WALAILAK JOURNAL
http://wjst.wu.ac.th Information Technology

Acquiring Sentiment from Twitter using Supervised Learning and Lexicon-based Techniques

Jitrlada ROJRATANAVIJIT[1,*], Preecha VICHITTHAMAROS[1] and Sukanya PHONGSUPHAP[2]

[1]School of Applied Statistics, National Institute of Development Administration, Bangkok 10240, Thailand
[2]Faculty of Information and Communication Technology, Mahidol University, Bangkok 10400, Thailand

([*]Corresponding author's e-mail: jitrlada.roj@stu.nida.ac.th)

Received: 23 June 2016, Revised: 29 November 2016, Accepted: 28 December 2016

Abstract

The emergence of Twitter in Thailand has given millions of users a platform to express and share their opinions about products and services, among other subjects, and so Twitter is considered to be a rich source of information for companies to understand their customers by extracting and analyzing sentiment from Tweets. This offers companies a fast and effective way to monitor public opinions on their brands, products, services, etc. However, sentiment analysis performed on Thai Tweets has challenges brought about by language-related issues, such as the difference in writing systems between Thai and English, short-length messages, slang words, and word usage variation. This research paper focuses on Tweet classification and on solving data sparsity issues. We propose a mixed method of supervised learning techniques and lexicon-based techniques to filter Thai opinions and to then classify them into positive, negative, or neutral sentiments. The proposed method includes a number of pre-processing steps before the text is fed to the classifier. Experimental results showed that the proposed method overcame previous limitations from other studies and was very effective in most cases. The average accuracy was 84.80 %, with 82.42 % precision, 83.88 % recall, and 82.97 % F-measure.

Keywords: Twitter, sentiment analysis, social media content, opinion mining, social media mining

Introduction

The increasing usage of social media has undoubtedly meant it has become even more important as a source of qualitative and quantitative information and has become a most vital source of news and opinions on a wide variety of topics. Social media has become important for the communication and exchange of information. Moreover, social media is a potent feedback channel for companies to use to understand consumers, especially the huge amount of user-generated content steadily increasing on social networking sites [1]. The emergence of social media tools has created a wealth and diversity of textual data, which contain hidden knowledge for businesses to leverage for a competitive edge. In addition, the large amount of feedback on social media coming directly from customers has become a new source from which to mine what is referred to as competitive intelligence. In particular, marketers are able to sift through huge amounts of social media data to discover brand popularity and patterns of interest, so as to achieve a competitive advantage for companies over their competitors [2]. The explosion in the amount of text data has meant that text mining has become a popular method to use to deal with it, and is a helpful tool for companies to gain insight on their customers from social media content. Applications that leverage unstructured data from online public communications to support marketing intelligence and business intelligence are divided in 3 categories: early alerting, buzz tracking, and sentiment mining [3].

Sentiment mining of social media content has become increasingly popular. Researchers from diverse fields have analyzed social media content to generate specific knowledge for their respective subject domains. For example, Gaffney [4] analyzed Tweets with the hashtag #iranElection using histograms, user networks, and the frequencies of top keywords to quantify online activism. A similar study has been conducted on a natural disaster event, Hurricane Sandy. Dong [5] explored the causality correlation between an approaching hurricane and the sentiment of the public towards it. Other research used to gauge business real world outcomes, such as competitive analysis, have been carried out in the pizza industry [2], predicting box office revenues [6], and analyzing business performance [7].

Twitter, one of the most popular social media tools, claims that it has more than 550 million clients, out of which more than 271 million are dynamic [8]. Twitter allows people to broadcast and share real time short messages made of 140 characters, called *Tweets*, which correspond to thoughts or ideas. Many people use it to send updates about their activities, as a tool for conversation, and to share information and report news [9]. Twitter has also become popular in Thailand, with 4.5 million users, which ranks the country at 17th in the world of global Twitter users [10]. Tweets may include one or more entities and reference places in their content. Tweet entities include user mentions (@), hashtags (#), URLs, and media that may be associated with a Tweet, and places are locations in the real world that may be attached to a Tweet [8]. The competitive pricing of smart phones has been an important factor for the widespread use of social media networking by Thais. As a result, Tweets have rapidly become a gold mine of information for companies to monitor their brands and more readily understand their customers by extracting and analyzing sentiment from them. Based on the reasons above, the challenge in this area of research is how to acquire sentiment from the social media content generated by the online social activity of Thais.

Khan, Bashir, and Qamar [11] proposed a new Twitter Opinion Mining (TOM) framework to categorize the polarity of Tweets into positive, negative, or neutral sentiments by applying a variant of the techniques used for Twitter feed analysis and classification. This involved pre-processing steps and a hybrid scheme of classification algorithms. The proposed pre-processing steps included: the removal of URLs, the hashtag symbols, usernames, and special characters; spelling correction using a dictionary; the substitution of abbreviations and slang with expansions; lemmatization; and stop words removal. They proposed a classification algorithm incorporating a hybrid scheme using emoticon analysis, an improved polarity classifier using a list of positive and negative words, and SentiWordNet analysis, as shown in **Figure 1**. In their research, the average accuracy of the TOM framework was 85.7 %, with 85.3 % precision and 82.2 % recall. However, the final results may have been contaminated with news and other information.

For the Thai language, Haruechaiyasak and Kongthon [12] proposed a framework for constructing a Thai language resource for feature-based opinion mining obtained from hotel reviews. They constructed a set of patterns from a tagged corpus, and then automatically extracted patterns and collected more sub-features and polar words from an untagged corpus. Later, Haruechaiyasak *et al.* [13] proposed *S-Sense*, a framework for analyzing sentiment from Thai social media content. They collected data from Twitter posts and the Pantip web board in mobile service domains. Then, they applied the Naïve Bayes algorithm to identify the classifiers models. They manually labeled texts with appropriate intension and sentiment classes. The Lexicon consisted of general terms from the dictionary and clue terms which helped to identify the intension and sentiment. The intension analysis experiment involved training a binary classification model with 2 classes, related and other, to analyze 4 different intensions (announcement, request, question, and sentiment). For sentiment analysis, they trained a binary classification model with 2 classes, positive and negative, and the accuracy was 91.64 %. However, there was the possibility that content could be neutral, which was not considered in the study.

```
Begin
    Input QueryString
    Until the data is retrived from Twitter Streaming API, Do
    Filter English Language Tweets
    Remove Duplicates
    For each tweet, Do
        Procedure Pre-process (tweet)
            Remove URL
            Remove Hashtags
            Remove Username
            Spell Check & Correction
            Replace Slangs
            Replace Abbreviations
            Remove Stop Words
            Lemmatization
            Remove Special Characters
        End Procedure
        Procedure Classification (Refined tweet)
            Classify refined tweet using Enhanced Emoticon Classifier
            IF tweet is classified NEUTRAL
                Classify refined tweet using Enhanced Polarity Classifier
            END IF
            IF tweet is classified NEUTRAL
                Classify refined tweet using SentiWordNet Classifier
            END IF
            Write the classification result to file
        End Procedure
    End Until
End
```

Figure 1 The polarity classification algorithm of the TOM framework.

Materials and methods

In this study, we propose a technique of sentiment analysis of Twitter data generated by Thais, with the main focus being on Tweet classification and on solving data sparsity issues. The challenges are: classification accuracy, sarcasm, word usage variation, and data sparsity problems [11,13]. The reason for these issues is the variety of slang words and other abbreviations used, because of the limit of a Tweet (140 characters). The main idea is to pre-process the raw data and operate variant transformations to deal with the slang, transliterated words, abbreviations, and other noise. Additionally, there are no spaces between words in Thai language, so they must be segmented before being fed to the classifier. For the classification process, we propose a mixed method of supervised learning techniques and lexicon-based techniques to filter Thai opinions and classify them into positive, negative, or neutral sentiments.

There are significant differences between written Thai and English. English has 26 letters, whereas Thai has 44 consonant letters (Thai: พยัญชนะ, *phayanchana*), fifteen vowel symbols (Thai: สระ, *sara*), and 4 tone diacritics (Thai: วรรณยุกต์ or วรรณยุต, *wannayuk* or *wannayut*) [14]. In English, a space is used between words to separate them, and there is punctuation, such as a period (.) to indicate the end of a sentence. In Thai, there are no spaces between words; spaces in Thai content demonstrate the end of a clause or sentence. Therefore, existing text mining and sentiment analysis techniques cannot be directly applied to the Thai language.

Procedure of the proposed method

In this paper, a Thai opinion mining method based on the techniques for Twitter feed analysis and classification is applied. The process is subdivided into 3 modules: (1) data collection, (2) data pre-processing, and (3) classification and evaluation. Tweets are obtained from the Twitter search API [15] using query strings. The data pre-processing is used to extract the Tweets, for text pre-processing, and for

Thai word segmentation. For the classification and evaluation module, the main objective is to identify the polarity of Thai opinion Tweets. The proposed method is shown in **Figure 2**.

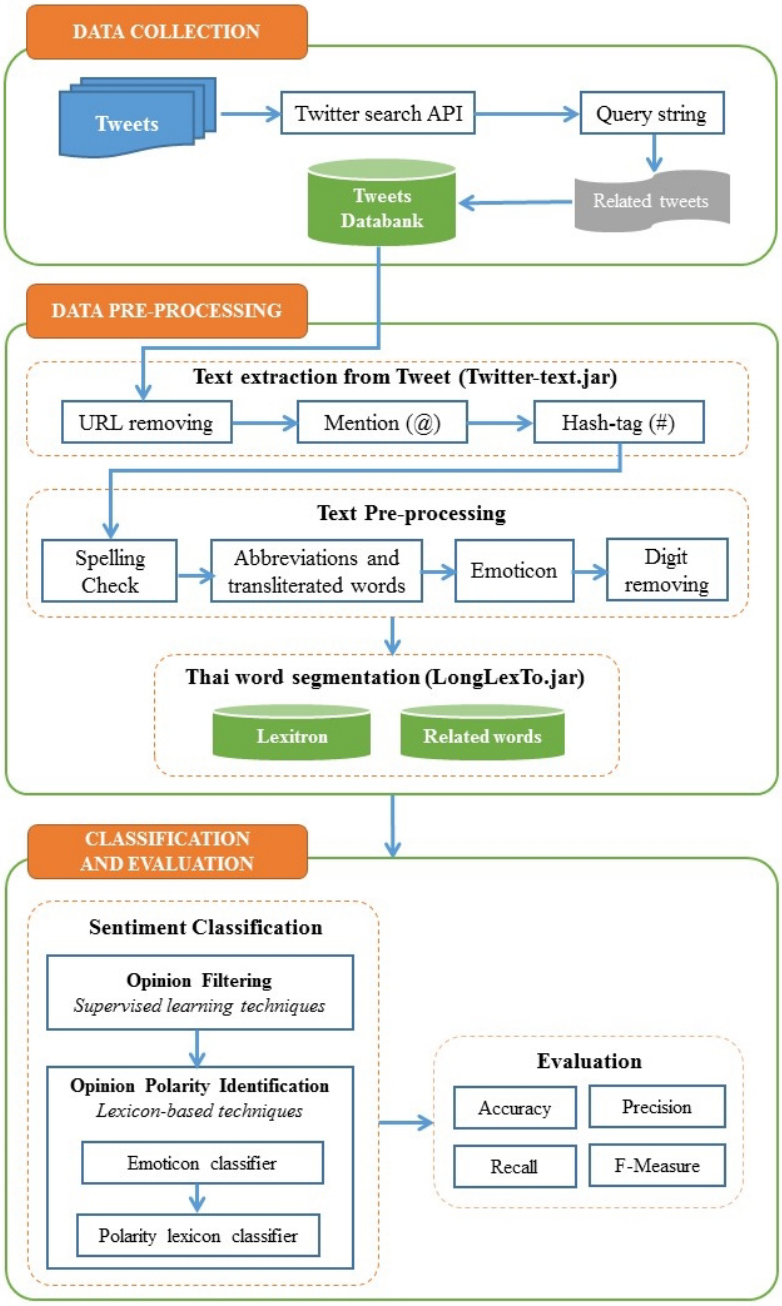


Figure 2 Acquiring Sentiment from Twitter System (ASTS).

Data collection module

A collection program has been developed to search for Tweets of interest keywords from the Twitter feed using the Twitter Search API [15], which allows queries against the indices of recent or popular Tweets. The Twitter Search API is used for this purpose and is configured to extract only Thai language Tweets by setting the language parameter "lang=th" and excluding reTweets. After this, the Tweets is kept in the Tweets databank and acts as input to the data pre-processing module.

Data pre-processing module

Tweets usually include text possibly with special symbols, such as the user mentions (@), hashtags (#), URLs (http links), and so on. Additionally, because of the difference of writing systems between Thai and English (e.g., there being no spaces between words in the Thai language), existing text pre-processing techniques cannot be directly applied to our Thai sentiment classification system. Moreover, a Thai Tweet has slang words, transliterated words, and emoticons, so data pre-processing is necessary before classification of the opinions. Our pre-processing is subdivided into 3 steps, as follows:

1) Text extraction from Tweet

In this step, twitter-text java library [16] is applied to extract identities of Tweets from text messages. There are URLs (http links), user mentions (@) and hashtags (#). We remove all of the URLs (http links). For user mentions, @ is used to indicate a user account; we remove the @ sign with these user mentions, except for those that match keywords. We merely eliminate the # sign, and keep all the hashtag texts.

Example:

Input Tweet: "รู้สึกว่าจะเจอจุดอับสัญญาณเน็ต @TrueMoveH ตรงท่าเรือด่วนคลองแสนแสบ http://t.co/2JHzG5sKKv" ("I feel that there is no signal at Port Saensaeb. @TrueMoveH http://t.co/2JHzG5sKKv")

Output: "รู้สึกว่าจะเจอจุดอับสัญญาณเน็ต TrueMoveH ตรงท่าเรือด่วนคลองแสนแสบ" ("I feel that there is no signal at Port Saensaeb. TrueMoveH")

2) Text pre-processing

In the text pre-processing step, we define 4 types of words as abbreviations, transliterated words, slang words, and misspelled words, the requisite steps being to gather and organize words into their types. We use 1,500 collected Tweets on 3 brands: AIS, DTAC, and TRUEMOVEH as an input source. After this, the method is subdivided into 3 steps, as follows:

(1) Create a new list file for each type.
(2) Read and examine the text in the Tweet. If a word from one of the 4 types is found, it is added to the appropriate list file and assigned the original word.
(3) Continue until 1,500 Tweets have been processed.

The texts that pass from the previous step are automatically checked with the words which are defined as abbreviations, transliterated words, slang words, and misspelling words. Then, they are replaced by expansions or the original words. **Table 1** contains a sample and the number of words of each type discovered in this step. Emoticons are domain and language independent [11], and have become an important token for social media content, since they can express the feelings of the writer in the form of icons [17]. For each Tweet, emoticons are assigned with token labels, as shown in **Table 2**. Then, the final step is to remove all of the digits.

Table 1 Example list of words in text pre-processing step.

Words in Tweet	Original Word	Type	Number of words
พนง., พนง	พนักงาน (employee)	Abbreviations	19
สมาร์ทโฟน, สมาร์ตโฟน	Smartphone	Transliterated words	279
ปังมาก	ดีมาก (very good)	Slang words	217
ใบเส็ด, ใบเส็จ	ใบเสร็จ (receipt)	Misspelling words	17

Table 2 Examples of positive and negative emoticon sets.

Positive Emoticons			Negative Emoticons		
Emoticons	**Meaning**	**Token Label**	**Emoticons**	**Meaning**	**Token Label**
:-) , :) , :D , :o) , :]	Happy	ehappyw	:-(, :(, >:[, :<	Sad	esadw
(^v^) , (^u^) , (^o^) , ^-^	Happy	ehappye	T-T , T^T , '_' , =_=	Sad	esade
:-D, 8-D, XD, =3, B^D	Laugh	elaughw	:'-(, :'(	Cry	ecryw

3) Thai word segmentation

The LongLexTo library was developed by the National Electronics and Computer Technology Center (NECTEC), Thailand [18]. This library has been constructed with a dictionary-based approach, using the longest matching technique. Input text is scanned from left to right, and then the longest match with a word in the dictionary is selected, along with any other matching words, to improve the accuracy of word segmentation. For the Thai word segmentation process, we have modified the LongLexTo java library with a total of 42,833 words: 42,221 words from the Lexitron data dictionary [18], and 612 words from related words in the domains of telecommunication and sentiment [19,20]. Texts passed on from the text-preprocessing step are automatically split into word tokens. For any English words, often included in Thai Tweets, conversion to lowercase is carried out. Lastly, other symbols are removed.

The flexibility of being able to add new words to the dictionary, including English words in common use in Tweet, help to improve the accuracy of the segmentation immensely. Moreover, Tweets are short (140 characters), and so the results of segmentation are better than if applied to longer texts.

Classification and evaluation module

Sentiment classification

In sentiment classification, the main intention is to identify the polarity of the opinions. Our classification system includes the following processes:

- Opinion filtering
- Opinion polarity identification

In this research, the sentiment is divided into positive, neutral, and negative. We bring the procedure classification for sentiment analysis from the TOM framework [11] so as to be able to apply it to Thai Tweets. However, retrieved Tweets from the Twitter Search API are often combined with customer opinions and news. Subsequently, the final results of customer's sentiment may be contaminated with

news and other information. Therefore, we have added an opinion filtering process to classify Tweets that were really opinions from customers before using the procedure classification by the TOM framework.

1) Opinion Filtering

Supervised learning techniques were applied in this task. We have created a classifier to classify Tweets based on opinions and non-opinions by using the WEKA java library [21]. At first, we developed an opinion filtering model by using 1,000 messages (500 opinions, 500 non-opinions) for the training set. In this process, we used emoticons to improve accuracy and then converted strings to word vectors by setting parameter TF-IDF (term frequency-inverse document frequency) [22], as shown in Eq. (1) to Eq. (3), and removing stop words [23]. Next, to construct our classification model, we have used the Multinomial Naïve Bayes (MNB) [24] and the Support Vector Machines (SVM) [25] techniques to create the opinion filtering model.

TF Transform in WEKA library is defined as Eq. (1). It is a measure of information on the frequency of the appearance of a word in a document.

$$\mathrm{TF} = \log_2(1 + f_{ij}) \tag{1}$$

where f_{ij} is the frequency of word i in document (instance) j.

IDF Transform in WEKA library is defined as Eq. (2). It is a measure of how much information the word provides, that is, whether the term is common or rare across all documents.

$$\mathrm{IDF} = f_{ij} * \log_2\left(\frac{num\ of Docs}{num\ of\ Docs\ with\ word(i)}\right) \tag{2}$$

where f_{ij} is the frequency of word i in document (instance) j.

Then TF-IDF is the product of 2 statistics, Term Frequency (TF) and Inverse Document Frequency (IDF), as shown in Eq. (3).

$$\text{TF-IDF} = \mathrm{TF} * \mathrm{IDF} \tag{3}$$

2) Opinion polarity identification

In order to identify the polarity of an opinion, we have modified the procedure classification sentiment in the TOM framework [11] by using 2 classifiers: the enhanced emoticon classifier, and the improved polarity classifier. Due to the limitations of the SentiWordNet classifier for Thai language, as mentioned in [26], we do not use it in this work. The details of our method for opinion polarity identification are as follows:

2.1) The enhanced emoticon classifier

An emoticon is a short sequence of letters and symbols, usually written to express a person's feelings or mood, and the classification of an emoticon is based on sets of positive and negative emoticons. The emoticon is replaced by an emoticon token in a data-preprocessing module. We have used a total of 140 emoticons from Wikipedia [17], 80 of which are tagged as positive and 60 tagged as negative, with each emoticon token having the same weight. Positive and negative emoticon tokens in Tweets are counted and the sum is calculated. Firstly, the sentiment score has an assigned value of 0. Each time a positive emoticon token is found, the score is incremented by 1. On the other hand, if the emotion token is found in the negative set, the score is decreased by 1. The sentiment of opinion is dependent on the sum sentiment score. If the sum sentiment score is greater than zero, it constitutes a positive opinion, and if the sum is less than zero, it comprises a negative opinion. If the sum is zero, it signifies a neutral opinion, and then passes to the polarity lexicon classifier step.

2.2) The improved polarity lexicon classifier

The polarity lexicon classifier uses a 'bag of words' approach, where the set of positive and negative words have been created from the Lexitron dictionary [18], Wiktionary [19], and Thai researchers [20]. The total word count list is 506, which comprises 76 positive words and 430 negative words (sample sets of which are shown in **Table 3**), with each word having the same weight. Each word in a Tweet is checked for both positive and negative word sets in order to calculate the Tweet sentiment score. In the first step, the sentiment score has an assigned value of 0. Each time a positive word is found, the score is incremented by 1. On the other hand, the discovery of a negative word means that the score is decreased by 1. At the end of the process, if the total sentiment score is greater than zero, then the opinion is marked as positive; if less than zero, the opinion is marked as negative; and a total score of zero classifies the opinion as neutral.

Table 3 Positive and negative words samples.

Positive words		Negative words	
ชอบ (like)	ดีมาก (very good)	แย่ (bad)	ห่วย (poor)
ประทับใจ (Impress)	ชมเชย (commend)	แย่มาก (very bad)	กาก (dregs)
รัก (love)	ปลื้ม (delight)	เกลียด (hate)	เฮงซวย (inferior)

Evaluation

Confusion matrix, precision, recall, F-measure, and accuracy are used as measures to evaluate the performance of the proposed method.

Confusion matrix

A confusion matrix contains information about the actual and the predicted class obtained using a classification system. It is a specific table layout that allows the visualization of the performance of an algorithm. **Table 4** shows the confusion matrix, with each column representing the instances in a predicted class, while each row represents the instances in an actual class.

Table 4 Confusion matrix.

Dataset		Predicted Class		Total
		Class A	Class B	
Actual Class	Class A	tpA	eAB	tacA
	Class B	eBA	tpB	tacB
	Total	tpcA	tpcB	N

The entries in the confusion matrix have the following meaning in the context of our study:

- tpA and tpB are the numbers of correct classifications, and are in the diagonal elements in the confusion matrix.
- tacA and tacB are the total number of actual instances of Class A and Class B, respectively.
- tpcA and tpcB are the total number of instances predicted as Class A and Class B, respectively.
- eBA and eAB are the numbers of incorrect classifications.

- N is the total number of instances.

Precision

Precision is defined as the fraction of true positives against all positive results (both true positives and false positives). The equation for calculating the precision of Class A is defined as Eq. (4);

$$\text{Precision of A} = \frac{tpA}{tpA+eBA} \quad (4)$$

where *tpA* is the number of true positives for Class A, and *eBA* is false positives for Class A.

Recall

Recall is defined by the fraction of true positives against all actual classified positives (true positives + false negatives). The equation for calculating recall of Class A is defined as Eq. (5);

$$\text{Recall of A} = \frac{tpA}{tpA+eAB} \quad (5)$$

where *tpA* is the number of true positive for Class A, and *eAB* is the number of false negatives for Class A.

F-measure

The F-measure is defined as a harmonic mean between precision and recall, as shown in Eq. (6).

$$\text{F-measure} = \frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

Accuracy

The equation for calculating accuracy is defined by the proportion of true results (both true positives and true negatives) from all of the given data, as shown in Eq. (7).

$$\text{Accuracy} = \frac{tpA+tpB}{N} \quad (7)$$

Experiment

As mentioned in the data collection section, datasets are collected using the Twitter search API, which allows queries against the indices of recent or popular Tweets [15]. It has been configured to extract only Thai language Tweets by setting the parameter "lang=th" and excluding reTweets. Words relating to mobile network operators are used as keywords for searching, as shown in **Table 5**. Tweets are collected from 1 October 2014 to 31 March 2015 (6 months), with 72,661 Tweets in total, and are kept in a Tweets databank. One thousand (1,000) random Tweets from 1 October 2014 to 30 December 2014 are used to train the model, and 1,500 random Tweets from 1 January 2015 to 31 March 2015 are used to test the model. Tweets that contain more than one brand are excluded, because the message may include a comparative sentence, which is outside the scope of this study.

Table 5 Related keywords for mobile network operators in Thailand.

Mobile Operator Company	Keyword for Query
1. Advanced Info Service Public Company Limited (AIS)	AIS, AIS_Privilege, AIS_Thailand,12Call, เอไอเอส
2. Total Access Communication Public Company Limited (DTAC)	DTAC, Trinet, ดีแทค
3. True Move H Universal Communication Company Limited (TRUEMOVEH)	TRUEMOVEH, TRUEMOVE, ทรูมูฟ
4. Others	3G, 4G, Edge, Wi-Fi

Experiments are conducted on 3 different test datasets of random Tweets on various companies being considered for analysis. **Table 6** shows some examples of positive, negative, neutral, and non-opinion Tweets. The number Tweets for each dataset are shown in **Table 7**.

Table 6 Examples of Tweets.

Sentiment	Keyword	Tweet
Positive	AIS	ลอง net ใหม่ ais เร็วขึ้นนิดนึงยิ้มอ่อน
Negative	DTAC	threeg dtac คือ กากสุดเปิดแล้ว no service
Neutral	TRUEMOVEH	truemoveh ปรับ package fourg ใหม่ให้ มี fup เหมือนเดิมแล้วแฮะ
Non-opinion	DTAC	พนักงาน dtac ต้อนรับซีอีโอคนใหม่อย่างอบอุ่น iphonedroid

Table 7 Sample datasets.

Dataset	Company brand	No. of Tweets for Analyzing Sentiment Classification
Dataset 1	AIS	500
Dataset 2	DTAC	500
Dataset 3	TRUEMOVEH	500

The overall dataset for each class is given in **Figure 3**. Dataset 1 is shown in **Figure 3(a)**; we use a total of 500 Tweets, classified as 63 positive, 226 negative, 111 neutral, and 100 non-opinion. **Figure 3(b)** shows the distribution of dataset 2; we use a total of 500 Tweets, classified as 56 positive, 248 negative, 96 neutral, and 100 non-opinion. Dataset 3 is shown in **Figure 3(c)**; we use a total of 500 Tweets, classified as 72 positive, 256 negative, 72 neutral, and 100 non-opinion. **Figure 3(d)** shows the distribution of the overall dataset; we use a total of 1,500 Tweets, classified as 191 positive, 730 negative, 279 neutral, and 300 non-opinion.

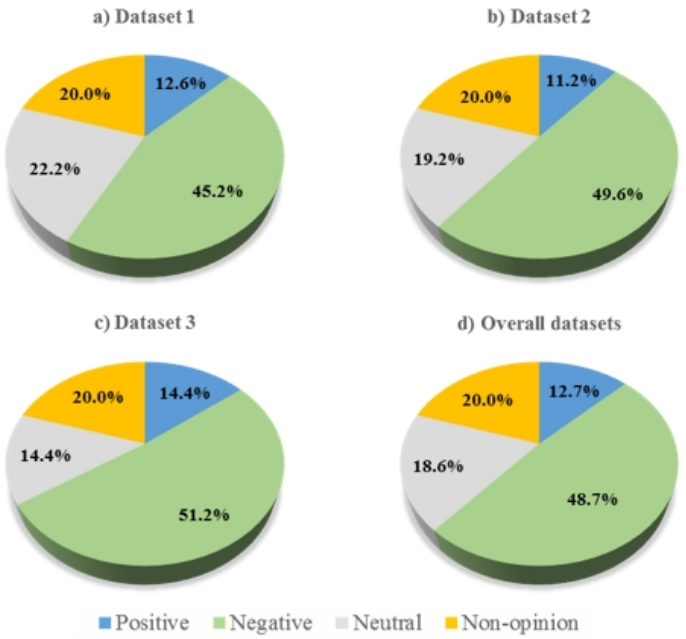


Figure 3 Distribution of overall datasets in non-opinion, positive, negative, and neutral sentiments: a) Dataset 1, b) Dataset 2, c) Dataset 3 and d) Overall datasets.

Results and discussion

The experimental results were evaluated using a confusion matrix for precision, recall, F-measure, and accuracy.

Results of the first experiment

Supervised learning techniques were applied to perform opinion filtering classification. MNB and SVM techniques, running under the WEKA java library environment, were used to identify the classification model and evaluate the performance of the classification. Additionally, previous researches related to Thai opinion did not use emoticons for analysis [12-13,27]. In this research, we showed the advantages of using emoticons. We applied the same training and testing data sets to MNB and SVM techniques, and trained by using 2 binary classification models for each technique with 2 classes, opinion and non-opinion. In the first model, we removed all symbol characters, and in the other, we used emoticon tokens and removed any other symbol characters. Then, we converted strings to word vectors by setting parameter TF-IDF and removing stop words [19]. For training the Tweet set, we prepared random Tweets, 500 opinions and 500 non-opinions, from the Tweet databank, and manually assigned them. We took a random sample of Tweets and categorized them into 500 opinions and 500 non-opinions for testing performance.

The test results were based on 10-fold cross validation, and are shown in **Tables 8**. The overall results of technique testing show that the MNB technique is better than the SVM technique, as seen in **Figure 4.** Additionally, the experimental results show that adding emoticon tokens (Model 2) into the term feature can improve the accuracy of opinion classification with both techniques. For the MNB technique, the accuracy improvement was 2.50 %, and it was able to classify opinion and non-opinion with 91.10 % accuracy. For the SVM technique, the accuracy improvement was 1.60 %, and the accuracy was 86.60 %.

Table 8 Results of opinion and non-opinion classification with the MNB and SVM techniques.

Test Set (500 Opinions and 500 Non-opinions)		Total	MNB technique			SVM technique		
			Confusion Matrix		Accuracy	Confusion Matrix		Accuracy
			Opinion	Non-opinion		Opinion	Non-opinion	
Model 1 (no emoticons)	Non-opinion	500	463	37	88.60 %	420	80	85.00 %
	Opinion	500	77	423		70	430	
	Total	1,000	540	460		490	510	
Model 2 (emoticons)	Non-opinion	500	459	41	91.10 %	408	92	86.60 %
	Opinion	500	48	452		42	458	
	Total	1,000	507	493		450	550	

Result of the second experiment

The second experiment was to classify the sentiment of Tweets. From the first experiment, the MNB technique showed the better accuracy for opinion classification than the SVM technique. Therefore, the MNB technique was selected to be used in the process of opinion filtering, and then to identify the polarity of opinions. **Figure 5** shows our process, which was extended from the TOM method (**Figure 5(a)**). The differences were in the classification module, in which we included the opinion filtering module by using the MNB algorithm (see **Figure 5(b)**).

The classification was run over the test datasets and each Tweet processed (the number of Tweets in each dataset are shown in **Table 7**). Each Tweet in the dataset was classified as either positive, negative, neutral, or non-opinion, of which the distributions of data are shown in **Figure 3**. We applied the same training and testing datasets to evaluate our proposed method and TOM.

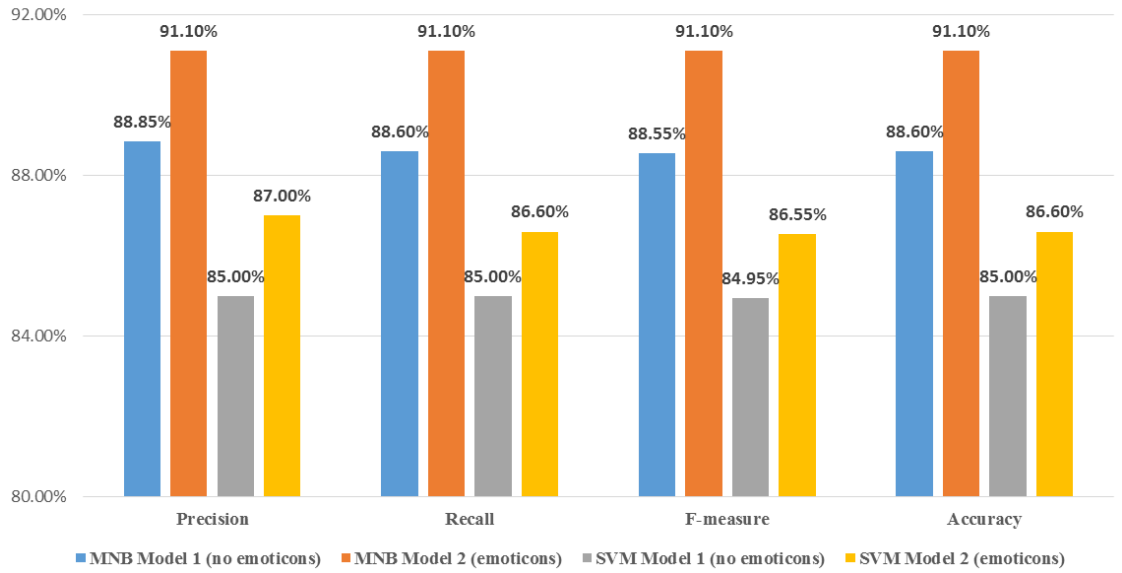
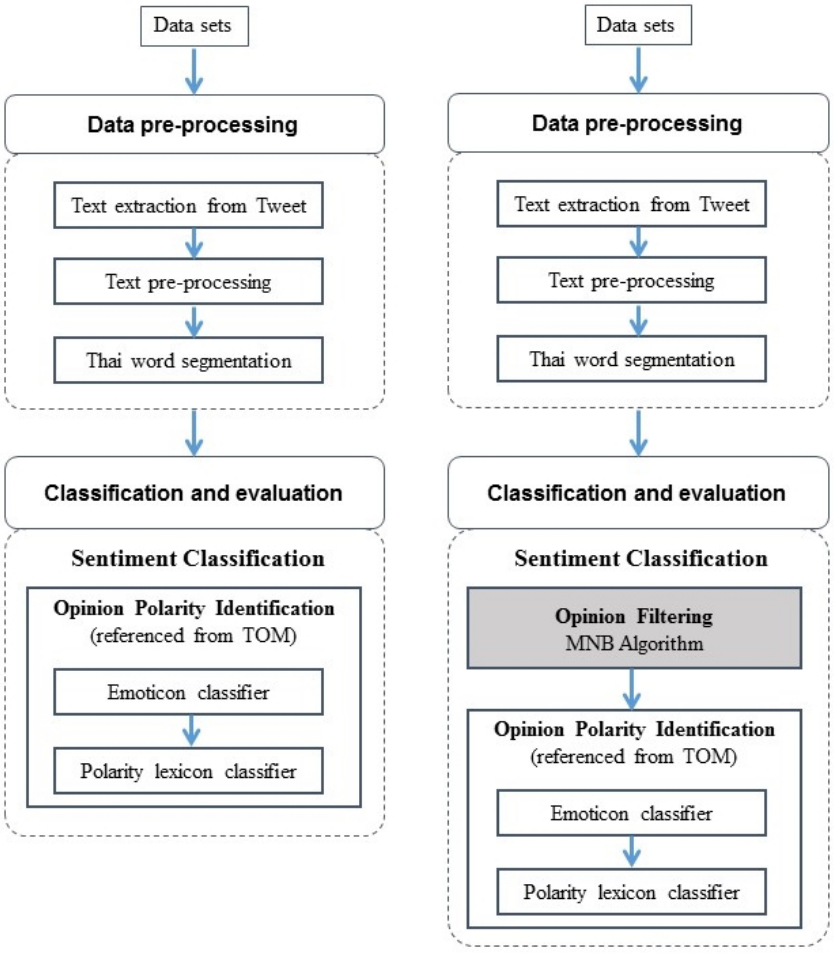


Figure 4 Precision, recall, and F-measure of MNB and SVM techniques.



(a) Base on TOM concept (b) Proposed method

Figure 5 Processes of sentiment classification by TOM and the proposed method.

Tables 9 - 11 show the results for all 3 datasets. They demonstrate that the proposed method showed better performance for classification. The results of average precision, recall, F-measure, and accuracy of the proposed method for each dataset is shown in **Figure 6**. From the results, the proposed method was very effective in most cases, with 84.80 % accuracy (an improvement of 18.67 %), 82.42 % precision, 83.88 % recall, and 82.97 % F-measure.

From our deep analysis, we found that there were non-opinion Tweets (errors) spread in the positive, negative and neutral categories, which means a low accuracy for the original TOM, with 66.13 % accuracy. An example of a non-opinion Tweet is "AIS จับมือ CIMB เปิดตัวบริการใหม่ beat banking เพิ่มความสะดวก ในการทำธุรกรรมการเงิน ผ่าน MPAY" ("AIS collaborates with CIMB bank to launch the beat banking service for creating convenient financial transactions through MPAY").

The results confirm that our method can significantly improve the original TOM based method. This comes from our main contribution of adding the opinion filtering module. A graphical representation of the improvements of the results is illustrated in **Figure 7**, which clearly shows that opinion filtering helps to analyze Tweets more accurately. In addition, we can make use of the filtering results, which are

roughly classified into 2 groups of opinion Tweets and non-opinion Tweets, for other purposes or applications; for example, Tweets that are relevant to a particular company can be applied to brand monitoring and other business indicators.

Table 9 Results of sentiment classification for Dataset 1.

Dataset 1		Confusion matrix				Total	Results			
		Positive	Negative	Neutral	Non-opinion		Precision	Recall	F-measure	Accuracy
TOM Concept	Positive	47	4	12	N/A	63	51.60 %	74.60 %	61.00 %	65.20 %
	Negative	7	192	27	N/A	226	88.50 %	85.00 %	86.70 %	
	Neutral	12	13	86	N/A	111	44.80 %	77.50 %	56.80 %	
	Non-opinion	25	8	67	N/A	100	N/A	N/A	N/A	
	Total	91	217	192	N/A	500				
Proposed Method	Positive	47	4	12	0	63	70.10 %	74.60 %	72.30 %	84.00 %
	Negative	7	192	27	0	226	91.90 %	85.00 %	88.30 %	
	Neutral	12	13	86	0	111	67.20 %	77.50 %	72.00 %	
	Non-opinion	1	0	3	96	100	100.00 %	96.00 %	98.00 %	
	Total	67	209	128	96	500				

Error analysis

To analyze errors, we examine the test instances which are misclassified. We can summarize 3 major causes of errors as being word sense ambiguity, new slang words, and sarcasm. The first problem appears when a word contains many meanings, depending on the context. For example, the word "เร็ว" (quick), when used with "สัญญาณ" (signal), will give positive polarity. On the other hand, when used with "ทวงเงิน" (debt collecting), it will give negative polarity. To solve this problem, associated words will be considered in order to identify the polarity of the opinion.

The second problem is the making of new slang words, which is a new trend for Thais using social media. This problem has more of an effect in Thai language, because there are no spaces between words. For example, the word "แรงงง" ("strong") will be split into "แรง งง" ("strong confuse"). Moreover, they give original words a new meaning. For example, the word "หอย", which originally means "shellfish", is given the meaning of "lower speed" in another context. The solution for this problem is to more often add and update new words in the database of the Thai word segmentation and polarity lexicon. In addition, the contexts in a business domain will be considered.

The third problem is sarcasm, in which is difficult to detect the polarity of opinions. It is always composes with 2 sentences, but with different polarities. For example, Tweet "AIS 3G ครอบคลุมทุกจังหวัด แต่ที่บ้านเรา ไม่มีสัญญาณสักขีด" ("AIS 3G cover all provinces, but there is no signal at my home") is considered to be a sarcastic sentence. In this situation, the Tweet will be mostly classified as a neutral opinion. However, it is still difficult to solve, and is a challenging task in sentiment analysis [13].

Table 10 Results of sentiment classification for Dataset 2.

Dataset 2		Confusion matrix				Total	Results			
		Positive	Negative	Neutral	Non-opinion		Precision	Recall	F-measure	Accuracy
TOM Concept	Positive	47	3	6	N/A	56	54.70 %	83.90 %	66.20 %	65.20 %
	Negative	6	213	29	N/A	248	88.80 %	85.90 %	87.30 %	
	Neutral	9	21	66	N/A	96	37.90 %	68.80 %	48.90 %	
	Non-opinion	24	3	73	N/A	100	N/A	N/A	N/A	
	Total	86	240	174	N/A	500				
Proposed method	Positive	47	3	6	0	56	74.60 %	83.90 %	79.70 %	84.00 %
	Negative	6	213	29	0	248	89.90 %	85.90 %	87.80 %	
	Neutral	9	21	66	0	96	62.30 %	68.80 %	65.30 %	
	Non-opinion	1	0	5	94	100	100.00 %	94.00 %	96.90 %	
	Total	63	237	106	94	500				

Table 11 Results of sentiment classification for Dataset 3.

Dataset 3		Confusion matrix				Total	Results			
		Positive	Negative	Neutral	Non-opinion		Precision	Recall	F-measure	Accuracy
TOM concept	Positive	63	1	8	N/A	72	58.30 %	87.50 %	70.00 %	68.20 %
	Negative	6	225	25	N/A	256	91.80 %	87.90 %	89.80 %	
	Neutral	6	13	53	N/A	72	36.10 %	73.60 %	48.40 %	
	Non-opinion	33	6	61	N/A	100	N/A	N/A	N/A	
	Total	108	245	147	N/A	500				
Proposed method	Positive	63	1	8	0	72	78.80 %	87.50 %	82.90 %	86.20 %
	Negative	6	224	25	1	256	93.70 %	87.50 %	90.50 %	
	Neutral	6	13	53	0	72	59.60 %	73.60 %	65.80 %	
	Non-opinion	5	1	3	91	100	98.90 %	91.00 %	94.80 %	
	Total	80	239	89	92	500				

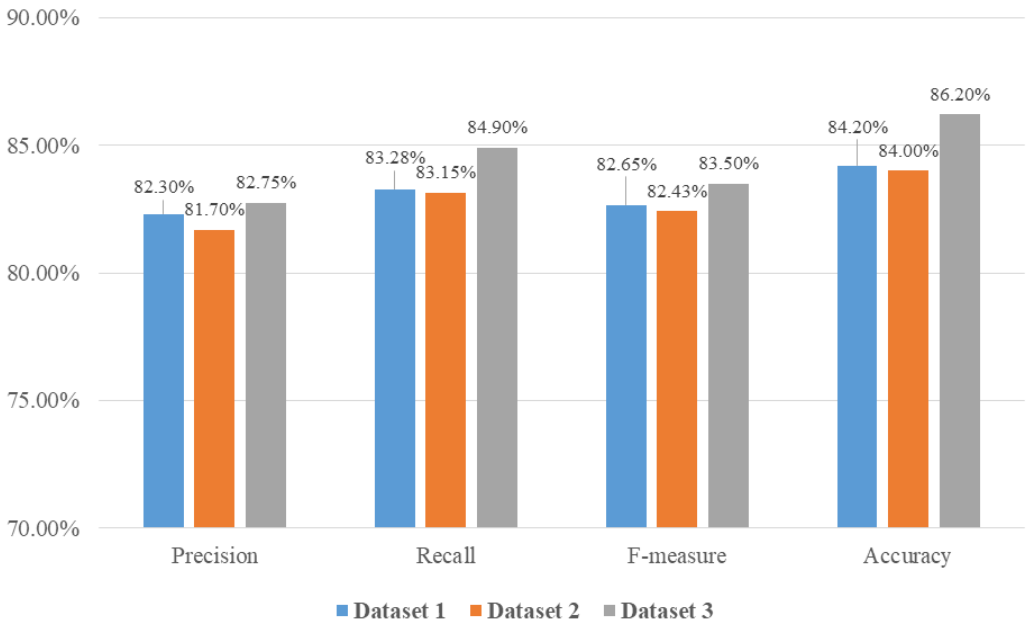


Figure 6 Average of precision, recall, F-measure, and accuracy of the proposed method.

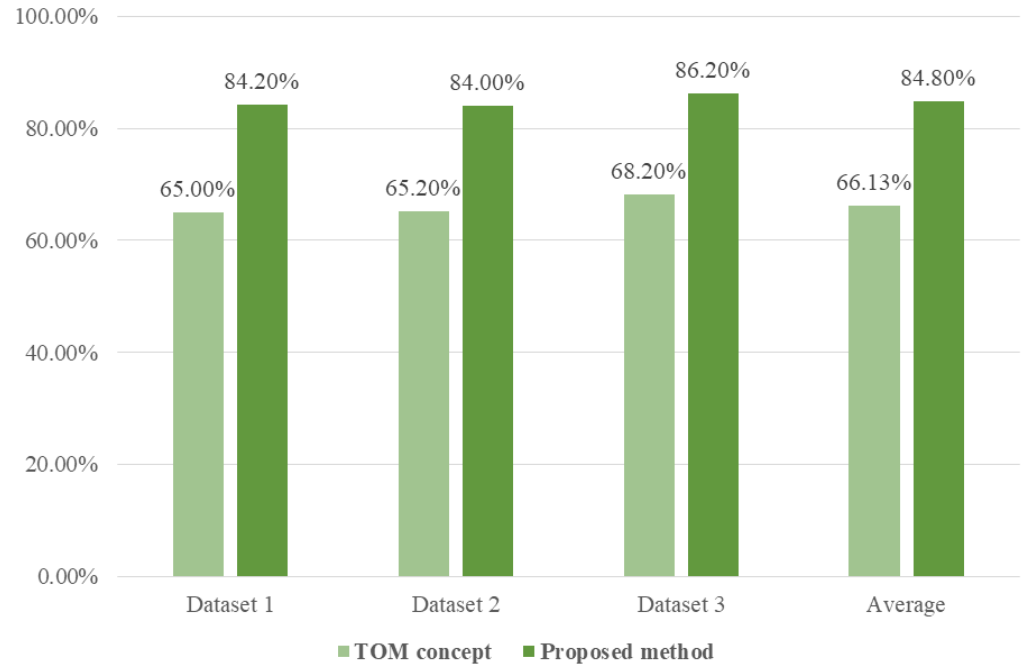


Figure 7 Performance of the proposed method for sentiment classification and its improvement from the TOM method.

Conclusions

In this research, we proposed a new method for twitter sentiment analysis by using both supervised learning techniques and lexicon-based techniques. Experiments were conducted on social media data, Tweets, in the domains of mobile network operators obtained from Twitter search API, focused on Thai. The results of testing the proposed method show significant improvement of the basic concept of using the TOM framework. We achieved an average accuracy of 84.80 %. This shows great improvement (an improvement of 18.67 %) from the original TOM framework, with 66.13 % accuracy. In particular, it clearly shows that opinion filtering helps to analyze Tweets more accurately. Moreover, we can make use of the filtering results for other applications. For example, Tweets that are relevant to a particular company will be useful for various applications, such as brand monitoring, campaign monitoring, competitive analysis, and customer engagement. However, there are a number of limitations, which also leads to many possible directions for future works, such as analysis of comparative sentences which contain more than one brand. In addition, applying this proposed method in other business domains is challenging.

References

[1] C Zinner and C Zhou. *Social Media and the Voice of the Customer*. *In:* N Smith, R Wollan and C Zhou (eds.). The Social Media Management Handbook: Everything You Need to Know to Get Social Media Working in Your Business. John Wiley & Sons, New Jersey, 2011, p. 67-70.

[2] W He, S Zha, and L Li. Social media competitive analysis and text mining: A case study in the pizza industry. *Int. J. Inform. Manag.* 2013; **33**, 464-72.

[3] N Glance, M Hurst, K Kigam, M Siegler, R Stockton and T Tomokiyo. Deriving marketing intelligence from online discussion. *In:* Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Washington DC, 2005, p. 419-28.

[4] D Gaffney. #iranElection: Quantifying online activism. *In*: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, North Carolina, 2010.

[5] H Dong. 2013, Social Media Data Analytics applied to Hurricane Sandy. Master's Thesis. University of Maryland, Maryland, USA.

[6] S Asur and BA Huberman. Predicting the future with social media. *In:* Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Washington DC, 2010, p. 492-9.

[7] J Paniagua and J Sapena. Business performance and social media: Love or hate? *Bus. Horizons* 2014; **57**, 719-28.

[8] Wikipedia: "Twitter", Available at http://en.wikipedia.org/wiki/Twitter, accessed November 2014.

[9] A Java, X Song, T Finin and B Tseng. Why we Twitter: Understanding microblogging. *In:* Proceedings of the Joint 9th WebKDD and 1st SNA-KDD 2007 Workshop, San Jose, California. 2007, p. 56-65.

[10] S Sakawee. Thailand Social Media Stats, Available at https://www.techinasia.com/thailand-social-media-stats-28-million-facebook-45-million-twitter-17-million-instagram, accessed October 2014.

[11] F H Khan, S Bashir and U Qamar. TOM: Twitter opinion mining framework using hybrid classification scheme. *Decis. Support Syst.* 2014; **57**, 245-57.

[12] C Haruechaiyasak and A Kongthon. Constructing Thai opinion mining resource: a case study on hotel reviews. *In:* Proceedings of the 8th Workshop on Asian Language Resources, Beijing, China. 2010, p. 64-71.

[13] C Haruechaiyasak, A Kongthon, P Palingoon and K Trakultaweekoon. S-Sense: A sentiment analysis framework for social media sensing. *In:* Proceedings of the Workshop on Natural Language Processing for Social Media, Nagoya, Japan, 2013, p. 6-13.

[14] Wikipedia: "Thai alphabet (in Thai)", Available at https://en.wikipedia.org/wiki/Thai_alphabet, accessed January 2016.

[15] Twitter Developers: "Twitter Developer Documentation", Available at https://dev.twitter.com/rest/public, accessed August 2014.
[16] C Goncalves. GitHub Inc: Twitter-text Library, Available at https://github.com/twitter/twitter-text, accessed January 2015.
[17] Wikipedia: "List of Emoticon", Available at https://en.wikipedia.org/wiki/List_of_emoticons, accessed January 2015.
[18] NECTEC: "LexTo - Thai Lexeme Tokenizer (in Thai)", Available at http://www.sansarn.com/lexto, accessed August 2014.
[19] Wiktionary: "The Free Dictionary (in Thai)", Available at https://th.wiktionary.org, accessed August 2015.
[20] O Chinakarapong. Conceptual metaphor of Thai curse words (*in Thai*). *J. Hum. Fac. Hum. Naresuan Univ.* 2014; **11**, 57-76.
[21] WEKA: "Data Mining Software in Java", Available at http://www.cs.waikato.ac.nz/ml/weka, accessed March 2015.
[22] WEKA: "Text categorization with WEKA", Available at https://weka.wikispaces.com/Text+categorization+with+WEKA, accessed March 2015.
[23] V Kasorn. 2010, Similarity Measurement of Thai Document using Natural Language Processing (*in Thai*). Independent Study. Chiang Mai University, Chiang Mai, Thailand.
[24] A Bifet and E Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. *In*: Proceedings of 13th International Conference on Discovery Science, Canberra, Australia. 2010, p. 1-15.
[25] B Liu. *Sentiment Analysis and Opinion Mining, Draft*. Morgan & Claypool Publishers, 2012, p. 31.
[26] AsianWordNet Project: "Thai WordNet", Available at http://awn.iisilab.org, accessed January 2016.
[27] W Wunnasri, T Theeramunkong and C Haruechaiyasak. Solving unbalanced data for Thai sentiment analysis. *In:* Proceedings of the 10th International Joint Conference on Computer Science and Software Engineering, Mahasarakham, Thailand, 2013, p. 200-5.