Automatically Classifying Sentence into Question Categories on Thai Text

Saranlita Chotirat
Information Technology Program,
Department of Information Technology, Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok,
Bangkok, Thailand
saranlita.c@email.kmutnb.ac.th

Phayung Meesad
Department of Information Technology Management,
Faculty of Information Technology and Digital Innovation,
King Mongkut's University of Technology North Bangkok,
Bangkok, Thailand
pym@kmutnb.ac.th

***Abstract*— The purpose of this study was to study question classification that automates define the category of wh-questions from Thai text. This was achieved compared efficiently classify text to a wh - question class of simple sentences and interrogative sentence trained through natural language processing (NLP) which considerate the top 5 and top 10 POS (Part of Speech Tagging) and used classification models (Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbors, and neural network which employs long short-term memory (LSTM) to classify the question categories. The experimental results showed that the average classification accuracy up to 78.00%, precision 81.76%, recall 78.00% and F1 score 79.84%, which suggests that filter words only Top5 POS can solve the problem of classifying Thai text to effectively question classification.**

Keywords—Question Classification, Sentences Classification, POS tagging, Analysis Thai Sentence

I. INTRODUCTION

Questions are asked to fulfill the informational needs. A question is a linguistic expression used to make a request for information which may be provided with an answer. In the recent years, how to seek answers from huge amounts of data has become a popular topic in information technology research. The Question Classification (QC) is one of the effective ways to solve this problem. Question Classification is a strong signal for Answer Selection [1].

In an analysis, Thai text has a which is a crucial problem caused by the following reasons: (1) it does not have boundary indicators to separate words in sentences; (2) an interrogative sentence contains the following feature word – "อะไร: a rai (what)", "ที่ไหน: thee nai (where)", "เมื่อไหร่: meuua rai (when)", "ใคร: khrai (who)", "เท่าไหร่: thao rai (how much/many)" that is located in the front or the end of the sentences; and (3) it is structural ambiguities as a part of speech depends on word order. One word may have more than one meaning. For example, the word "กำลัง: gam lang" has two meanings. It means "power" if it is a noun and it means "in the act of" if it is the preposition [2]. Thus, considering a sequence of words in sentences is significant with an analysis of Thai sentences.

In this work, we aim to compare the efficiency classification from a simple sentence to question categories that the considerate frequency of Part of Speech tagging (POS) in sentences especially simple sentences to classify wh-question category.

In Section II we present a review of some related works on question classification and part of speech tagging. In Section III, we give details for the main steps in the preparing dataset, data preprocessing, and classification. In Section IV, we present our results on the classification performance of Thai sentences. Section V gives conclusion remarks and future work.

II. RELATED WORK

A. Question classification

Question classification (QC) is an essential part of question answering (QA) systems, which can be defined to match a question to one or several classes to determine the answer type [3]. There are two groups of question classification methods. The first group is based on traditional methods. For previous studies on traditional classification of question and sentences, Support vector Machine (SVM) [4, 5] is the most used algorithm. Some researchers also used other machine learning algorithms such as Naïve Bayes[6], Logistic Regression [7] and K-Nearest Neighbors [8]. The second groups is based on emerging techniques like deep learning [9], which has become popular methods of question classification now a day.

B. Part of Speech Tagging

Part of Speech Tagging (POS) known as entity extraction, that is one of the essential elements in Natural Language Processing (NLP) task for extracting features and marks the word in a text with labels. POS identify words as nouns, verb, adverb, etc. in that context. Many researches applied POS in pre-process task for improved accuracy text classification. In 2019, Silva, Bittencourt, and Maldonado [5] surveyed studies directly and specifically involved in question classification. They found almost 90 percent (88.75 percent) used some extraction/selection mechanism on automatic question classifiers. Pre-processing method can improves the precision and the accuracy of the classification rises to 87.6% when they keep only nouns, verbs, and adjectives; while when all words are kept, the precision decreases [10].

Based on the problem statement above about question classification, in this research we proposed a methodology for automatically sentence classifying of question classification. Only inputs Thai sentences are considered in this research. The main technique uses natural language processing for pre-processing considers with POS, classification model to the wh-question category, and compares accuracy classify when we keep some words, which are filtered by POS.

III. METHODOLOGY

In order to implement the model, we used Python 3.7.0. The DL model used Keras API. The experiments were carried on a laptop machine with the following specifications: Intel (R) core (TM) i7-8750H CPU with 2.20 GHz, 16GB

RAM with Windows 10, 64 bit operating system. A question classification of Thai text consists of pre-processing, classification, and evaluation stages that is detailed as follows.

A. Preparing dataset

In this study, two different datasets have been prepared. The first dataset is a simple sentence in Thai text amount of 1,000 sentences from Thai Wikipedia. The second dataset is an interrogative sentence, which was generated from the first dataset amount of 1,000 sentences. We manually categorized into 5 wh-question categories (Who, What, Where, When, and How) and labeled. The annotation statistics were the following: Who (425 sentences), What (231 sentences), Where (226 sentences), When (74 sentences), and How (44 sentences). We randomly split the dataset with 80:20 ratio, that is 800 training sentences and 200 testing sentences both simple sentences and interrogative sentences.

Some samples of data

Simple Sentence: กรุงเทพมหานครเป็นเมืองหลวงของประเทศไทย :
groong thaehp ma haa na khaawn bpen meuuang luaang khaawng bpra thaeht thai (Bangkok is the capital city of Thailand.) → Where

Interrogative sentence: เมืองหลวงของประเทศไทยคือที่ไหน :
meuuang luaang khaawng bpra thaeht thai kheuu thee nai (Where is the capital city of Thailand?) → Where

B. Pre-processing

Pre-processing consists of three main steps:

- Tokenization: Thai text does not have boundary indicators to separate words, not same as English. Our word tokenization method is a dictionary-based word segmentation combined with DeepCut model [11].
- Considered words: We keep the frequency of each word and part of speech tagging that found in the corpus which applied a NLP toolkit for Thai (PyThaiNLP) [12].

  Pre-process the sentences to difference 3 datasets including (1) all words are kept, (2) some words are filtered by POS, only words in Top5 POS ranked are kept and deleted other words, and (3) some words are filtered by POS, only words in Top10 POS ranked are kept and deleted other words. The top 10 POS ranked by the frequency that is the most frequently occurs in the Interrogative sentence's dataset.

Table I. Top10 Part of Speech Tagging (POS) Ranked

Rank	Part of Speech	Examples
1	NCMN (Common noun)	book, building
2	PUNC(Punctuation)	(,), “, ‘, :
3	RPRE (Preposition)	under, on, in
4	VSTA (Stative verb)	see, know
5	VACT (Active verb)	walk, run
6	PNTR (Interrogative pronoun)	who, where, what
7	NPRP (Proper noun)	iPhone, Pepsi
8	JSBR (Subordinating conjunction)	if, because
9	PREL (Relative pronoun)	which, that
10	VATT (Attributive verb)	good, beautiful

Table II. Amount Token in sentence

	Sentence	average
All words	Interrogative sentence	27.15
	Simple sentence	39.53
Top5 POS	Interrogative sentence	20.23
	Simple sentence	34.93
Top10 POS	Interrogative sentence	24.62
	Simple sentence	36.71

- Transform the text into a vector; We use the TF-IDF algorithm to extract the test text features and transform the test text into a vector. For neural network have 3 steps in pre-process phases: tokenization, mapping dictionary, and zero padding.

For Example (all words)

Tokenization: [“กรุงเทพมหานคร”, “เป็น”, “เมืองหลวง”, “ของ”, “ประเทศไทย”]

Mapping dictionary: [2012, 1134, 491, 1401, 1241]

Zero padding: [0, 0, 0, …, 2012, 1134, 491, 1401, 1241]

C. Classification

We choose five commonly used classification models. They are Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and neural network which employs long short-term memory (LSTM).

Model Implementation, we use dropout layer to prevent overfitting. The dropout rate is chosen from {0.1,0.2,0.3, 0.4, and 0.5}. Our experiment, its peak performance when the dropout rate is 0.3. Two dense layers were subsequently added of size 64 and 6 with the activation function being *ReLu*. We train model using the Adam optimizer with batch size 24 for a total of 20 epochs and selected only the best result.

D. Evaluation Criteria

In this study, we evaluate the results based on the confusion matrix to calculate the accuracy (A), precision (P), and F1-Score (F1). The calculation formula is as follows,

$$A\text{= Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$P\text{= Precision} = \frac{TP}{TP+FP} \quad (2)$$

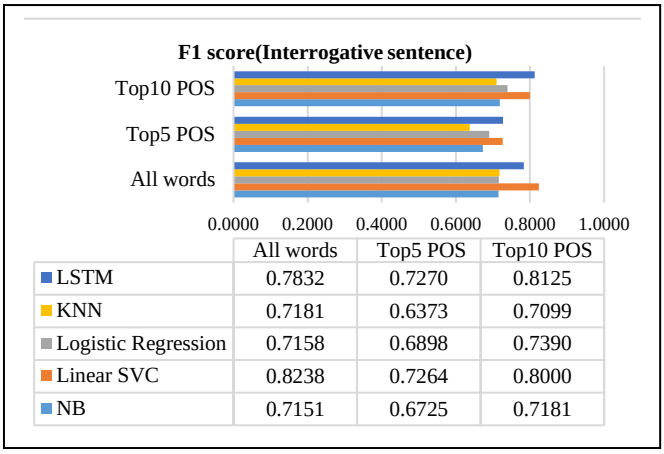
$$R\text{= Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F_1\text{= F}_1\text{-score} = \frac{2PR}{P+R} \quad (4)$$

IV. RESULT

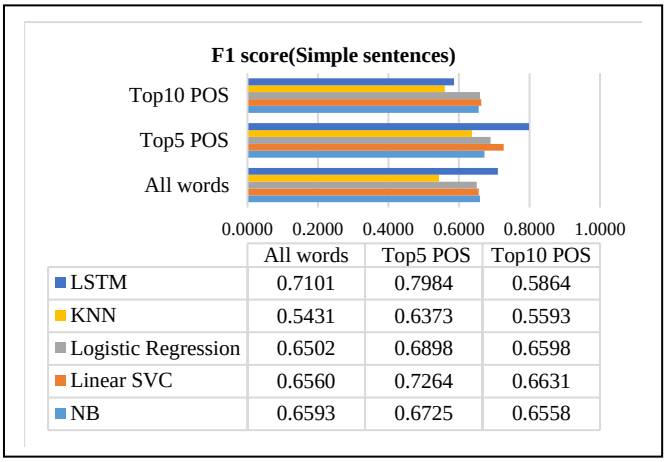
We tested our experiment that different datasets and classification models. The result classifies wh-question from Thai text which is an interrogative sentence dataset with an average accuracy of 71.20%, but simple sentence dataset only 56.80% for experiment all words in sentences. We found that apply only Top5 POS words could improve accuracy increasing up to 6.11%, with an average accuracy of 65.20% and improve F1-score increasing up to 8.40% for classifying the simple sentences to question category. The LSTM the highest score with an accuracy of 78.00%, precision 81.76%, recall 78.00%, and F1-score 79.84%, as shown in Table II.

Office of the Permanent Secretary (OPS), MHESI Thailand
(Ministry of Higher Education, Science, Research and Innovation)



	All words	Top5 POS	Top10 POS
LSTM	0.7832	0.7270	0.8125
KNN	0.7181	0.6373	0.7099
Logistic Regression	0.7158	0.6898	0.7390
Linear SVC	0.8238	0.7264	0.8000
NB	0.7151	0.6725	0.7181

Fig. 1. F-measure obtained for the label confusion (Interrogative sentence)



	All words	Top5 POS	Top10 POS
LSTM	0.7101	0.7984	0.5864
KNN	0.5431	0.6373	0.5593
Logistic Regression	0.6502	0.6898	0.6598
Linear SVC	0.6560	0.7264	0.6631
NB	0.6593	0.6725	0.6558

Fig. 2. F-measure obtained for the label confusion (Simple sentence)

Tables I to III present a comparison of the classification results when POS is applied. The results showed in terms of accuracy, recall, and precision.

TABLE I. THE CLASSIFICATION RESULT (ALL WORDS)

ALL WORD	Interrogative sentence			Simple sentences		
	A	P	R	A	P	R
NB	0.6050	0.8741	0.6050	0.5300	0.8719	0.5300
Linear SVC	0.8050	0.8436	0.8050	0.6100	0.7095	0.6100
Logistic Regression	0.6300	0.8287	0.6300	0.5200	0.8675	0.5200
KNN	0.7100	0.7263	0.7100	0.5150	0.5745	0.5150
LSTM	0.8100	0.7582	0.8100	0.6650	0.7618	0.6650
Average	0.7120	0.8062	0.7120	0.5680	0.7570	0.5680

TABLE II. THE CLASSIFICATION RESULT (APPIED TOP5 POS)

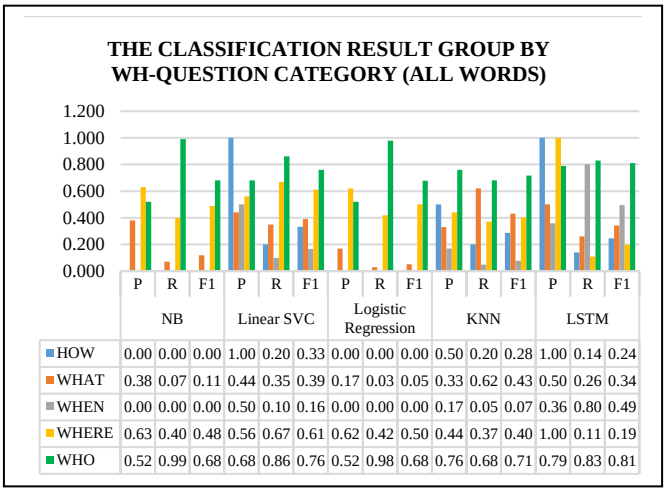
Top5 POS	Interrogative sentence			Simple sentences		
	A	P	R	A	P	R
NB	0.5700	0.8198	0.5700	0.5700	0.8198	0.5700
Linear SVC	0.7000	0.7549	0.700	0.7000	0.7549	0.7000
Logistic Regression	0.5850	0.8402	0.5850	0.5850	0.8402	0.5850
KNN	0.6250	0.6500	0.6250	0.6250	0.6500	0.6250
LSTM	0.7300	0.7240	0.7300	0.7800	0.8176	0.7800
Average	0.6420	0.7578	0.6420	0.6520	0.7765	0.6520

TABLE III. THE CLASSIFICATION RESULT (APPIED TOP10 POS)

Top10 POS	Interrogative sentence			Simple sentences		
	A	P	R	A	P	R
NB	0.6100	0.8727	0.6100	0.5300	0.8600	0.5300
Linear SVC	0.7750	0.8267	0.7750	0.6200	0.7126	0.6200
Logistic Regression	0.6600	0.8394	0.6600	0.5350	0.8606	0.5350
KNN	0.7000	0.7200	0.7000	0.5400	0.5801	0.5400
LSTM	0.8050	0.8201	0.8050	0.6850	0.7005	0.5043
Average	0.7100	0.8150	0.7100	0.5820	0.7428	0.5459

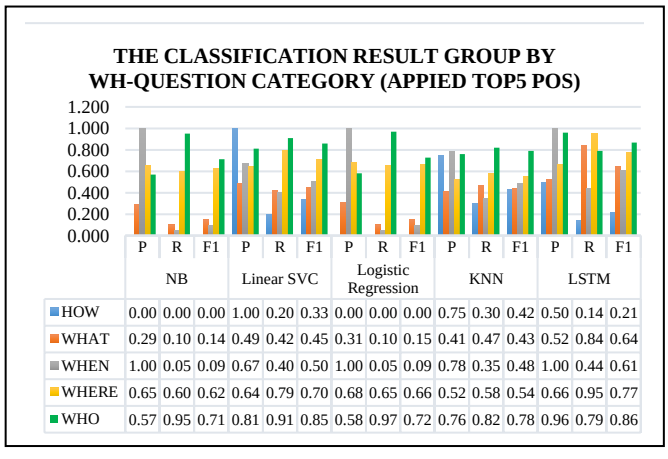
In this study, we focus on classified simple sentence to the wh-question categories on Thai texts when considering POS of each word in the sentence. Results from Fig. 3 to Fig. 5 show a comparison results of the classification performance based on precision, recall, and F1-score.

The best results for classifying sentences into question categories are achieved when kept only words (Top5 POS). For "Who" category is the most accurately predicted with the precision, recall, and F1-score are 96.00%, 79.00%, and 86.67%, respectively. Notwithstanding, for "How" category is the least than the other category with precision, recall, and F1-score only 50.00%, 14.00%, and 21.90%, respectively. However, the result has near both the result of the experiment with all words in a sentence and applied kept only Top10 POS in the sentence.



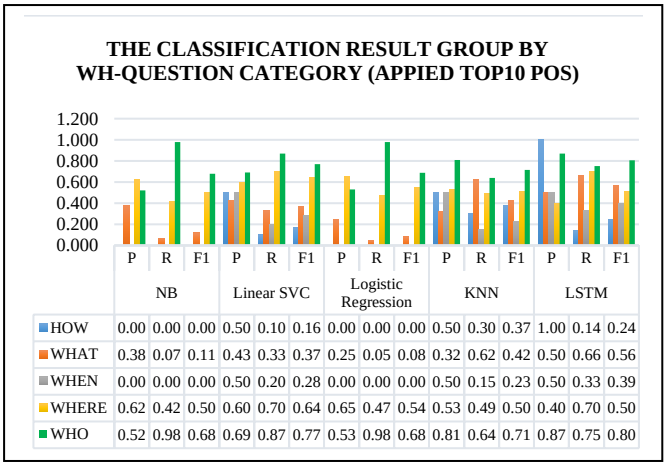
	NB			Linear SVC			Logistic Regression			KNN			LSTM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HOW	0.00	0.00	0.00	1.00	0.20	0.33	0.00	0.00	0.00	0.50	0.20	0.28	1.00	0.14	0.24
WHAT	0.38	0.07	0.11	0.44	0.35	0.39	0.17	0.03	0.05	0.33	0.62	0.43	0.50	0.26	0.34
WHEN	0.00	0.00	0.00	0.50	0.10	0.16	0.00	0.00	0.00	0.17	0.05	0.07	0.36	0.80	0.49
WHERE	0.63	0.40	0.48	0.56	0.67	0.61	0.62	0.42	0.50	0.44	0.37	0.40	1.00	0.11	0.19
WHO	0.52	0.99	0.68	0.68	0.86	0.76	0.52	0.98	0.68	0.76	0.68	0.71	0.79	0.83	0.81

Fig. 3. The classification result group by wh-question category (All words)



	NB			Linear SVC			Logistic Regression			KNN			LSTM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HOW	0.00	0.00	0.00	1.00	0.20	0.33	0.00	0.00	0.00	0.75	0.30	0.42	0.50	0.14	0.21
WHAT	0.29	0.10	0.14	0.49	0.42	0.45	0.31	0.10	0.15	0.41	0.47	0.43	0.52	0.84	0.64
WHEN	1.00	0.05	0.09	0.67	0.40	0.50	1.00	0.05	0.09	0.78	0.35	0.48	1.00	0.44	0.61
WHERE	0.65	0.60	0.62	0.64	0.79	0.70	0.68	0.65	0.66	0.52	0.58	0.54	0.66	0.95	0.77
WHO	0.57	0.95	0.71	0.81	0.91	0.85	0.58	0.97	0.72	0.76	0.82	0.78	0.96	0.79	0.86

Fig. 4. The classification result group by wh-question category (Applied Top5 POS)



	NB			Linear SVC			Logistic Regression			KNN			LSTM		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
HOW	0.00	0.00	0.00	0.50	0.10	0.16	0.00	0.00	0.00	0.50	0.30	0.37	1.00	0.14	0.24
WHAT	0.38	0.07	0.11	0.43	0.33	0.37	0.25	0.05	0.08	0.32	0.62	0.42	0.50	0.66	0.56
WHEN	0.00	0.00	0.00	0.50	0.20	0.28	0.00	0.00	0.00	0.50	0.15	0.23	0.50	0.33	0.39
WHERE	0.62	0.42	0.50	0.60	0.70	0.64	0.65	0.47	0.54	0.53	0.49	0.50	0.40	0.70	0.50
WHO	0.52	0.98	0.68	0.69	0.87	0.77	0.53	0.98	0.68	0.81	0.64	0.71	0.87	0.75	0.80

Fig. 5. The classification result group by wh-question category (Applied Top10 POS)

Fig. 6 and 7 show accuracy and loss over 20 epochs for training with the dataset (applied Top5 POS) for classifying simple sentences to the wh-question category.

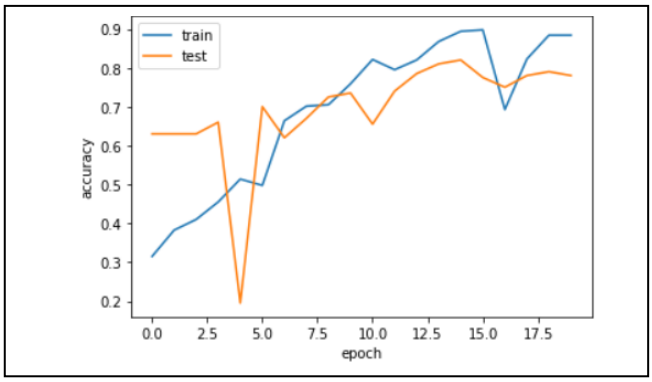


Fig. 6. Training and validation accuracy

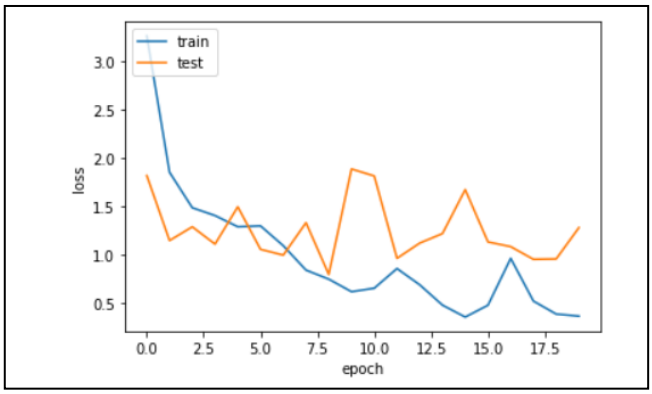


Fig. 7. Training and validation loss

V. Conclusions and Future Work

This work proposes an automated Thai sentence classification to a wh-question category using pre-process which considers the word in Top5 and Top10 POS (Part-Of-Speech) in the sentence. The experimental results show that the proposed pre-process (applied Top5 POS) can improve accurately classify the wh-question category with an accuracy of 78.00% (LSTM) and 65.20% (on average) for simple sentence. Thus, filter word only Top5 POS could improve accuracy, precision, recall, and F1 score for classifying the wh-question category ("WHO", "WHAT", "WHERE", and "WHEN") from simple sentences but not "HOW category". Thus, making it possible to identify a type of question from a simple sentence with considered POS. In addition, it should more study other features and effective accuracy for all categories.

Future work would involve comparing the performance of deep learning using extract feature type and comparison between difference method in many phases such as embedded words, pre-trained, and propose a deep neural network with a multi-layer with an attention mechanism to predict the possible category wh-question and possible answer from Thai text. Besides, studying question classification to improve the question generated automatic performance.

Acknowledgment

The researcher is grateful for financial support from Office of the Permanent Secretary (OPS), MHESI Thailand and technology support from King Mongkut's University of Technology North Bangkok, Thailand.

References

[1] H. Tayyar Madabushi, M. Lee, and J. Barnden, "Integrating Question Classification and Deep Learning for improved Answer Selection," in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3283–3294.

[2] T. Nomponkrang and K. Woraratpanya, "Thai-sentence classification using conceptual graph," in 2010 International Conference on Educational and Information Technology, 2010, pp. V2-479-V2-483, doi: 10.1109/ICEIT.2010.5607620.

[3] R.A. Anbuselvan Sangodiah, W.F.W. Ahmad, "Taxonomy based features in question classification using support vector machine," Journal of Theoretical and Applied Information Technology, 2017, pp. 2814-2823.

[4] A.W. Haryanto, E.K. Mawardi, and Muljono, "Influence of Word Normalization and Chi-Squared Feature Selection on Support Vector Machine (SVM) Text Classification, " in 2018 International Seminar on Application for Technology of Information and Communication, 2018, pp. 229-233, doi: 10.1109/ISEMANTIC.2018.8549748.

[5] V.A. Silva, I.I. Bittencourt, and J.C. Maldonado, "Automatic Question Classifiers: A Systematic Review," IEEE Transactions on Learning Technologies, 2019. 12(4), pp. 485-502.

[6] D. Phuc and N.T.K. Phung, "Using Naïve Bayes Model and Natural Language Processing for Classifying Messages on Online Forum, " in 2007 IEEE International Conference on Research, Innovation and Vision for the Future, 2007, pp. 247-252, doi: 10.1109/RIVF.2007 .369164.

[7] T. Pranckevičius and V. Marcinkevičius, "Application of Logistic Regression with part-of-the-speech tagging for multi-class text classification, " in 2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE). 2016.

[8] H. Li, H. Jiang, D. Wang and B. Han, "An Improved KNN Algorithm for Text Classification, " in 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 2018, pp. 1081-1085, doi: 10.1109/IMCCC. 2018.00225.

[9] K. Sundus, F. Al-Haj, and B. Hammo, "A Deep Learning Approach for Arabic Text Classification, " in 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), 2019, pp. 1-7, doi: 10.1109/ICTCS.2019.8923083.

[10] B. Billal, A. Fonseca, and F. Sadat, "Efficient natural language pre-processing for analyzing large data sets, " in 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 3864-3871, doi: 10.1109/BigData.2016.7841060.

[11] R. Kittinaradorn, T.A., K. Chaovavanich, K. Srithaworn, C. Kaewkasi, T. Ruangrong, and K. Oparad, "DeepCut: A Thai word tokenization libary using Deep Neural Network," Sept. 2019, Available from: URL http://doi.org/10.5281/zenodo. 3457707

[12] W. Phatthiyaphaibun and K. Chaovavanich. PyThaiNLP: a Python NLP package for Thai. 2016, Available from: https://www.thainlp.org /pythainlp/docs/2.0/api/tag.html.