Enhancement of Social Media Text Classification

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

[1] School of ICT, Sirindhorn International Institute of Technology, Thammasat University,
Pathum Thani, Thailand
phat.jotikabukkana @stud.siit.tu.ac.th
virach@siit.tu.ac.th
[2] Tokyo Institute of Technology, Ookayama Campus, Ookayama Meguro-ku, Tokyo, Japan
oku@pi.titech.ac.jp

[3] National Electronics and Computer Technology Center, Thailand Science Park,
Pathum Thani, Thailand
choochart.haruechaiyasak@nectec.or.th

Abstract. In this information age, social media is a powerful online communication tool for people to present their expression such as the real-time events report, personal information including their emotions. Social media text significantly demonstrate information in a current society, also indicate a trend of the dynamic social movement. However, these text characteristics are in a form of informal and unstructured language, i.e. using abbreviation, short text message, slang and argot. It is difficult to classify and extract the key information. There are many techniques proposed to categorize this kind of text information. The social media text classification by using Term Frequency-Inverse Document Frequency (TF-IDF) weighting with Word Article Matrix (WAM) and initial keywords from the well-formed source, i.e. online news article, is one of a productive technique. It generates a set of the sufficient keyword terms to represent text categories. In this paper, we discuss about the proper iteration number of the WAM updating process which generate the most effective word vector matrix that can classify the social media text effectively.

Keywords: social media text; Term Frequency-Inverse Document Frequency (TF-IDF) weighting; Word Article Matrix (WAM); sufficient keywords.

1 Introduction

Nowadays, social media become a very useful tool for people to communicate and share their information. This online tool creates a huge virtual community in the cyber space. In this virtual world, people feel free to propose their opinions and emotions. Twitter is one of the most popular social media application. Refer to recent statistics, there are 4.5 million twitter users in Thailand with nearly 2 million active users/day [1]. As the result, you will found many verbal text in the tweets, the Twitter text message. This is the main reason that doing a classification in the social media text is hard to conduct. The concept of classifying this kind of text messages by utilizing a well-formed source, i.e. online news articles [2], has proposed. This technique extracts a set of main keywords from the online news article that present in the written format and already categorized by publishers. This concept produce a promising result, the text from social media can be classified by a suitable word vectors in Word Article Matrix (WAM) table. However, there is an unidentified issue related to the performance improvement of this technique. We need to define the iteration number of the modified WAM modification for producing the most effective output, the result is converged to the steady state nearly 100% accuracy.

In this paper, section 2 explains the related works with main techniques such as web crawling, word segmentation, Term Frequency-Inverse Document Frequency (TF-IDF), and WAM. Section 3 explains our approach and experiment, while section 4 shows and discusses the experiment result. Finally, section 5 is conclusion.

2 Related Works

2.1 Web Crawling

In this experiment, we need to retrieve news article from the online news website. A web crawler, also known as a robot or a spider [3], is a main module to get access to the well-formed data source. First, we have to verify structure of our targeted websites, most of them are implemented with the Hyper Text Markup Language (HTML), the Extensible Markup Language (XML), and the Cascading Style Sheet (CSS). Second, specifying the Uniform Resource Locator (URL) and the news article part are needed as main parameters. Afterwards, we apply these parameters through the XML Path (XPath) query technique to retrieve our demanded data, online news articles. In this experiment we use the RapidMiner software [4] as a main web crawler module.

2.2 Word Segmentation

Word segmentation is a crucial step in text mining. We conduct experiment in the social media text written in the Thai language. The Thai language is written without spaces between words. In this paper, we use a word segmentation module applying the maximal matching algorithm to determine the word boundary [2]. We also updated the recent word lists in the dictionary before conduct the research. Consequently, the segmentation result is acceptable to determine the essential words for further processing in keyword identification.

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

TF-IDF is the widely used technique to extract the keywords from documents. It is composed of 2 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. It can defines as a counting function [5] (1).

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

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):

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

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

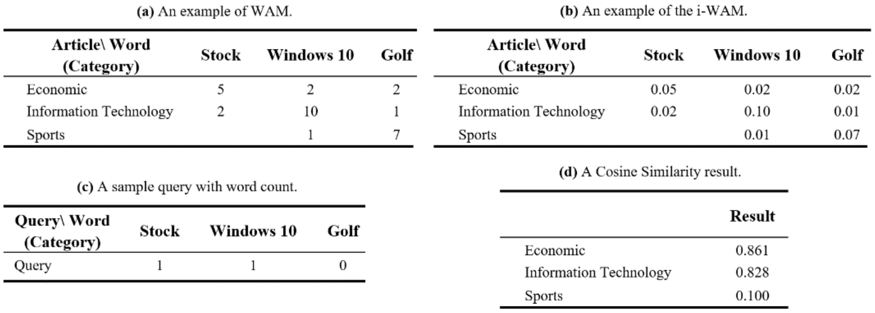
$$IDF(t) = log \frac{|D|}{1+|\{d:t \in d\}|} \tag{3}$$

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

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

2.4 Word Article Matrix (WAM)

In the Generic Engine for Transpose Association (GETA), WAM is a significant data structure. [6]. It models 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 Fig.1(a).We generate the initial WAM (i-WAM) by using the TF value of each word. For example, if we consider 10 documents as a training set with 100 words, total number of words of all documents. The i-WAM with the TF values will be shown in Fig.1(b). 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: "Microsoft stock got a small boost from the launch of Windows 10". This query is converted into a model of word vectors shown in Fig.1(c).



(a) An example of WAM.

Article\ Word (Category)	Stock	Windows 10	Golf
Economic	5	2	2
Information Technology	2	10	1
Sports		1	7

(b) An example of the i-WAM.

Article\ Word (Category)	Stock	Windows 10	Golf
Economic	0.05	0.02	0.02
Information Technology	0.02	0.10	0.01
Sports		0.01	0.07

(c) A sample query with word count.

Query\ Word (Category)	Stock	Windows 10	Golf
Query	1	1	0

(d) A Cosine Similarity result.

	Result
Economic	0.861
Information Technology	0.828
Sports	0.100

Fig. 1. An example of WAM, i-WAM, and sample result

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

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

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 $d2$.

Afterwards, we calculate the cosine similarity values and get a result of an example query as shown in Fig.1(d). As the weight of a word "Stock" in economic category is high, 0.5.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 0.861.

3 Our Approach and Experiment

Our approach focus on the iteration number of the modified WAM (m-WAM) modification. First, we retrieve news articles from our representative well-formed source, Dailynews online news [8]. We got written style document with proper grammar and appropriately classified by publishers. There are 7 categories of the news articles, economic, entertainment, foreign, information technology (IT), politics, regional, and sports. Then, we extract the keywords from these data by using the Thai word segmentation module and the TF-IDF weighting technique. The keywords with high TF-IDF values are selected as the word to implement the i-WAM as shown in Fig. 2.

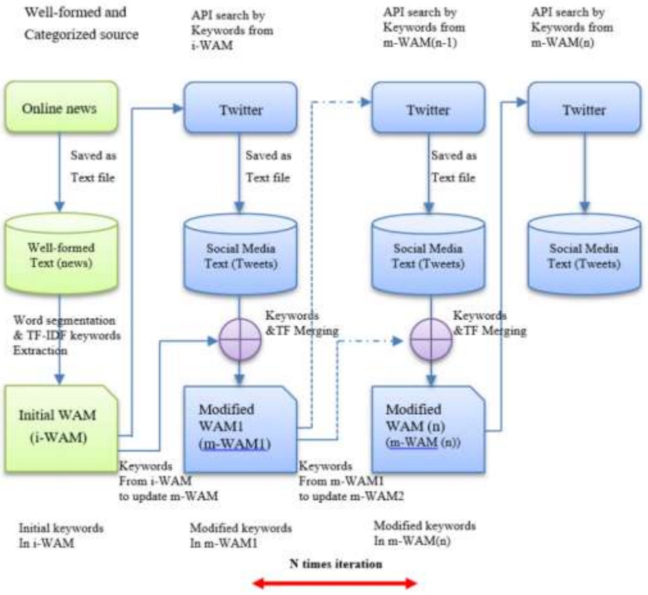


Fig. 2. The m-WAM implementation, and n times iteration of the m-WAM modification.

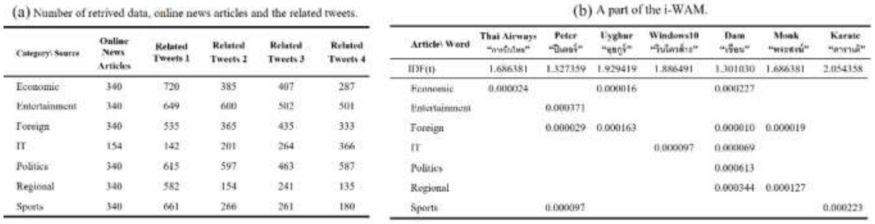
Then we use these keywords to collect the related tweets through Twitter search API. API allows to collect the related social media text which a search index has a 7-day limit search back [9]. We got a heap of tweets saved in the text file format. Afterwards, we conduct the same process to extract keywords by using Thai word segmentation and 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 the social media.

For the 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 the i-WAM is recomputed additional count. The newly found words with their TF values are added into the table. All TF values of words in the i-WAM and newly found words from the related tweets are normalized by using the L2-normoalization, Euclidian norm, as the word vector normalization process.

The evaluation of social media text classification is conducted manually. The collected tweets related to category are evaluated by human judging. The Precision, Recall, F-measure, and accuracy value are calculated. We do the same process of the m-WAM modification, iterate this procedure until we get the result of Precision, Recall, F-measure, and accuracy value are in the steady state, nearly 100%. 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.

4 Experiment Result

After crawling online news data by using web crawler module, we have got around 2,200 online news articles as shown in Fig.3(a). We extract a set of the keywords by selecting the words with highest TF-IDF score and generate the initial-WAM (i-WAM) as shown in Fig.3(b). We add a row to show an IDF value of each keyword to identify the important weight of each keyword. The word "Thai Airways"/"การบินไทย", "Windows10"/"วินโดวส์10", and "Karate"/"คาราเต้" are a sample of the specific keywords which effectively represent their category of economic, information technology (IT), and sports respectively. There are some common words which appear in more than one category, such as "Peter"/"ปีเตอร์", "Uyghur"/"อุยกูร์" and "Dam"/"เขื่อน". However, they can be a representative of their category when consider their normalized TF values in the i-WAM table.



(a) Number of retrived data, online news articles and the related tweets.

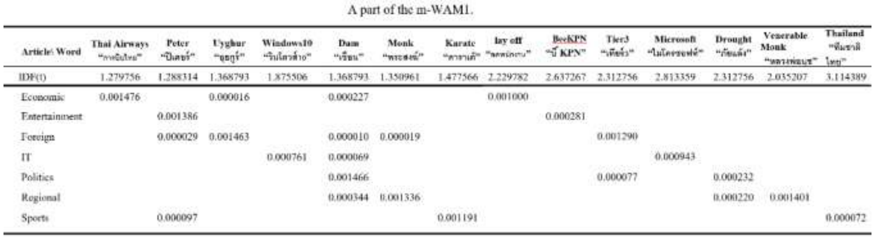
Category\ Source	Online News Articles	Related Tweets 1	Related Tweets 2	Related Tweets 3	Related Tweets 4
Economic	340	720	385	407	287
Entertainment	340	649	600	502	501
Foreign	340	535	365	435	333
IT	154	142	201	264	366
Politics	340	615	597	463	587
Regional	340	582	154	241	135
Sports	340	661	266	261	180

(b) A part of the i-WAM.

Article\ Word	Thai Airways "การบินไทย"	Peter "ปีเตอร์"	Uyghur "อุยกูร์"	Windows10 "วินโดวส์10"	Dam "เขื่อน"	Monk "พระสงฆ์"	Karate "คาราเต้"
IDF(t)	1.686381	1.327359	1.929419	1.886491	1.301030	1.686381	2.054358
Economic	0.000024		0.000016		0.000227		
Entertainment		0.000371					
Foreign		0.000029	0.000163		0.000010	0.000019	
IT				0.000097	0.000069		
Politics					0.000613		
Regional					0.000344	0.000127	
Sports		0.000097					0.000223

Fig. 3. Number of the retrieved data, related tweets, and a part of the i-WAM.

Then, we select a set of new keywords from each category to search the related tweets. We have got around 4,000 twitter messages, and extract the newly found keywords. Then we generate the m-WAM1 from these new specific keywords from the social media and merge with the existing keywords from the i-WAM, well-formed text source keywords. The newly found words in the m-WAM1 show the significant result. For example, "Thai Airways"/"การบินไทย", the sample keywords from the i-WAM lead us to found a new keyword, "lay off"/"ลดพนักงาน", which scope down the words vector for the economic category. Another category keywords also generated a promising result as shown in Fig.4.

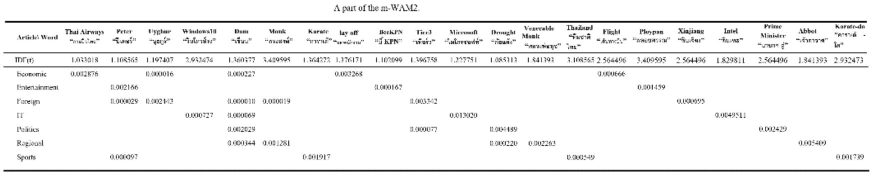


A part of the m-WAM1.

Article\ Word	Thai Airways "การบินไทย"	Peter "ปีเตอร์"	Uyghur "อุยกูร์"	Windows10 "วินโดวส์10"	Dam "เขื่อน"	Monk "พระสงฆ์"	Karate "คาราเต้"	lay off "ลดพนักงาน"	BeeKPN "บี KPN"	Tier3 "เทียร์3"	Microsoft "ไมโครซอฟต์"	Drought "ภัยแล้ง"	Venerable Monk "[illegible]"	Thailand "ทีมชาติไทย"
IDF(t)	1.279756	1.288314	1.368793	1.875506	1.368793	1.350961	1.477566	2.229782	2.637267	2.312756	2.813359	2.312756	2.035207	3.114389
Economic	0.001476		0.000016		0.000227			0.001000						
Entertainment		0.001386							0.000281					
Foreign		0.000029	0.001463		0.000010	0.000019				0.001290				
IT				0.000761	0.000069						0.000943			
Politics					0.001466					0.000077		0.000232		
Regional					0.000344	0.001336						0.000220	0.001401	
Sports		0.000097					0.001191							0.000072

Fig. 4. A part of the m-WAM1.

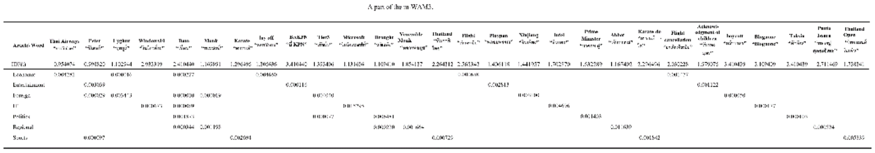
Afterwards, we conduct the same process, the m-WAM modification. We select the high potential keywords from the m-WAM, words with highest TF-IDF score (top 5) in their own category, to gather all related tweets. From this technique, we found more specific keywords, less of common words, which effectively represent their category. So, we can generate the new m-WAM which be a productive model for the social media text classification. We repeat this procedure until we found the steady state of the Precision, Recall, F-measure, and accuracy result. Finally, the number of the iteration of the m-WAM modification which satisfied the best performance of social media text classification is 3, the i-WAM, the m-WAM1, the m-WAM2 and the m-WAM3. All of the result can show as Fig.5., and Fig.6.



A part of the m-WAM2.

Article\Word	Thai Airways "การบินไทย"	Peter "ปีเตอร์"	Uyghur "อุยกูร์"	Windows10 "วินโดวส์10"	Dam "เขื่อน"	Monk "พระสงฆ์"	Karate "คาราเต้"	lay off "เลิกจ้าง"	BeeKPN "บี้ KPN"	Tier3 "เทียร์3"	Microsoft "ไมโครซอฟท์"	Drought "ภัยแล้ง"	Venerable Monk "หลวงพ่อพุธ"	Thailand "ทีมชาติไทย"	Flight "เที่ยวบิน"	Ploypan "พลอยพรรณ"	Xinjiang "ซินเจียง"	Intel "อินเทล"	Prime Minister "นายกฯ ตู่"	Abbot "เจ้าอาวาส"	Karate-do "คาราเต้-โด"
IDF(t)	1.033018	1.108565	1.197407	2.932474	1.360377	3.409595	1.364272	1.376171	1.102099	1.396758	1.227751	1.085313	1.841393	3.108565	2.564496	3.409595	2.564496	1.829811	2.564496	1.841393	2.932473
Economic	0.002876		0.000016		0.000227			0.003268							0.000666						
Entertainment		0.002166							0.000167							0.001459					
Foreign		0.000029	0.002443		0.000010	0.000019				0.003342							0.000695				
IT				0.000727	0.000069						0.013020							0.0049511			
Politics					0.002029					0.000077		0.004489							0.002429		
Regional					0.000344	0.001281						0.000220	0.002263							0.005409	
Sports		0.000097					0.001917							0.000549							0.001739

Fig. 5. A part of the m-WAM2.



A part of the m-WAM3.

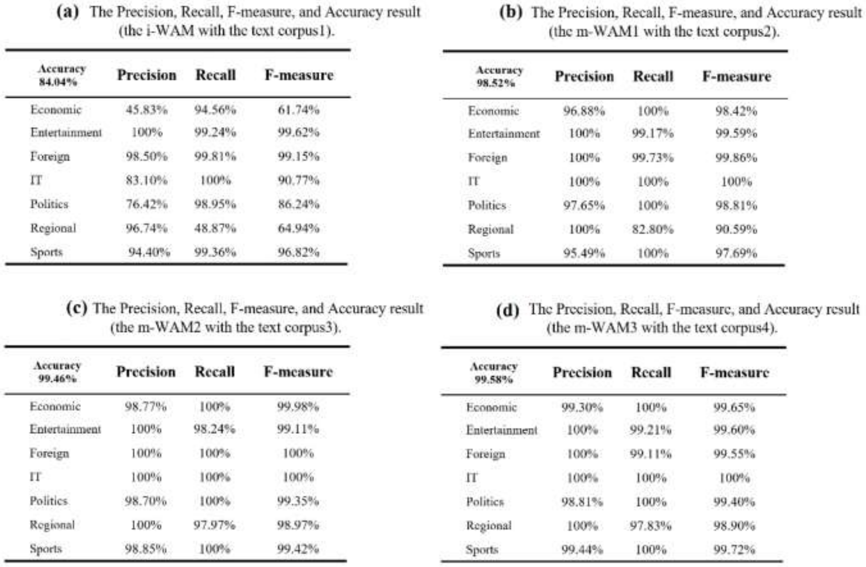
Article\Word	Thai Airways "การบินไทย"	Peter "ปีเตอร์"	Uyghur "อุยกูร์"	Windows10 "วินโดวส์10"	Dam "เขื่อน"	Monk "พระสงฆ์"	Karate "คาราเต้"	lay off "เลิกจ้าง"	BeeKPN "บี้ KPN"	Tier3 "เทียร์3"	Microsoft "ไมโครซอฟท์"	Drought "ภัยแล้ง"	Venerable Monk "หลวงพ่อพุธ"	Thailand "ทีมชาติไทย"	Flight "เที่ยวบิน"	Ploypan "พลอยพรรณ"	Xinjiang "ซินเจียง"	Intel "อินเทล"	Prime Minister "นายกฯ ตู่"	Abbot "เจ้าอาวาส"	Karate-do "คาราเต้-โด"	Flight cancellation "ยกเลิกเที่ยวบิน"	Acknowledgement of children "รับรองบุตร"	boycott "คว่ำบาตร"	Blognone "Blognone"	Taksin "ทักษิณ"	Punta Joana [illegible]	Thailand Open "ไทยแลนด์ โอเพ่น"
IDF(t)	0.954074	0.598820	1.102944	2.933319	2.410440	1.168891	1.296405	1.305636	3.410440	1.393406	1.131656	1.109410	1.854137	2.264812	2.565342	1.406118	1.441937	1.702570	1.632389	1.167403	2.296456	2.080228	1.979076	3.410439	3.109409	3.410439	2.711469	1.738341
Economic	0.004283		0.000016		0.000227			0.004650							0.000646							0.001177						
Entertainment		0.003059							0.000115							0.002815							0.001122					
Foreign		0.000029	0.005413		0.000010	0.000019				0.003350							0.000100							0.000070				
IT				0.000673	0.000069						0.015758							0.004656							0.000137			
Politics					0.001843					0.000077		0.005481							0.001403							0.000108		
Regional					0.000344	0.001193						0.000220	0.001654							0.011639							0.000524	
Sports		0.000097					0.002694							0.000725							0.001542							0.005339

Fig. 6. A part of the m-WAM3.

For the evaluation process, Fig.7(a)-(d) 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-WAM1: The modified WAM based on the i-WAM
The m-WAM2: The modified WAM based on the m-WAM1
The m-WAM3: The modified WAM based on the m-WAM2

Text corpus1: Tweets collected by terms from the i-WAM
Text corpus2: Tweets collected by terms from the m-WAM1
Text corpus3: Tweets collected by terms from the m-WAM2
Text corpus4: Tweets collected by terms from the m-WAM3



(a) The Precision, Recall, F-measure, and Accuracy result (the i-WAM with the text corpus1).

Accuracy 84.04%	Precision	Recall	F-measure
Economic	45.83%	94.56%	61.74%
Entertainment	100%	99.24%	99.62%
Foreign	98.50%	99.81%	99.15%
IT	83.10%	100%	90.77%
Politics	76.42%	98.95%	86.24%
Regional	96.74%	48.87%	64.94%
Sports	94.40%	99.36%	96.82%

(b) The Precision, Recall, F-measure, and Accuracy result (the m-WAM1 with the text corpus2).

Accuracy 98.52%	Precision	Recall	F-measure
Economic	96.88%	100%	98.42%
Entertainment	100%	99.17%	99.59%
Foreign	100%	99.73%	99.86%
IT	100%	100%	100%
Politics	97.65%	100%	98.81%
Regional	100%	82.80%	90.59%
Sports	95.49%	100%	97.69%

(c) The Precision, Recall, F-measure, and Accuracy result (the m-WAM2 with the text corpus3).

Accuracy 99.46%	Precision	Recall	F-measure
Economic	98.77%	100%	99.98%
Entertainment	100%	98.24%	99.11%
Foreign	100%	100%	100%
IT	100%	100%	100%
Politics	98.70%	100%	99.35%
Regional	100%	97.97%	98.97%
Sports	98.85%	100%	99.42%

(d) The Precision, Recall, F-measure, and Accuracy result (the m-WAM3 with the text corpus4).

Accuracy 99.58%	Precision	Recall	F-measure
Economic	99.30%	100%	99.65%
Entertainment	100%	99.21%	99.60%
Foreign	100%	99.11%	99.55%
IT	100%	100%	100%
Politics	98.81%	100%	99.40%
Regional	100%	97.83%	98.90%
Sports	99.44%	100%	99.72%

Fig. 7. The Precision, Recall, F-measure, and Accuracy

Due to there are more common keywords in the i-WAM, the Precision, Recall, and F-measure score are too low, especially in the economic and politics category. However, when we update the m-WAM with more specific keywords from the related social media information, all of the evaluation factor are increased dramatically. Finally, the value of Precision, Recall, F-measure, and accuracy are nearly converged to 100% after we update the m-WAM3 as shown in Fig. 8.

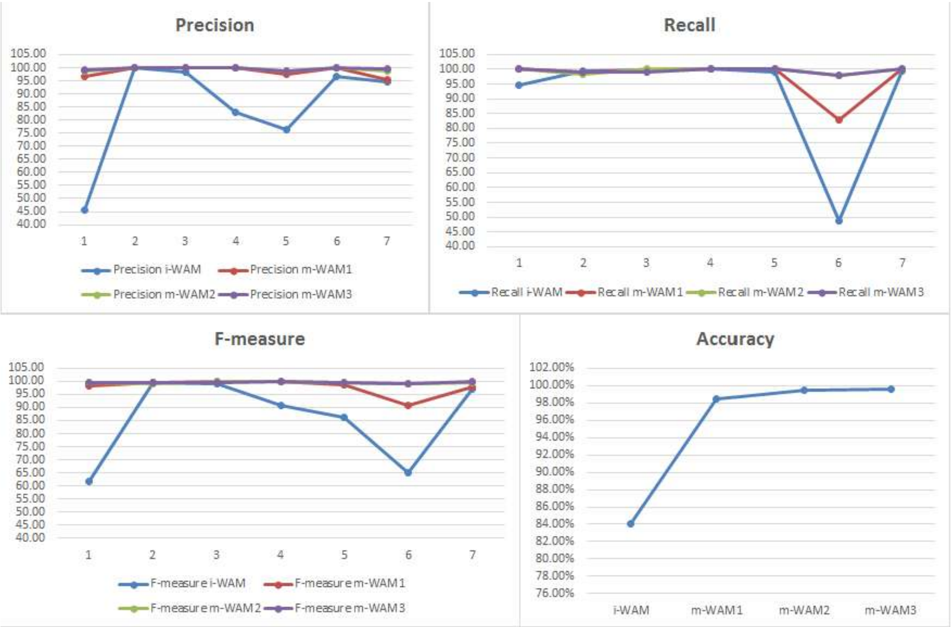


Fig. 8. The result of Precision, Recall, F-measure, and accuracy, after modified the m-WAM "3 times".

5 Conclusion

The social media text classification by using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting technique and the Word Article Matrix (WAM) is so effective. We can categorize text from social media with a sense of human familiarity, utilizing the online news categories which already categorized by the publishers. We can expect the good result from the proper modified WAM (m-WAM) for social media text classification after we have updated it for 3 times, the suitable iteration number of the m-WAM modification. However, the great result is based on the performance of the Thai word segmentation module also. Another productive Thai word segmentation, such as Name Entity Recognition (NER) [10], can generate a proper word boundary for conducting the other processes. We can generate more accurate keywords and the model's accuracy will be improved significantly. The last concerned issue is the limitation of the Twitter search API, 7 days search back. It is a time limitation for the researchers who conduct the experiment related to the Twitter message, the result of keywords will be missed by shifting of the searching timeframe. The researchers should manage their experiment period to produce the accurate and reliable result.

6 Acknowledgement

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.

References

1. Digital, Social & Mobile Worldwide in 2015, http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/
2. Phat J., Virach S., Okumura M., Choochart H., : Effectiveness of Social Media Text Classification by Utilizing the Online News Category: The 2015 International Conference On Advanced Informatics. In: Concepts, Theory And Application (ICAICTA2015)
3. Christopher O., Marc N.,: Web Crawling. In: Foundation and Trends in Information Retrieval. Vol. 4, No. 3(2010), pp. 175-246
4. RapidMiner, https://rapidminer.com/
5. Ho C.W., Robert W.P.L., Kam F.W., Kui L.K. :Interpreting TF-IDF term weights as making relevance decisions. In: Association for Computing Machinery Transactions on Information Systems, 2008. doi: 10.1145/1361684.1361686
6. Kobkrit V., Susumu K., Mizuhito O.,: A Comparison of Four Association Engines in Divergent Thinking Support System on Wikipedia. In: A Comparison of Four Association Engines in Divergent Thinking Support System on Wikipedia (KICSS'10).
7. Tf-Idf and Cosine similarity, https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/
8. Daily news, http://www.dailynews.co.th
9. Get Search/Tweets, https://dev.twitter.com/rest/reference/get/search/tweets
10. Tepdang S., Haruechaiyasak C., Kongkachandra R.,: Improving Thai word segmentation with Named Entity Recognition. In: 10th International Symposium on Communications and Information Technologies, 2010. doi: 10.1109/ISCIT.2010.5665124