Social Media Text Classification by Enhancing Well-Formed Text Trained Model

Phat Jotikabukkana[1,*], Virach Sornlertlamvanich[1], Okumura Manabu[2] & Choochart Haruechaiyasak[3]

[1]School of ICT, Sirindhorn International Institute of Technology,
Thammasat University, Pathum Thani, Thailand
[2]Tokyo Institute of Technology, Ookayama Campus,
Ookayama Meguro-ku, Tokyo, Japan
[3]National Electronics and Computer Technology Center,
Thailand Science Park, Pathum Thani, Thailand
*E-mail: phat.jotikabukkana@studentmail.siit.tu.ac.th

Abstract. Social media are a powerful communication tool in our era of digital information. The large amount of user-generated data is a useful novel source of data, even though it is not easy to extract the treasures from this vast and noisy trove. Since classification is an important part of text mining, many techniques have been proposed to classify this kind of information. We developed an effective technique of social media text classification by semi-supervised learning utilizing an online news source consisting of well-formed text. The computer first automatically extracts news categories, well-categorized by publishers, as classes for topic classification. A bag of words taken from news articles provides the initial keywords related to their category in the form of word vectors. The principal task is to retrieve a set of new productive keywords. Term Frequency-Inverse Document Frequency weighting (TF-IDF) and Word Article Matrix (WAM) are used as main methods. A modification of WAM is recomputed until it becomes the most effective model for social media text classification. The key success factor was enhancing our model with effective keywords from social media. A promising result of 99.50% accuracy was achieved, with more than 98.5% of Precision, Recall, and F-measure after updating the model three times.

Keywords: *online news; semi-supervised learning; social media text; well-formed text; Term Frequency-Inverse Document Frequency (TF-IDF) weighting; Word Article Matrix (WAM).*

1 Introduction

In these days, social media have a huge impact on our life in many dimensions: socialization, business, politics, etc. It is the most popular digital communication tool to spread ideas and information. The present number of active social media accounts is around 2 billion out of all 7.2 billion people in the world [1]. This novel source of data is very attractive for researchers and

Received October 29th, 2015, 1st Revision March 23rd, 2016, 2nd Revision July 1st, 2016, Accepted for publication July 1st, 2016.
Copyright © 2016 Published by ITB Journal Publisher, ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2016.10.2.6

decision makers. In our research, we focused on Twitter as a social media text source for two reasons. Firstly, Twitter is one of the most popular social media applications and it is the fastest growing one. Referring to recent statistics, there are 316 million monthly active users who generate 500 million Twitter messages (tweets) per day [2]. Twitter had 95% growth in active users and 35% growth in members through 2014 [3]. These are significant numbers, which means that we will gain a lot of benefit if we can extract the essence from this data type. Secondly, the data structure of Twitter and their support API are convenient for researchers to operate with. Tweets are text files limited to 140 characters presented in JSON file format [4] and a Twitter search API can be used to retrieve tweets with a rate limited to 180 queries per 15-minute window [5] and a seven-day search back [6].

The main challenge is to analyze social media text. Since tweets are short text messages, they look like colloquial text compared to written documents. The data stream contains a large amount of noisy and unstructured information, informal language, slang and missing words. This makes text classification in order to distinguish categories before extracting useful information very difficult. In our experiment we therefore applied a technique to extract keywords using Term Frequency-Inverse Document Frequency (TF-IDF) and Word Article Matrix (WAM) to expand the set of keywords reflecting the nature of the texts retrieved from Twitter. We collected data and created word vectors from an online news source consisting of well-formed text, which was already categorized beforehand by publishers to extract keywords and classify information from Twitter. Semi-supervised machine learning can solve self-assigned classes labeling for topic classification problems. The computer can automatically extract categories from the news website for use as proper class-labels with a sense of human familiarity, such as Economic, Entertainment, Foreign, Information Technology (IT), Politics, Regional, Sports, etc. Finally, we get a productive set of keywords from the official site and Twitter, which can be representative for text categories. New words, abbreviations and argot that never appear in well-formed documents are extracted from the tweet messages, which can then be used as the main keywords that reflect interesting topics in society at a certain moment in time.

This paper is organized as follows. Related research works are discussed in Section 2. In Section 3, we explain our approach and the main techniques that were used. The experimental results are shown in Section 4. Finally, in Section 5, we draw the overall conclusions from the experiment, including a brief discussion of future work regarding the effectiveness of social media text classification.

2 Related Works

A number of recent papers address social media text classification. Irfan, *et al.* [7] reviewed different text mining techniques to discover various textual patterns from the social web. Text mining using classification with various machine learning-based and ontology-based algorithms and a hybrid approach were reviewed. There is no algorithm that performs best for all kinds of data sets. For better performance of the hybrid approach, several parameters need to be defined in advance. Patel, *et al.* [8] reviewed different types of classifiers for text classification with an eye on their advantages and disadvantages. Six different algorithms were reviewed: Bayesian Classifier, Decision Tree, K-nearest neighbor (K-NN), Support Vector Machine (SVM), Neural Network, and Rocchio's. The common disadvantage of all algorithms is their performance limitations. Some of them are easy to implement but their performance is very poor. Some of them perform greatly but need extra time for training and parameter tuning. Lee, *et al.* [9] classified Twitter Trending Topics with two approaches for topic classification: the well-known Bag-of-Words approach for text classification and network-based classification. They identified 18 classes and classified trending topics into these categories. In the final result, the network-based classifier performed significantly better than the text-based classifier. Kateb, *et al.* [10] discuss methods that overcome problems in classifying short texts from streaming data in social media. In the classification techniques section they present some common issues that are useful in general to address before conducting text classification: 1. Define the research goal. 2. Does speed matter? 3. What is the size of the data? On the basis of these simple considerations we selected a suitable technique (classification, regression or clustering) and suitable algorithm for conducting our experiment.

Chirawichitchai, *et al.* [11] compared six methods of feature weighting in a Thai document categorization framework. They found that ltc weighting with SVM yielded the best performance for Thai document categorization. Theeramunkong, *et al.* [12] proposed a multidimensional framework for classifying text documents. Classifying text documents based on a multidimensional category model by using multidimensional-based and hierarchy-based classifications beat flat-based classification. Viriyayudhakorn, *et al.* [13] compared four divergent thinking support engines using associative information extracted from Wikipedia. They used Word Article Matrix (WAM) to compute the association function. This is a useful and effective technique for divergent thinking support. Sornlertlamvanich, *et al.* [14] proposed a new method for fine-tuning a model trained with some known documents containing richer context information. They used WAM to classify text and track keywords from social media to understand social developments. WAM with cosine similarity measure is an effective method of text classification.

The related literature revealed that there are many different techniques for social media text classification. All algorithms still have complex issues related to their performance. This inspired us to adapt some useful techniques in a novel, simple way to effectively classify social media text with a sense of human familiarity.

3 Experiment

Viriyayudhakorn, *et al.* [13] and Sornlertlamvanich, *et al.* [14] used WAM only for their specific purposes (divergent thinking support and keyword tracking), while the present study focused on using WAM to classify social media text with additional techniques of text class self-learning (semi-supervised learning) and enhancing the WAM model with specific keywords from social media until it becomes the most suitable model for social media text classification.

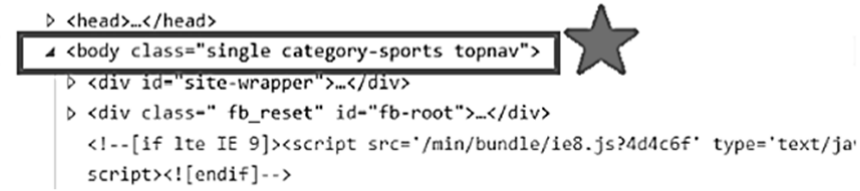
The rest of this paper consists of two main parts: first, the main techniques that were used are discussed. Second, our approach for effective social media text classification is explained.

3.1 Main Techniques

3.1.1 Web Crawler

For the initial state, we need to retrieve news articles from an online news source consisting of well-formed text. A web crawler, also known as a robot or a spider [15], is the main module to get access to the data source. Because most websites today are implemented with hypertext markup language (HTML), extensible markup language (XML) and cascading style sheet (CSS), the structure of the targeted website must be verified. Then, the uniform resource locator (URL), the news category and the news article part, which are needed as main parameters, have to be specified. Afterwards, these parameters are applied through the XML Path (XPath) query technique to retrieve the demanded data, i.e. the online news articles. The RapidMiner software application [16] was used as the main web crawler module.

As an example, the XPath query command was used to automatically extract the news category (*sports*) from the part "<body class= "single category-sports topnav">" as shown in Figure 1. This is an effective semi-supervised learning method. The computer can extract the news category even if the publisher decides to change the category label related to an article.



```
▷ <head>…</head>
◢ <body class="single category-sports topnav">
  ▷ <div id="site-wrapper">…</div>
  ▷ <div class=" fb_reset" id="fb-root">…</div>
    <!--[if lte IE 9]><script src='/min/bundle/ie8.js?4d4c6f' type='text/ja
    script><![endif]-->
```

Figure 1 Example code of an online news article (from http://dailynews.co.th/) [17].

3.1.2 Word Segmentation

The experiment in this study was conducted on social media text written in the Thai language. Word segmentation is a crucial factor in text mining. However, the Thai language is written without spaces between words. Therefore, a word segmentation module was used, applying the maximal matching algorithm to determine the word boundaries [18]. The recent word list in the dictionary was updated before the research was conducted. Consequently, the segmentation result was acceptable for determining the essential words for further processing in keyword identification.

3.1.3 Term Frequency-Inverse Document Frequency (TF-IDF)

At present, there are many weighting schemes for text mining: Boolean weighting, Term Frequency (TF) weighting, TF-IDF weighting, tfc weighting, ltc weighting, and Entropy weighting [11]. TF-IDF is the most widely used technique to extract keywords from documents. It is composed of 2 steps: Term Frequency (TF) and Inverse Document Frequency (IDF). TF is computed from the number of times a word appears in a document, divided by the total number of words in that document. It can be defined as a counting function in Eq. (1) [19].

$$\mathrm{TF}(\mathrm{t},\mathrm{d}) = \sum_{\mathrm{x} \in \mathrm{d}} \mathrm{fr}(\mathrm{x},\mathrm{t}) \tag{1}$$

$\mathrm{TF}(\mathrm{t},\mathrm{d})$ is actually the total number of term t appearing in document d and $\mathrm{fr}(\mathrm{x},\mathrm{t})$ is a simple function defined as Eq.(2):

$$\mathrm{fr}(\mathrm{x},\mathrm{t}) = \begin{cases} 1, & \text{if } \mathrm{x} = \mathrm{t} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

IDF is defined as the logarithm of the number of all documents in a collection divided by the number of documents in which the observed term appears in Eq. (3).

$$\mathrm{IDF(t)} = \log\frac{|\mathrm{D}|}{1+|\{\mathrm{d:t} \in \mathrm{d}\}|} \tag{3}$$

$1 + |\{\mathrm{d: t} \ \in \ \mathrm{d}\}|$ is the number of documents in which the term t appears. When the term-frequency function satisfies $\mathrm{TF(t, d)} \neq 0$, we apply "1 +" to avoid divide by zero. Then, the TF-IDF formula can be defined as (4):

$$\mathrm{TF\text{-}IDF(t)} = \mathrm{TF(t, d)} \ \times \ \mathrm{IDF(t)} \tag{4}$$

3.1.4 Term Frequency Merging (TF-Merging)

Term Frequency (TF) is one of the weighting techniques that can be used to identify the importance weight of words related to their corpus. However, we cannot directly merge TF values of words that appear in more than one corpus. Each corpus contains a different total number of words with different weights. For example, the word "AAA" in corpus 1 has a TF value of 0.5, while the total number of words in Corpus 1 is 100 words. The word "AAA" in Corpus 2 has a TF value of 0.4, while the total number of words in Corpus 2 is 1,000 words. We cannot find the TF value of the word "AAA" by adding their 2 TF values directly because we are considering two different corpus weights. The way to solve this problem is using word vector normalization (TF normalization) (5), (6) and TF-merging (7).

$$||\vec{\mathrm{V}}||_2 = \sqrt{\mathrm{V_1}^2 + \mathrm{V_2}^2 + \mathrm{V_3}^2 + \cdots + \mathrm{V_n}^2} \tag{5}$$

$||\vec{\mathrm{V}}||_2$ is the L2-normalization factor, Euclidean norm. $V_1, V_2, \ldots, V_n$ are the term of occurrence of each word (i) for (n) words in the corpus.

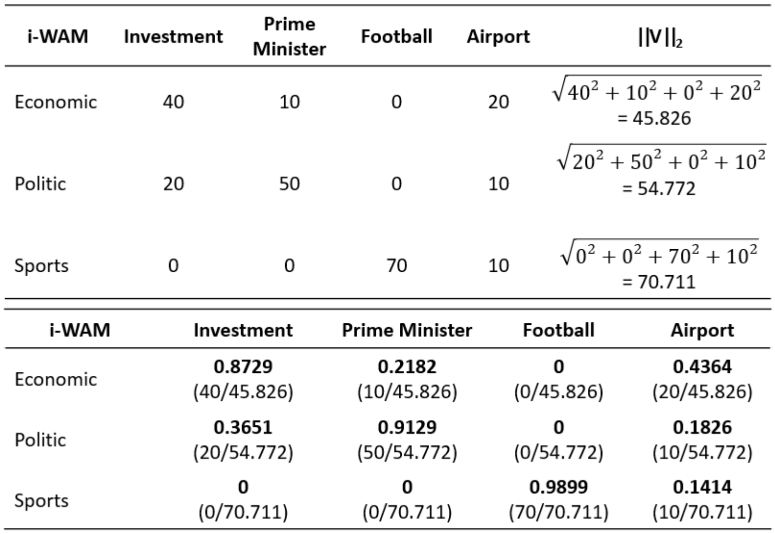
$$\mathrm{TF_{word(i)}} = \frac{\mathrm{V_i}}{||\vec{\mathrm{V}}||_2} \tag{6}$$

$TF_{word(i)}$ is the normalized TF value of word (i), while V_i is the term of occurrence of word (i) and $||\vec{V}||_2$ is the L2-normalization factor of the corpus.

$$TF_{word(i)total} = \sum\nolimits_{j=0}^{n} TF_{word\,(i)\,(j)} \tag{7}$$

$TF_{word(i)\,total}$ is the final TF value of word (i) when considering n as the number of words in the corpus. While $TF_{word\,(i)\,(j)}$ is the normalized TF value of word (i) in corpus (j), start from 1 to n.

For example, in Figure 2, if Corpus 1 (Economic category) contains 3 words in total: "investment", "prime minister" and "airport" with their terms of occurrence at 40, 10, and 20 respectively. We can calculate the L2-normalization factor of this corpus as the square root of ((40 power 2) + (10 power 2) + (20 power 2)), which is equal to 45.826. Then we can find that the TF value of the word "investment" in Corpus 1 (Economic category) is 0.8729.

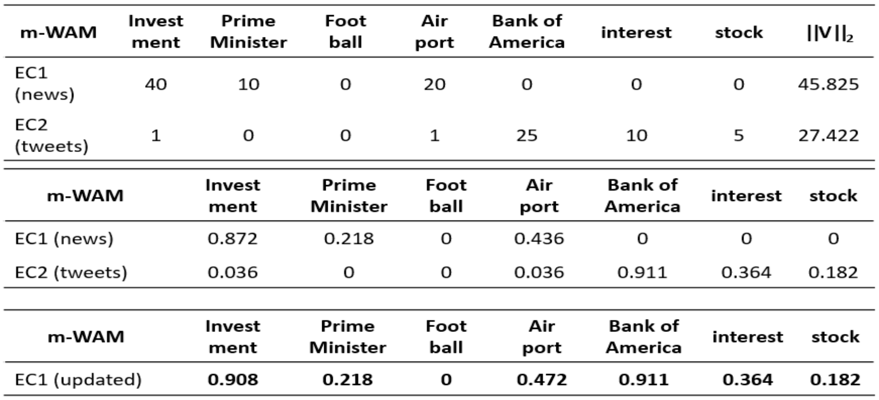


i-WAM	Investment	Prime Minister	Football	Airport	$\|V\|_2$
Economic	40	10	0	20	$\sqrt{40^2+10^2+0^2+20^2}$ = 45.826
Politic	20	50	0	10	$\sqrt{20^2+50^2+0^2+10^2}$ = 54.772
Sports	0	0	70	10	$\sqrt{0^2+0^2+70^2+10^2}$ = 70.711

i-WAM	Investment	Prime Minister	Football	Airport
Economic	**0.8729** (40/45.826)	**0.2182** (10/45.826)	**0** (0/45.826)	**0.4364** (20/45.826)
Politic	**0.3651** (20/54.772)	**0.9129** (50/54.772)	**0** (0/54.772)	**0.1826** (10/54.772)
Sports	**0** (0/70.711)	**0** (0/70.711)	**0.9899** (70/70.711)	**0.1414** (10/70.711)

Figure 2 Example of TF value (normalized) calculation.

For an example of the TF-Merging operation, see Figure 3. Here, the normalized TF value of the word "investment" in Corpus 1 (Economic category from online news articles) is 0.872. The normalized TF value of the word "investment" in Corpus 2 (Economic category from related tweets) at 0.036. Finally, we see that the total TF value of the word "investment" is 0.872 + 0.036 = 0.908, which can be representative for the Economic category.



m-WAM	Invest ment	Prime Minister	Foot ball	Air port	Bank of America	interest	stock	$\|V\|_2$
EC1 (news)	40	10	0	20	0	0	0	45.825
EC2 (tweets)	1	0	0	1	25	10	5	27.422

m-WAM	Invest ment	Prime Minister	Foot ball	Air port	Bank of America	interest	stock
EC1 (news)	0.872	0.218	0	0.436	0	0	0
EC2 (tweets)	0.036	0	0	0.036	0.911	0.364	0.182

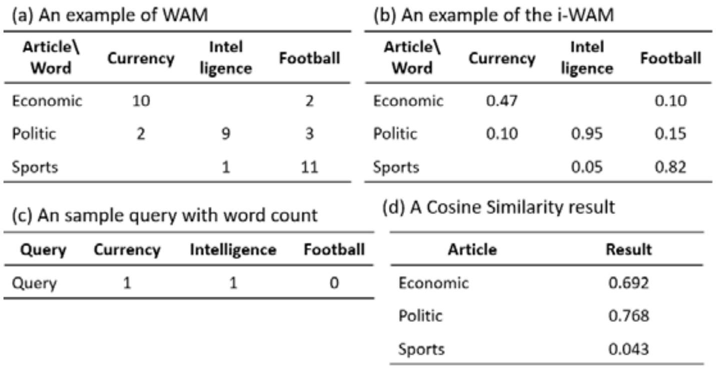
m-WAM	Invest ment	Prime Minister	Foot ball	Air port	Bank of America	interest	stock
EC1 (updated)	**0.908**	**0.218**	**0**	**0.472**	**0.911**	**0.364**	**0.182**

Figure 3 Example of TF-Merging operation.

3.1.5 Word Article Matrix (WAM)

WAM is a significant data structure [13] in the Generic Engine for Transpose Association (GETA). It creates a large matrix of weighted relationships between documents and keywords in which the rows are indexed by names of documents (articles) and the columns are indexed by keywords from the documents. The keywords in the documents are counted to fill in the table as shown in Figure 4(a). The initial WAM (i-WAM) is generated by using the normalized TF value of each word. The i-WAM with normalized TF values is shown in Figure 4(b). The documents and words are represented in the form of vectors. The value in each row is the vector of the words representing a document.

Assume there is a query: “You can run the Business Intelligence Wizard to create currency conversion calculations.” This query is converted into a model of word vectors, as shown in Figure 4(c).



(a) An example of WAM

Article\ Word	Currency	Intel ligence	Football
Economic	10		2
Politic	2	9	3
Sports		1	11

(b) An example of the i-WAM

Article\ Word	Currency	Intel ligence	Football
Economic	0.47		0.10
Politic	0.10	0.95	0.15
Sports		0.05	0.82

(c) An sample query with word count

Query	Currency	Intelligence	Football
Query	1	1	0

(d) A Cosine Similarity result

Article	Result
Economic	0.692
Politic	0.768
Sports	0.043

Figure 4 Example of WAM.

The set of documents in a corpus is viewed as a set of vectors in a vector space. Each term will have its own axis. Using the cosine similarity technique [20] we can find out the similarity between any two documents (8).

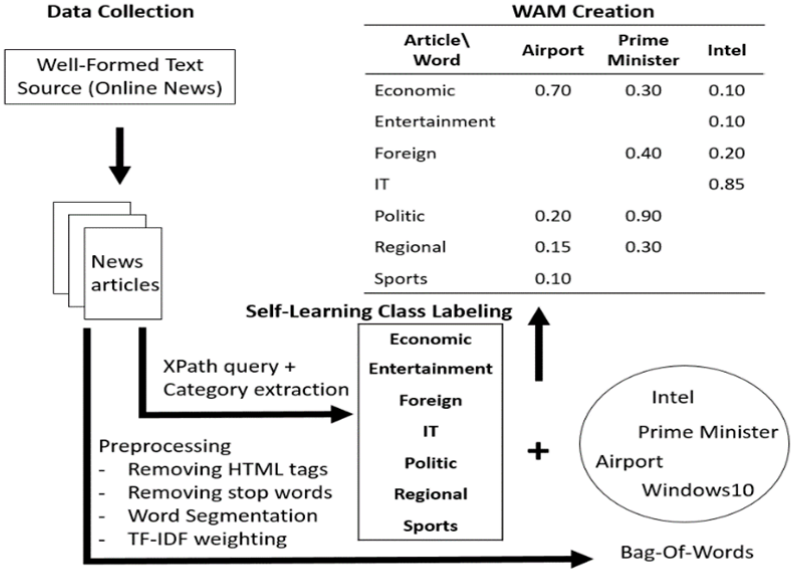
$$Cosine\ Similarity(d1, d2) = \frac{d1.d2}{||d1||*||d2||} \tag{8}$$

The $Cosine\ Similarity(d1, d2)$ is the similarity between document $d1$ and $d2$, where $d1.d2$ is the dot product of document vectors $d1$ and $d2$. $||d1|| * ||d2||$ is the Euclidean length of document vectors $d1$ and $d2$.

Lastly, we calculate the cosine similarity values and get the result of the example query as shown in Figure 4(d). As the weight of the word “intelligence” in the Information Technology (IT) category is high, 0.95, the result of the operation shows that the query is more likely to be for a document about IT, which produced the highest cosine similarity score at 0.768.

3.2 This Study's Approach

We propose a semi-supervised learning technique with the utilization of a well-formed text source, as shown in Figure 5. This is the first step. An online news source is used as the main source to collect data from, which gives access to a well-formed document with appropriate grammar that is properly categorized by the publishers. The online news article was retrieved from the Dailynews website, http://www.dailynews.co.th/ [17], published by a popular newspaper in Thailand.



Article\ Word	Airport	Prime Minister	Intel
Economic	0.70	0.30	0.10
Entertainment			0.10
Foreign		0.40	0.20
IT			0.85
Politic	0.20	0.90	
Regional	0.15	0.30	
Sports	0.10		

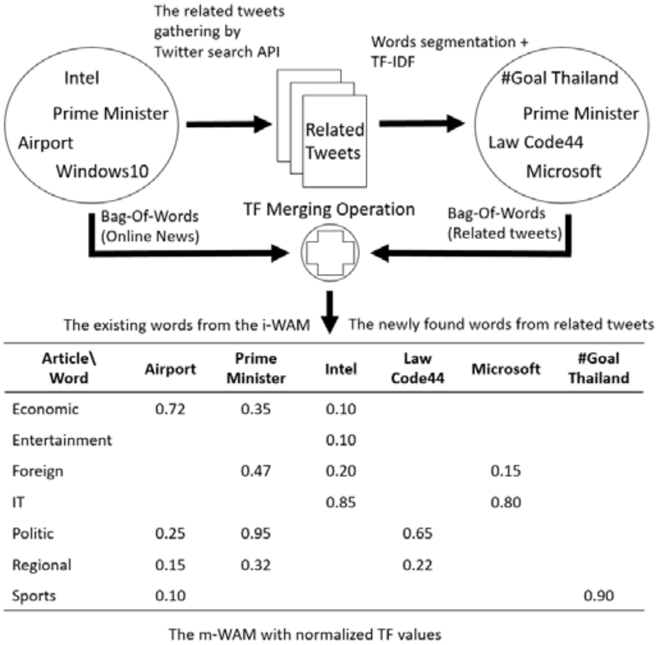
Figure 5 Initial WAM implementation.

A total of 13,085 news articles were collected, as shown in Table 1. The news categories that will become class labels can be extracted automatically. The news articles can be extracted related to their category, after executing the following preprocessing steps: removing HTML tags, removing stop words, word stemming. Then we used Thai word segmentation and the TF-IDF weighting technique to extract a bag of keywords from each news category. Afterwards, we generated the initial WAM (i-WAM) from the set of extracted keywords. The top six of the terms with the highest TF-IDF score were selected as the keywords for each category to search Twitter to enhance the model.

Table 1 Numbers of retrieved online news documents.

Category	Number of Documents
Economic	1630
Entertainment	1450
Foreign	1550
IT	1500
Politics	2755
Regionals	2550
Sports	1650

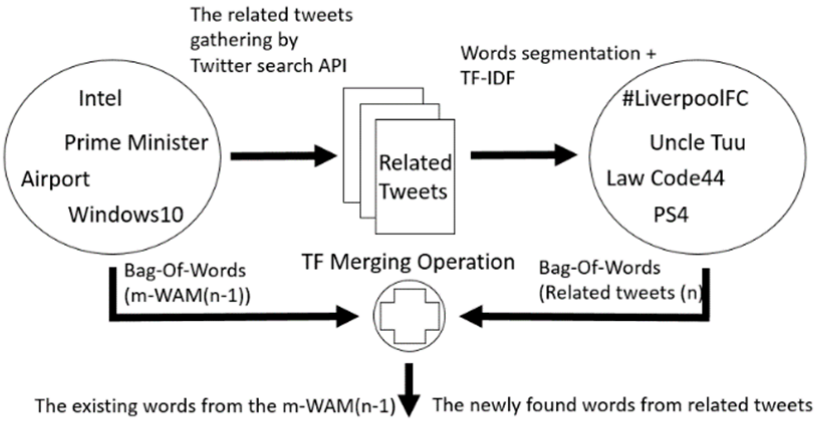
Then, we used the keyword set from i-WAM to collect related tweets through a Twitter search API, as shown in Figure 6. The API allows collecting related social media text, where the search index has a searchback limit of 7 days. After collecting a heap of tweets, around twenty thousand, they were saved in text file format. Subsequently, the same process as described before was used to extract keywords by using Thai word segmentation and the TF-IDF technique. Additional terms were selected according to their TF-IDF value. The result was a new set of keywords indicating specific categories that are potentially used in social media.


The related tweets gathering by Twitter search API
Words segmentation + TF-IDF
Intel
Prime Minister
Airport
Windows10
Related Tweets
#Goal Thailand
Prime Minister
Law Code44
Microsoft
Bag-Of-Words (Online News)
TF Merging Operation
Bag-Of-Words (Related tweets)
The existing words from the i-WAM
The newly found words from related tweets
The m-WAM with normalized TF values

Article\ Word	Airport	Prime Minister	Intel	Law Code44	Microsoft	#Goal Thailand
Economic	0.72	0.35	0.10			
Entertainment			0.10			
Foreign		0.47	0.20		0.15	
IT			0.85		0.80	
Politic	0.25	0.95		0.65		
Regional	0.15	0.32		0.22		
Sports	0.10					0.90

Figure 6 Modified WAM (1) implementation.

In the implementation of m-WAM, the Term Frequency merging (TF merging) technique is used, which is generated by updating i-WAM. The TF of existing words in i-WAM is recomputed with additional counting. The newly found words with their TF values are added into the table. As shown in Figure 7, the m-WAM process is repeated, iterating the procedure until a result is achieved in which Precision, Recall, F-measure, and accuracy are in steady state at nearly 100%. Finally, m-WAM is modified to fit social media text. This m-WAM will be an effective model containing terms that can represent a text category and reflect social developments.

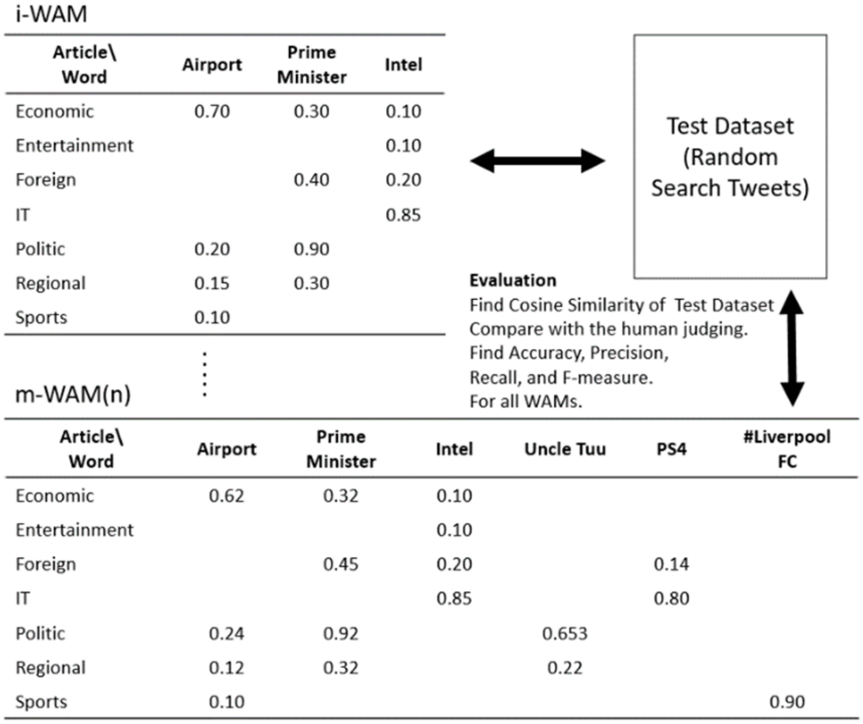

The related tweets gathering by Twitter search API
Intel
Prime Minister
Airport
Windows10
Related Tweets
Words segmentation + TF-IDF
#LiverpoolFC
Uncle Tuu
Law Code44
PS4
Bag-Of-Words (m-WAM(n-1))
TF Merging Operation
Bag-Of-Words (Related tweets (n)
The existing words from the m-WAM(n-1)
The newly found words from related tweets

Article\ Word	Airport	Prime Minister	Intel	Uncle Tuu	PS4	#Liverpool FC
Economic	0.62	0.32	0.10			
Entertainment			0.10			
Foreign		0.45	0.20		0.14	
IT			0.85		0.80	
Politic	0.24	0.92		0.653		
Regional	0.12	0.32		0.22		
Sports	0.10					0.90

The m-WAM(n) with normalized TF values

Figure 7 Modified WAM (n) implementation.

As can be seen in Figure 8, the evaluation of social media text classification is conducted manually. The training data set is used for building the model while the testing data set is searched from Twitter randomly and used for evaluating the model. The retrieved tweets are evaluated by human judging. The testing data set evaluates all models, from i-WAM to m-WAM (n). Finally, accuracy rate, Precision, Recall, and F-measure value are determined.



i-WAM

Article\ Word	Airport	Prime Minister	Intel
Economic	0.70	0.30	0.10
Entertainment			0.10
Foreign		0.40	0.20
IT			0.85
Politic	0.20	0.90	
Regional	0.15	0.30	
Sports	0.10		

m-WAM(n)

Article\ Word	Airport	Prime Minister	Intel	Uncle Tuu	PS4	#Liverpool FC
Economic	0.62	0.32	0.10			
Entertainment			0.10			
Foreign		0.45	0.20		0.14	
IT			0.85		0.80	
Politic	0.24	0.92		0.653		
Regional	0.12	0.32		0.22		
Sports	0.10					0.90

Test Dataset (Random Search Tweets)

Evaluation
Find Cosine Similarity of Test Dataset
Compare with the human judging.
Find Accuracy, Precision,
Recall, and F-measure.
For all WAMs.

Figure 8 Evaluation process.

4 Experiment Result

After retrieving online news data by using the web crawler module and extracting a set of keywords, we selected the words with the highest TF-IDF score and generated the initial-WAM (i-WAM), as shown in Table 2. We added a row to show the IDF value of each keyword to identify their importance weight. The words “financial budget”/ “งบการเงิน”, “Gubgib”/ “กุ๊บกิ๊บ”, “refugee”/ “ผู้อพยพ”, “Windows 10”/ “วินโดวส์10”, “politician”/ “นักการเมือง”, “artificial rain”/ “ฝนหลวง”, “karate”/ “คาราเต้” are examples of keywords with their TF value in each category (Economic, Entertainment, Foreign, IT, Politics, Regional and Sports) respectively.

The keywords that were extracted from the online news source showed a significant result, especially the keywords from the Entertainment category

(“Gubgib”/ “กุ๊บกิ๊บ”: the name of a popular actress in Thailand), IT category (“Windows 10”/ “วินโดวส์10”), and sports category (“karate”/ “คาราเต้”). However, their TF values can identify their text categories when we consider the word vector cosine similarity. Then, these keyword terms were used to search Twitter through the Twitter search API.

Table 2 Part of the i-WAM.

Article\Word	Word1	Word2	Word3	Word4	Word5	Word6	Word7
	financial budget ‘งบการเงิน’	**gubgib** ‘กุ๊บกิ๊บ’	**refugee** ‘ผู้อพยพ’	**Windows 10** ‘วินโดวส์10’	**politician** ‘นักการเมือง’	**artificial rain** ‘ฝนหลวง’	**karate** ‘คาราเต้’
IDF (t)	**2.23044**	**1.83250**	**1.68638**	**1.88649**	**1.38535**	**2.23044**	**2.05435**
Economic	0.00013				0.00001		
Entertainment		0.00023					
Foreign	0.00001		0.00044		0.00004		
IT				0.00009			
Politics					0.00035	0.00001	
Regionals						0.00012	
Sports			0.00001				0.00022

Around twenty thousand tweets were collected as our data source from to extract a new set of the Twitter keywords. The m-WAM1 was generated from these new specific keywords and merged with the existing keywords from i-WAM (TF merging operation), well-formed text source keywords. The newly found words in the m-WAM1 showed a significant result. For example, in Table 3, “refugee”/“ผู้อพยพ” (sample keyword from i-WAM) led to finding a new keyword, “Tier3”/ “เทียร์3”, which scoped down the word vectors for the Foreign category. Other category keywords also generated promising results.

Subsequently, the same m-WAM process was repeated. Keywords with high potential were selected from the m-WAM(n-1), i.e. the words with the highest TF-IDF score (top 5) in their own category, to gather all related tweets. From this technique, more specific keywords were found – less common words – which can effectively represent their category. Hence, we could generate a new m-WAM(n), which can be a productive model for social media text classification. This procedure was repeated until the Precision, Recall, F-measure, and accuracy results reached steady state. Finally, it was found that the iteration number of m-WAM that satisfied the best performance of social media text classification is 3 (n = 3), i.e. i-WAM, m-WAM1, m-WAM2 and m-WAM3. The rest of the results are shown in Tables 4 and 5.

Table 3 Part of m-WAM1

Article\Word	**Word1**	**Word2**	**Word3**	**Word4**	**Word5**	**Word6**	**Word7**
	financial budget ‘งบการเงิน’	**Gubgib** ‘กุ๊บกิ๊บ’	**refugee** ‘ผู้อพยพ’	**Windows 10** ‘วินโดวส์ 10’	**politician** ‘นักการเมือง’	**artificial rain** ‘ฝนหลวง’	**karate** ‘คาราเต้’
IDF (t)	**1.59151**	**1.28831**	**1.57032**	**1.87550**	**1.27553**	**1.61838**	**1.47756**
Economic	0.00077				0.00001		
Entertainment		0.00156					
Foreign	0.00001		0.00116		0.00004		
IT				0.00076			
Politics					0.00180	0.00001	
Regionals					0.00008	0.00098	
Sports			0.00001				0.00119

Article\ Word	**Word8**	**Word9**	**Word10**	**Word11**	**Word12**	**Word13**	**Word14**
	layoff ‘ลดพนักงาน’	**BeeKPN** ‘บี๋KPN’	**Tier3** ‘เทียร์3’	**Microsoft** ‘ไมโครซอฟท์’	**drought** ‘ภัยแล้ง’	**Venerable Monk** ‘หลวงพ่อนุช’	**#Team Thailand** ‘#ทีมชาติไทย’
IDF (t)	**2.22978**	**2.63726**	**2.31275**	**2.81335**	**2.31275**	**2.03520**	**3.11438**
Economic	0.0010						
Entertainment		0.00028					
Foreign			0.00129				
IT				0.00094			
Politics			0.00007		0.00023		
Regionals					0.00022	0.00140	
Sports							0.00007

Table 4 Part of m-WAM2.

Article\Word	**Word1**	**Word2**	**Word3**	**Word4**	**Word5**	**Word6**	**Word7**
	financial budget ‘งบการเงิน’	**Gubgib** ‘กุ๊บกิ๊บ’	**refugee** ‘ผู้อพยพ’	**Windows 10** ‘วินโดวส์ 10’	**politician** ‘นักการเมือง’	**artificial rain** ‘ฝนหลวง’	**karate** ‘คาราเต้’
IDF (t)	**1.40959**	**1.08944**	**1.40527**	**2.93247**	**1.10856**	**1.45535**	**1.36427**
Economic	0.00130				0.00001		
Entertainment		0.00261					
Foreign	0.00004		0.00171		0.00004		
IT				0.00072			
Politics					0.00280	0.00001	
Regionals					0.00008	0.00154	
Sports			0.00001				0.00191

Article\Word	Word8	Word9	Word10	Word11	Word12	Word13	Word14
	layoff ‘ลดพนักงาน’	**BeeKPN** ‘บี้KPN’	**Tier3** ‘เทียร์3’	**Microsoft** ‘ไมโครซอฟท์’	**drought** ‘ภัยแล้ง’	**Venerable Monk** ‘หลวงพ่อนุช’	**#Team Thailand** ‘#ทีมชาติไทย’
IDF (t)	**1.37617**	**1.10209**	**1.39675**	**1.22775**	**1.08531**	**1.84139**	**3.10856**
Economic	0.00326						
Entertainment		0.00016					
Foreign			0.00334				
IT				0.01302			
Politics			0.00007		0.00448		
Regionals					0.00022	0.00226	
Sports							0.00054

Article\Word	Word15	Word16	Word17	Word18	Word19	Word20	Word21
	Flight ‘เส้นทางบิน’	**Ploypan** ‘พลอยพรรณ’	**Xinjiang** ‘ซินเจียง’	**Intel** ‘อินเทล’	**Prime minister Tuu** ‘นายกฯ ตู่’	**Abbot** ‘เจ้าอาวาส’	**Karate-do** ‘คาราเต้-โด’
IDF (t)	**2.56449**	**3.40959**	**2.56449**	**1.82981**	**2.56449**	**1.84139**	**2.93247**
Economic	0.00066						
Entertainment		0.00146					
Foreign			0.00069				
IT				0.00495			
Politics					0.00242		
Regionals						0.00540	
Sports							0.00173

As an interesting point, some new terms were also added because they occurred very frequently on Twitter rather than in the online news document, for example more variations of abbreviations and trendy terms such as “#TeamThailand” / “#ทีมชาติไทย”, a specific hash tag created by some social media users for Thailand sports fans, and “Prime Minister Tuu” / “นายกฯตู่”, “Tuu”/ “ตู่” being the nick name of the present Prime Minister of Thailand and “นายกฯ” is an abbreviation of “Prime Minister” in the Thai language. This is a common phenomenon on Twitter, referring to the nature of this useful communication tool, free and open for opinion sharing with a 140-character limitation. In addition, when the series of all category keywords is considered, we can see a real-time reflection of interesting social issues. For example, from the keywords in the Foreign category in Table 5 – word 3 (“Refugee”/ “ผู้อพยพ”), word 10 (“Tier3”/ “เทียร์3”), word 17 (“Xinjiang”/ “ซินเจียง”), and word 24 (“Boycott”/ “คว่ำบาตร”) – it can be seen that there was a topic related to the Uyghur refugees with some political issues shared on social media during the period covered by the experiment.

Table 5 Part of m-WAM3.

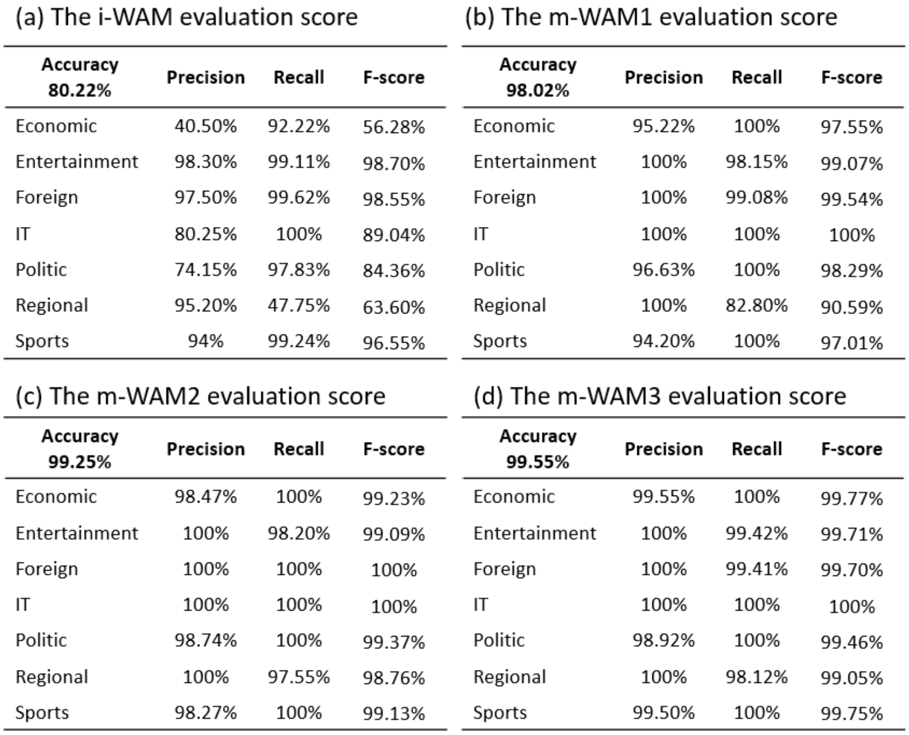
Article\Word	Word1	Word2	Word3	Word4	Word5	Word6	Word7
	financial budget ‘งบการเงิน’	**Gubgib** ‘กุ๊บกิ๊บ’	**refugee** ‘ผู้อพยพ’	**Windows 10** ‘วินโดวส์ 10’	**politician** ‘นักการเมือง’	**artificial rain** ‘ฝนหลวง’	**karate** ‘คาราเต้’
IDF (t)	**1.41044**	**1.10941**	**3.41044**	**2.93331**	**3.41044**	**1.43731**	**1.29649**
Economic	0.00174				0.00001		
Entertainment		0.00349					
Foreign	0.00005		0.00151		0.00004		
IT				0.00067			
Politics					0.00376	0.00001	
Regionals					0.00008	0.00144	
Sports			0.00001				0.00269

Article\Word	Word8	Word9	Word10	Word11	Word12	Word13	Word14
	layoff ‘ลดพนักงาน’	**BeeKPN** ‘บี๋KPN’	**Tier3** ‘เทียร์3’	**Microsoft** ‘ไมโครซอฟท์’	**drought** ‘ภัยแล้ง’	**Venerable Monk** ‘หลวงพ่อนุช’	**#Team Thailand** ‘#ทีมชาติไทย’
IDF (t)	**1.30663**	**3.41044**	**1.39340**	**1.13168**	**1.10941**	**1.85413**	**2.26431**
Economic	0.00465						
Entertainment		0.00011					
Foreign			0.00404				
IT				0.01529			
Politics			0.00007		0.00648		
Regionals					0.00022	0.00166	
Sports							0.00072

Article\Word	Word15	Word16	Word17	Word18	Word19	Word20	Word21
	flight ‘เส้นทางบิน’	**Ploypan** ‘พลอยพรรณ’	**Xinjiang** ‘ซินเจียง’	**Intel** ‘อินเทล’	**prime minister Tuu** ‘นายกฯ ตู่’	**Abbot** ‘เจ้าอาวาส’	**karate-do** ‘คาราเต้-โด’
IDF (t)	**2.56534**	**1.40611**	**1.44195**	**1.70287**	**1.63228**	**1.16740**	**2.29649**
Economic	0.00065						
Entertainment		0.00281					
Foreign			0.00310				
IT				0.00469			
Politics					0.00140		
Regionals						0.01163	
Sports							0.00184

Article\Word	**Word22**	**Word23**	**Word24**	**Word25**	**Word26**	**Word27**	**Word28**
	flight cancelation ‘ยกเลิกเที่ยวบิน’	**acknowledgement of children** ‘รับรองบุตร’	**boycott** ‘คว่ำบาตร’	**Blognone** **‘Blognone’**	**Taksin** ‘ทักษิณ’	**Putta Issara** ‘หลวงปู่พุทธอิสระ’	**Thailand Open** ‘ไทยแลนด์ โอเพ่น’
IDF (t)	**2.03022**	**1.97907**	**3.41044**	**3.10941**	**3.41044**	**2.71147**	**1.73834**
Economic	0.00147						
Entertainment		0.00112					
Foreign			0.00005				
IT				0.00017			
Politics					0.00010		
Regionals						0.00052	
Sports							0.00533

As for the evaluation process, Figure 9 (a)-(d) shows the Precision, Recall and F-measure values. All models were evaluated with a testing data set, which was randomly extracted from Twitter.



(a) The i-WAM evaluation score

Accuracy 80.22%	Precision	Recall	F-score
Economic	40.50%	92.22%	56.28%
Entertainment	98.30%	99.11%	98.70%
Foreign	97.50%	99.62%	98.55%
IT	80.25%	100%	89.04%
Politic	74.15%	97.83%	84.36%
Regional	95.20%	47.75%	63.60%
Sports	94%	99.24%	96.55%

(b) The m-WAM1 evaluation score

Accuracy 98.02%	Precision	Recall	F-score
Economic	95.22%	100%	97.55%
Entertainment	100%	98.15%	99.07%
Foreign	100%	99.08%	99.54%
IT	100%	100%	100%
Politic	96.63%	100%	98.29%
Regional	100%	82.80%	90.59%
Sports	94.20%	100%	97.01%

(c) The m-WAM2 evaluation score

Accuracy 99.25%	Precision	Recall	F-score
Economic	98.47%	100%	99.23%
Entertainment	100%	98.20%	99.09%
Foreign	100%	100%	100%
IT	100%	100%	100%
Politic	98.74%	100%	99.37%
Regional	100%	97.55%	98.76%
Sports	98.27%	100%	99.13%

(d) The m-WAM3 evaluation score

Accuracy 99.55%	Precision	Recall	F-score
Economic	99.55%	100%	99.77%
Entertainment	100%	99.42%	99.71%
Foreign	100%	99.41%	99.70%
IT	100%	100%	100%
Politic	98.92%	100%	99.46%
Regional	100%	98.12%	99.05%
Sports	99.50%	100%	99.75%

Figure 9 Accuracy rate, Precision, Recall and F-measure values.

Because there were more common keywords in i-WAM, its Precision, Recall, and F-measure scores were low, especially in the Economic and Politics categories. However, when m-WAM was updated with more specific keywords from related social-media text, all of the evaluation factors increased dramatically. Finally, the value of Precision, Recall, F-measure, and accuracy nearly converged to 100% after reaching m-WAM3, as shown in Figure 10.

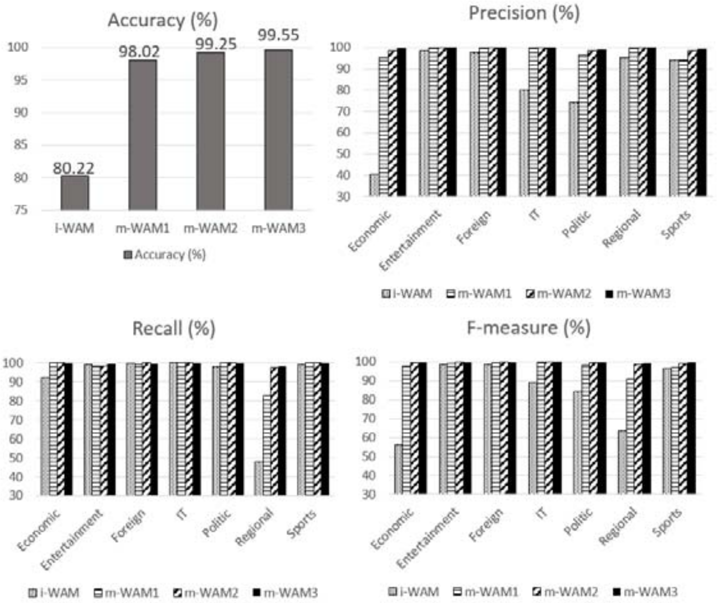


Figure 10 Graphs of accuracy rate, Precision, Recall, and F-measure values.

5 Conclusions and Future Work

The growth and information power of social media text are remarkable. Keywords collected from social media can be a prediction tool of social developments. A holistic decision support system can be developed according to interesting topics collected from the dynamic social media environment, which is a factor of concern today and will be in the future. Social media text classification using Term Frequency-Inverse Document Frequency (TF-IDF) weighting and Word Article Matrix (WAM) is very effective. Text from social media can be categorized with a sense of human familiarity by utilizing online news categories that have already been indicated by the publishers. Good results can be expected from the proper modified WAM (m-WAM) for social media text classification after updating it for 3 times, the suitable iteration number of m-WAM modifications. This modified WAM can be a suitable model for social media text classification and the set of keyword terms can be representative of interesting social topics during the time of monitoring. However, a good result also depends on the performance of the Thai word segmentation module. Alternative Thai word segmentation programs, such as Name Entity Recognition (NER), can generate proper word boundaries for conducting other processes, so keywords can be generated more accurately and the model's accuracy will be improved significantly. Deep learning could also be a good choice for conducting experiments related to natural language processing and text mining.

Acknowledgements

This research was financially supported by Thailand Advanced Institute of Science and Technology (TAIST), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology, Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU) under the TAIST Tokyo Tech Program, and partly supported by the National Research University Project of the Office of the Higher Education Commission and Thammasat University.

We would also like to express our gratitude to Dr. Eakasit Pacharawongsakda for his assistance with data-mining techniques and his comments that greatly improved the manuscript during the course of this research.

References

[1] Simon, K., *Digital, Social & Mobile Worldwide in 2015*, We Are Social Ltd., http://wearesocial.net/tag/statistics/ (21 January 2015).

[2] Twitter, *Twitter Usage/Company Facts*, Twitter, Inc., https://about.twitter.com/company (30 June 2015).

[3] Dave, C. *Global Social Media Research Summary 2015*, Smart Insights (Marketing Intelligence), Ltd., http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research (13 October 2015).

[4] Twitter, *Entities in Objects*, Twitter, Inc., 2015. https://dev.twitter.com/overview/api/entities-in-twitter-objects (13 October 2015).

[5] Twitter, *API Rate Limits*, Twitter, Inc., 2015. https://dev.twitter.com/rest/public/rate-limiting (13 October 2015).

[6] Episod, *Search API is limited to the last 7 days?*, Twitter, Inc., https://twittercommunity.com/t/search-api-is-limited-to-the-last-7-days/11603 (17 July 2013).

[7] Irfan, R., King, C.K., Grages, D., Ewen, S., Khan, S.U., Madani, S.A., Kolodziej, J., Wang, L., Chen, D., Rayes, A., Tziritas, N., Xu, C.Z., Zomaya, A.Y., Alzahrani, A.S. & Li, H., *A Survey on Text Mining in*

Social Networks, Cambridge Journal, The Knowledge Engineering Review, **30**(2), pp. 157-170, 2015.

[8] Patel, P. & Mistry, K., *A Review: Text Classification on Social Media Data*, IOSR Journal of Computer Engineering, **17**(1), pp. 80-84, 2015.

[9] Lee, K., Palsetia, D., Narayanan, R., Patwary, Md.M.A., Agrawal, A. & Choudhary, A.S, *Twitter Trending Topic Classification*, in Proceeding of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDW'11, pp. 251-258, 2011.

[10] Kateb, F. & Kalita, J., *Classifying Short Text in Social Media: Twitter as Case Study*, International Journal of Computer Applications, **111**(9), pp. 1-12, 2015.

[11] Chirawichitichai, N., Sanguansat, P. & Meesad, P., *A Comparative Study on Feature Weight in Thai Document Categorization Framework*, 10th International Conference on Innovative Internet Community Services (I2CS), IICS, pp. 257-266, 2010.

[12] Theeramunkong, T. & Lertnattee, V., *Multi-Dimension Text Classification*, SIIT, Thammasat University, 2005.http://www.aclweb.org /anthology/C02-1155 (25 October 2015).

[13] Viriyayudhakorn, K., Kunifuji, S. & Ogawa, M., *A Comparison of Four Association Engines in Divergent Thinking Support Systems on Wikipedia*, Knowledge, Information, and Creativity Support Systems, KICSS2010, Springer, pp. 226-237, 2011.

[14] Sornlertlamvanich, V., Pacharawongsakda, E. & Charoenporn, T., *Understanding Social Movement by Tracking the Keyword in Social Media*, in MAPLEX2015, Yamagata, Japan, February 2015.

[15] Olston, C. & Najork, M., *Web Crawling*, Foundation and Trends in Information Retrieval, **4**(3), pp. 175-246, 2010.

[16] RapidMiner, *The Open Source Platform of Choice*, Rapid Miner, 2015. https://rapidminer.com/ (15 October 2015).

[17] Dailynews, *Online News*, Dailynews web, Ltd., 2015, http://www.dailynews.co.th/ (15 October 2015).

[18] Meknavin, S., Charoenpornsawat, P. & Kijsirikul, B., *Feature-based Thai Word Segmentation*, National Electronics and Computer Technology Center, 1997, http://www.cs.cmu.edu/~paisarn/papers/nlprs97.pdf (15 October 2015).

[19] Wu, H.C., Luk, R.W.P., Wong, K.F. & Kwok, K.L., *Interpreting TF-IDF Term Weights as Making Relevance Decisions*, ACM Transactions on Information Systems, **26**(3), Article 13, pp. 1-37, 2008.

[20] Vembunarayanan, J., *Tf-Idf and Cosine Similarity*, https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/ (27 October 2013).