Effectiveness of Social Media Text Classification by Utilizing the Online News Category

Phat Jotikabukkana
School of ICT,
Sirindhorn International Institute of Technology,
Thammasat University,
Pathum Thani 12121,
Thailand

phat.j@mod.go.th

Virach Sornlertlamvanich
School of ICT,
Sirindhorn International Institute of Technology,
Thammasat University,
Pathum Thani 12121,
Thailand

virach@siit.tu.ac.th

Okumura Manabu
Tokyo Institute of Technology, Ookayama Campus, Ookayama
Meguro-ku
Tokyo, 152-8550
Japan

oku@pi.titech.ac.jp

Choochart Haruechaiyasak
National Electronics and Computer Technology Center, Thailand Science Park,
Pathum Thani 12120,
Thailand

choochart.haruechaiyasak@nectec.or.th

Abstract— **Social media text can illustrate significant information of our real social situation. It can show the direction of real-time social movement. However, it has its own characteristics such as using short text and informal language, many unstructured information and argot. This kind of text is hard to classify and difficult to analyze to extract the useful information. In this paper, we propose an effective technique to classify the social media text by utilizing the initial keywords from well-formed sources of data, such as online news. Term frequency–inverse document frequency weighting technique (TF-IDF) and Word Article Matrix (WAM) are used as main methods in this research. We use the extracted keywords from the well-formed source as a main factor to do experiment on Twitter messages. We found a set of the social media keywords can represent the essence of social events and can be used to classify the text effectively.**

Keywords- social media; Twitter; keywords extraction; term frequency–inverse document frequency (TF-IDF); Word Article Matrix (WAM)

I. INTRODUCTION

Nowadays, people use social media as an important channel to communicate to each other. They usually use it to broadcast their ideas, their feeling, and their information to the cyber space. There are around one third of number of people in the world are indicated as an active social media accounts value [1]. Unsurprisingly, many social network data such as tweets from Twitter could reflect current the real social events. We can get main idea of information, the real-time social situation, from reading the social media text as same as we can get them from newspapers. In many cases, the social media can conduct the real social direction by its information power. In example, many companies use the social media as a main channel to promote their products and use it to create products interesting. This means that if we can utilize this data type, we will gain a lot of benefit from them. The main challenge is analyzing the social media text. It is the data stream which contains a lot of noisy and unstructured information, informal languages, slang, and absent words. It is so difficult to do text classification to group it before extracting the useful information. Our experiment, we focus on Twitter as the social media text source, refer to recent statistics, there are 4.5 million twitter users in Thailand with nearly 2 million active users/day [2]. It is a significant value which can present that if we can filter the related tweets with the productive keywords we can notice the society relationship. We analyze the times series of tweets with specific period to prove our technique that can extract keywords, categorize social media text productively, and can illustrate the evolution of social behavior on a happening.

The Twitter message, a tweet, is a short message text. Twitter limits the tweet length to 140 characters [3]. It looks like colloquialism text, compared with written document. Consequently, in our experiment, we apply a technique to extract keywords using term frequency–inverse document frequency (TF-IDF) [4], and Word Article Matrix (WAM) to expand the set of keywords reflecting the nature of the text from Twitter [5]. We collect data and create word vector from well-formed text, online news, which already initially categorized by publishers to extract keywords and categorize information from Twitter which is a hardly classified information source. Finally, we get a productive set of keywords from the official site and the social media which can be a representative of text categories. We found a new words, abbreviation and argot that never appear in the well-formed documents, from the Tweet messages which become the main keywords that can reflect the interesting topics in the real social at that time.

II. Related Research Works

A. Word Segmentation

Word segmentation is an essential step in natural language processing (NLP) [6]. We conduct experiment in the social media text written in the Thai language. The Thai language is written without spaces between words. In this research, we use a word segmentation module [7] applying the maximal matching algorithm [8] to determine the word boundary. The segmentation result is acceptable to determine the essential words for further processing in keyword identification.

B. Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF weight is often used in text mining [4]. This technique can identify keywords of a document in a corpus. It consists of two terms, Term Frequency (TF) and Inverse Document Frequency (IDF). The TF is computed from the number of times a word appears in a document, divided by the total number of words in that document. The IDF is computed from the logarithm of the number of all documents in a collection divided by the number of documents which the observed term appears. The TF can defines as a counting function [9] (1).

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

The $TF(t,d)$ is actually the total number of the term t that appears in the document d, and the $fr(x,t)$ is a simple function defined as (2):

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

The IDF (inverse document frequency) is defined as (3):

$$IDF(t) = log\frac{|D|}{|\{d: t \in d\}|} \qquad (3)$$

The $|\{d: t \ \in \ d\}|$ is the number of documents where the term t appears, when the term-frequency function satisfies $TF(t,d) \neq 0$.Then, the TF-IDF formula is defined as (4):

$$TF - IDF(t) = TF(t,d) \times \ IDF(t) \qquad (4)$$

C. Word Article Matrix (WAM)

WAM is a key data structure in the Generic Engine for Transposable Association (GETA) [5]. It constructs a large matrix of weighted relation between document and keyword which rows are indexed by names of documents (articles) and columns are indexed by words, keywords from the documents. Keywords in a document are counted to fill in the table as shown in Table I.

TABLE I. An Example of WAM

	investment	**Prime Minister**	**football**	**airport**
Economic	4	1	0	2
Politic	2	5	0	1
Sports	0	0	7	1

In this experiment, we weight the value of WAM by the number of total words in the documents collection. For example, we have five documents with fifty words, total number of words of all documents. We can extract a set of keywords, and count their value to fill in the table as shown in Table I. Afterwards, we divide all value by fifty, total number of words. As a result, we can generate initial WAM with TF values as shown in Table II.

TABLE II. An Example of the initial WAM

	investment	**Prime Minister**	**football**	**airport**
Economic	0.08	0.02	0	0.04
Politic	0.04	0.1	0	0.02
Sports	0	0	0.14	0.02

The documents and words are represented in the form of vector. The values in each row is the vector of words to represent a document. Assuming that there is a query: “Prime Minister welcomes new £50m airside investment at Edinburgh Airport”. This query is converted into a model of word vector shown in Table III.

TABLE III. AN EXAMPLE QUERY, WORD VECTOR

	investment	**Prime Minister**	**football**	**airport**
query	1	1	0	1

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 [10] we can find out the similarity between any two documents (5).

$$Cosine\ Similarity(d1,d2) = \frac{d1.d2}{||d1||*||d2||} \qquad (5)$$

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

After a calculation of the cosine similarity value, we get a result of an example query as shown in Table IV.

TABLE IV. A COSINE SIMILARITY RESULT

	Result
Economic	0.881917
Politic	0.843274
Sports	0.08165

The result of operation shows that the query is more likely to be for the document of economic, which produces the highest cosine similarity score of 0.881917.

III. OUR APPROACH AND EXPERIMENT

Our proposed technique, as shown in Fig. 1, use the online news as a main source to collect the data. This is a channel to get access to the well-formed document with appropriate grammar and properly categorized by publishers. There are seven categories that we can extract from our representative news source, Thairath online news [11]. Those are economic, entertainment, foreign, lifestyle, politic, social, and sports.

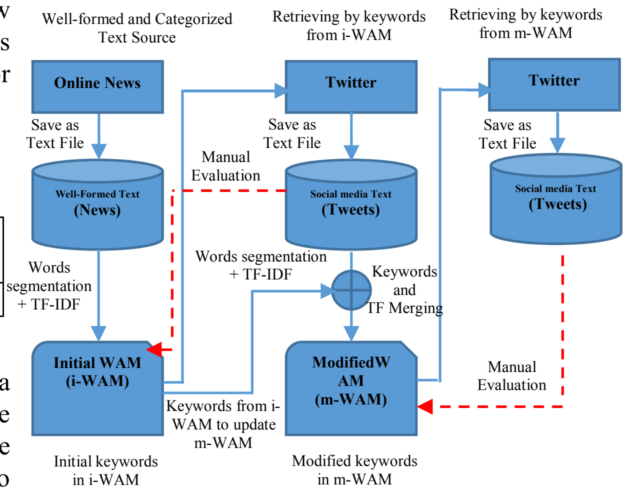
Well-formed and Categorized Text Source
Retrieving by keywords from i-WAM
Retrieving by keywords from m-WAM
Online News
Twitter
Twitter
Save as Text File
Save as Text File
Save as Text File
Manual Evaluation
Well-Formed Text (News)
Social media Text (Tweets)
Social media Text (Tweets)
Words segmentation + TF-IDF
Words segmentation + TF-IDF
Keywords and TF Merging
Initial WAM (i-WAM)
Keywords from i-WAM to update m-WAM
ModifiedWAM (m-WAM)
Manual Evaluation
Initial keywords in i-WAM
Modified keywords in m-WAM

Figure 1. The Modified WAM (m-WAM) implementation

After retrieving, we totally collected 13,014 news articles as shown in Table V. This data will be extracted to get the keywords related to category by using the Thai word segmentation module [9] and the TF-IDF terms weighting technique. Terms with highest TF-IDF score, top six, will be selected as the keyword for each category.

TABLE V. ONLIE NEWS DOCUMENTS

	Number of document
Economic	1600
Entertainment	1410
Foreign	1545
Lifestyle	1600
Politic	2677
Social	2547
Sports	1635

We generate the initial WAM (i-WAM) from the set of extracted keywords. Then we use these keywords to collect the related tweets through Twitter search API. API allows to collect the related social media text which a search index has a 7-day limit search back. We got a heap of tweets, around twenty thousand tweets, saved in text file format. Afterwards, we conduct the same process to extract keywords by using Thai word segmentation and the TF-IDF technique. We select the additional terms according to their TF-IDF value. We get a new set of keywords which indicated specific category and potentially used in social media.

In the part of m-WAM implementation, we use term frequency merging technique (TF merging), and generate it from updating the i-WAM. The TF of existing words in i-WAM is recomputed additional count. The newly found words with their TF values are added into the table. Finally, m-WAM modified to the social media text is generated. This m-WAM will be an effective model which contains terms that can present a text category and can reflect the social media.

As shown in Fig. 1., the evaluation of social media text classification is conducted manually. The retrieved tweets by category are evaluated by human judging. Finally, Precision, Recall, and F-measure value are determined.

IV. Experiment Result

After we retrieve online news data and extract a set of the keywords, we can generate initial-WAM as shown in Table VI. We add a row to show an IDF value of each keyword to identify the important weight of each keyword. The word “Bank of Thailand”/“ธนาคารแห่งประเทศไทย”, “DJ.Pueak”/“ดีเจ เผือก”, “Virus”/“ไวรัส”, “Lose weight”/“ลดความอ้วน”, “Prime Minister”/“นายกรัฐมนตรี”, “Law code”/“มาตรา”, “Sea Games”/“ซีเกมส์” are a sample of keywords with their TF value in each category of economic, entertainment, foreign, lifestyle, politic, social, and sports.

TABLE VI. A part of the i-WAM

	Bank of Thailand ‘ธนาคารแห่งประเทศไทย’	DJ. Pueak ‘ดีเจ เผือก’	Virus ‘ไวรัส’	Lose weight ‘ลดความอ้วน’
IDF(t)	**1.87999**	**2.50324**	**1.99809**	**3.20221**
Economic	0.10400	0	0	0
Entertainment	0	0.01784	0	0
Foreign	0	0	0.03747	0
Lifestyle	0	0	0	0.00325
Politic	0	0	0	0
Social	0	0	0.022810	0
Sports	0	0	0	0

	Prime Minister ‘นายกรัฐมนตรี’	Law code ‘มาตรา’	Sea Games ‘ซีเกมส์’	
IDF(t)	**0.84810**	**1.28840**	**1.82200**	
Economic	0.08320	0.00891	0	
Entertainment	0	0	0.00973	
Foreign	0.05059	0	0	
Lifestyle	0.01466	0.00488	0	
Politic	0.20022	0.09474	0.00097	
Social	0.01184	0.00921	0.00087	
Sports	0.00206	0	0.07239	

The keywords which we extracted from online news show a valuable result, especially the keyword from economic category (“Bank of Thailand”/“ธนาคารแห่งประเทศไทย”), entertainment category (“DJ.Pueak”/“ดีเจเผือก”: a name of the popular disc jockey in Thailand), and lifestyle category (“Lose weight”/“ลดความอ้วน”). They are a specific words which can be a representative for their category effectively. While, other keywords, “Virus”/“ไวรัส”, “Prime Minister”/“นายกรัฐมนตรี”, “Law code”/“มาตรา”, and “Sea Games”/“ซีเกมส์”, are seem to be common words that appear in more than one categories. However, their TF values can identify their text categories when we consider their word vector cosine similarity.

Afterwards, we use these keyword terms to search Twitter and get around twenty thousand tweets as our data collection to be extracted a new set of the Twitter keywords. Tables VII., shows a part of the m-WAM which are modified according the related tweets. In the m-WAM, the values of the TF are much more confirmed to the terms extracted from the related tweets. In addition, some new terms are also added because they occur very frequent in Twitter rather than the online news document for example more variation of abbreviation, and trendy terms as shown in Table VIII. For example, “Governor of the Bank of Thailand”/“ผู้ว่าการธนาคารแห่งประเทศไทย”, and “Law Code Number 44”/“มาตรา 44”, these terms are official format, while we found “Governor of the BOT”/“ผู้ว่าการธปท.”, and “Law Code No. 44”/“ม.44” as the keywords from social media, the abbreviation. This is an ordinary event that we often found in the tweets, because of their limitation, 140-characters length. In addition, we found the keywords from a specific hash tag, an argot that created by some social media users, such as “HowtoPerfect” from “#HowtoPerfect, and “goalthailand” from “#goalthailand” which can reflect the hot topic in the social media and the real social at that period. Finally, we found the specific keywords which can show the relationship of the existing keywords from the i-WAM and the newly found keywords from the m-WAM such as “DJ.Pueak”/“ดีเจเผือก” and “Han River”/“แม่น้ำฮัน”, this popular disc jockey begs his girlfriend for the marriage at the Han river in the South Korea. This is an interesting entertainment issue in the online news and the social media in our experiment period. The “Virus”/“ไวรัส” and “MERS”/“ไวรัสเมอร์ส”, and “Prime Minister”/“นายกรัฐมนตรี” and “Counterrevolution”/“ปฏิวัติซ้อน” are also the good results.

TABLE VII. A PART OF THE M-WAM, EXISTING WORDS FROM THE I-WAM

	Bank of Thailand ‘ธนาคารแห่งประเทศไทย’	DJ. Pueak ‘ดีเจ เผือก’	Virus ‘ไวรัส’	Lose weight ‘ลดความอ้วน’
IDF(t)	**2.66875**	**3.73570**	**1.29427**	**1.61733**
Economic	0.13373	0	0	0
Entertain ment	0	0.01296	0	0
Foreign	0	0	0.90476	0
Lifestyle	0	0	0.00104	0.42016
Politic	0	0	0.00169	0
Social	0	0	0.01760	0
Sports	0	0	0	0
	Prime Minister ‘นายกรัฐมนตรี’	Law code ‘มาตรา’	Sea Games ‘ซีเกมส์’	
IDF(t)	**1.33892**	**1.41625**	**1.00870**	
Economic	0.05349	0.00573	0	
Entertain ment	0	0	0.00707	
Foreign	0.02961	0	0	
Lifestyle	0.00940	0.00313	0	
Politic	0.40258	0.04119	0.01486	
Social	0.01049	0.22754	0.00067	
Sports	0.00053	0	0.52594	

TABLE VIII. A PART OF THE M-WAM, NEWLY FOUND WORDS FROM TWITTER.

	Governor of the BOT ‘ผู้ว่าการ ธปท.’	Han River ‘แม่น้ำฮัน’	MERS ‘ไวรัสเมอร์ส’	How to Perfect ‘#Howto Perfect’
IDF(t)	**3.91179**	**3.43467**	**3.43467**	**3.20221**
Economic	0.01528	0	0	0
Entertain ment	0	0.00707	0	0
Foreign	0	0	0.24675	0.00109
Lifestyle	0	0	0.00104	0.23412
Politic	0	0	0	0
Social	0	0	0.01354	0.00067
Sports	0	0	0.00160	0
	Counter revolution ‘ปฏิวัติซ้อน’	Law code No. 44 ‘ม.44’	goal thailand ‘#goal thailand’	
IDF(t)	**1.35730**	**1.82901**	**1.82200**	
Economic	0	0.00668	0	
Entertain ment	0	0	0	
Foreign	0	0	0	
Lifestyle	0	0.00104	0	
Politic	0.31595	0.02590	0	
Social	0.00067	0.08465	0	
Sports	0	0	0.18550	

Tables IX. and Table X., show the Precision, Recall, and F-measure of the result of the cosine similarity from different criteria as listed below.

The i-WAM: The initial WAM

The m-WAM: The modified WAM based on the i-WAM

Text corpus1: Tweets collected by terms from the i-WAM

Text corpus2: Tweets collected by terms from the m-WAM

TABLE IX. THE PRECISION, RECALL, F-MEASURE RESULT (THE I-WAM WITH THE TEXT CORPUS1)

	Precision	**Recall**	**F-measure**
Economic	57.58%	57.58%	57.58%
Entertainment	94.84%	46.13%	62.07%
Foreign	32.38%	28.12%	30.10%
Lifestyle	93.81%	41.11%	57.17%
Politic	90.10%	90.10%	90.10%
Social	75.71%	52.54%	62.03%
Sports	71.78%	71.78%	71.78%

TABLE X. THE PRECISION, RECALL, F-MEASURE RESULT (THE M-WAM WITH THE TEXT CORPUS2)

	Precision	**Recall**	**F-measure**
Economic	93.45%	93.45%	93.45%
Entertainment	97.67%	97.67%	97.67%
Foreign	85%	85%	85%
Lifestyle	100%	90.56%	95.04%
Politic	100%	100%	100%
Social	97.44%	75.90%	85.33%
Sports	100%	100%	100%

The Precision, Recall and F-measure are improved when using m-WAM in all cases. This result reflect the proper technique we used to extract keywords from the Twitter text. The result in Table 7. A part of the m-WAM, existing words from the i-WAM, keywords from the i-WAM, extracted from online news, have a lot of common words. So, the Precision, and Recall value is too low especially in 'Foreign' category. However, when we collected the Twitter data, and extract the keywords, using the TF/IDF technique. It showed the most specific keywords related to their categories. This is a main reason that when we update the i-WAM to the m-WAM with more specific keywords, the result was showed promisingly.

V. Conclusion and Future Work

The Thai word segmentation tool is so important for doing experiments related to the text which written in Thai. And, the term frequency–inverse document frequency (TF-IDF) weighting is an effective technique to extract keywords. While, the Words Article Matrix (WAM) with the cosine similarity measure is a productive method to classify the text. The initial WAM can be generated from any official sources, already classified documents. Then, the modified WAM can be created through the specific keyword terms from the social media, Twitter. This modified WAM could be a suitable model for the social media text classification, and the set of keyword terms could be a representative of the real-time social interesting topics at the monitoring time. The growth and the information power of the social media text are remarkable. The keywords from the social media can be a prediction of the real social movement. We can do a holistic decision support system according to the interesting issues from the dynamic social media which is a concerned factor for today and in the future.

Finally, due to our research's time limitation, we can do just a proof of concept of the modified WAM efficiency. For more result accuracy, the important parts which can be developed are the Thai word segmentation tool and the number of iteration to update the modified WAM. The Thai word segmentation tool with more accuracy in scanning Thai word boundary can be selected, using the probabilistic model or the name entity recognition. And, the number of the repetition of the modified WAM updating should be identified, until the result of the Precision, Recall, and F-measure are nearly 100%.

Acknowledgment

This research is 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.

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

References

[1] K. Simon. (2015, Jan 21). Digital, Social & Mobile Worldwide in 2015[Online]. Available: http://wearesocial .net/blog/2015/01/digital-social-mobile-worldwide-2015/

[2] V. Monlamai. (2015, Jan 13). Digital, Social & Mobile Worldwide in 2015[Online]. Available:http://syndacast .com/infographic-online-marketing-thailand-the-state-of-social-media/

[3] Twitter, Inc. (2015). Character Counting [Online]. Available:https://dev.twitter.com/overview/api/counting-characters

[4] H. Wu et al., "Interpreting TF-IDF term weights as making relevance decisions" in Association for Computing Machinery Transactions on Information Systems, 2008. doi: 10.1145/1361684.1361686

[5] S. Virach et al., "Understanding Social Movement by Tracking the Keyword in Social" in MAPLEX2015, Yamagata, Japan, 2015.

[6] K.Canasai et al., "A Word and Character-Cluster Hybrid Model for Thai Word Segmentation" in InterBEST2009, Thailand, 2009.

[7] S.Vee. (2013, April 7). Thai Word Segmentation Tool using PHP [Online].Available:https://veer66.wordpress .com/2013/04/07/thai_word_breaker_in_php/

[8] M.Surapant et al., "Feature-based Thai Word Segmentation" in the Natural Language Processing Pacific Rim Symposium, Phuket, Thailand, 1997.

[9] C.S.Perone. (2011, September 18). Machine Learning Text feature extraction (tf-idf) [Online].Available:http://blog. christianperone.com/?p=1589

[10] [V.Jana. (2013, October 27). Tf-Idf and Cosine similarity [Online]. Available:https://janav.wordpress.com/2013/ 10/27/tf-idf-and-cosine-similarity/

[11] Vacharaphol Co.,Ltd. (2015, June). Thairath Online News [Online]. Available:http://www.thairath.co.th