The Comparison of Algorithms for Thai-Sentence Classification

Thanyarat Nomponkrang and Charun Sanrach

***Abstract*—The development of an online discussion board in collaborative learning makes available of tracking behavior and collaboration among students. The classification of discussion sentences on the board is an essential mechanism used to describe students' interaction patterns. This paper proposes feature extraction module that is used to extract the feature of Thai-sentence according to the sentence function. The module extracts two main features which are term binary (TB) of key phrase and term frequency (TF) of part-of-speech (POS).The TB term is used to indicate the presence of key phrase that cannot identified by POS and the TF term is used to calculate the term frequency-inverse document frequency (TF-IDF).Each feature is extracted from Thai-sentence to perform the data setswhichare1) TF of POS 2) TB and TF of POS 3) TF-IDF of POS and 4) TB and TF-IDF of POS. The performance of all data set is compared using 4 classification algorithms including: Decision Tree, Naïve Baye, K-nearest neighbor (k-NN) and Support vector machine (SVM).In this experiment shows the result in two dimensions which are the appropriate algorithm and the appropriate features to classify the Thai-sentence. The result is SVM algorithm is the optimal model on the dataset that have key phrase and TF-IDF term of POS.**

***Index Terms*—Thai-sentence classification, feature extraction, classification performance, Decision Tree, Naïve Baye, k-NN, SVM.**

I. Introduction

Online collaboration learning (CL) promotes problem solving and critical thinking skills, facilitates the development of a professional learning community, and supports student learning via interactions and co-construction of knowledge with others [1]. Discussion board is an important place for social support beside task and information exchange because group members encourage each other such as ask questions, explain and justify their opinions, articulate their reasoning, and elaborate and reflect upon their knowledge, thereby motivating and improving learning.

In educational contexts, especially in CL, the effective collaborative learning behavior does not guarantee by only placing students in a group and assigning them a task. The better understanding of students' behavior leads to easier improvement of students' learning and effective collaborative learning team. Thus, many researchers have been reported with students' participation and cooperation based on social support database. N. Jahng, W. S. Nielsen, and E. K. H. Chan [2] investigated the participation pattern in small-group collaborative learning. During the period of social analysis method, they categorized text message in students' learning process into three types: cognitive, social, and managerial. The frequency of students' text message can be used to indicate the participation pattern of each small group. This method can ultimately identify the level of collaboration and successful group achievement.It was also consistent with the study done by A. Soller, B. Goodman, F. Linton, and R Gaimari [3] that introduced the sentence opener when students begin their statements which are composed of argue, request, inform, task, motivate, mediate, maintenance, or acknowledge. According to the simulation results, all types of statement can be utilized to determine students' role within the group and skills he/she need to practice.W. Chamlertwat, P. Bhattarakosol, and T. Rungkasiri [4] proposed the Micro-blocked Sentiment Analysis System (MBSA) that discovers customer insight whether it is a positive or negative sentiment through conversations and thoughts related to smartphones on Twitter. To achieve the high performance analysis, the system combined both machine learning-based and lexicon-based approaches.

From the mentioned researches, the sentence classification on online discussion board is an interest procedure to extract the collaboration of students. In this experimental proposes the comparison of algorithms for Thai-sentences classification using text mining process. In our study, we performed the following tasks; feature selection and classification model construction.

The remainder of this paper is organized as follows; Section II describes the problem description and background of Thai-sentences. Section III states the theory of text classification. Section IV, the feature extraction module for Thai-sentences is proposed. The performance of classification algorithms are shown in Section V. Section VI, the experimental results are reported and discussed. Finally, the conclusion and future works are presented in Section VII.

II. Problem Description

Generally, the sentence function can be classified into four types: declarative, negative, interrogative, and imperative sentences. In Thai language the sentence are classified as follows: 1) a declarative sentence is comprised of noun phrases (as subject), verb phrases, and noun phrases (as object); 2) a negative sentence begins with "ไม่ (non)" or

Manuscript received February 28, 2015; revised June 19, 2015. This work was supported in part by the Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

The authors are with the Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand (e-mail: thanyaryt@fte.kmutnb.ac.th).

other words to deny something; 3) an interrogative sentence contains the following feature words-such as "ทำไม (why)," "เมื่อไหร่ (when)," "อะไร (what)," "ใคร (who)," "ที่ไหน (where)," or "อย่างไร (how)"- placed in the front or the end of sentences; and 4) an imperative sentence begins with a verb and is called a verb phrase such as "ช่วย,กรุณา (Please)." To classify a type of sentence, a key sentence feature which is known as a sentence function should be clearly identified, so that the sentence classification can accurately perform.

Furthermore, parsing Thai sentences is a crucial problem caused by the following reasons. Firstly, it cannot simply identify a sentence containing more than one key feature word as demonstrated in Table I (Case 1).

Secondly, a Thai phrase or sentence is composed of con-catenated words without explicit word delimiters or blank space between them, as shown in Table I (Case 2), when com- pare with English sentences. This leads to why Thai-sentence should be segmented correctly to generate word token as shown in Table I (Case 3).

Thirdly, structural ambiguities often arise in Thai language, since a part of speech depends on word orders, as shown in Table I (Case 4). The order of word "กำลัง" in sentences a) and b) makes its meaning different. In sentence a), it is a noun; its meaning is "power or capacity."On the other hand, if it is a preposition, its meaning is "in the act of." In sentence b), it is a preposition, since the order of "กำลัง" precedes "ทำงาน." Thus, an analysis of Thai sentences needs to consider the meaning of word according to the order of words in sentence.

Fourthly, there are inconsistencies in ordering relations within and across phrasal categories in Thai language. In Table I (Case 5), "ใคร (who)" is a key feature word that can be placed in either the beginning or the end of a sentence. In this case, the meaning of two sentences is the same; no matter which position of "ใคร" is in. Therefore, a syntactic analysis of Thai sentences needs a clear identification.

Lastly, discussion sentences often use informal words that are not in Thai dictionary, especially in interrogative sentence. They uses many informal words, such as "เหรอ," "ใช่ป่าว," "มั้ง," "จิงป่าว," "ใช่มั้ย,"and "มั้ย.". In Table I (Case 6), all sentences have the same meaning although they are use difference words at the end; only the first sentence is a formal sentence. As a result, we must define the function to these words.

In order to overcome these obstacles, the features extraction for text classification is a solution, since it is able to denote important feature in a sentence.

III. Theory

A. Text Mining

Text Mining is the discovery of new and useful information by automatically extracting information from textual document repositories [5]. One of the main challenges for text mining process is the high dimensionality of text data which have to be transformed from unstructured data to structured data in text pre-processing.

S. Jusoh and H. M. Alfawareh [6] presented foundation methods for conducting text mining that include natural language processing (NLP) and information extraction (IE). NLP is a technique that concerns with natural language generation (NLG) and natural language understanding (NLU). NLG uses some level of underlying linguistic representation of text to make sure that the generated text is grammatically correct. NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic. NLU consists of at least one of the following components; tokenization, morphological or lexical analysis, syntactic analysis, and semantic analysis. In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol. Morphological or lexical analysis is a process where each word is tagged with its part of speech (POS). The complexity arises in this process when it is possible to tag a word with more than one part of speech. Syntactic analysis is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence.

TABLE I: Example of Thai-Sentences

Case 1:	
	(ไม่)รู้ว่าจะไปหาข้อมูล(ที่ไหน)ดี ไม่(Non): Negative key phrase ที่ไหน (where) : Interrogative key phrase
Case 2:	
	ไม่รู้ว่าจะไปหาข้อมูลที่ไหนดี [I don't know. Where can I find this information?]
Case 3:	
	ไม่/รู้/ ว่า /จะ /ไป /หา /ข้อมูล /ที่ไหน/ดี NEG/VSTA/JSBR/XVBM/XVAM/VACT/NCMN/PDMN/ADVN
Case 4:	
	a) เครื่องยนต์ทำงานเต็ม(กำลัง) เครื่องยนต์ / ทำงาน / เต็ม(กำลัง) The engine is / run / at full capacity b) เครื่องยนต์(กำลัง)ทำงาน เครื่องยนต์ / (กำลัง)ทำงาน The engine is / running
Case 5:	
	a) (ใคร)เป็นหัวหน้ากลุ่ม / b) หัวหน้ากลุ่มเป็น(ใคร) [Who is the group leader ?]
Case 6:	
	a) เครื่องยนต์ทำงานปกติใช่ไหม b) เครื่องยนต์ทำงานปกติใช่มั้ย c) เครื่องยนต์ทำงานปกติใช่ป่าว [The engine runs normally. Right?]

To discover the information from any text, Information Extraction (IE) is the necessary part. It deals with the extraction of specified entities, events, and relationships from the text sources. The extracted key information, from original text is mapped to be predefined, structured representation, and stored into a structured database. Fig. 1 shows the relationship between NLP and IE in pre-processing process, which is also known as feature extraction.

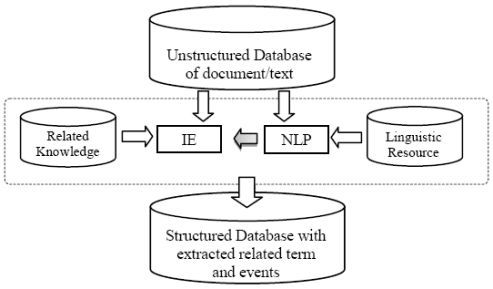


Fig. 1. Feature extraction process.

B. Feature Extraction

Feature extraction is the first step of pre-processing that is used to recognize and classify significant vocabulary items in unrestricted natural language texts [7]. The steps of the feature extraction are [8]:

1) Tokenization: A document is treated as a string, and then partitioned into a list of tokens.
2) Removing stop word: Stop word is the insignificant words which need to be removed.
3) Stemming word: The different word forms are converting into similar form.

Part of speech (POS) is the basic type of word, which used to show characteristics and annotation. In order to perform semantic analysis in text mining, POS is used because some words may have multiple meanings. POS tagging is token to each word after tokenization step as shown in Fig. 2, which is example of POS in Thai-sentence.

Source text : การจำแนกประเภทข้อมูลด้วยเทคนิค SVM
Feature extraction :
- Tokenization:
การ | จำแนก | ประเภท | ข้อมูล | ด้วย | เทคนิค || SVM
- POS tagging
FIXN| VACT | NCMN| NCMN | RPRE| NCMN||NPRP
- Removing stop word
การ | จำแนก | ประเภท | ข้อมูล | ด้วย | เทคนิค | SVM
-Feature selection :
Vector space=[การ , จำแนก , ประเภท , ด้วย, ข้อมูล , เทคนิค,SVM]
Vector space=[FIXN,VACT,NCMN,NCMN,RPRE,NCMN,NPRP]

Fig. 2. Example of pre-processing process.

C. Feature Selection

Feature selection (FS)is process to construct vector space of triggering items. Term is condition to consider the importance of the triggering items. The terms are marked as highest score according to predetermined measure such as 1) term binary (TB), which checks if a particular words/tokens appear in the document, 2) term frequency (TF), which computes the number of repetitions of a words/tokens in a document, or3) term frequency-inverse document frequency (TF-IDF), which determines the relative frequency of words/tokens in a specific document through an inverse proportion of the word over the entire document corpus. Each formula as shown in following [8]:

$$TB = b_{ij} = \begin{cases} 1 & \text{if the word appears in document} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$TF = f_{ij} = \quad frequency\ of\ term\ i\ in\ document\ j \quad (2)$$

$$TF - IDF = tf_{ij} idf_{ij} = tf_{ij} \times log_2 \frac{N}{df_i} \quad (3)$$

where b_{ij} is the binary digit of term j, tf_{ij} is term frequency of term $i$in document j, $tf_{ij}idf_{ij}$ is term frequency-inverse document frequency of term i in document j, N is the number of document in the collection and df_i is the document frequency of term i in the collection.

Fig. 2 shows the example of feature extraction and feature selection to transform unstructured data to structured data. Firstly, text is tokenized into list of words and thenremovedall control characters, space between words, dot, commas, and similar characters. Lastly, feature selection is performed, which is putting all of words into vector space to select features according to the weight of each word.

D. Text Classification

Text classification is an area of text mining which automatically assigns a given document to a set of predefined categories based on its textual content and extracted features. There are many algorithms used for text classification. However, 4 algorithms will be explained including: Decision Tree, Naïve Bayes, k-NN, and SVM respectively.

The decision tree algorithm is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret. The goal is to create a classification model that predicts the value of a target attribute (often called class or label) based on several input attributes of the dataset [9].

Naïve Bayes algorithm is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. In simple terms, a Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature [9].

K-nearest neighbor (k-NN) is used to test the degree of similarity between document and k training data and to store a certain amount of classification data, thereby determining the category of test documents. The training data is stored in n-dimensional (i.e. n-attribute) pattern space. When given an unknown document, this algorithm searches the pattern space for k-training data that closet to the unknown. This method is an instant-based learning algorithm that categorized object based on closest feature space in the training set [9].

Support vector machines (SVMs) algorithm is supervised learning method for classification to find out the linear separating hyper plane which maximize the margin, i.e. the optimal separating hyper plane (OSH) and maximize the margin between the two data sets. The model is representation of the examples as points in the space, mapped so that the examples of separate categories are divide by clear margin. New examples are then mapped into that same space and predicted to belong to a category based on which side of the

margin they fall on [9].

After the texts or documents have been processed on feature extraction and feature selection process. The extracted features are used for training a model. To find the optimal model, the researchers must compare the performance of the model between several classification algorithms.

S. Wakade, C. Shekar, K. J. Liszka, and C. C. Chan [10] conducted an experiment on Twitter Data using Decision Tree and Naïve Baye for sentiment analysis tweets about iPhone and Microsoft. The features were created from the list of sentiment words plus emotions. For example, the positive words are "beautiful," "easy," "popular," and the negative words are "fragile," "grumpy," "stressed." This study shown decision tree classifiers out-perform Naïve Baye classifier.

S. Tan and J. Zhang [11] employed k-NN, Naïve Bayes and SVM for sentiment classification on Chinese documents. They experimented on Chinese documents to categorize the positive and negative sentiment. They studied in 2 dimensions: 1) the performance of feature selection methods and 2) the performance of classifiers. The experimental results indicated that Information Gain performs the best for sentimental terms and SVM exhibits the best performance for sentiment classification.

N. Jindal and B. Liu [12] compared Naïve Bayes and SVM to identify the comparative sentences in text documents. Their experimental result shown that the performance of Naïve Baye is better than SVM. The dataset, that used to classify, are generated from: 1) POS tags, 2) keywords, 3) Class Sequence Rules(CSRs), and 4)manual rules.

E. Performance Evaluation

To evaluate the performance of text classification algorithms, we use accuracy, recall, precision and F-measure, as following equations:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (4)$$

$$\text{Recall} = \frac{\text{Number of correct positive predictions}}{\text{Number of positive examples}} \quad (5)$$

$$\text{Precision} = \frac{\text{Number of correct positive predictions}}{\text{Number of positive predictions}} \quad (6)$$

$$\text{F-measure} = \frac{2}{\text{Recall+Precision}} \quad (7)$$

IV. Feature Extraction Module

In this experiment, we classified Thai sentence by the function of sentence as depicted in Table II. Each type represents students' purpose on discussion board.

TABLE II: Sentence Type by Function

Sentence Type (Class)	Function
Compound	state the idea more than one
Declarative	state an idea
Negative	state an idea as conflicting
Interrogative	show strong emotions
Imperative	give orders or directions

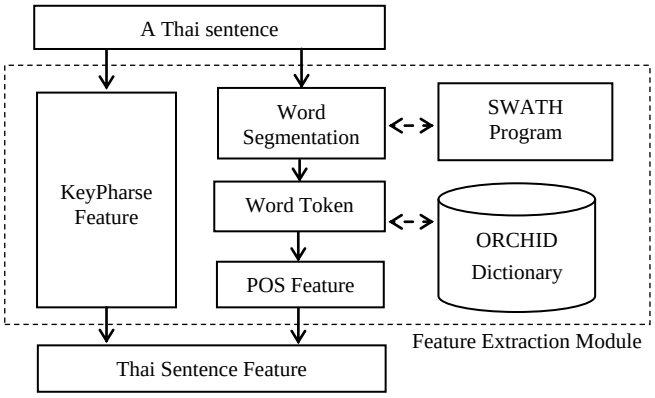


Fig. 3. The Architecture of data pre-processing module.

The architecture of feature extraction module is show in Fig. 3. The module is extracting the feature in 3 terms: 1) the key phrase, 2) the term frequency of Thai POS, and 3) the term frequency-inverse document frequency.

TABLE III: Key Pharse of Thai-Senetece Feature

Type of sentence	Key Pharse		Position	
			Head	End
Imperative	ขอ	[request for]	●	
	ขอโทษ	[excuse me]	●	
	ควร	[should]	●	
	ช่วย	[help]	●	
	กรุณา	[please]	●	
	โปรด	[please]	●	
	เชิญ	[invite]	●	
	หน่อย	[help]		●
	อย่า	[Don't]	●	
	ต้อง	[must]	●	
	ต้องการ	[want]	●	
	ห้าม	[Nix!]	●	
	!			●
Interrogative	เหรอ	[or not]		●
	มั้ย	[yes or no]		●
	รึ	[yes or no]		●
	หรือ	[yes or no]	●	●
	กี่	[how much/how many]		●

TABLE IV: The Features of 4 Datasets

	Term Binary of Key Phrase		Term Weighting
	Imperative	Interrogative	
Dataset A			TF in 47 POS features
Dataset B	●	●	TF in 47 POS features
Dataset C			TF-IDF in 47 POS features
Dataset D	●	●	TF-IDF in 47 POS features

The key phrase module is need because, on discussion board, students may use informal Thai sentence which causes of 2 problems. Firstly, the beginning POS of imperative and declarative sentences are verb phrase. Thus, we must define weather the words are at the beginning or the end position in imperative sentence as illustrate in Table III. Secondly, the problem of interrogative sentence, as described in problem description section case 6, is feature extraction module that

has identified informal word as shown in Table III. All key phrases are stored in database. Feature extraction module will search for existing of key phrase from the beginning or the end of original sentence.

Before counting the frequency of POS in POS feature module, the original sentence is segmented into a sequence of words by the automated Thai word segmentation algorithm, SWATH[13], developed by National Electronics and Computer Technology Center (NECTEC). After that, word token module is applied to the sequence of words in order to generate a token for each word. The token is mapped by a Thai part-of-speech (Thai POS), which is a part of ORCHID dictionary in SWATH program as shown in appendix. The result of the word token module is sequence of Thai POS. In next step, the TF of Thai POS is defined by counting the number of each Thai POS in the sequence. Lastly, the dataset of TF-IDF is computed from TF. The purpose of this experiment is to compare the performance of classification on 4 datasets as depicted in Table IV.

V. Feature Extraction Module

A. Data Preparation

In this experiment, the proposed module is tested with 7,900 sentences from www.thailanguage.com. These samples are manually divided into 5 classes consisting of: 1,911compoundsentences, 3,782 declarative sentences, 697 interrogative sentences, 1,120 negative sentences, and 390 imperative sentences. Adages and greeting sentences are not considered in this experiment.

B. Dataset and Experiment

We use feature extraction system to extract 7,900 collected Thai-sentences into 4 datasets. After data pre-processing, we tried each dataset with 4 algorithms: Decision Tree, Naïve Bayes, k-NN and SVM, using WEKA as a tool. In addition, we performed 2-dimensions comparison. The first dimension is accuracy comparison between 4 classification algorithms on 4 datasets. The second one is classification performance comparison of each class on 4 datasets according to classification algorithms as defined. The results of all dimensions are shown in the next section.

C. Comparison of Classification Algorithms

1) Accuracy

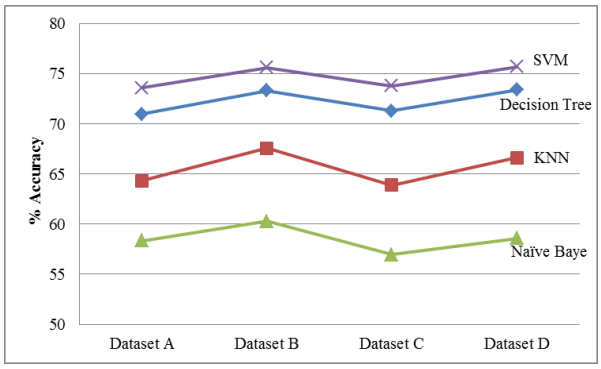
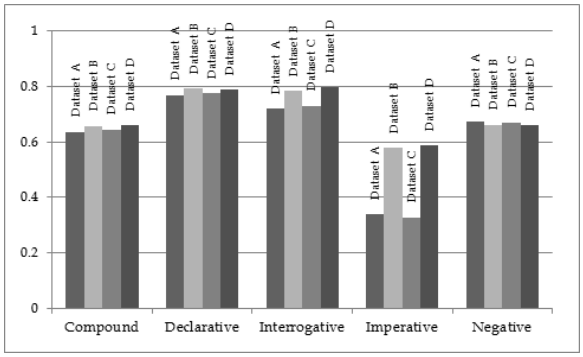
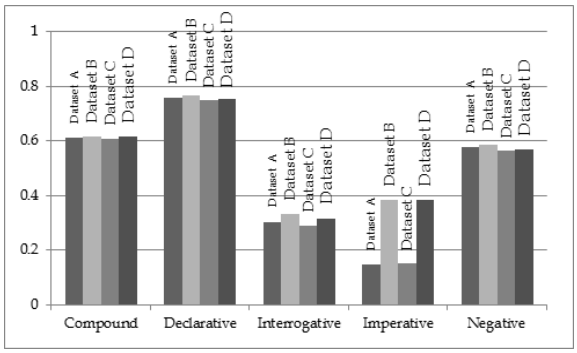


Fig. 4. The comparison of accuracy on 4 algorithms.

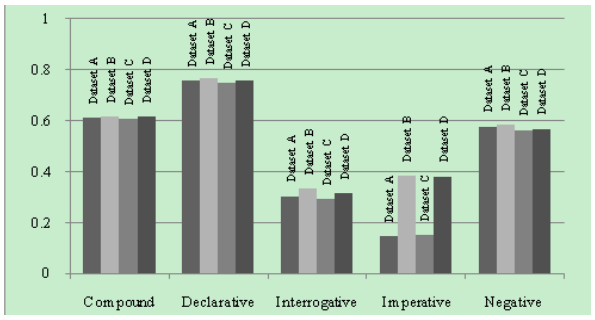
2) The precision of each class



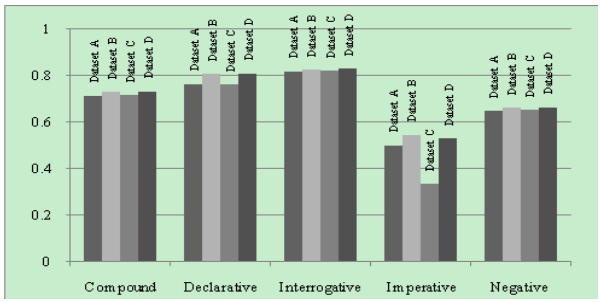
(a) The precision of decision tree algorithm.



(b) The precision of naïve baye algorithm.



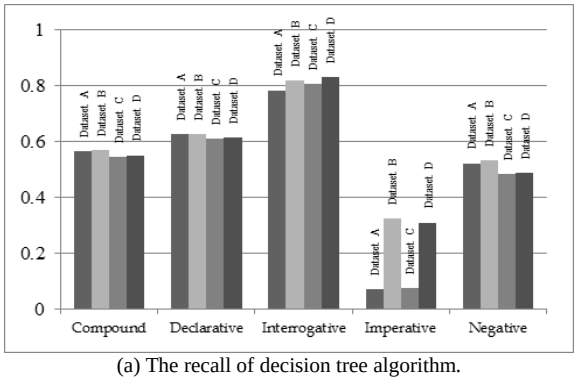
(c) The precision of k-NN algorithm.



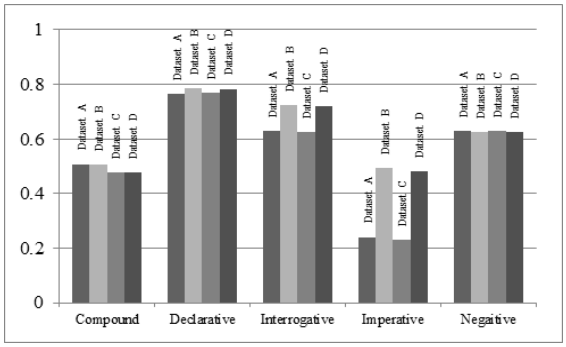
(d) The precision of SVM algorithm.

Fig. 5. The precision of each algorithm.

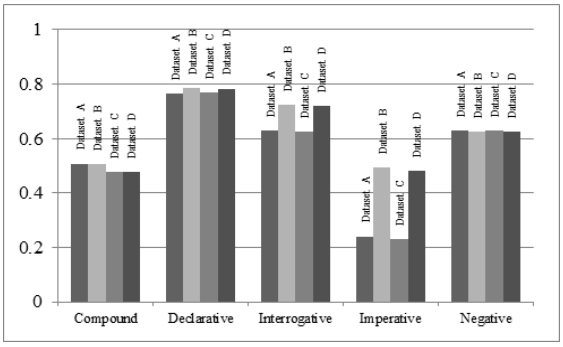
3) The recall of each class



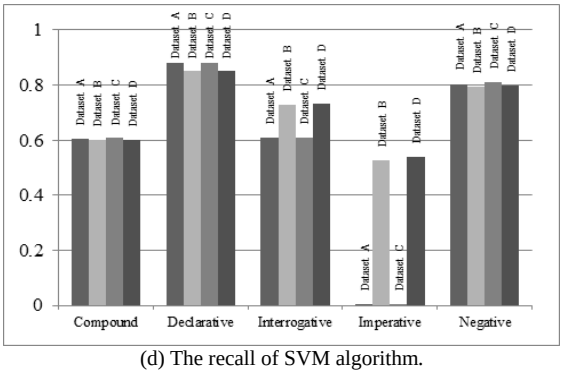
(a) The recall of decision tree algorithm.



(b) The recall of naïve baye algorithm.



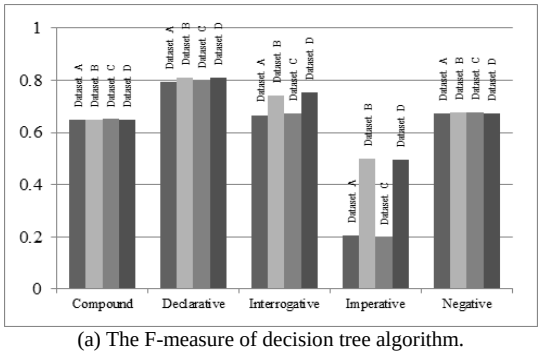
(c) The recall of k-NN algorithm.



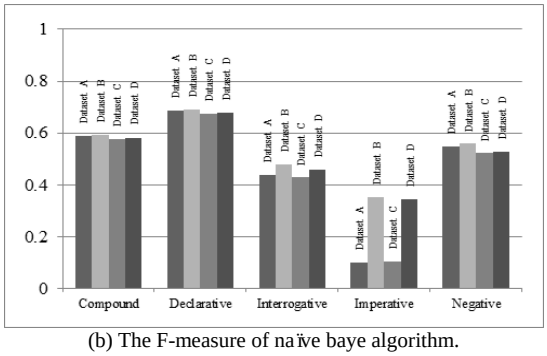
(d) The recall of SVM algorithm.

Fig. 6. The recall of each algorithm.

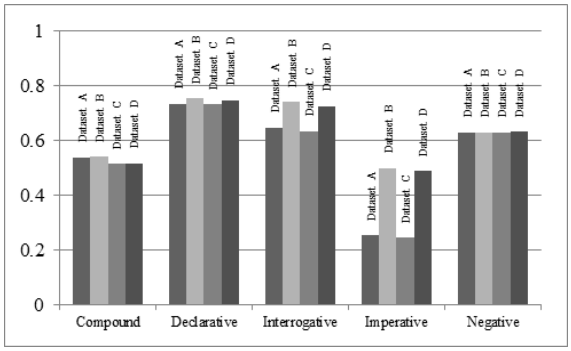
4) The F-measure of each class



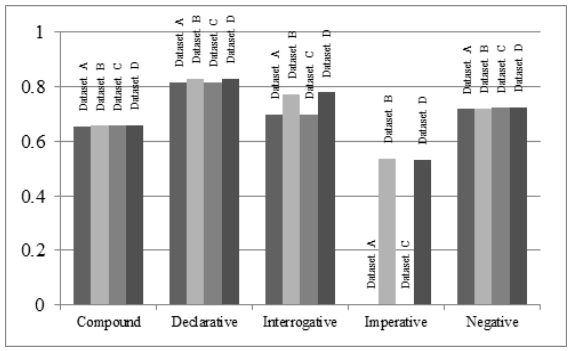
(a) The F-measure of decision tree algorithm.



(b) The F-measure of naïve baye algorithm.



(c) The F-measure of k-NN algorithm.



(d) The F-measure of SVM algorithm.

Fig. 7. The F-measure of each algorithm.

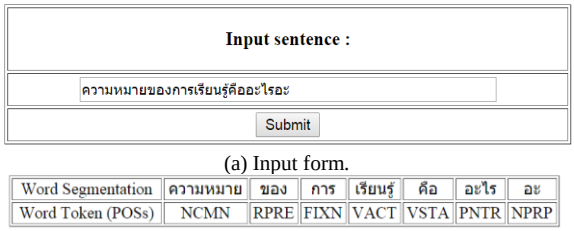
VI. Experimental Results

In this experiment, we compared the performance of classification algorithm in 2 dimensions. The first dimension is the accuracy of classification algorithms as shown in Fig. 4. The result shows that the value in Dataset-B and Dataset-D, which have key phrase, are higher than the values in Dataset-A and Dataset-C that have only TF/TF-IDF. As the value are: SVM-76% and 74%, Decision Tree-73% and 71%, k-NN-67% and 64%, naïve baye-59% and 57%, respectively. According to the accuracy, all algorithms perform better on data sets that have key phrase than data sets that only have TF/TF-IDF.

The second dimension is performance of 5 classes on 4 datasets according to 4 classification algorithms. Firstly, the precision of 5 classes as depicted in Fig. 5. The results are: 1) the precisions of compound class are 0.60-0.70 similar for every dataset and every algorithm 2) the precisions of declarative class are 0.70-0.80.The values on datasets that have key phrase are slightly higher than datasets that only have TF/TF-IDF 3) the precisions of interrogative class in SVM, Decision Tree, and k-NN are 0.60-0.80 but Naïve baye is only 0.30. The values in Decision Tree, Naïve baye, and k-NN on datasets that have key phrase are higher than datasets that only have TF/TF-IDF 4) the precisions of imperative class are 0.20-0.60. The values of every algorithm on datasets that have key phrase and data sets that only have TF/TF-IDF are very different. 5) the precision of negative class is about 0.60 and the values on 4 datasets are similar, as shown in Fig. 5.

Secondly, Fig. 6 depicted the recall of 5 classes, as the results are: 1) the recalls of compound class are 0.50-0.65. The results in each algorithm are different. In Decision Tree, the recalls on datasets that have only TF/TF-IDF are higher than datasets that have key phrase. While the values of Naïve

baye, and k-NN are in the same direction, i.e. the recalls on datasets that have TF (Dataset-A, and Dataset-B) are higher than datasets that have TF-IDF (Dataset-C, and Dataset-D). In SVM, the values are similar on all datasets. 2) the recalls of declarative class are 0.60-0.80. The values of Decision Tree, and k-NN on datasets that have key phrase are slightly higher than the datasets that only have TF/TF-IDF. In Naïve Baye, the recalls on datasets that have TF are higher than datasets that have TF-IDF. But the results in SVM are different from the others because the recalls on datasets that have key phrase are less than datasets that have only TF/TF-IDF. 3) the recalls of interrogative class are 0.60-0.80. The values on datasets that have key phrase are higher than datasets that have only TF/TF-IDF in every algorithm. 4) the recalls of imperative class are 0.00-0.50. By all algorithms, the values on datasets that have key phrase are very higher than datasets that have only TF/TF-IDF, especially in SVM. 5) the recalls of negative class are more than 0.60 and the values on all datasets in Decision Tree, k-NN, and SVM are similar. Only Naïve baye is less than 0.60 and the values on datasets that have TF are higher than the values on datasets that have TF/IDF.



(a) Input form.

(b) The result of prediction.

Fig. 8. Screenshot of prediction result.

Lastly, the F-measure of 5 classes as illustrated in Fig.7. The results are: 1) the F-measures of compound class are 0.50-0.65 which are similar for all datasets. 2) the F-measures of declarative class are about 0.65-0.80. In each algorithm, the values on datasets that have key phrase are slightly higher than data sets that only have TF/TF-IDF. 3) the F-measures of interrogative class in Decision tree, k-NN, and SVM are 0.65-0.70 but Naïve baye is only 0.40-0.50. The values of datasets that have key phrase are higher than data sets that only have TF/TF-IDF. 4) the F-measures of imperative class are 0.00-0.50. The value of data sets that only have TF/TF-IDF in SVM is very low whereas the values on datasets that have key phrase are very high. 5) the F-measures of negative class is 0.55-0.70 which are similar for all datasets in each algorithm.

Fig. 8 shows a screenshot of the Thai-sentence classification program. After inputting the original sentence, as depicted in Fig. 8(a), the feature extraction module extracts features of sentence and sends them to SVM model. For example, the feature of an input sentence, "ความหมายของการเรียนรู้คืออะไรอะ (What is the meaning of learning?)" was extracted. The result of prediction is interrogative sentence with the probability of 0.888 as shown in Fig. 8(b).

VII. Conclusion and Future Works

This paper proposes feature extraction module of Thai discussion sentence to create training dataset used to train a model to classify Thai sentence into 5 classes in order to track students' behavior on online discussion board. The experimental results show SVM algorithm is the optimal model on the dataset that have key phrase and TF-IDF term of POS.

For future work, the model will be further implemented to automatically tracking students' behavior in online discussion board.

Appendix

TABLE V: Thai Part-of-Speech

POSs	Description
ADVN	Adverb with normal form
FIXN	Nominal prefix
NPRP	Proper noun
NCNM	Cardinal number
NONM	Ordinal number
NCMN	Common noun
PPRS	Personal pronoun
PDMN	Demonstrative pronoun
PNTR	Interrogative pronoun
RPRE	Preposition
VACT	Active verb
VSTA	Stative verb
VATT	Attributive verb
XVAM	Pre-verb auxiliary, after negator "ไม่"
XVMM	Pre-verb, before or after negator "ไม่"
XVBM	Pre-verb auxiliary, before negator
DDAQ	Definite determiner, following quantitative expression
DIAQ	Indefinite determiner, following quantitative expression
DCNM	Determiner, cardinal number expression
DONM	Determiner, ordinal number expression
CNIT	Unit classifier
CLTV	Collective classifier
CMTR	Measurement classifier
CFQC	Frequency classifier
JCRG	Coordinating conjunction
JCMP	Comparative conjunction
JSBR	Subordinating conjunction
EITT	Ending for interrogative sentence
NEG	Negator

References

[1] V. Zygouris-Coe, "Collaborative learning in an online teacher education course: Lessons learned," in *Proc. International Conference on Information Commuication Technologies in Education*, pp. 332-342, 2012.

[2] N. Jahng, W. S. Nielsen, and E. K. H. Chan, "Collaborative learning in an online course: A comparison of communication patterns in small and whole group activities," *IJEDE*, vol. 24, no. 2, pp. 39-58, 2010.

[3] A. Soller, B. Goodman, F. Linton, and R Gaimari, "Promoting effective peer interaction in an intelligent collaborative learning system," in *Proc. the 4th International Conference on Intelligent Tutoring Systems*, pp. 186-195, USA: Springer, 1998.

[4] W. Chamlertwat, P. Bhattarakosol, and T. Rungkasiri, "Discovering consumer insight from Twitter via sentiment analysis," *J. UCS*, vol. 18, no. 8, pp. 973-992, April 2012.
[5] R. S. Segall, Q. Zhang, and M. Cao, "Web-based text mining of hotel customer comments using SAS® text miner and megaputer polyanalyst® ," in *Proc. SWDSI 2009 Annual Conference*, pp. 141-152, USA, 2009.
[6] S. Jusoh and H. M. Alfawareh, "Techniques, applications and challenging issue in text mining," *IJCSI*, vol. 9, no. 2, pp. 431-436, November 2012.
[7] J. Dörre, P. Gerstl, and R. Seiffert, "Text mining: Finding nuggets in mountains of textual data," in *Proc. the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 398-401, USA: ACM, 1999.
[8] B. Baharudin, L. H. Lee, and K. Khan, "A review of machine learning algorithms for text-documents classification," *Journal of Advances in Information Technology*, vol. 1, no. 1, pp. 4-20, 2010.
[9] Rapidminerperator reference. [Online]. Available: http://docs.rapid-miner.com/studio/operators/
[10] S. Wakade, C. Shekar, K. J. Liszka, and C. C. Chan, "Text mining for sentiment analysis of Twitter data," presented at the 2012 World Congress in Computer Science, Computer Engineering, and Applied Computing, USA, 2012.
[11] S. Tan and J. Zhang, "An empirical study of sentiment analysis for Chinese documents," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2622-2629, 2008.
[12] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *Proc. the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 244-251, USA: ACM, 2006.
[13] S. Meknavin and P. Charoenpornsawat, "Feature-based thai word segmentation," presented at NLPRS, 1997

Thanyarat Nomponkrang received the MS degree in computer technology from King Mongkut's University of Technology North Bangkok (KMUTNB), Bangkok, Thailand in 2005. Currently, she is a computer education PhD candidate at KMUTNB. Her research interests include data mining and applying technology for learning and teaching. Currently, she is with the Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

Charun Sanrach received the Ph.D. degree in diplôme de docteur (informatique) INPL, France. His research interests include applying artificial intelligence in computer education, applying technology for learning and teaching, problem-based learning, project-based learning, and constructivist. Currently, he is with the Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.