การคัดเลือกคุณลักษณะที่มีประสิทธิภาพในการจำแนกความคิดเห็นสำหรับการปรับปรุงหลักสูตร

THE EFFECTIVE FEATURES SELECTION THROUGH OPINION CLASSIFICATION FOR CURRICULUM ADJUSTMENT

เบนจามิน ชนะคช[1]*** *** จรัญ แสนราช[2]

Benjamin Chanakot[1]*****, *Charun Sanrach*[2]

ภาควิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok.

***Corresponding author, e-mail:** b.benjamin.ch@gmail.com*

***Received:** 20 February 2019; **Revised:** 21 May 2019; **Accepted:** 7 June 2019*

บทคัดย่อ

งานวิจัยนี้นำเสนอการจำแนกความคิดเห็นด้วยเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อสร้างโมเดลในการจำแนกความคิดเห็น ด้วยการเปรียบเทียบเพื่อคัดเลือกคุณลักษณะที่มีประสิทธิภาพในการจำแนกความคิดเห็น สำหรับการปรับปรุงหลักสูตร จากแบบสอบความคิดเห็นที่มีต่อหลักสูตร จำนวน 1,575 ชุด โดยทำการคัดเลือกคุณลักษณะด้วยวิธีทีเอฟ-ไอดีเอฟ ไคสแควร์ และอินฟอร์เมชันเกน และทดสอบประสิทธิภาพในการจำแนกความคิดเห็น (Opinion Classification) ด้วยอัลกอริทึม นาอีฟเบย์ เคเนียร์เรสเนเบอร์ ซัพพอร์ตเวกเตอร์แมชชีน

จากการทดลองพบว่า เมื่อกำหนดค่า Threshold มากกว่าหรือเท่ากับ 1 ในการจำแนกความคิดเห็น ด้วยวิธีทีเอฟ-ไอดีเอฟ พบว่าอัลกอริทึม ซัพพอร์ตเวกเตอร์แมชชีน ให้ความถูกต้องมากที่สุด โดยมีค่า Accuracy เท่ากับ 90% ดังนั้นสามารถสรุปได้ว่า เมื่อเปรียบเทียบผลลัพธ์การคัดเลือกคุณลักษณะ ผลลัพธ์จากทีเอฟ-ไอดีเอฟมีค่าผลการจำแนกความคิดเห็นต่อหลักสูตรที่ถูกต้องและมีประสิทธิภาพมากกว่าวิธีอื่น ซึ่งผลจากการทดสอบนี้สามารถใช้เป็นแนวทางในการพัฒนาระบบแนะนำการปรับปรุงหลักสูตรด้วยเหมืองข้อความที่มีประสิทธิภาพต่อไป

คำสำคัญ: จำแนกความคิดเห็น ทีเอฟ-ไอดีเอฟ ไคสแควร์ อินฟอร์เมชันเกน

Abstract

This research presents the opinion classification based on machine learning techniques for creating opinion classification model. We compared for finding the effective feature selection in the opinion classification for curriculum improvement. Based on the questionnaire of 1,575 sets, Feature Selection based on Term Frequency-Inverse Document Frequency method, Chi-Square method and Information

Gain method. The testing of effective on opinion classification with Naïve Bayes Algorithm, K-Nearest Neighbors, Support Vector Machine.

It was found that when the value of Threshold was greater than or equal to 1 for opinion classification, TF-IDF method indicated that the Support Vector Machine algorithm was the most accurate, with an accuracy of 90%. In conclusion, the result for comparison of features selection results from the TF-IDF method provided more accurate and effective for curriculum adjustment than other methods. This research result could be the guideline for effectiveness of the development on the recommendation system for adjust curriculum with text mining.

Keywords: Opinion Classification, TF-IDF, Chi-Square, Information Gain

Introduction

The competition in the 21st century is caused for changing in culture, economy, society and technology, which affected the higher education because higher education institutions is important mechanism in order to produce quality and knowledgeable graduates who compatible with national needs and labor market. Therefore, the improvement of the curriculum to be up-to-date in the 21st century is very necessary. Popular tools for crawling course updates from curriculum stakeholders are comments and suggestions questionnaire. The comments and suggestions from the questionnaire not only reflex on the quality of the current course as well as being used as a guide to develop a quality curriculum and compatible with needs of curriculum's stakeholders but also opinion analysis from questionnaire is difficult. It must due to the large amount of obtainable information. It takes a long time for reading to understand the opinion questionnaire. However, computer usage offers analyze data but it is still a problem. The text used in the questionnaire is a natural language which complexed for interpretation. Sometimes the same sentence can convey to several things that depends on the situation you want to convey. Moreover, the natural language is constantly changing and always provide new words that the computer does not understand the meaning of the text because of ambiguity and ill-formedness. Due to limitations and problems, classification technique can help to analyze data for informing types of opinion and compatible with requirements. Opinion classification is an information analysis by using text mining to find patterns and relationships of information through statistics analysis and machine leaning to predict the results of the classification. It must be processed in natural language processing to create computer's understanding on meaning in the opinion. Feature Selection is required to reduce the size or feature of the data to be smaller by choosing effective features for the classification of opinions. It uses the weight values of the words that appear in the opinions as a feature to reduce the size of the opinion text to be smaller without losing any important feature of the opinion and the less accuracy of the processing results.

Trstenjak, Mikac, and Donko [1], indicated that KNN with TF-IDF Based Framework for Text Categorization by using the KNN algorithm with TF-IDF and Framework for text classification which focus on the speed and efficiency of text classification and Harish, Revanasiddappa [2], presented A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents by using a good feature selection to solve problems in text categorization due to Term Frequency-Inverse Document Frequency (TF-IDF), Information Gain (IG), Mutual Information (MI), Chi-Square, Ambiguity Measure (AM), Term Strength (TS), Term Frequency-Relevance Frequency (TF-RF) and Symbolic Feature Selection (SFS) and another research of Florence, Vaidya, Panchal, and Negi [3], presented document classification using NLP techniques to classify set of unknown documents into their accurately categories due to TF-IDF and Vector Space model.

Based on literature review found that there are many research projects aimed to solve problems about classification of text categorization by using algorithms to classify documents which each document will be represented in vector format. The vector is represented by the weight of the word that is calculated by the following methods: TF-IDF, Chi-square, Information gain (IG) etc. It uses the weight of word value as a guideline for classification of text categorization.

Therefore, this research presents the opinion classification based on machine learning techniques for creating opinion classification model by experiment to compare the effective of features selection. The feature selection is conducted on detail of opinion texts. Three technique can be obtained from Term Frequency-Inverse Document Frequency method, Chi-Square method and Information Gain method. The result of this experiment can find the significant words in each opinion for opinion classification and creation of opinion classification model. The model can be used to classify ideas for further curriculum adjustment.

Objectives

To compare efficiency of the feature selection of opinion classification for curriculum adjustments by Term Frequency-Inverse Document Frequency (TF-IDF), Chi-Square (CHI) and Information Gain (IG) method.

Methods

Opinion Classification

Opinion classification is a classification of opinions based on the detail of opinion text. Split Test method is obtained to use as classify the opinion by random into 2 part. The first part, 70% used training data to create opinion classification model. The second past, 30% used testing data to perform testing of effective of opinion classification model. It conducts machine learning technique for 3 algorithms such as Naïve Bayes, K-Nearest Neighbors, Support Vector Machine.

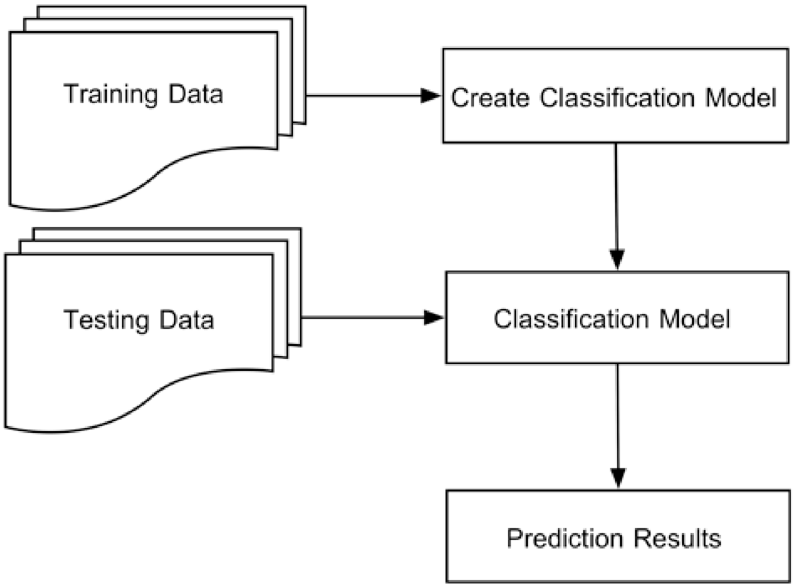

Figure 1. Opinion Classification based on Split Test Method.

Steps of Opinion Classification Development

The steps of opinion classification development by using machine learning technique consist of the following 6 steps:

1) Data Collection

2) Data Preparation

3) Feature Selection

4) Representation the opinion through Vector Space Model

5) Creation of opinion classification model

6) Model evaluation

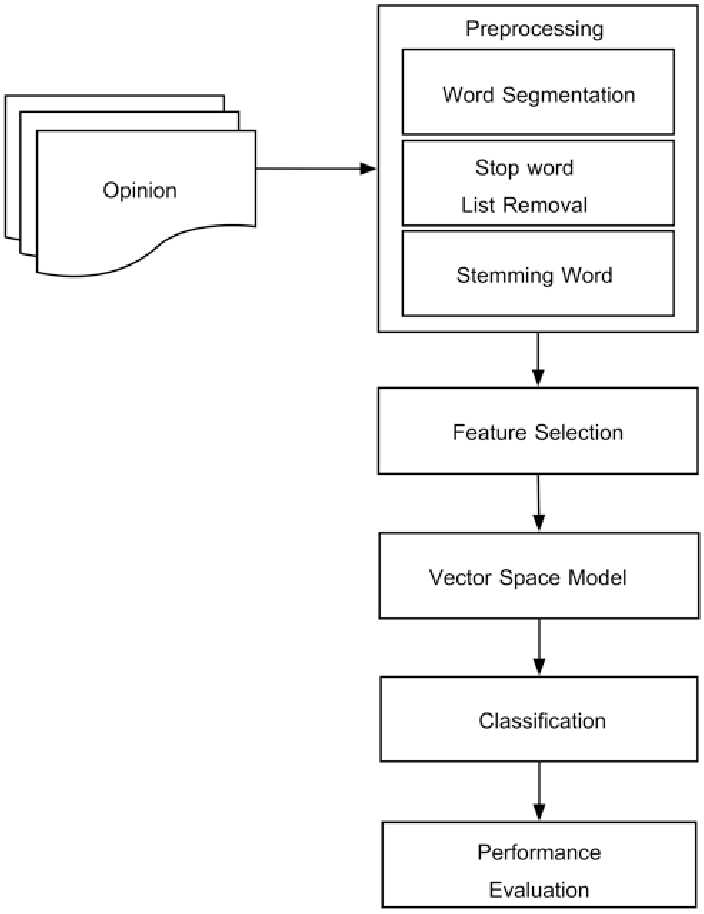

Figure 2. Steps of opinion classification development.

Data Collection

The collection of opinion for feature selection to classify the opinion by collection information from 1,557 sets of curriculum review questionnaire, Faculty of Management Technology, Rajamangala University of Technology Srivijaya. It divides opinion into 3 topics such as 1) curriculum suggestion 2) teaching and learning suggestion and 3) development student activities suggestion.

Data Preprocessing

1) Word Segmentation is the process of cutting apart of words with the longest matching technique. It is separates words from each other, which still has the correct meaning and integrity. It uses the comparison with the longest word [4-5] that find in the dictionary but if not find in the dictionary, it will reduce the length of the word by each 1 letter.

2) Stop Word Removal is a wrapping process on a frequent word. Those words are without the meaning on opinion. When cut off them from the opinion, it will not change the meaning of the opinion including prepositions, conjunctions, pronouns, adverbs and interjection, etc.

Therefore, the stop word removal process is a step that should be taken before create index. This will reduce the size of the index to be smaller size. Moreover, it reduces the storage space and faster processing time [6].

3) Steaming Word is the process of finding the original form of words or improve the words that have same meaning to be the same word. Finding the etymology of a word is a step that should be done before indexing. Finding the etymology of a word is a step that should be done before indexing because it could reduce the index size and improve efficiency of the classification [7].

Feature Selection

Feature selection is a reduce size process or a reduce feature of information to be smaller based on convert the opinion into the same format and replace the feature of the opinion with single syllables and phrases to choose an effective feature for opinion classification. It is obtained weight of the words that appear in the opinion t as feature. For the opinion classification, feature selection will also determine with Term Frequency-Inverse document frequency method (TFIDF), Chi-Square method (CHI) and Information Gain method (IG).

Therefore, feature selection is a reduce size process or a reduce feature of information to be smaller based on choose an effective feature for opinion classification. It is obtained weight of the words that appear in the opinion t as feature. For the opinion classification, feature selection will also determine with Term Frequency-Inverse document frequency method (TFIDF), Chi-Square method (CHI) and Information Gain method (IG).

Weighting by TF-IDF Method (Term Frequency – Inverse Document Frequency)

TF-IDF is a statistical procedure for finding the significance of a word from an opinion in a message set. The finding values of TF-IDF is the most popular method for describing documents in the Vector Space Model. The term frequency of words in an opinion is called the frequency. TF and IDF frequency can be calculate the frequency of the general words or words that less appear in an opinion. It is calculated from the total of opinions divided by the number of words appearing in the opinion. Then, it discovers values with the logarithm and multiply by the term frequency of the words that appear in the opinion. Lastly, it indicated the ratio of the number of times the word appears in the opinion. This is the frequency of all the words that appear in the opinion [8].

TF-IDF is weight finding to determine which word provide the significant meaning for opinion. If it conducted a high weight, that words will be a significant in the opinion sentence.

The weight can be derived as calculate based on equation 1:

$$TF - IDF = TF \times log\left(\frac{N}{DF}\right) \qquad (1)$$

Where:

$\boldsymbol{N}$ is total of number of opinions

$\boldsymbol{DF}$ is number of appearing words in all opinion

$\boldsymbol{TF}$ is number of frequency of words appearing in opinion

Weighting by Information Gain Method (IG)

Information Gain method (IG) is Information Gain (IG) is evaluation of the value of the specific characteristic by measuring from IG value. The increment of value of key word represent as "t" that can be obtained as calculate based on equation 2 [9].

$$IG(t) = E(X) - \sum_{i=1}^{m} p(s_i)E(x_{Si}) \qquad (2)$$

Where: $IG(t)$ is the increment of value of key word t for attribute S

$E(X)$ is entropy of target attribute

$S \in S$ is subset of attribute S when $i \in \{1,2,3,\ldots m\}$

m are all possible subsets of the attribute S

$p(S)$ is probability of subset i for attribute S

Weighting by Chi-Square

Chi-Square (Chi) is a comparison the relationship between variables as comparison is presented the opinion measurement in term of frequency that can classify the opinion. In the classification of opinion, Chi-Square lead to calculate the weight of words from the group of opinion based on comparing the relationship between word and the type of opinion. It will consider both words that appear and do not appear in the opinion. It can be indicated as calculate based on equation 3 [10].

$$x^2(t_k, c_i) = \frac{N[P(t_k, c_i)P(\overline{t_k}, \overline{c_i}) - P(t_k, \overline{c_i})P(\overline{t_k}, c_i)]^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})} \qquad (3)$$

Where: t_k is term

c_j is topic or group of opinion

$P(t_k, c_i)$ is probability of opinion in topic of c_i when appear in term t_k

$P(t_k, \overline{c_i})$ is probability of opinion no in topic of c_i when appear in term t_k

$P(\overline{t_k}, c_i)$ is probability of opinion in topic of c_i when disappear in term t_k

$P(\overline{t_k}, \overline{c_i})$ is probability of opinion no in topic of c_i when disappear in term t_k

Representation the Opinion Through Vector Space Model

Representation the opinion through Vector Space Model is an algorithm that used for opinion classification. Each of opinion will be replace in term of vector format by considering of opinion content from bag words but avoid grammar. It contains each of vector attribute is represented by weight value of term. The weight value is used the frequency of appeared word in opinion based on method calculation that consist of Term frequency inverse document frequency method (TFIDF), Chi-Square method (CHI) and Information Gain method (IG). The weight value will be a guideline for opinion classification in the future.

The opinion text is replaced by the vector format. The size of the vector is based on the number of appeared words in the opinion. The representation the opinion through Vector Space Model must pass through the stop word list removal step and stemming word step before finding the frequency of the unique word. Then, it can be represented text of the opinion into a vector [11]. The value of weight equaled 0 means that the word does not appear in any opinion but if opinion text contained many of unique words that will be expand the vector size however it passed the stop word removal process before.

Thus, it should reduce the size of the vector to be a smaller size with choose only the key words that can be represented of each opinion group.

Table 1. Opinion Representation Model by Word Frequency

Opinion	Weight of Words				
O1	t1	t2	t3	.....	tn
	w1	w2	w3	.....	wn

From table1, the opinion O1 consist of word t1, t2, t3....tn where n is the number of unique words that appeared in the opinion O1 and w is the weight of each words [10].

Opinion 1: I would like to provide more teaching and learning in Python.

Opinion 2: Python should provide to teaching in programming subject.

Opinion 3: A good in programming should practice.

Opinion 4: I would like to provide more practice in programming subject.

Figure 3. Example of opinion.

Table 2. Frequency of appeared word in each opinion

	leaning	teaching	python	programming	practice
Opinion 1	1	1	1	0	0
Opinion 2	0	1	1	1	0
Opinion 3	0	0	0	1	1
Opinion 4	0	1	0	1	1

From Table 2 indicated the frequency of each word appears in the opinion. It was provided frequency value in term of vector as followed; opinion 1 [1 1 1 0 0], opinion 2 [0 1 1 1 0], opinion 3 [0 0 0 1 1], opinion 4 [0 1 0 1 1]. The weight of a word as equaled of 0 means that the word does not appear in any opinion. When analyzed various of the number of opinions, it will create numerus of unique words as well as a result, the size of the vector is large even though pass the stop word removal process and the stemming word process. Thus, it should be selected only keywords that are important and can be as representative of opinion. It will reduce the size of the vector to be a smaller size.

Creation of Opinion Classification Model

This research examines 3 algorithms of machine learning that consist of Naïve-Bayes, KNN: K-Nearest Neighbors, and Support Vector Machine.

1) NAÏVE-Bayes

Naïve-Bayes is a learning algorithm that uses the principle of probability. The model is created in form of probability. Naïve-Bayes is an uncomplicated algorithm which is provided probability for classify the previous knowledge information. It represents P(ai|vj) where ai is the attribute i and vi is the attribute aj. In conclusion, when want to know that which data can be classified to what is the class, the probability is calculated as highest amount that obtained closely the probability of all classes and multiplication of probability for previous knowledge information feature. As calculate based on equation 4 [12].

$$V_{NB} = argmaxP(v_j)\times \prod_{i=1} P\left(a_i|v_j\right) \qquad (4)$$

2) K-Nearest Neighbors

K-Nearest Neighbors is an algorithm used to classify data that provide distribution feature. It will be found the distance of the data, feature that need to forecast with the previous result. For data classification is considered the k value that will be consider the data feature close to the k value. It could be calculation of the distance of the data. As calculate based on equation 5 [12-13].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(X_i) - a_r(X_j))^2} \qquad (5)$$

where: x_i is a new data set

x_j is training data set as j

$a_r(x_i)$ is attribute value of data x_i at r position

$a_r(x_j)$ is attribute value of data x_j at r position

3) Support Vector Machine

Support Vector Machine is an algorithm for creating linear equations to divide the two fields by creating a centerline between the data, the range of the two groups is maximized. Mapping function is obtained to move data from the Input Space to the Feature Space and create a measure function similar to the Kernel Function. As can be seen in Figure 4 [14].

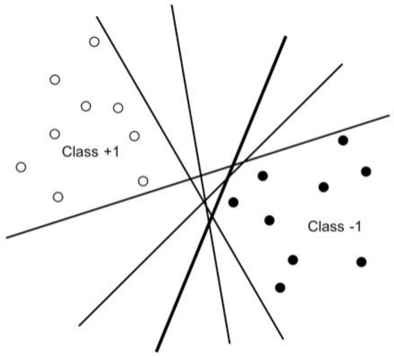

Figure 4. Support Vector Machine Concept.

Model Evaluation

Model Evaluation is a measurement of model performance to predict the opinion classification that consist of 1) Precision 2) Recall 3) F-measure Model

Table 3. Confusion Matrix

Predicted/Actual	Yes	No
Yes	TP	FP
No	FN	TN

1) Precision is measured the precision of the model by class. As calculate based on equation 6.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad (6)$$

2) Recall is a measure of class accuracy. As calculate based on equation 7.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad (7)$$

3) F-measure is measure Recall and Precision simultaneously one by one. As calculate based on equation 8.

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (8)$$

Results

Based on the experiments, the classification of opinions by features selections are obtained Term Frequency-Inverse Document Frequency method, Chi-Square method and Information Gain method. They used 3 algorithms that consist of Naïve Bayes algorithm, K-Nearest Neighbors algorithm, and Support

Vector Machine algorithm based on Split Test technique. Split Test method is obtained to use as classify the opinion by random into 2 part. The first part, 70% used training data to create opinion classification model. The second past, 30% used testing data to perform testing of effective of opinion classification model with compare among Precision, Recall and F-measure. The result is indicated in table 4.

Table 4. The result of effective performance model for opinion classification based on machine learning technique

Threshold	Feature Select	Evaluation	Naïve Bayes	k-NN	SVM
1	TF-IDF	Accuracy	86.67	83.33	90.00
		Precision	78.95	91.67	83.33
		Recall	100.00	73.33	100.00
		F-measure	88.24	81.48	90.91
	Chi Squared	Accuracy	86.67	76.67	86.67
		Precision	82.35	75.00	82.35
		Recall	93.33	80.00	93.33
		F-measure	87.50	77.42	87.50
	Information Gain	Accuracy	83.33	80.00	86.67
		Precision	75.00	80.00	78.95
		Recall	100.00	80.00	100.00
		F-measure	85.71	80.00	88.24

Table 4 indicated that the results of the feature selection by using machine learning techniques and with a threshold value is greater or equal to 1. Threshold value is value of word frequency used to calculate the weight of word due to adjusting the threshold value for finding the most suitable value. In this experiment found that threshold value is greater or equal to 1. It was calculating of the weight of words with Term Frequency-Inverse Document Frequency (TF-IDF). It had an accuracy of 90% and contained a recall value of 100 %. It means that TF-IDF method is the most effective for opinion classification when compare with Chi-Square method and Information Gain method.

Conclusions and Discussion

This research is an experiment research to find the effective features selection based on the calculation the weight of words through Term Frequency-Inverse Document Frequency method (TF-IDF), Chi-Square method (CHI) and Information Gain method (IG). It was considering on the factors affecting the features selection such as the number of comments and threshold value that used in determining the weight of the word. It selects the features of the word group in the opinion that can be used as a representative of the opinion. The representative of the opinion is obtained by vector space mode where

each attribute of vector is represented by the weight of the word. It used machine learning technique such as Naïve Bayes algorithm, K-Nearest Neighbors algorithm and Support Vector Machine algorithm.

The results of the research indicated that the features selection, the method of calculating the weight of a word with Term Frequency-Inverse Document Frequency (TF-IDF) method, Chi-Square method and Information Gain method. When the Threshold value is greater than or equal to 1 for selecting word to define a feature based on the TF-IDF method, the support vector machine algorithm was the most accurate, with an accuracy of 90%. The Chi-square method was 86.67%, and the Information Gain method was 86.67% when compare the result of feature selection. In conclusion, TF-IDF is better than other methods. The number of comments used in learning that affecting on efficiency of features selection. When the number of comments has increased, validity value or feature selection will be increased as well.

References

[1] Bruno Trstenjak; Sasa Mikac, & Dzenana Donko. (2014). KNN with TF-IDF Based Framework for Text Categorization. *Procedia Engineering, 69*, 1356-1364. Retrieved May 9, 2019, from https://www.sciencedirect.com/science/article/pii/S1877705814003750

[2] B.S. Harish, & M.B. Revanasiddappa. (2017, April). A Comprehensive Survey on Various Feature Selection Methods to Categorize Text Document. *International Journal of Computer Applications*, *164*(8), 1-7.

[3] Angelin Florence; Chinmay Vaidya; Devendra Panchal, & Lokesh Negi. (2018, April). Document Classification Using NLP Techniques.*International Journal for Research in Applied Science & Engineering Technology*, *6*(4), 1222-1224.

[4] Rachawit Tipsena; Chatklaw Jareanpon, & Gamgarn Somprasertsri. (2014, September-October). Automatic Question Classification on Web board Using Text Mining Techniques. *Journal of Science and Technology Mahasarakham University*, *33*(5), 493-502.

[5] Supaporn Kurdkit; Aongart Aun a nan, & Phayung Meesad. (2015, July-December). Classification of Reliable Content on Cancer Thai Website using CancerDic+. *Journal of Information Science and Technology*, *5*(2), 34-43.

[6] Ghayda A. Al-Talib, & Hind S. Hassan (2013, October). A Study on Analysis of SMS Classification Using TF-IDF Weighting. *International Journal of Computer Networks and Communications Security*, *1*(5), 189-194.

[7] Rawisuda thetmuang, & Nivet Chirawichitchai. (2017, January-June). Thai Sentiment Analysis of Product Review Online Using Support Vector Machine. *Engineering Journal of Siam University*, *18*(34), 1-12.

[8] Vijayarani S.; Ilamathi j. & Nithay. (2015, February-March). Preprocessing Techniques for Text Mining-An Overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

[9] Pilapan Phonarin; Supot Nitsuwat, & Choochart Haruechiyasak. (2011, April-September). Improvement of Classification for Agriculture Bibliographic Data. *International Journal of Applied Computer Technology and Information Systems*, *1*(1), 60-66.

[10] Ponrudee Netisopakul. (2011). *News Topic Identification using TFIDF and Zipf's Law in Supervised Learning.* Bangkok: King Mongkut's Institute of Technology Ladkrabang.

[11] Vuttichai Vichianchai. (2017, March-April). A Proper Method for a Courses Transfer by Semantic Similarity. *Journal of Science and Technology Mahasarakham University*, *36*(2), 260-264.

[12] Ronnakon Kumpetch, & Wararat Songpan. (2017). Analysis news of rubber prices using text mining technique. *National Conference on Information Technology*, *9*, 80-86. Nakhon Pathom: Mahidol University.

[13] Paramet Tanwanont; Chaiyakorn Yingseree; Worapol Phongphet, & Thanapat Kangkachit. (2017, January-June). Predict stock price trends in Stock Exchange of Thailand Using Ensemble Model. *Journal of information science and technology*, *7*(1), 12-21.

[14] Chalita Chareonnet; Jaree Thongkam, & Sittichai Budsmun. (2014, May-June). Comparison of Data Mining Techniques in Face Recognition. *Journal of Science and Technology Mahasarakham University*, *34*(3), 263-269.