp. 6387-6396, 2022.

[21] T. S. N. Ayutthaya and K. Pasupa. "Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features." *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP),* pp. 1-6, 2018.

[22] C. Jitboonyapinit. "Development of Sentiment Analysis Model Based on Thai Social Media Using Deep Learning Techniques." *Huachiew Chalermprakiet Science and Technology Journal,* Vol. 2, 2022.

[23] S. Khruahong, O. Surinta, and S. C. Lam. "Sentiment Analysis of Local Tourism in Thailand from YouTube Comments Using BiLSTM." *Mult-disciplinary Trends in Artificial Intelligence,* pp. 169-177, 2022.

[24] P. Klairith and S. Tanachutiwat. "Thai Clickbait Detection Algorithms Using Natural Language Processing with Machine Learning Techniques." *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST),* pp. 1-4, 2018.

[25] S. S. Birunda and R. K. Devi. "A review on word embedding techniques for text classification." *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA,* pp. 267-281, 2021.

[26] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul, and P. Chormai. *PyThaiNLP: Thai Natural Language Processing in Python,* Available Online at https://pythainlp.github.io/docs/2.3/api/tag.html, accessed on 29 September 2022.

[27] T. Kurita. "Principal component analysis (PCA)." *Computer Vision: A Reference Guide,* pp. 1-4, 2019.

[28] S. Jayalakshmi and A. Sheshasaayee. "Question Classification: A Review of State-of-the-Art Algorithms and Approaches." *Indian Journal of Science and Technology,* Vol. 8, 2015.

[29] A. Chugh, *Deep Learning | Introduction to Long Short Term Memory.* Available Online at https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory, accessed on 10 August 2022.

[30] A. Taparia, *Bidirectional LSTM in NLP.* Available Online at https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/, accessed on 5 August 2022.

[31] V. Jakkula. "Tutorial on support vector machine (svm)." *School of EECS, Washington State University*, Vol. 37, No. 2.5, pp. 3, 2006.

[32] K. Jearanaitanakij, N. Kueakool, P. Limwanichsin, T. Kullawan, and C. Yongpiyakul. "LCS-based Thai Trending Keyword Extraction from Online News." *Naresuan University Engineering Journal,* Vol. 17, No. 2, pp. 54-61, 2022.

[33] P. Prakrankamanant. *Data augmentation for Thai natural language processing using different tokenization,* M.S. Thesis, Chulalongkorn University, Bangkok, Thailand, 2021.

[34] M. Jaiwai, K. Shiangjen, S. Rawangyot, S. Dangmanee, T. Kunsuree, and A. Sa-nguanthong. "Automatized educational chatbot using deep neural network." *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering,* pp. 5-8, 2021.