SOCIAL MEDIA TEXT CLASSIFICATION BY ENHANCING THE WELL-FORMED TEXT TRAINED MODEL

BY

PHAT JOTIKABUKKANA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF MASTER OF
ENGINEERING (INFORMATION AND COMMUNICATION
TECHNOLOGY FOR EMBEDDED SYSTEMS)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2015

SOCIAL MEDIA TEXT CLASSIFICATION BY ENHANCING THE WELL-FORMED TEXT TRAINED MODEL

BY

PHAT JOTIKABUKKANA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF MASTER OF
ENGINEERING (INFORMATION AND COMMUNICATION
TECHNOLOGY FOR EMBEDDED SYSTEMS)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2015

SOCIAL MEDIA TEXT CLASSIFICATION BY ENHANCING THE WELL-FORMED TEXT TRAINED MODEL

A Thesis Presented

By

PHAT JOTIKABUKKANA

Submitted to

Sirindhorn International Institute of Technology

Thammasat University

In partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING (INFORMATION AND COMMUNICATION TECHNOLOGY FOR EMBEDDED SYSTEMS)

Approved as to style and content by

Advisor and Chairperson of Thesis Committee ____________________

(Dr. Virach Sornlertlamvanich)

Committee Member ____________________

(Prof. Manabu Okumura)

Committee Member ____________________

(Prof. Dr. Thanaruk Theeramunkong)

Committee Member and
Chairperson of Examination Committee ____________________

(Dr. Choochart Haruechaiyasak)

MAY 2016

Acknowledgements

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.

I would also like to show my gratitude to Dr. Eakasit Pacharawongsakda, Ms.Nungnut Rodsri, and Ms.Kansinee Panprasert for assistance with data mining techniques and comments that greatly improved the manuscript during the course of this research.

Abstract

SOCIAL MEDIA TEXT CLASSIFICATION BY ENHANCING THE
WELL-FORMED TEXT TRAINED MODEL
by

PHAT JOTIKABUKKANA

Bachelor of Engineering (Computer Engineering), Mahidol University, 2000
Master of Engineering (Information and Communication Technology for Embedded System), Sirindhorn International Institute of Technology, Thammasat University, 2015

Social media mining is an important process to extract beneficial information from social media. Classification is a major task for this process while it is so difficult to deal with vast and noisy data like social media text, many slangs, argots, and absent words. I focused on utilizing a vector space model, and well-formed text source to generate an initial model for the social media text classification. This kind of data source like online news articles are primarily proofed and classified by publishers. A bag-of-words which contained high quality and variety of words could generate well and stable word vector. Moreover, machine be learned classes of online news as self-learning classes of the model automatically. Term Frequency Inverse Document Frequency (TF-IDF) and Word Article Matrix (WAM) are used as main techniques to extract keywords and build the beginning vector space model. Set of keywords be used to search Twitter message text to enhance the model by newly found words from the social media. The TF-IDF merging with terms weighting parameter is considered to grant more weight for the new terms. I iterated this procedure until I found the proper state of the model which generated the promising result, Precision, Recall, and F-measure score are stable nearly 100%. Finally, the model was updated three times and found the accuracy of the model at 99.79%, average Precision at 99.31%, average Recall at 99.28%, and average F-measure at 99.28%.

Keywords: social media text, well-formed text, online news, self-learning, Term Frequency-Inverse Document Frequency (TF-IDF), Word Article Matrix (WAM).

Table of Contents

Chapter **Title** **Page**

List of Tables

Tables **Page**

List of Figures

Figures **Page**

Chapter 1
Introduction

"The golden age of Social Media Mining is now. A large amount of useful data could be dug freely. However, a classification task for this kind of data is arduous to deal with. The utilization of well-formed text source as primary state and model enhancing by the social media text itself, relevant update, is an interesting concept to discover as the effective solution."

1.1 Objective

Nowadays, a number of social media active users is soared up dramatically, more than 2 billion accounts compared with a number of people in the world at 7.2 billion [1]. People have used this kind of tool to express their opinion freely. In addition, many mass media already used this channel to broadcast their news and information. We can see a large amount of useful information and comments from the social media which can generate a huge impact in the real social in many dimensions: business, politics, socialization, disaster awareness, etc. Many researchers and decision makers have aimed their attraction to this novel data source. Twitter is one of the most popular social media communication tools. There are more than 300 million active users with around 350,000 Twitter message (tweets) per minute, 500 tweets per day [2], and average 95% growth in active users through 2014 [3]. Moreover, the tweets data structure and support API are very helpful and convenient for researchers to deal with. The tweet is created as JSON file format with 140 characters limited text file [4]. Furthermore, the Twitter search API is a very useful programming interface to collect our required information from Twitter with 180 queries per 15 minutes windows [5] and last seven day search back as its limitation [6]. So, these are the main reason that I decided to use Twitter as main data source to do experiment related to the social media.

Social media mining is a morning star of information science field in this digital information era. The major challenge is analyzing short message text like tweets. It looks like colloquialism text, compared with written document, a lot of informal languages, slang, and absent words. So, classification process for this kind of data is very difficult to handle. I am concerned about this problem and would like to find an effective algorithm to classify the social media text productively.

1.2 Scopes

In my experiment, I decided to utilize a vector space model and well-formed text source to build an initial model for classification. Quality of words, variety of words, and coverage area of words in each considered category are important factors of model efficiency. Online news article, well-formed text, is a very useful data source. It contains written format text which already proofed and classified by publishers. We can extract all terms, and news category into bag-of-words related to its category. The online news categories become model classes with a sense of human familiarity automatically as a self-learning concept, e.g. economic, entertainment, foreign, lifestyle, politic, social, sports. Term Frequency–Inverse Document Frequency (TF-IDF) is used to glean all terms from the online news articles with their significant values to create Word Article Matrix (WAM), one kind of vector space model, as the beginning model. The quality and variety of words from the well-formed text source generates efficient word vectors. A relative update with terms from Twitter, search by keywords from the well-formed text, is used to enhance the model in part of the coverage area expansion for each class. The significant word from the social media which never appear in the formal text, abbreviation, argot, slang, hashtag, is the important factor to increase the model efficiency. The iteration of model update can also increase the model accuracy for the social media text classification until we found the stable state of the model, accuracy, precision, recall, and F-measure are nearly 100%.

1.3 Limitations

As social media text, tweets, and online news articles are the time series information, terms and keywords will be changed by the real social events and interestings in any research period. And refer to a seven days search back is a main limitation of the Twitter Search API [6], all processes in this experiment need to be done within 7 days period. Another solution is capturing all tweets and online news articles as snap shot databases. Then data analyzing and evaluating could be done precisely.

Another concern is a specification of the computer to run a model in this experiment. As a vector space model, bag-of-words, concept, the model will generate a very large matrix of terms and articles with a lot of floating point computation. The computer memory, and the CPU speed are the first and second concerns respectively. More main memory, 8GB or higher, and high CPU speed could reduce the computational time dramatically.

In Section 2, related research works are explained. In Section 3, my approach and main techniques are described. Then, in section 4, the experiment result is illustrated and discussed. Finally, in Section 5, a conclusion of this experiment and a discussion of a future work is shown.

Chapter 2
Literature Review

"Many techniques are proposed for Social Media Text Classification task. The Vector Space Model, Word Article Matrix (WAM), is one of the most effective algorithm to handle this difficult job."

A number of recent papers have addressed the social media text classification. Irfan et al. [7] review different text mining techniques to discover various textual patterns form the social web. In social media, people always use informal conversation, wrong spelling and inaccurate grammatical sentence. Using unstructured or semi-structured language may leads to different types of ambiguities such as lexical, syntactic, and semantic [8]. Therefore, a critical task is extracting logical pattern with precise information from this kind of unstructured data. Text mining become more complex as compared to datamining due to unstructured and fuzzy nature of natural language text [9]. Many text mining techniques are purposed in the past few year to extract significant text pattern from online data sources. However, many of existing research papers did not mentioned the pre-processing phase. Irfan et al. [7] mentioned the pre-processing in the text mining as an important phase for the simplification of text mining process. The "garbage in garbage out" problem may ocuur if the text has not been considered carefully [10]. Unstructured text may leads to poor text analysis which affects the accuracy of the model [11]. Two basic methods of the text pre-processing: (a) feature extraction and (b) feature selection, are reviewed. The Feature Extraction (FE) can be separated as Morphological Analysis (MA), Syntactical Analysis (SA) and Semantic Analysis (SA). Dealing with individual worfs which represented in a text with tokenization, removing stop word and word stemming are main process of MA. The Sysntactical Analysis deals with Part-Of-Speech tagging (POS) and parsing which usually used to add contextually related grammatical knowledge of a single word in a sentence. The POS tagging based on dictionarires [12], rule based morphological analysis and stochastic model such as Hidden Markov Model (HMM), is the most promising approach. While, the WordNet [13] and the SentiWordNet [14] are used to find meaning, synonym, and emotion measuring as a keyword spotting technique. In addition, the semantic network is introduced as a new paradigm to overcome the limitation of keyword spotting technique to achieve true understanding [15]. The Feature Selection (FS) is used to eliminate unrelated information from the considerd text. FS selects important features by scoring the words. The importance of the word in the document is represented by the identified score [16]. Term Frequency-Inverse Document Frequency (TFIDF) is a widely used technique to calculate feature vectors and relevancy of word in a document as a vector space model, text document [17]. Laten Semantic Indexing (LSI) and Random Mapping (RM) are another two commonly used technique to improve the lexical matching and similar words extraction [18]. The text mining using classification with various algorithm, machine learning based and ontology based, also hybrid approach are reviewed. The machine learning based text classification consists of Rocchio Algorithm (RA), Instance Based Learning Algorithm, Decision Tree (DT) and Support Vector Machine (SVM), Artificial Neural Network (ANN), and Genetic Algorithm (GA). The ontology based text classification introduces explicating of conceptualization based on concepts, descriptions and the semantic relashionship [19]. It is categorized as (a) Domain Ontology (DO), concepts and relashionship in particular domain, and (b) Ontology Instance (OI), related with automatic generation of web pages [20]. I found that there is no algorithm perform as the best one for all kind of data set. For better performance of the hybrid approach, several parameters need to be defined in advance.

Patel et al. [21] review the different types of classifiers used for text classification and having an eye on their advantages and disadvantages. The definite task in the classification is text representation, represented by collecting the set of features. Bag-of-words is represented as the set of owrds presence in the documents and their significant score, allied frequency of weights [22]. Document Frequency Threshold, Information gain, Mutual information, and Chi-square statistics are used as the feature selection methods. Six different algorithm are reviewed: Bayesian Classifier (BC), Decision Tree (DT), K-nearest neighbor (K-NN), Support Vector Machine (SVM), Neural Network (NN), and Rocchio's. BC is the most commonly used technique. The main idea is to find a probability of which class that document will belong to. This technique is used for anti spam filtering and it works well in supervised leraning environment. DT designes a hierarchical decomposition of the data space. Single attribute split, Similarity-based multi-attribute split, and Dimensional-based multi-attribute split are kinds of splits in the DT which are implemented in the text context tend to be small variations compared to ID3, C4.5 for the purpose of the text classification [22]. K-NN calculates the Eucledian distance or Cosine similarity between test document and each neighbor. It is a case-based learning algorithm which the optimal k value is very difficult to find out. SVM will identify the linear hyper plane that maximize the margin. The representatives of document which are nearest to the decision surface be called as the support vector. NN consists of a large number of neurons nodes, processing elements which working accordingly to slove any specific problem. This technique can slove nonlinear separable case effectively. Rocchio's is implemented by using relevant feedback method. Synomym can be interpreted by manipulating the document using the relevant feedback method, an iterative process. Finally, I found that the common disadvantage of all algorithm is performance. Some of them are easy to implement while their performance is very poor. Some of them perform greatly but they needs more time to train and tune parameters.

Lee et al. [23] classify 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 themselves and classify the Trending Topics into these categories. In text-based classification, word vectors are constructed with trending topic definition and tweets. The TF-IDF weighting technique is used to classify the topics using a Multinominal Naïve Bayes (NB) [24]. In network-based classification, top 5 similar topics based on the number of influential users are identified. The categories of the similar topics and the number of common influential users between the given topic using C5.0 decision tree learner. As the final result, network-based classifier perform significantly better than text-based classification.

Kateb et al. [25] discuss methods which overcome stream data problem of classify short text in social media. In classification techniques section, they present a useful issue that need to be considered before text classification in general; 1. Define the research goal. 2. Does speed matter? 3. What is the size of the data? The algorithm should be fast and efficient. It is difficult to achieve both of them at the same time with stream data. The incoming data need to bemeasured in order to select the best fit algorithm [26]. In the challenge of short text classification section, they discuss about advantages and drawbacks of current approaches which working on Twitter data. Many researchers combine many tweets as a single document to extract the significant keywords and summarize the document. Some researchers collect tweets based on time frame as a daily document to measure the topic's popularity, and perform time series analysis to track topic appearance in each document [27]. Another approaches, single tweet is considered as one document. This concept is used for a sentiment analysis purpose. They also discuss the way to overcome difficulties in classifying short text like tweets. Enriching the Twitter posts is the method to make short text to be longer one. The related content is needed to be added into the target tweets. Enriching procedure with using an internal sources can be done by extracting word synonyms from tweets. WordNet and word semantic analysis are the main technique to handle this case. In contrast, enriching procedure with an external sources can be done by using the external content such as news articles. This approach links a tweets to the content of news articles found at the URL in the tweets itself. The purpose is to understand the meaning of hashtags or ambiguous content in the tweets. They measure the similarity between a tweet and news article with TF-IDF score [28]. From these concerned issues, I can locate a suitable technique (classification, regression or clustering) and suitable algorithm for conducting our experiment.

Chirawichitchai et al. [29] compare six methods of feature weight in Thai document categorization framework. Boolean, TF, TF-IDF, TFC, ltc, and Entropy weighting are evalated. The evaluation is done by using Thai news articles corpus with three supervised learning classifiers, SVM, DT, and NB. They found ltc weighting with SVM yielded the best performance for Thai document categorization.

Theeramunkong et al. [30] propose a multi-dimensional framework for classifying text documents. This framework classifies each text document in a collection using multiple predefined sets of categories. Each set corresponds to a dimension. This approach can slove a text document problem with a large number of classes or a large hierarchy. KNN, NB, and centroid-based classifiers are used to evaluate this framework. Finally, classifying text documents based on a multi-dimensional category model by using the multi-dimensional-based and hierarchy-based classifications beat the flat-based classification.

Viriyayudhakorn et al. [31] compare four divergent thinking support engines using the associative information extracted from the Wikipedia. They use Word Article Matrix (WAM), a vector space model, to compute association function. WAM is commonly a very large sparse matrix of which columns are indexed by words and row are indexed by names of documents. The similarity function, inner product, is used to verify the associated calss of any queries. It is the useful and effective technique for divergent thinking support.

Sornlertlamvanich et al. [32] propose a new method to fine tune the model trained from some known documents containing richer context information. They use WAM to classify text and track keywords from social media to understand the social movement. TF-IDF is used to assigned words score, and find the keywords from tweets. WAM with cosine similarity measure is an effective way for text classification.

After reviewing related literatures, there are a lot of techniques for social media text classification. While all algorithms still have some complex issues related to their performance. This reason inspire me to adapt some useful techniques to be a novel simple way to effectively classify social media text with a sense of human familiarity. Viriyayudhakorn et al. [31] and Sornlertlamvanich et al. [32] just use WAM for their specific purposes, the divergent thinking support and keywords tracking, while I focus on using WAM to classify the social media text with the additional techniques of text classes self-learning (semi-supervised learning) and enhancing WAM model by the specific keywords from the social media until it come to be the most suitable model for the social media text classification [33] as I describe as follows.

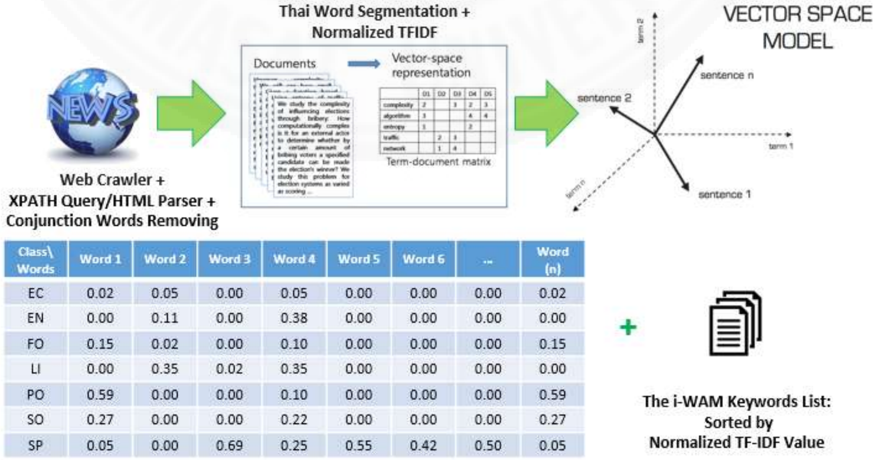
Chapter 3
Experiment

“The quality of bag-of-words: correctness, variety, coverage area of words, is considered as the main important success factor for the Vector Space model. The initial step with the online news articles utilization with self-learning class concept, the relevant updating to enhance model productivity, and iteration of model updating until model be the most suitable one for the social media text classification are proposed as the complete solution.”

There are three main parts of this experiment. First, main concept, a whole idea to conduct the experiment. Second, main techniques which I have used. Third, my approach model for the efficient social media text classification.

3.1 Main Concept

The quality of bag-of-words: correctness, variety, coverage area of words, is an important factor to build the effective vector space model for classification process. Main idea of this experiment is utilizing the well-formed text, online news articles, as a primary data source to create a productive beginning Word Article Matrix (WAM), one kind of the vector space model as shown in Figure 1.

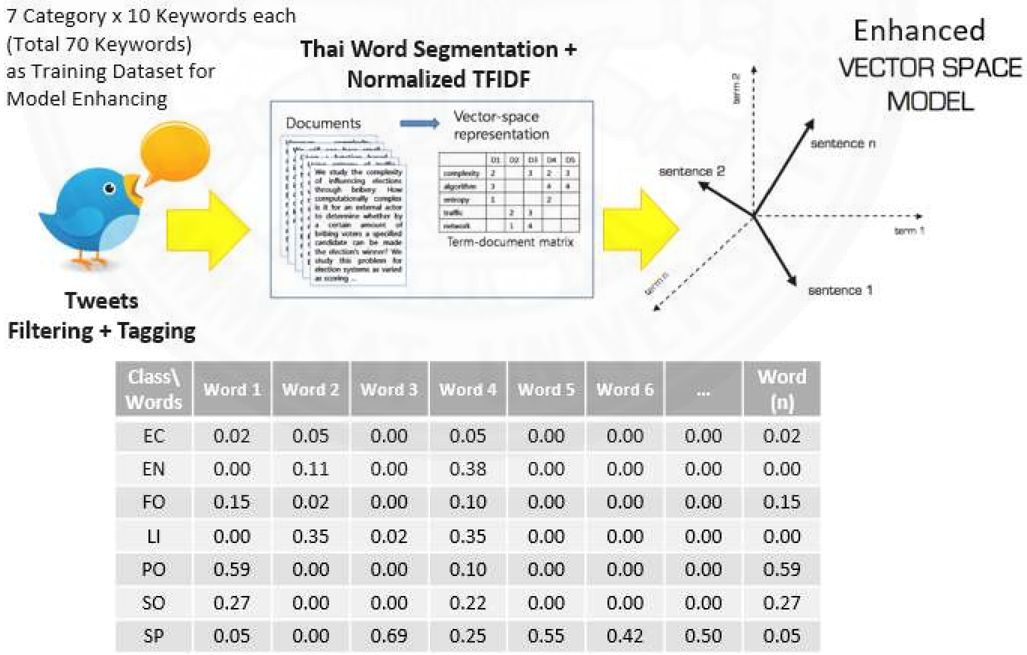


Class\ Words	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	...	Word (n)
EC	0.02	0.05	0.00	0.05	0.00	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.38	0.00	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.10	0.00	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.35	0.00	0.00	0.00	0.00
PO	0.59	0.00	0.00	0.10	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.22	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.25	0.55	0.42	0.50	0.05

Figure 1 The initial WAM from the online news articles.

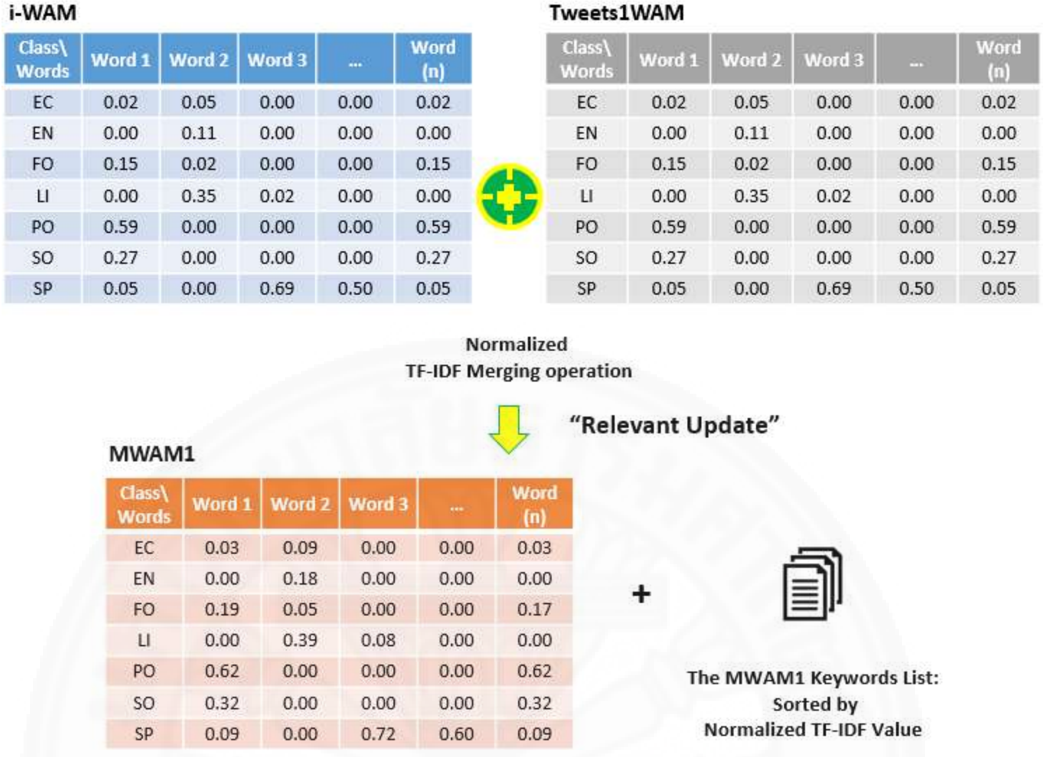
A bag-of-words of each news category, model classes, such as economic, entertainment, foreign, lifestyle, politic, social, sports, etc., generate a stable document vector/ news category vector for its class. There is a group of specific terms, terms with high TF-IDF score, in each category which should be a good representative for its category when we consider the similarity value. For example, terms “soccer” must have highest TF-IDF score in “Sports” class when “Prime Minister” must have the highest score in “Politic” class. The model efficiency depends on numbers of specific words/terms in each class which we call “coverage area of words”.

Then, the model be enhanced by more specific and suitable words, coverage area expansion, from related tweets which searched by keywords from the initial WAM, top 10 keywords with highest TF-IDF score per each class: totally 70 keywords. We call this process as “Relevant update” by using “Normalized TF-IDF Merging Operation with Specific Terms Weighting Technique”. The model be updated to be a Modified WAM (MWAM) as shown in Figure 2 and Figure 3.



Class\ Words	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	...	Word (n)
EC	0.02	0.05	0.00	0.05	0.00	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.38	0.00	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.10	0.00	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.35	0.00	0.00	0.00	0.00
PO	0.59	0.00	0.00	0.10	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.22	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.25	0.55	0.42	0.50	0.05

Figure 2 The tweets WAM from the related tweets.



i-WAM

Class\ Words	Word 1	Word 2	Word 3	...	Word (n)
EC	0.02	0.05	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.00	0.00
PO	0.59	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.50	0.05

Tweets1WAM

Class\ Words	Word 1	Word 2	Word 3	...	Word (n)
EC	0.02	0.05	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.00	0.00
PO	0.59	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.50	0.05

MWAM1

Class\ Words	Word 1	Word 2	Word 3	...	Word (n)
EC	0.03	0.09	0.00	0.00	0.03
EN	0.00	0.18	0.00	0.00	0.00
FO	0.19	0.05	0.00	0.00	0.17
LI	0.00	0.39	0.08	0.00	0.00
PO	0.62	0.00	0.00	0.00	0.62
SO	0.32	0.00	0.00	0.00	0.32
SP	0.09	0.00	0.72	0.60	0.09

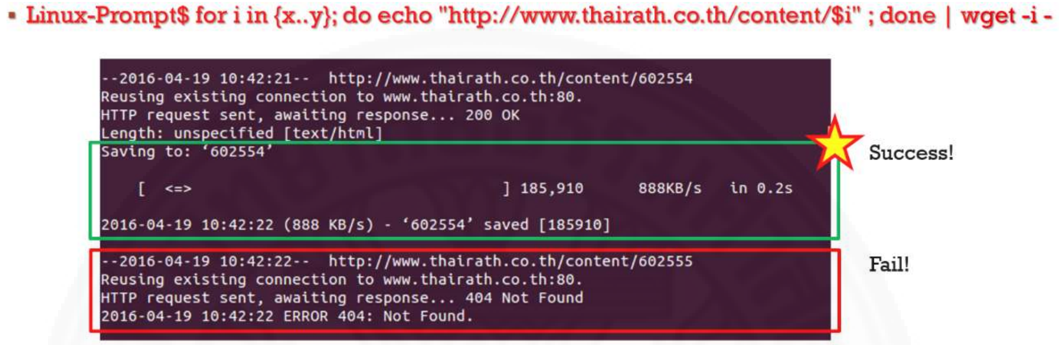
Figure 3 A Relevant Update, the model enhancing process.

The iteration process of the MWAM updating be done until the model can classify the social media text, tweets, efficiently. For evaluation process and total amount of Training Dataset, Test Dataset, I will describe in an Approach Modelsection, after Main Techniques section, as follows.

3.2 Main Techniques

3.2.1 Web Crawler

Retrieving news articles from online news websites, we need a main operation like a web crawler, robot or spider [34]. At present, there are many web crawler techniques for researchers or developers to deploy such as complete crawler application, crawler API/libraries, open-sources, etc. Moreover, all websites are used HTML, XML, and CSS framework as a main structure. We need to verify the target web's structure before choosing the tools and extract our required information. In this experiment, I decided to use a “wget” command in Linux to collect the online news articles in HTML format as shown in Figure 4. Then, I used the XML Path (XPath) Query Technique, HTML Parser Libraries, in Python programming coding to extract the online news articles and news content in pure text format. Finally, the online news articles be filtered into their category properly.



```
▪ Linux-Prompt$ for i in {x..y}; do echo "http://www.thairath.co.th/content/$i" ; done | wget -i -
```

```
--2016-04-19 10:42:21--  http://www.thairath.co.th/content/602554
Reusing existing connection to www.thairath.co.th:80.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: ‘602554’

    [  <=>                                 ] 185,910      888KB/s   in 0.2s

2016-04-19 10:42:22 (888 KB/s) - ‘602554’ saved [185910]

--2016-04-19 10:42:22--  http://www.thairath.co.th/content/602555
Reusing existing connection to www.thairath.co.th:80.
HTTP request sent, awaiting response... 404 Not Found
2016-04-19 10:42:22 ERROR 404: Not Found.
```

Figure 4 A simple Linux command, “wget”, for web crawling.

3.2.2 Word Segmentation

A crucial step for text mining is Word Segmentation. Words boundary should be verified before using them as the input of machine learning. Dealing with Thai language in the online news articles and tweets, we need a Thai Word Segmentation tool because Thai language does not use space between words in sentence. I decided to use Thai Word Segmentation with maximal matching in dictionary technique [35]. I also updated a dictionary file with recent words, important person names, present events, major places and point of interests. Therefore, I got the promising result of word segmentation, high quality of words, before utilizing them as the input of the model.

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

Nowadays, there are many weighting schemes to consider such as TF-IDF weighting, Term Frequency (TF) weighting, Boolean weighting, tfc weighting, ltc weighting, and Entropy weighting [29]. The 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 [36] (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):

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

From this useful technique, I can extract the essential keyword terms from each news category to be the keyword set to search the related tweets for model enhancing. In addition, these terms with TF-IDF score should be the representative of their class/category when we consider the similarity value of any queries.

3.2.4 Normalized Term Frequency-Inverse Document Frequency Merging (Normalized TF-IDF Merging) with Specific Terms Weighting Technique

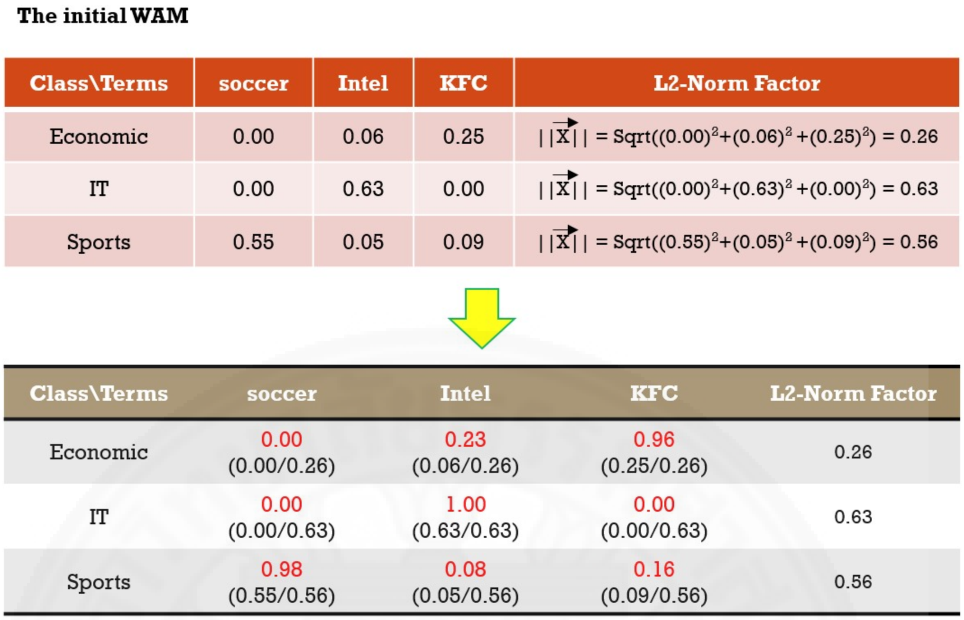
The problem when dealing with different weight corpuses is finding an exact terms weight. The standard solution for this problem is normalization, unit vector consideration as L2-normalization or also known as Euclidean normalization. As I used the TF-IDF weighting technique to specify significant terms weight values, normalization value of TF-IDF of each term should be calculated. For example, term "soccer" from the online news articles has the TF-IDF value at 0.55, in a class "Sports". This value be normalized by using word vector normalization (5) (6), and final result is shown in Figure 5, L2-normalization factor is 0.56 ,the final normalized TF-IDF of term "soccer" is 0.98.

$$||\vec{X}||_2 = \sqrt{{X_1}^2 + {X_2}^2 + {X_3}^2 + \cdots + {X_n}^2} \qquad (5)$$

The $||\vec{X}||_2$ is the Euclidean norm factor, L2-normalization. The X_1, X_2, … , X_n are the TF-IDF value of terms (1) to terms (n) in the corpus.

$$TF - IDF_{term(i)} = \frac{X_i}{||\vec{X}||_2} \qquad (6)$$

The $TF - IDF_{term(i)}$ is the normalized TF-IDF value of the term (i). While X_i is the TF-IDF value of term (i) and $||\vec{X}||_2$ is the L2-normalization factor of the corpus.



The initial WAM

Class\Terms	soccer	Intel	KFC	L2-Norm Factor
Economic	0.00	0.06	0.25	$\|\|\vec{X}\|\| = Sqrt((0.00)^2+(0.06)^2+(0.25)^2) = 0.26$
IT	0.00	0.63	0.00	$\|\|\vec{X}\|\| = Sqrt((0.00)^2+(0.63)^2+(0.00)^2) = 0.63$
Sports	0.55	0.05	0.09	$\|\|\vec{X}\|\| = Sqrt((0.55)^2+(0.05)^2+(0.09)^2) = 0.56$

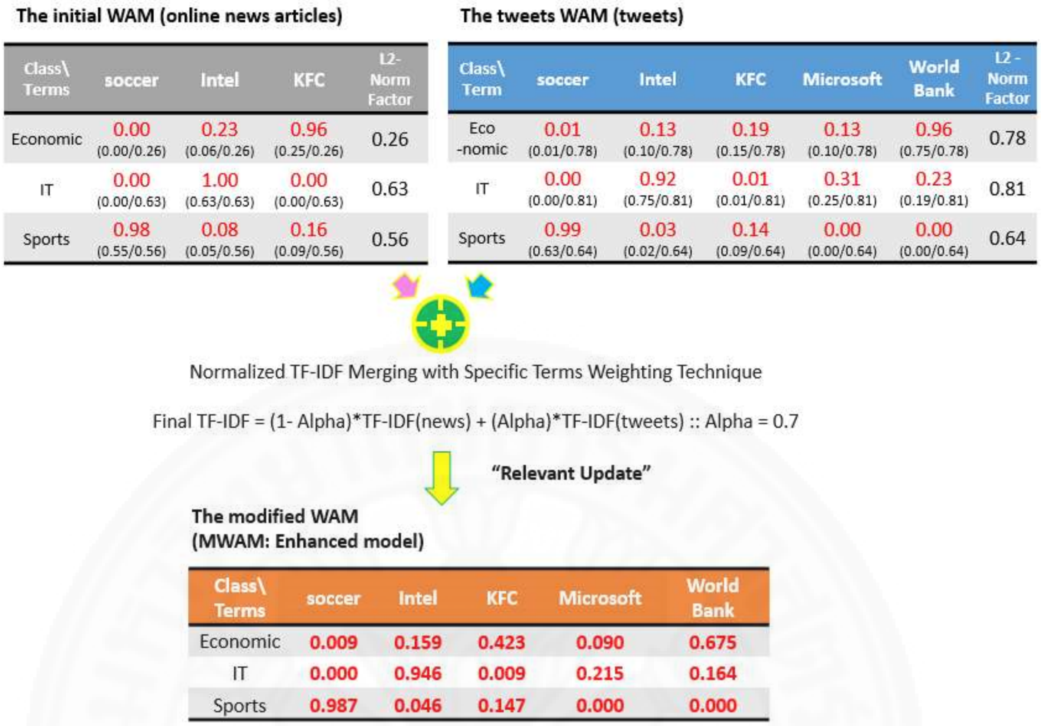
Class\Terms	soccer	Intel	KFC	L2-Norm Factor
Economic	0.00 (0.00/0.26)	0.23 (0.06/0.26)	0.96 (0.25/0.26)	0.26
IT	0.00 (0.00/0.63)	1.00 (0.63/0.63)	0.00 (0.00/0.63)	0.63
Sports	0.98 (0.55/0.56)	0.08 (0.05/0.56)	0.16 (0.09/0.56)	0.56

Figure 5 A normalized TF-IDF calculation example.

After, the tweets WAM is created by related keywords Twitter searching, we can find the exact significant TF-IDF value of term "soccer" by using "Normalized TF-IDF Merging with Specific Terms Weighting Technique". The Alpha weighting parameter is deployed as (7).

$$\begin{aligned} Final\ TF-IDF\ term(i) \\ &= (1-Alpha)*Normalied\ TF-IDF\ term(i{:}\,online\ news) \\ &+ (Alpha)*Normalized\ TF-IDF\ term(i{:}\,tweets) \qquad (7) \end{aligned}$$

Main idea of using Alpha weighting parameter is we need to have more focus on terms from tweets, because we are now focusing on the social media text classifying. Terms from the social media should matched and proper to enhance the model efficiency. As an example, normalized TF-IDF value of term "soccer" in the initial WAM is 0.98, and normalized TF-IDF value of term "soccer" in the tweets WAM is 0.99. I set the Alpha value at 0.7, proper value of Alpha: I will discuss again in Experiment Result and Discussion section. Finally, the final TF-IDF value of term "soccer" when I merged these two WAMs into the modified WAM is 0.987, [(1-0.7)*0.98 + (0.7)*0.99] as shown in Figure 6.



The initial WAM (online news articles)

Class\ Terms	soccer	Intel	KFC	L2-Norm Factor
Economic	0.00 (0.00/0.26)	0.23 (0.06/0.26)	0.96 (0.25/0.26)	0.26
IT	0.00 (0.00/0.63)	1.00 (0.63/0.63)	0.00 (0.00/0.63)	0.63
Sports	0.98 (0.55/0.56)	0.08 (0.05/0.56)	0.16 (0.09/0.56)	0.56

The tweets WAM (tweets)

Class\ Term	soccer	Intel	KFC	Microsoft	World Bank	L2 - Norm Factor
Eco -nomic	0.01 (0.01/0.78)	0.13 (0.10/0.78)	0.19 (0.15/0.78)	0.13 (0.10/0.78)	0.96 (0.75/0.78)	0.78
IT	0.00 (0.00/0.81)	0.92 (0.75/0.81)	0.01 (0.01/0.81)	0.31 (0.25/0.81)	0.23 (0.19/0.81)	0.81
Sports	0.99 (0.63/0.64)	0.03 (0.02/0.64)	0.14 (0.09/0.64)	0.00 (0.00/0.64)	0.00 (0.00/0.64)	0.64

Normalized TF-IDF Merging with Specific Terms Weighting Technique

Final TF-IDF = (1- Alpha)*TF-IDF(news) + (Alpha)*TF-IDF(tweets) :: Alpha = 0.7

"Relevant Update"

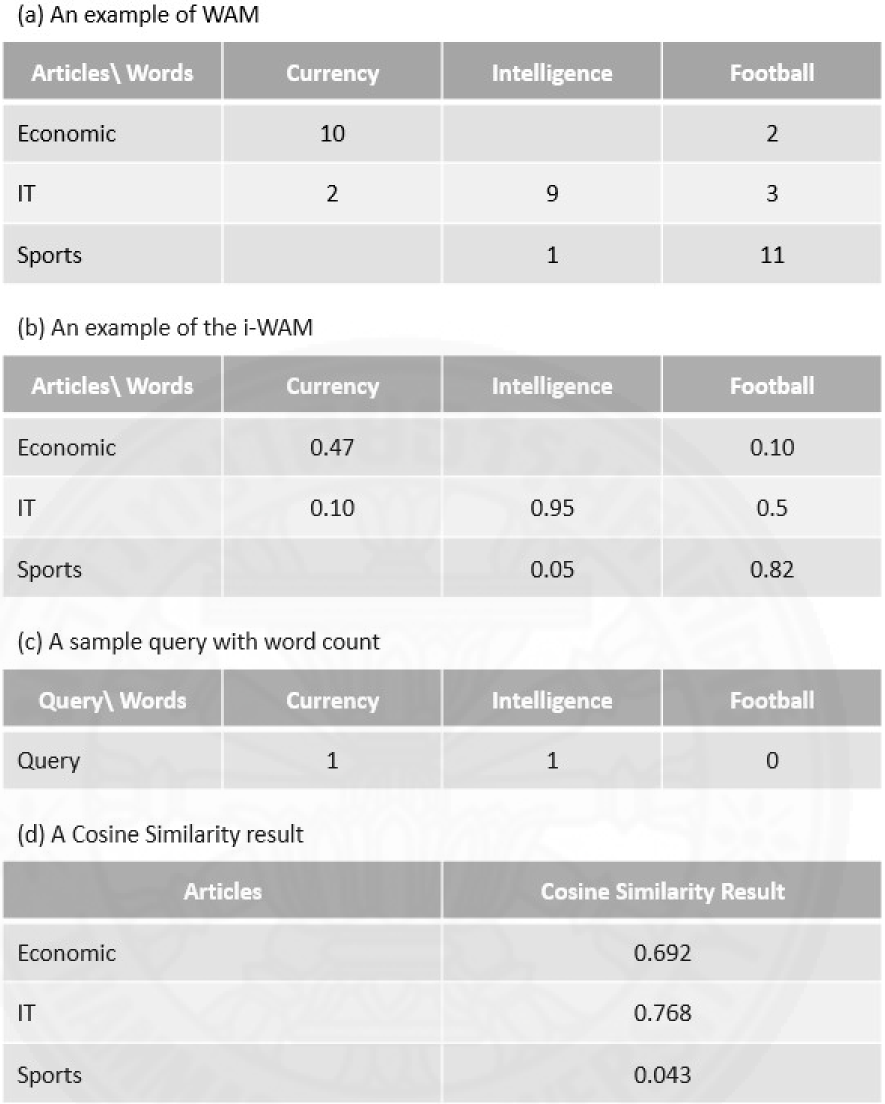
The modified WAM
(MWAM: Enhanced model)

Class\ Terms	soccer	Intel	KFC	Microsoft	World Bank
Economic	0.009	0.159	0.423	0.090	0.675
IT	0.000	0.946	0.009	0.215	0.164
Sports	0.987	0.046	0.147	0.000	0.000

Figure 6 An example of Normalized TF-IDF Merging with Specific Terms Weighting Technique.

3.2.5 Word Article Matrix (WAM)

WAM is a significant data structure [31] in the Generic Engine for Transpose Association (GETA). It creates a large matrix of weighted relation between documents and keywords which rows are indexed by names of documents (articles) and columns are indexed by words(terms), keywords from the documents. Keywords in a document are counted to fill in the table as shown in Figure 7(a).We generate the initial WAM (i-WAM) by using the normalized TF-IDF value of each word. The i-WAM with the normalized TF-IDF values will be shown in Figure 7(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: "You can run the Business Intelligence Wizard to create currency conversion calculations". This query is converted into a model of word vectors shown in Figure 7(c).



(a) An example of WAM

Articles\ Words	Currency	Intelligence	Football
Economic	10		2
IT	2	9	3
Sports		1	11

(b) An example of the i-WAM

Articles\ Words	Currency	Intelligence	Football
Economic	0.47		0.10
IT	0.10	0.95	0.5
Sports		0.05	0.82

(c) A sample query with word count

Query\ Words	Currency	Intelligence	Football
Query	1	1	0

(d) A Cosine Similarity result

Articles	Cosine Similarity Result
Economic	0.692
IT	0.768
Sports	0.043

Figure 7 A Word Article Matrix's example.

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 [37] we can find out the similarity between any two documents (8).

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

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 $d2$.

Lastly, we calculate the cosine similarity values and get a result of an example query as shown in Figure 7(d). As the weight of a word "Intelligence" in Information Technology (IT) category is high, 0.95, the result of operation shows that the query is more likely to be for the document of IT, which produces the highest cosine similarity score at 0.768.

3.3 Approach Model

As the quality of bag-of-words: correctness, variety, coverage area of words, is considered as the main important success factor for the Vector Space model. The initial step with the online news articles utilization with self-learning class concept, the relevant updating to enhance model productivity, and iteration of model updating until model be the most suitable one for the social media text classification are proposed as the complete solution as shown in Figure 8.

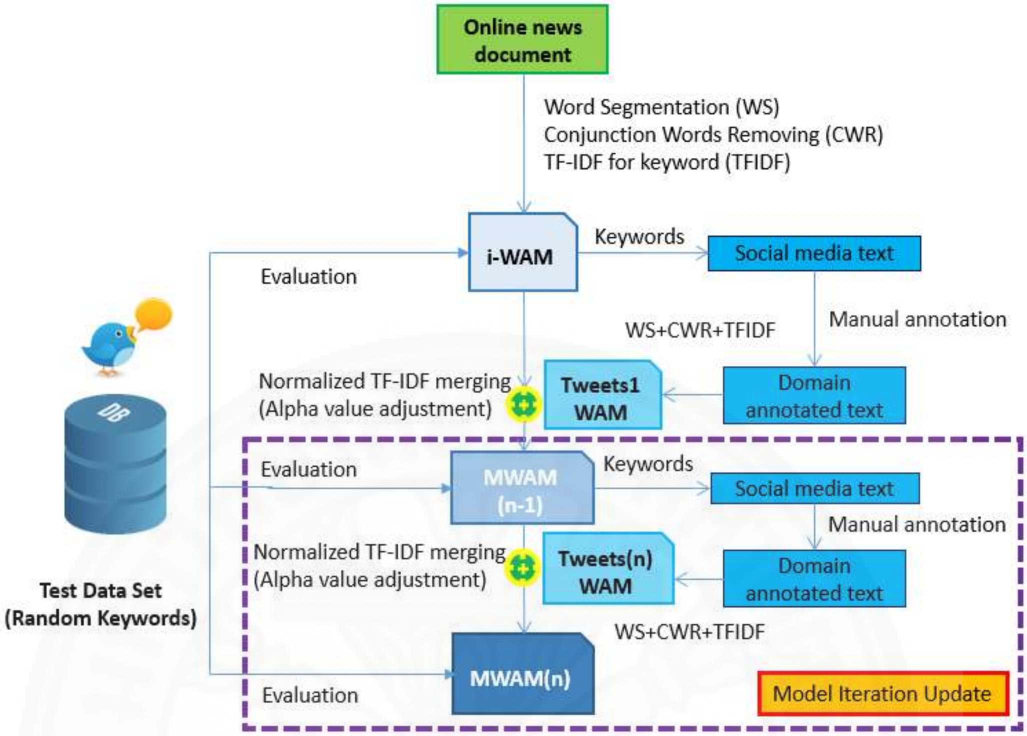


Figure 8 A full approach model.

3.3.1 The Online News Articles Retrieving

As a primary training dataset of the model, the online news articles are crawled and extract into their seven categories automatically: economic, entertainment, foreign, lifestyle, politic, social, and sports. Thairath Online News Website [38] is the main data source for this experiment. Totally, I got 3,548 news articles, economic: 507 articles, entertainment: 501 articles, foreign: 505 articles, lifestyle: 502 articles, politic: 515 articles, social: 514 articles, sports: 504 articles as shown in Table 1.

Table 1 A number of retrieved online news articles.

Online News Articles	Number of Articles	Number of Words	Number of unique Words
Economic	507	55,485 →	22,527
Entertainment	501		
Foreign	505		
IT	502		
Politics	515		
Regionals	514		
Sports	504		

From this training dataset, Word Segmentation tool, and TF-IDF technique are used to extract totally 55,485 words, 22,527 unique words, no duplicated bag-of-words, for building the primary model, the i-WAM. The conjunction word removing algorithm is use to eliminate Thai conjunction words such as “ถ้า”, “ก็”, “เช่น”, “นี้”, “นั้น”, etc. The main characteristic of most Thai conjunction words is less than 4 characters combining. So, I used this value as a criteria to remove the conjunction words in my coding. Finally, the i-WAM is created with a bag-of-words of each news category (model’s class) which be the stable 7 document vectors, EC: economic, EN: entertainment, FO: foreign, LI: lifestyle, PO: politic, SO: social, and SP: sports.

3.3.2 Related Tweets Retrieving

Top 10 terms with highest normalized TF-IDF score in each category, totally 70 keywords, be the keywords for related tweets searching, through Twitter Search API. This tweets be filtered and tagged into each class, 7 classes: EC, EN, FO, LI, PO, SO, and SP, by human annotation. Three staffs are applied to do this job to reduce a bias concerned issue. The Fleiss’ kappa statistics [39] is used to measure the reliability of tweets message, and highest vote score be an algorithm to determine the ambiguous tweets to tag into the right class. For example, tweets: “Thai Prime minister took selfie photos with N’Mae Rachanok. #NongMae#Champion#Badminton” (N’Mae or NongMae is a famous Thai’s female badminton player), the K value of Fleiss’ kappa statistics is calculated and show the value at 0.60, moderate agreement, and 2 annotators tagged is tweets into SP: sports class, while 1 annotator tagged this tweets into PO: politic. So, the final class of this tweets should be “SP: sports”.

In this experiment, related tweets searching is done for 4 times. First, the related tweets searching by the keywords set from the i-WAM, 21,000 tweets: 3,000 tweets per each class. Second, the related tweets searching by the keywords set from the MWAM1, 21,000 tweets: 3,000 tweets per each class. Third, the related tweets searching by the keywords set from the MWAM2, 21,000 tweets: 3,000 tweets per each class. The first three times of tweets search is used to build a Training Dataset to enhance the model efficiency as shown in Table 2.

Table 2 A number of retrieved tweets as Training Dataset.

Related tweets	Number of tweets1	Number of tweets2	Number of tweets3
Economic	3,000	3,000	3,000
Entertainment	3,000	3,000	3,000
Foreign	3,000	3,000	3,000
IT	3,000	3,000	3,000
Politics	3,000	3,000	3,000
Regionals	3,000	3,000	3,000
Sports	3,000	3,000	3,000
Total	21,000	21,000	21,000

Nature of tweets is diverse and noisy. We usually found duplicated tweets as re-tweets, and rude tweets or useless tweets as junk tweets. For example, junk tweets could be like this “@name: just boring” or “@name:!@#xx%^” or “@name: what the hell is going on!”. We need to sanitize or clean related tweets before using them as a good training dataset.

After tweets filtering and tagging process, duplicated tweets and junk tweets are cleared. Then, Word Segmentation, and TF-IDF technique are used to extract keywords and define the significant value of each terms again. Finally, the total number of clean tweets which suitable for model training, and number of words, no duplicated words are shown in Table 3. The number of tweets for each class are nearly the same, around 900 – 1,000 tweets per class. Balancing number of tweets per class is important to train the model into more reliable model.

Table 3 A number of clean tweets, number of words and number of unique words.

Related tweets	Number of clean tweets	Number of all words	Number of unique words
Tweets1	6,822	11,102	8,429
Tweets2	6,531	10,024	7,935
Tweets3	6,718	11,054	8,219

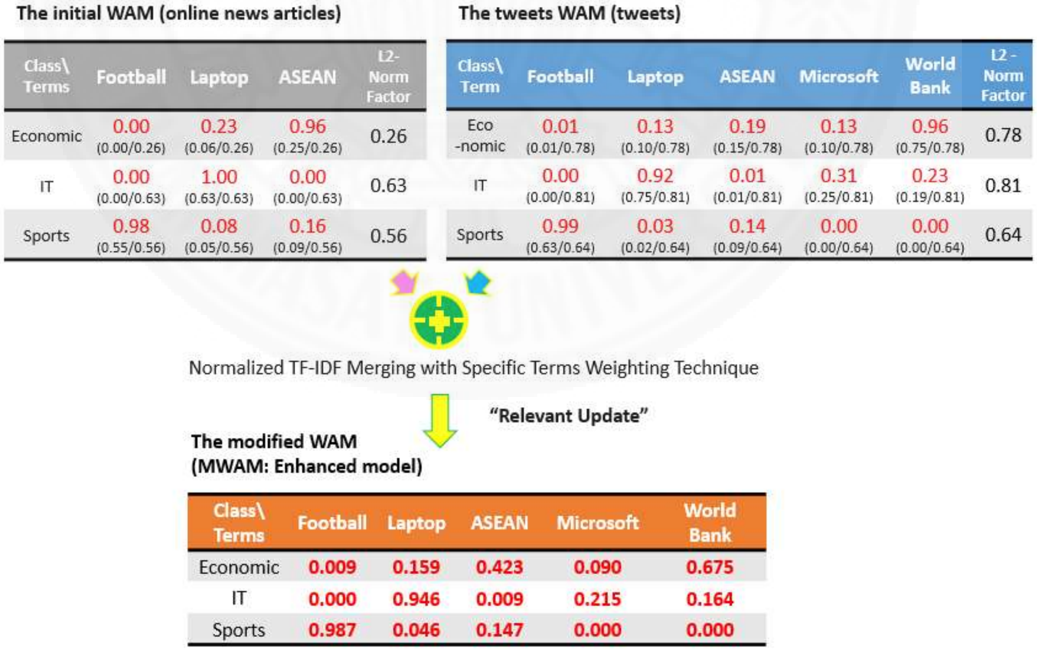
The last one, fourth, the related tweets searching by three annotators, any keywords they want to search to fit into 7 classes: no induction of keywords from the model, 20,000 tweets: as a “Test Dataset” for evaluation as shown in Table 3. The test dataset is separated into two groups. First, filtered test dataset: tweets which filtered all duplicated and junk. Second, no filtered tweets: real tweets contain retweets and junk for real test. I need to verify the model classifying capability compare with the human filtering skill. The result of this idea be shown in Result and Discussion section.

Table 4 A Test Dataset.

Test Dataset tweets	Number of tweets
Test Dataset1 (Filtered Tweets)	6,339
Test Dataset2 (Raw Tweets)	20,000

3.3.3 Model Enhancing, WAM Merging

The initial model, iWAM, be enhanced by merging between the iWAM and Tweets WAM1 to be the modified WAM1 (MWAM1). The normalized TF-IDF merging technique with specific terms weighting technique are used to update the model. Merging process uses all of words in the i-WAM and updates no duplicated words to fill into the MWAM. Then, the final TF-IDF value of each terms are calculated as shown in Figure 9.



The initial WAM (online news articles)

Class\ Terms	Football	Laptop	ASEAN	L2-Norm Factor
Economic	0.00 (0.00/0.26)	0.23 (0.06/0.26)	0.96 (0.25/0.26)	0.26
IT	0.00 (0.00/0.63)	1.00 (0.63/0.63)	0.00 (0.00/0.63)	0.63
Sports	0.98 (0.55/0.56)	0.08 (0.05/0.56)	0.16 (0.09/0.56)	0.56

The tweets WAM (tweets)

Class\ Term	Football	Laptop	ASEAN	Microsoft	World Bank	L2 - Norm Factor
Eco -nomic	0.01 (0.01/0.78)	0.13 (0.10/0.78)	0.19 (0.15/0.78)	0.13 (0.10/0.78)	0.96 (0.75/0.78)	0.78
IT	0.00 (0.00/0.81)	0.92 (0.75/0.81)	0.01 (0.01/0.81)	0.31 (0.25/0.81)	0.23 (0.19/0.81)	0.81
Sports	0.99 (0.63/0.64)	0.03 (0.02/0.64)	0.14 (0.09/0.64)	0.00 (0.00/0.64)	0.00 (0.00/0.64)	0.64

Normalized TF-IDF Merging with Specific Terms Weighting Technique
"Relevant Update"

The modified WAM (MWAM: Enhanced model)

Class\ Terms	Football	Laptop	ASEAN	Microsoft	World Bank
Economic	0.009	0.159	0.423	0.090	0.675
IT	0.000	0.946	0.009	0.215	0.164
Sports	0.987	0.046	0.147	0.000	0.000

Figure 9 A MWAM1 creation.

The MWAM1 is created by the i-WAM merging with the TweetsWAM1. The specific terms weighting formula will be (9). The Alpha value is tested from 0.5 to 0.9, the incremental step is 0.1, to find the best result.

$$\begin{aligned} &Final\ TF - IDF\ term(i) \\ &\qquad = (1 - Alpha) * Normalied\ TF - IDF\ term(i\colon online\ news) \\ &\qquad\quad + (Alpha) * Normalized\ TF - IDF\ term(i\colon tweets1) \end{aligned} \tag{9}$$

Then the iteration of model updating is performed until the most productive model is created, accuracy, precision, recall, and F-measure score are stable nearly 100%. Therefore, the MWAM(n) is created by the i-WAM merging with all tweets WAM, the TweetsWAM1+... +TweetsWAM(n). The specific terms weighting formula will be (10). The Alpha value is tested from 0.5 to 0.9, the incremental step is 0.1, to find the best result.

$$\begin{aligned} &Final\ TF - IDF\ term(i) \\ &\qquad = (1 - Alpha) * Normalied\ TF - IDF\ term(i\colon online\ news) \\ &\qquad\quad + (Alpha) * Normalized\ TF - IDF\ term(i\colon all\ tweets) \end{aligned} \tag{10}$$

3.3.4 Evaluation Process

A confusion matrix is used to evaluate the performance of my classification model, WAM. There are two test dataset, filtered tweets with 6,339 tweets, and raw tweets with 20,000 tweets to evaluate all model, 8 times evaluation. The evaluation concept is shown as Figure 10.

The test dataset1 (filtered tweets) evaluates the i-WAM.

The test dataset1 (filtered tweets) evaluates the MWAM1.

The test dataset1 (filtered tweets) evaluates the MWAM2.

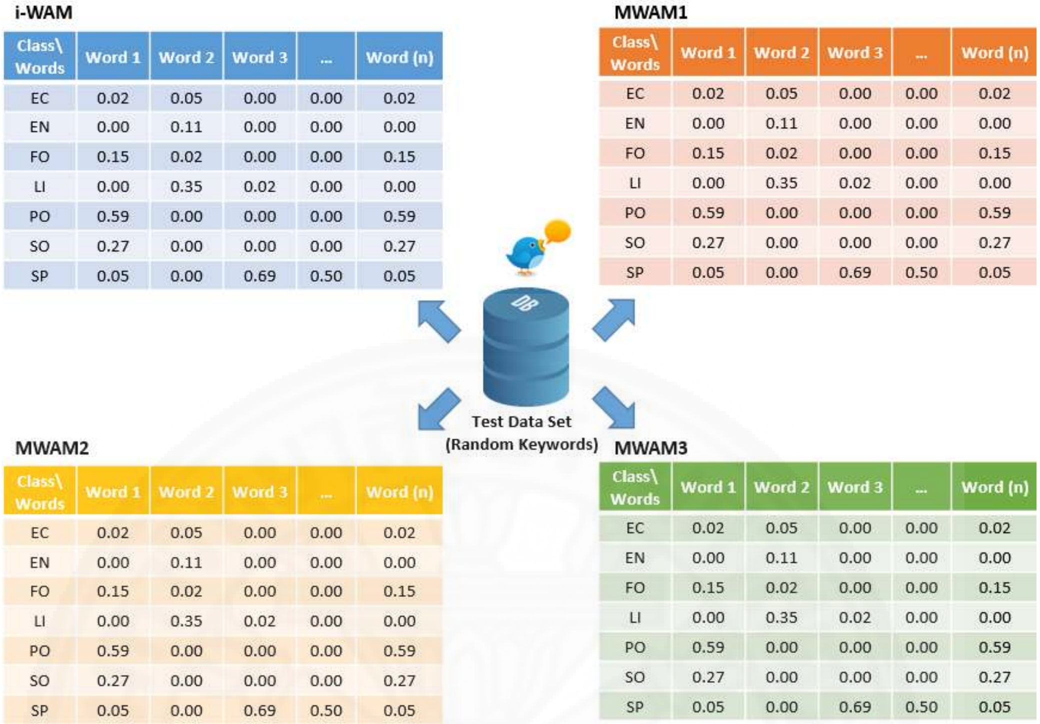
The test dataset1 (filtered tweets) evaluates the MWAM3.

The test dataset2 (raw tweets) evaluates the i-WAM.

The test dataset2 (raw tweets) evaluates the MWAM1.

The test dataset2 (raw tweets) evaluates the MWAM2.

The test dataset2 (raw tweets) evaluates the MWAM3.



i-WAM

Class\ Words	Word 1	Word 2	Word 3	...	Word (n)
EC	0.02	0.05	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.00	0.00
PO	0.59	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.50	0.05

MWAM1

Class\ Words	Word 1	Word 2	Word 3	...	Word (n)
EC	0.02	0.05	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.00	0.00
PO	0.59	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.50	0.05

MWAM2

Class\ Words	Word 1	Word 2	Word 3	...	Word (n)
EC	0.02	0.05	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.00	0.00
PO	0.59	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.50	0.05

MWAM3

Class\ Words	Word 1	Word 2	Word 3	...	Word (n)
EC	0.02	0.05	0.00	0.00	0.02
EN	0.00	0.11	0.00	0.00	0.00
FO	0.15	0.02	0.00	0.00	0.15
LI	0.00	0.35	0.02	0.00	0.00
PO	0.59	0.00	0.00	0.00	0.59
SO	0.27	0.00	0.00	0.00	0.27
SP	0.05	0.00	0.69	0.50	0.05

Figure 10 An evaluation concept.

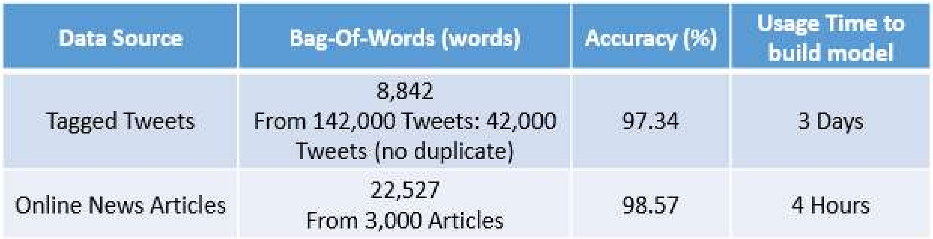
Chapter 4
Experiment Result and Discussion

"The initial WAM, primary model which built from the online news articles, with 22,527 unique terms generate a good result. Nevertheless, the modified WAM, enhanced model by relevant update of specific terms from the social media text itself, generate a promising result, especially when the iteration of model updating is done by 3 times to the MWAM3 with 39,952 unique terms."

4.1 Preliminary Testing

As the social media text classification is a challenge task, the remarkable question is raised when the experiment is conducted; "Is the beginning model created from tagged tweets be better than the beginning model created from the online news articles? Should words/terms from the social media text itself be more matched with the social media text classification job?" I do the preliminary testing to proof my main idea that the primary model created from the online news articles is better than the model created from tagged tweets.

After testing, 142,000 raw tweets are filtered duplicated and junk tweets to be 42,000 clean tweets, and tagged into 7 classes by manual labor. This clean tweets contain 8,842 terms as the bag-of-words of the model, and generated the primary model with 97.34% accuracy. All process of tagged tweets model building consumed a large amount of time, 24 hours: 3 days with 8 hours per day. Most of time consuming is filtering and tagging task. While, the primary model created from the online news articles, 3,000 articles with 22,527 unique terms, generated the result at 98.57% accuracy. Moreover, this process consumed only 4 hours. All details is shown in Figure 11.



Data Source	Bag-Of-Words (words)	Accuracy (%)	Usage Time to build model
Tagged Tweets	8,842 From 142,000 Tweets: 42,000 Tweets (no duplicate)	97.34	3 Days
Online News Articles	22,527 From 3,000 Articles	98.57	4 Hours

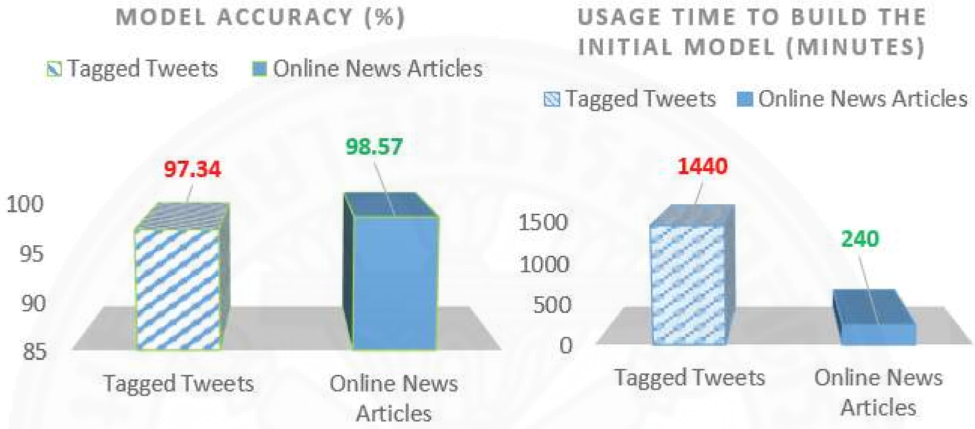
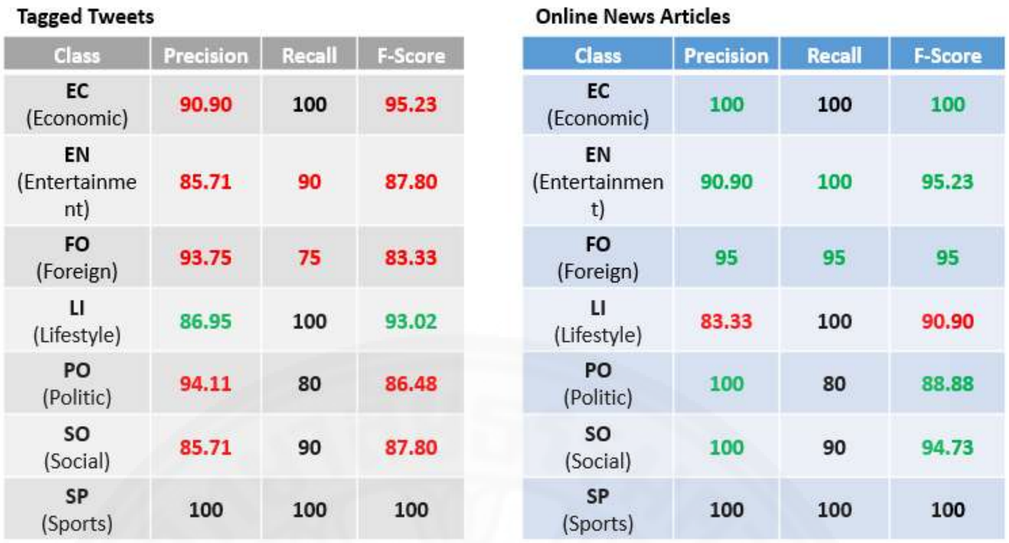


Figure 11 A performance comparison between a Primary Tagged Tweets WAM and a Primary Online News Articles WAM.

In addition, most evaluation score of the primary model created by the online news articles beat the primary model created by the tagged tweets as shown in Figure 12. These testing information is the reasonable support reason that the online news articles is a proper source to create the primary model of the social media text classification.



Tagged Tweets

Class	Precision	Recall	F-Score
EC (Economic)	90.90	100	95.23
EN (Entertainment)	85.71	90	87.80
FO (Foreign)	93.75	75	83.33
LI (Lifestyle)	86.95	100	93.02
PO (Politic)	94.11	80	86.48
SO (Social)	85.71	90	87.80
SP (Sports)	100	100	100

Online News Articles

Class	Precision	Recall	F-Score
EC (Economic)	100	100	100
EN (Entertainment)	90.90	100	95.23
FO (Foreign)	95	95	95
LI (Lifestyle)	83.33	100	90.90
PO (Politic)	100	80	88.88
SO (Social)	100	90	94.73
SP (Sports)	100	100	100

Figure 12 An evaluation score comparison between a Primary Tagged Tweets WAM and a Primary Online News Articles WAM.

4.2 The Most Suitable Specific Terms Weighting Parameter (Alpha) Value for the Model Enhancement

The specific terms weighting technique is used for the social media text terms concentration. The Alpha value is implemented as weighting parameter (10) which need to define the proper value for the model enhancement process. After testing, the most suitable Alpha value is 0.7 which generated the highest accuracy score as shown in Figure 13.

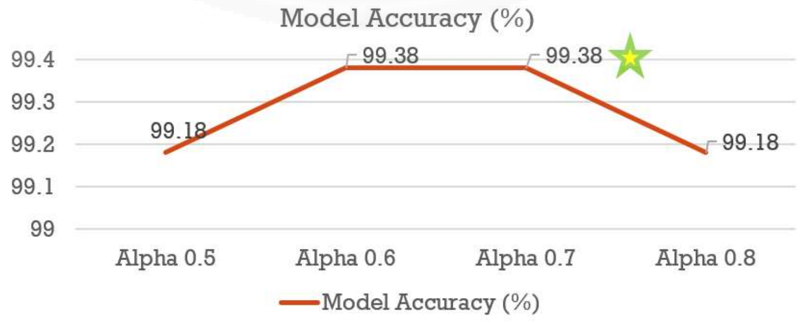


Figure 13 An accuracy model of vary Alpha score.

4.3 The Bag-Of-Words Quality of All Models

The initial WAM (i-WAM) is created from the online news articles, 3,548 articles with totally 22,527 unique terms in bag-of-words as shown in Table 5 and Figure 14. This is a good primary bag-of-words, most terms are written correctly and model contains more variety of words for each class, high coverage area of word as shown in Figure 15.

Table 5 All details of model's training dataset.

Training dataset\WAM	i-WAM	MWAM1	MWAM2	MWAM3
Number of Online News Articles	3,548	-	-	-
Number of Tweets	-	21,000	21,000	21,000
Number of clean tweets	-	6,822	6,531	6,718
Number of all words	55,485	11,102	10,024	11,054
Number of unique words	22,527	8,429	7,935	8,219
Number of relevant update words	-	4,476	9,516	3,433
Total terms in Bag-Of-Words	22,527	27,003	36,519	39,952

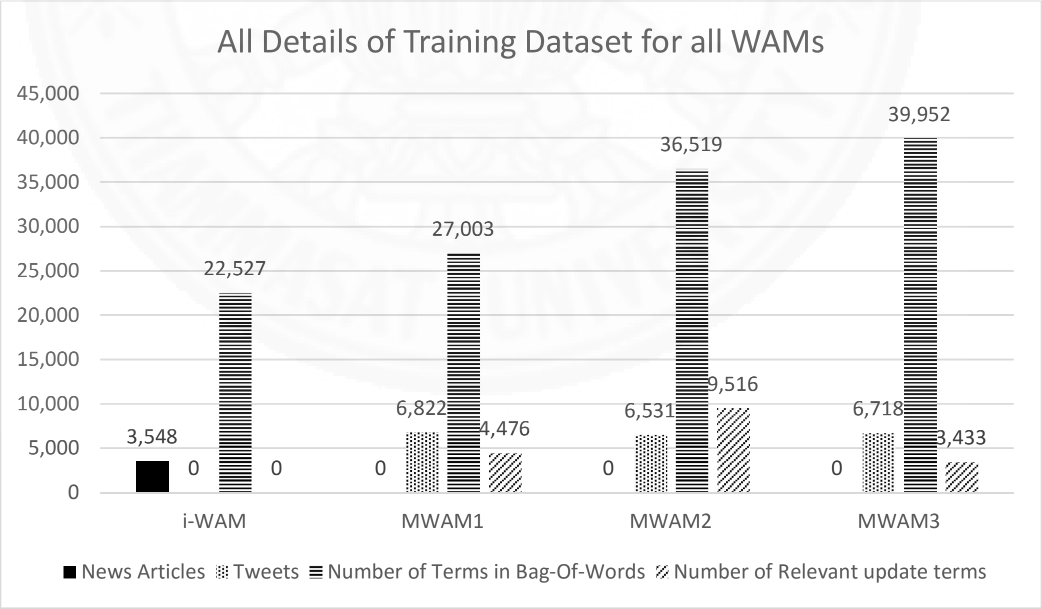
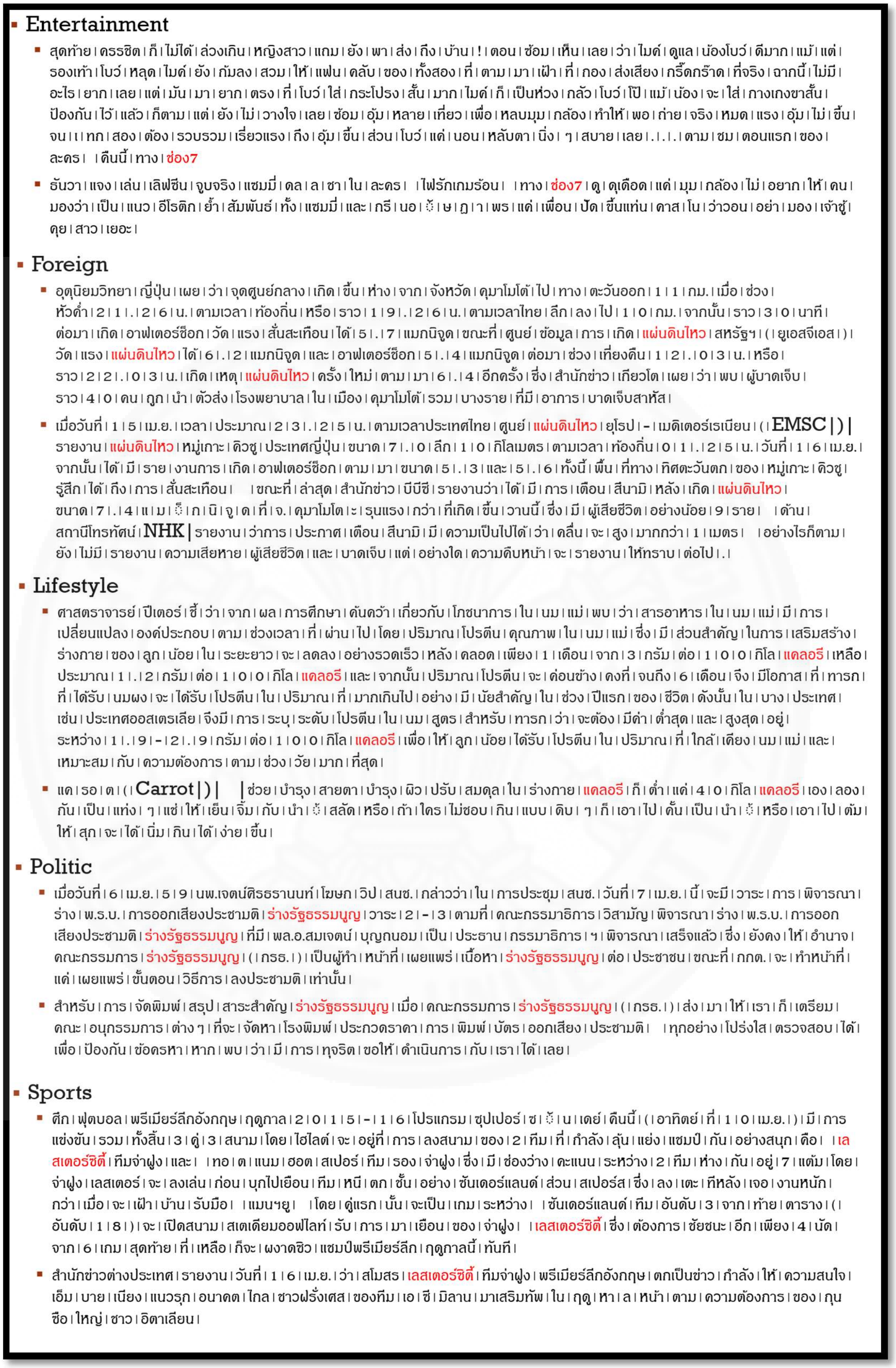


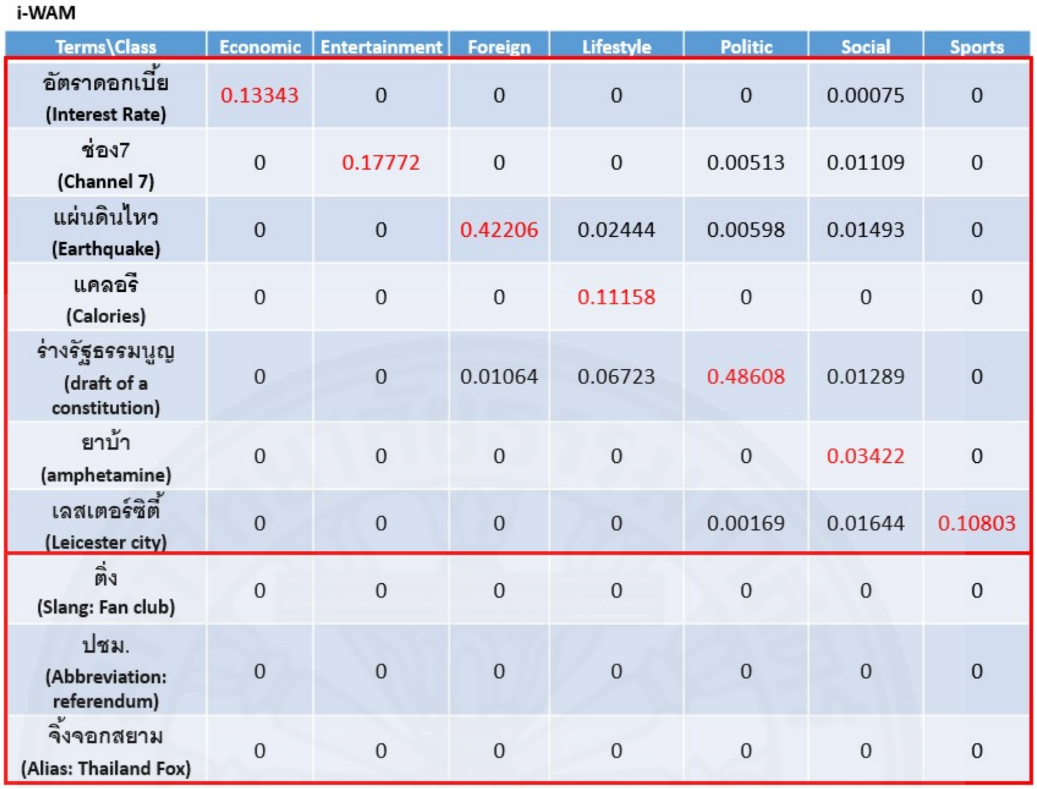
Figure 14 All details of training dataset for all WAMs.

The MWAM1 is created by the relevant update technique with 4,476 terms. This model contains 27,003 terms in bag-of-words. Form an example of the online news articles as shown in Figure 15, all significant terms are extracted to build the i-WAM as shown in Figure 16, and Top 10 keywords list, highest normalized TF-IDF score, of each class as shown in Figure 17. These keywords, totally 70 keywords, are used to search the related tweets to be a training dataset to enhance (relative update) model to be MWAM1, MWAM2, and MWAM3 respectively. From the retrieved related tweets, the relative words/terms are found, the terms that never appear in the online news, abbreviation, slang, alias as shown in Figure 18, Figure 19, and Figure 20.



- Entertainment
  - สุดท้าย|ครรชิต|ก็|ไม่ได้|ล่วงเกิน|หญิงสาว|แถม|ยัง|พา|ส่ง|ถึง|บ้าน|!|ตอน|ซ้อม|เห็น|เลย|ว่า|ไมค์|ดูแล|น้องโบว์|ดีมาก|แม้|แต่|รองเท้า|โบว์|หลุด|ไมค์|ยัง|ก้มลง|สวม|ให้|แฟน|คลับ|ของ|ทั้งสอง|ที่|ตาม|มา|เฝ้า|ที่|กอง|ส่งเสียง|กรี๊ดกร๊าด|ที่จริง|ฉากนี้|ไม่มี|อะไร|ยาก|เลย|แต่|มัน|มา|ยาก|ตรง|ที่|โบว์|ใส่|กระโปรง|สั้น|มาก|ไมค์|ก็|เป็นห่วง|กลัว|โบว์|โป๊|แม้|น้อง|จะ|ใส่|กางเกงขาสั้น|ป้องกัน|ไว้|แล้ว|ก็ตาม|แต่|ยัง|ไม่|วางใจ|เลย|ซ้อม|อุ้ม|หลาย|เที่ยว|เพื่อ|หลบมุม|กล้อง|ทำให้|พอ|ถ่าย|จริง|หมด|แรง|อุ้ม|ไม่|ขึ้น|จน|เ|ทก|สอง|ต้อง|รวบรวม|เรี่ยวแรง|ถึง|อุ้ม|ขึ้น|ส่วน|โบว์|แค่|นอน|หลับตา|นิ่ง| ๆ|สบาย|เลย|.|.|.|ตาม|ชม|ตอนแรก|ของ|ละคร| |คืนนี้|ทาง|ช่อง7
  - ธันวา|แจง|เล่น|เลิฟซีน|จูบจริง|แซมมี่|คล|ล|ชา|ใน|ละคร| |ไฟรักเกมร้อน| |ทาง|ช่อง7|ดู|ดุเดือด|แค่|มุม|กล้อง|ไม่|อยาก|ให้|คน|มองว่า|เป็น|แนว|อีโรติก|ย้ำ|สัมพันธ์|ทั้ง|แซมมี่|และ|กรี|นอ|◌|ษ|ฎ|า|พร|แค่|เพื่อน|ปัด|ขึ้นแท่น|คาส|โน|ว่าวอน|อย่า|มอง|เจ้าชู้|คุย|สาว|เยอะ|
- Foreign
  - อุตุนิยมวิทยา|ญี่ปุ่น|เผย|ว่า|จุดศูนย์กลาง|เกิด|ขึ้น|ห่าง|จาก|จังหวัด|คุมาโมโต้|ไป|ทาง|ตะวันออก|1|1|กม.|เมื่อ|ช่วง|หัวค่ำ|2|1|.|2|6|น.|ตามเวลา|ท้องถิ่น|หรือ|ราว|1|9|.|2|6|น.|ตามเวลาไทย|ลึก|ลง|ไป|1|0|กม.|จากนั้น|ราว|3|0|นาที|ต่อมา|เกิด|อาฟเตอร์ช็อก|วัด|แรง|สั่นสะเทือน|ได้|5|.|7|แมกนิจูด|ขณะที่|ศูนย์|ข้อมูล|การ|เกิด|แผ่นดินไหว|สหรัฐฯ|(|ยูเอสจีเอส|)|วัด|แรง|แผ่นดินไหว|ได้|6|.|2|แมกนิจูด|และ|อาฟเตอร์ช็อก|5|.|4|แมกนิจูด|ต่อมา|ช่วง|เที่ยงคืน|1|2|.|0|3|น.|หรือ|ราว|2|2|.|0|3|น.|เกิด|เหตุ|แผ่นดินไหว|ครั้ง|ใหม่|ตาม|มา|6|.|4|อีกครั้ง|ซึ่ง|สำนักข่าว|เกียวโด|เผย|ว่า|พบ|ผู้บาดเจ็บ|ราว|4|0|คน|ถูก|นำ|ตัวส่ง|โรงพยาบาล|ใน|เมือง|คุมาโมโต้|รวม|บางราย|ที่มี|อาการ|บาดเจ็บสาหัส|
  - เมื่อวันที่|1|5|เม.ย.|เวลา|ประมาณ|2|3|.|2|5|น.|ตามเวลาประเทศไทย|ศูนย์|แผ่นดินไหว|ยุโรป|-|เมดิเตอร์เรเนียน|(|EMSC|)|รายงาน|แผ่นดินไหว|หมู่เกาะ|คิวชู|ประเทศญี่ปุ่น|ขนาด|7|.|0|ลึก|1|0|กิโลเมตร|ตามเวลา|ท้องถิ่น|0|1|.|2|5|น.|วันที่|1|6|เม.ย.|จากนั้น|ได้|มี|ราย|งานการ|เกิด|อาฟเตอร์ช็อก|ตาม|มา|ขนาด|5|.|3|และ|5|.|6|ทั้งนี้|พื้น|ที่ทาง|ทิศตะวันตก|ของ|หมู่เกาะ|คิวชู|รู้สึก|ได้|ถึง|การ|สั่นสะเทือน| |ขณะที่|ล่าสุด|สำนักข่าว|บีบีซี|รายงานว่า|ได้|มี|การ|เตือน|สึนามิ|หลัง|เกิด|แผ่นดินไหว|ขนาด|7|.|4|แ|ม|◌|ก|นิ|จู|ด|ที่|จ.|คุมาโมโต|ะ|รุนแรง|กว่า|ที่เกิด|ขึ้น|วานนี้|ซึ่ง|มี|ผู้เสียชีวิต|อย่างน้อย|9|ราย| |ด้าน|สถานีโทรทัศน์|NHK|รายงาน|ว่าการ|ประกาศ|เตือน|สึนามิ|มี|ความเป็นไปได้|ว่า|คลื่น|จะ|สูง|มากกว่า|1|เมตร| |อย่างไรก็ตาม|ยัง|ไม่มี|รายงาน|ความเสียหาย|ผู้เสียชีวิต|และ|บาดเจ็บ|แต่|อย่างใด|ความคืบหน้า|จะ|รายงาน|ให้ทราบ|ต่อไป|.|
- Lifestyle
  - ศาสตราจารย์|ปีเตอร์|ชี้|ว่า|จาก|ผล|การศึกษา|ค้นคว้า|เกี่ยวกับ|โภชนาการ|ใน|นม|แม่|พบ|ว่า|สารอาหาร|ใน|นม|แม่|มี|การ|เปลี่ยนแปลง|องค์ประกอบ|ตาม|ช่วงเวลา|ที่|ผ่าน|ไป|โดย|ปริมาณ|โปรตีน|คุณภาพ|ใน|นม|แม่|ซึ่ง|มี|ส่วนสำคัญ|ในการ|เสริมสร้าง|ร่างกาย|ของ|ลูก|น้อย|ใน|ระยะยาว|จะ|ลดลง|อย่างรวดเร็ว|หลัง|คลอด|เพียง|1|เดือน|จาก|3|กรัม|ต่อ|1|0|0|กิโล|แคลอรี|เหลือ|ประมาณ|1|.|2|กรัม|ต่อ|1|0|0|กิโล|แคลอรี|และ|จากนั้น|ปริมาณ|โปรตีน|จะ|ค่อนข้าง|คงที่|จนถึง|6|เดือน|จึง|มีโอกาส|ที่|ทารก|ที่|ได้รับ|นมผง|จะ|ได้รับ|โปรตีน|ใน|ปริมาณ|ที่|มากเกินไป|อย่าง|มี|นัยสำคัญ|ใน|ช่วง|ปีแรก|ของ|ชีวิต|ดังนั้น|ใน|บาง|ประเทศ|เช่น|ประเทศออสเตรเลีย|จึงมี|การ|ระบุ|ระดับ|โปรตีน|ใน|นม|สูตร|สำหรับ|ทารก|ว่า|จะต้อง|มีค่า|ต่ำสุด|และ|สูงสุด|อยู่|ระหว่าง|1|.|9|-|2|.|9|กรัม|ต่อ|1|0|0|กิโล|แคลอรี|เพื่อ|ให้|ลูก|น้อย|ได้รับ|โปรตีน|ใน|ปริมาณ|ที่|ใกล้|เคียง|นม|แม่|และ|เหมาะสม|กับ|ความต้องการ|ตาม|ช่วง|วัย|มาก|ที่สุด|
  - แค|รอ|ต|(|Carrot|)| |ช่วย|บำรุง|สายตา|บำรุง|ผิว|ปรับ|สมดุล|ใน|ร่างกาย|แคลอรี|ก็|ต่ำ|แค่|4|0|กิโล|แคลอรี|เอง|ลอง|กัน|เป็น|แท่ง| ๆ|แช่|ให้|เย็น|จิ้ม|กับ|นำ|◌|สลัด|หรือ|ถ้า|ใคร|ไม่ชอบ|กิน|แบบ|ดิบ| ๆ|ก็|เอา|ไป|คั้น|เป็น|นำ|◌|หรือ|เอา|ไป|ต้ม|ให้|สุก|จะ|ได้|นิ่ม|กิน|ได้|ง่าย|ขึ้น|
- Politic
  - เมื่อวันที่|6|เม.ย.|5|9|นพ.เจตน์ศิรธรานนท์|โฆษก|วิป|สนช.|กล่าวว่า|ใน|การประชุม|สนช.|วันที่|7|เม.ย.|นี้|จะมี|วาระ|การ|พิจารณา|ร่าง|พ.ร.บ.|การออกเสียงประชามติ|ร่างรัฐธรรมนูญ|วาระ|2|-|3|ตามที่|คณะกรรมาธิการ|วิสามัญ|พิจารณา|ร่าง|พ.ร.บ.|การออกเสียงประชามติ|ร่างรัฐธรรมนูญ|ที่มี|พล.อ.สมเจตน์|บุญถนอม|เป็น|ประธาน|กรรมาธิการ|ฯ|พิจารณา|เสร็จแล้ว|ซึ่ง|ยังคง|ให้|อำนาจ|คณะกรรมการ|ร่างรัฐธรรมนูญ|(|กรธ.|)|เป็นผู้ทำ|หน้าที่|เผยแพร่|เนื้อหา|ร่างรัฐธรรมนูญ|ต่อ|ประชาชน|ขณะที่|กกต.|จะ|ทำหน้าที่|แค่|เผยแพร่|ขั้นตอน|วิธีการ|ลงประชามติ|เท่านั้น|
  - สำหรับ|การ|จัดพิมพ์|สรุป|สาระสำคัญ|ร่างรัฐธรรมนูญ|เมื่อ|คณะกรรมการ|ร่างรัฐธรรมนูญ|(|กรธ.|)|ส่ง|มา|ให้|เรา|ก็|เตรียม|คณะ|อนุกรรมการ|ต่าง ๆ|ที่จะ|จัดหา|โรงพิมพ์|ประกวดราคา|การ|พิมพ์|บัตร|ออกเสียง|ประชามติ| |ทุกอย่าง|โปร่งใส|ตรวจสอบ|ได้|เพื่อ|ป้องกัน|ข้อครหา|หาก|พบ|ว่า|มี|การ|ทุจริต|ขอให้|ดำเนินการ|กับ|เรา|ได้|เลย|
- Sports
  - ศึก|ฟุตบอล|พรีเมียร์ลีกอังกฤษ|ฤดูกาล|2|0|1|5|-|1|6|โปรแกรม|ซุปเปอร์|ซ|◌|น|เดย์|คืนนี้|(|อาทิตย์|ที่|1|0|เม.ย.|)|มี|การ|แข่งขัน|รวม|ทั้งสิ้น|3|คู่|3|สนาม|โดย|ไฮไลต์|จะ|อยู่ที่|การ|ลงสนาม|ของ|2|ทีม|ที่|กำลัง|ลุ้น|แย่ง|แชมป์|กัน|อย่างสนุก|คือ| |เลสเตอร์ซิตี้|ทีมจ่าฝูง|และ| |ทอ|ต|แนม|ฮอต|สเปอร์|ทีม|รอง|จ่าฝูง|ซึ่ง|มี|ช่องว่าง|คะแนน|ระหว่าง|2|ทีม|ห่าง|กัน|อยู่|7|แต้ม|โดย|จ่าฝูง|เลสเตอร์|จะ|ลงเล่น|ก่อน|บุกไปเยือน|ทีม|หนี|ตก|ชั้น|อย่าง|ซันเดอร์แลนด์|ส่วน|สเปอร์ส|ซึ่ง|ลง|เตะ|ทีหลัง|เจอ|งานหนัก|กว่า|เมื่อ|จะ|เฝ้า|บ้าน|รับมือ| |แมนฯยู| |โดย|คู่แรก|นั้น|จะเป็น|เกม|ระหว่าง| |ซันเดอร์แลนด์|ทีม|อันดับ|3|จาก|ท้าย|ตาราง|(|อันดับ|1|8|)|จะ|เปิดสนาม|สเตเดียมออฟไลท์|รับ|การ|มา|เยือน|ของ|จ่าฝูง| |เลสเตอร์ซิตี้|ซึ่ง|ต้องการ|ชัยชนะ|อีก|เพียง|4|นัด|จาก|6|เกม|สุดท้าย|ที่|เหลือ|ก็จะ|ผงาดซิว|แชมป์พรีเมียร์ลีก|ฤดูกาลนี้|ทันที|
  - สำนักข่าวต่างประเทศ|รายงาน|วันที่|1|6|เม.ย.|ว่า|สโมสร|เลสเตอร์ซิตี้|ทีมจ่าฝูง|พรีเมียร์ลีกอังกฤษ|ตกเป็นข่าว|กำลัง|ให้|ความสนใจ|เอ็ม|บาย|เนียง|แนวรุก|อนาคต|ไกล|ชาวฝรั่งเศส|ของทีม|เอ|ซี|มิลาน|มาเสริมทัพ|ใน|ฤดู|หา|ล|หน้า|ตาม|ความต้องการ|ของ|กุนซือ|ใหญ่|ชาว|อิตาเลียน|

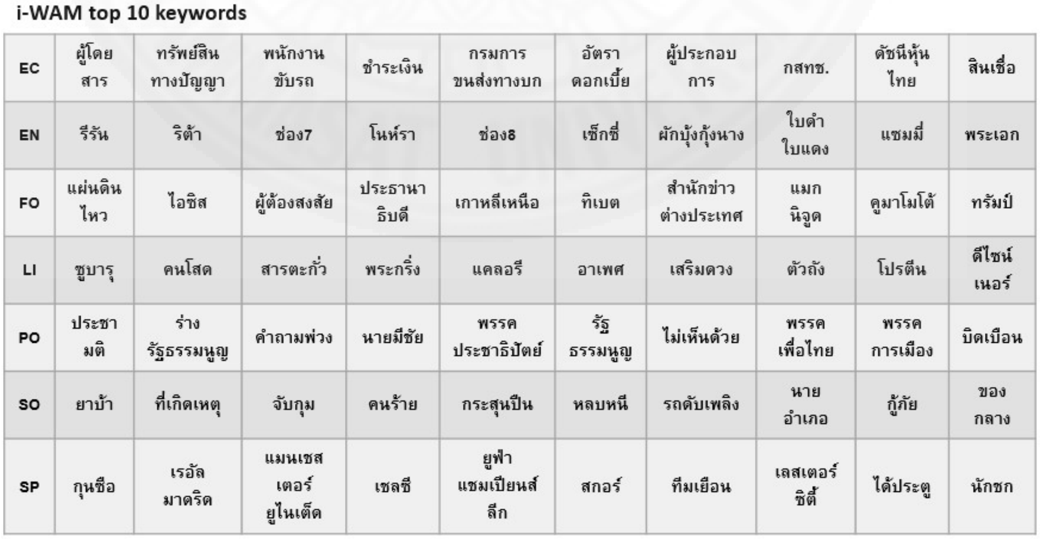
Figure 15 An example of the online news articles.



i-WAM

Terms\Class	Economic	Entertainment	Foreign	Lifestyle	Politic	Social	Sports
อัตราดอกเบี้ย (Interest Rate)	0.13343	0	0	0	0	0.00075	0
ช่อง7 (Channel 7)	0	0.17772	0	0	0.00513	0.01109	0
แผ่นดินไหว (Earthquake)	0	0	0.42206	0.02444	0.00598	0.01493	0
แคลอรี (Calories)	0	0	0	0.11158	0	0	0
ร่างรัฐธรรมนูญ (draft of a constitution)	0	0	0.01064	0.06723	0.48608	0.01289	0
ยาบ้า (amphetamine)	0	0	0	0	0	0.03422	0
เลสเตอร์ซิตี้ (Leicester city)	0	0	0	0	0.00169	0.01644	0.10803
ติ่ง (Slang: Fan club)	0	0	0	0	0	0	0
ปชม. (Abbreviation: referendum)	0	0	0	0	0	0	0
จิ้งจอกสยาม (Alias: Thailand Fox)	0	0	0	0	0	0	0

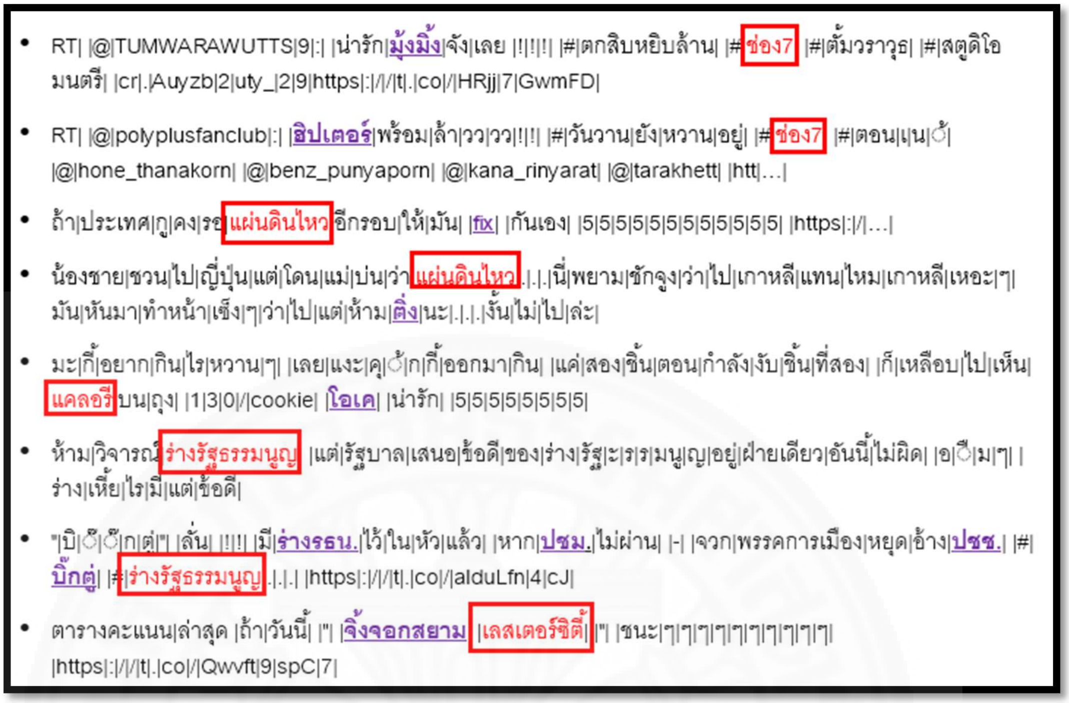
Figure 16 An example of the i-WAM.



i-WAM top 10 keywords

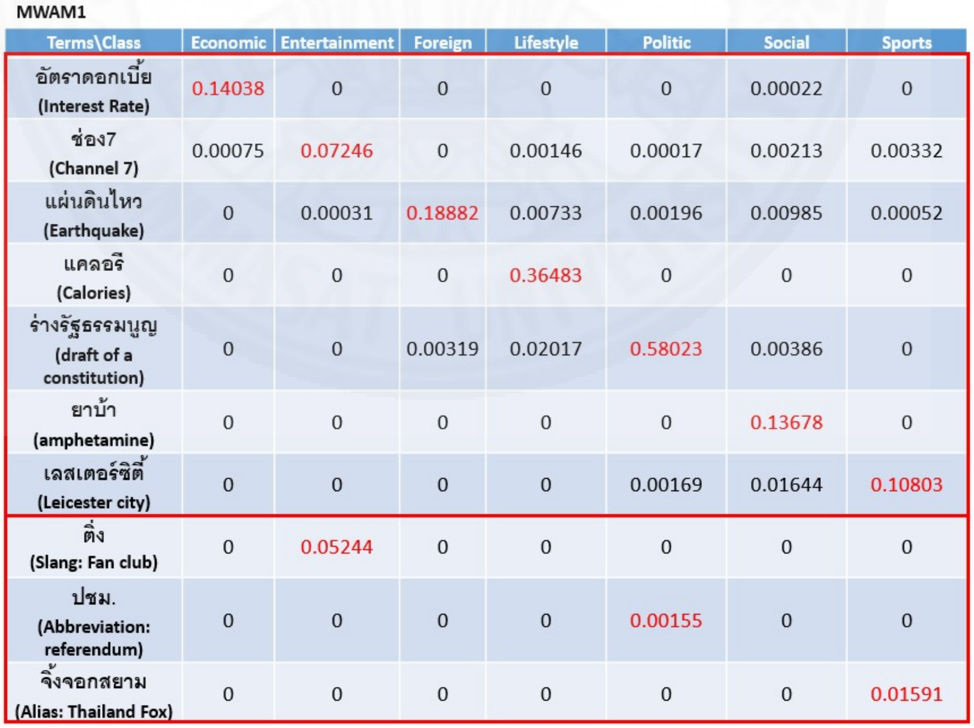
EC	ผู้โดยสาร	ทรัพย์สินทางปัญญา	พนักงานขับรถ	ชำระเงิน	กรมการขนส่งทางบก	อัตราดอกเบี้ย	ผู้ประกอบการ	กสทช.	ดัชนีหุ้นไทย	สินเชื่อ
EN	รีรัน	ริต้า	ช่อง7	โนห์รา	ช่อง8	เซ็กซี่	ผักบุ้งกุ้งนาง	ใบดำใบแดง	แชมมี่	พระเอก
FO	แผ่นดินไหว	ไอซิส	ผู้ต้องสงสัย	ประธานาธิบดี	เกาหลีเหนือ	ทิเบต	สำนักข่าวต่างประเทศ	แมกนิจูด	คูมาโมโต้	ทรัมป์
LI	ซูบารุ	คนโสด	สารตะกั่ว	พระกริ่ง	แคลอรี	อาเพศ	เสริมดวง	ตัวถัง	โปรตีน	ดีไซน์เนอร์
PO	ประชามติ	ร่างรัฐธรรมนูญ	คำถามพ่วง	นายมีชัย	พรรคประชาธิปัตย์	รัฐธรรมนูญ	ไม่เห็นด้วย	พรรคเพื่อไทย	พรรคการเมือง	บิดเบือน
SO	ยาบ้า	ที่เกิดเหตุ	จับกุม	คนร้าย	กระสุนปืน	หลบหนี	รถดับเพลิง	นายอำเภอ	กู้ภัย	ของกลาง
SP	กุนซือ	เรอัลมาดริด	แมนเชสเตอร์ยูไนเต็ด	เชลซี	ยูฟ่าแชมเปียนส์ลีก	สกอร์	ทีมเยือน	เลสเตอร์ซิตี้	ได้ประตู	นักชก

Figure 17 Top 10 keywords per class from the online news articles.



- RT| |@|TUMWARAWUTTS|9|:| |น่ารัก|มุ้งมิ้ง|จัง|เลย |!|!|!| |#|ตกสิบหยิบล้าน| |#|ช่อง7| |#|ตั้มวราวุธ| |#|สตูดิโอมนตรี| |cr|.|Auyzb|2|uty_|2|9|https|:|/|/|t|.|co|/|HRjj|7|GwmFD|
- RT| |@|polyplusfanclub|:| |ฮิปเตอร์|พร้อม|ล้า|ววว|วว|!|!| |#|วันวาน|ยัง|หวาน|อยู่| |#|ช่อง7| |#|ตอน|แน|◌้|
  |@|hone_thanakorn| |@|benz_punyaporn| |@|kana_rinyarat| |@|tarakhett| |htt|…|
- ถ้า|ประเทศ|กู|คง|รอ|แผ่นดินไหว|อีกรอบ|ให้|มัน| |fix| |กันเอง| |5|5|5|5|5|5|5|5|5|5|5|5| |https|:|/|…|
- น้องชาย|ชวน|ไป|ญี่ปุ่น|แต่|โดน|แม่|บ่น|ว่า|แผ่นดินไหว|.|.|.|นี่|พยาม|ชักจูง|ว่า|ไป|เกาหลี|แทน|ไหม|เกาหลี|เหอะ|ๆ|มัน|หันมา|ทำหน้า|เซ็ง|ๆ|ว่า|ไป|แต่|ห้าม|ติ่ง|นะ|.|.|.|งั้น|ไม่|ไป|ล่ะ|
- มะ|กี้|อยาก|กิน|ไร|หวาน|ๆ| |เลย|แงะ|คุ|◌้|ก|กี้|ออกมา|กิน| |แค่|สอง|ชิ้น|ตอน|กำลัง|งับ|ชิ้น|ที่สอง| |ก็|เหลือบ|ไป|เห็น|แคลอรี|บน|ถุง| |1|3|0|/|cookie| |โอเค| |น่ารัก| |5|5|5|5|5|5|5|5|
- ห้าม|วิจารณ์|ร่างรัฐธรรมนูญ |แต่|รัฐบาล|เสนอ|ข้อดี|ของ|ร่าง|รัฐ|ะ|ร|ร|มนูญ|อยู่|ฝ่ายเดียว|อันนี้|ไม่ผิด| |อ|◌ื|ม|ๆ| |ร่าง|เหี้ย|ไร|มี|แต่|ข้อดี|
- "|บิ|◌็|◌็|ก|ตู่|" |ลั่น| |!|!|!| |มี|ร่างรธน.|ไว้|ใน|หัว|แล้ว| |หาก|ปชม.|ไม่ผ่าน| |-| |จวก|พรรคการเมือง|หยุด|อ้าง|ปชช.| |#|บิ๊กตู่| |#|ร่างรัฐธรรมนูญ|.|.|.| |https|:|/|/|t|.|co|/|alduLfn|4|cJ|
- ตารางคะแนน|ล่าสุด |ถ้า|วันนี้| |"| |จิ้งจอกสยาม |เลสเตอร์ซิตี้| |"| |ชนะ|ๆ|ๆ|ๆ|ๆ|ๆ|ๆ|ๆ|ๆ|ๆ|ๆ|
  |https|:|/|/|t|.|co|/|Qwvft|9|spC|7|

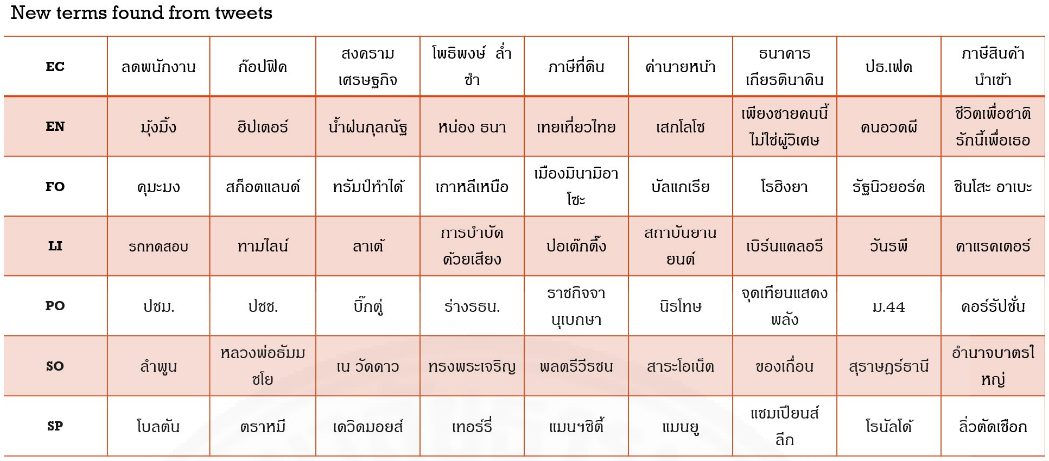
Figure 18 An example of retrieved related tweets with newly found keywords.



MWAM1

Terms\Class	Economic	Entertainment	Foreign	Lifestyle	Politic	Social	Sports
อัตราดอกเบี้ย (Interest Rate)	0.14038	0	0	0	0	0.00022	0
ช่อง7 (Channel 7)	0.00075	0.07246	0	0.00146	0.00017	0.00213	0.00332
แผ่นดินไหว (Earthquake)	0	0.00031	0.18882	0.00733	0.00196	0.00985	0.00052
แคลอรี (Calories)	0	0	0	0.36483	0	0	0
ร่างรัฐธรรมนูญ (draft of a constitution)	0	0	0.00319	0.02017	0.58023	0.00386	0
ยาบ้า (amphetamine)	0	0	0	0	0	0.13678	0
เลสเตอร์ซิตี้ (Leicester city)	0	0	0	0	0.00169	0.01644	0.10803
ติ่ง (Slang: Fan club)	0	0.05244	0	0	0	0	0
ปชม. (Abbreviation: referendum)	0	0	0	0	0.00155	0	0
จิ้งจอกสยาม (Alias: Thailand Fox)	0	0	0	0	0	0	0.01591

Figure 19 An example of the MWAM1.

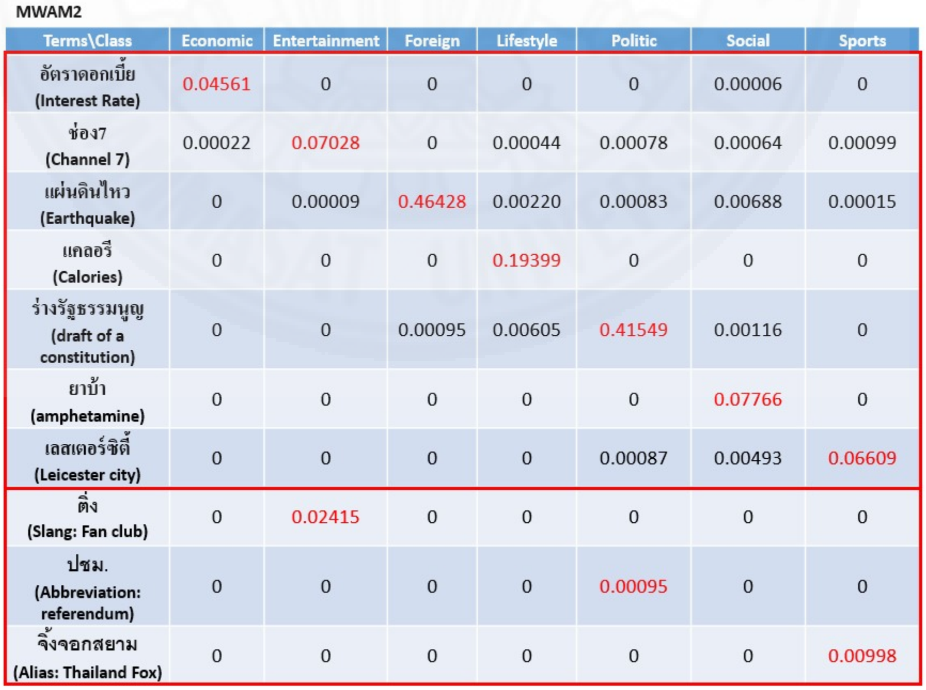


New terms found from tweets

EC	ลดพนักงาน	ก๊อปฟิค	สงครามเศรษฐกิจ	โพธิพงษ์ ล่ำซำ	ภาษีที่ดิน	ค่านายหน้า	ธนาคารเกียรตินาคิน	ปธ.เฟด	ภาษีสินค้านำเข้า
EN	มุ้งมิ้ง	ฮิปเตอร์	น้ำฝนกุลณัฐ	หน่อง ธนา	เทยเที่ยวไทย	เสกโลโซ	เพียงชายคนนี้ไม่ใช่ผู้วิเศษ	คนอวดผี	ชีวิตเพื่อชาติรักนี้เพื่อเธอ
FO	คุมะมง	สก็อตแลนด์	ทรัมป์ทำได้	เกาหลีเหนือ	เมืองมินามิอาโซะ	บัลแกเรีย	โรฮิงยา	รัฐนิวยอร์ค	ชินโสะ อาเบะ
LI	รถทดสอบ	ทามไลน์	ลาเต้	การบำบัดด้วยเสียง	ปอเต๊กติ๊ง	สถาบันยานยนต์	เบิร์นแคลอรี	วันรพี	คาแรคเตอร์
PO	ปชม.	ปชช.	บิ๊กตู่	ร่างรธน.	ราชกิจจานุเบกษา	นิรโทษ	จุดเทียนแสดงพลัง	ม.44	คอร์รัปชั่น
SO	ลำพูน	หลวงพ่อธัมมชโย	เน วัดดาว	ทรงพระเจริญ	พลตรีวีรชน	สาระโอเน็ต	ของเถื่อน	สุราษฎร์ธานี	อำนาจบาตรใหญ่
SP	โบลตัน	ตราหมี	เดวิดมอยส์	เทอร์รี่	แมนฯซิตี้	แมนยู	แชมเปียนส์ลีก	โรนัลโด้	ลิ่วตัดเชือก

Figure 20 Newly found keywords from the related tweets.

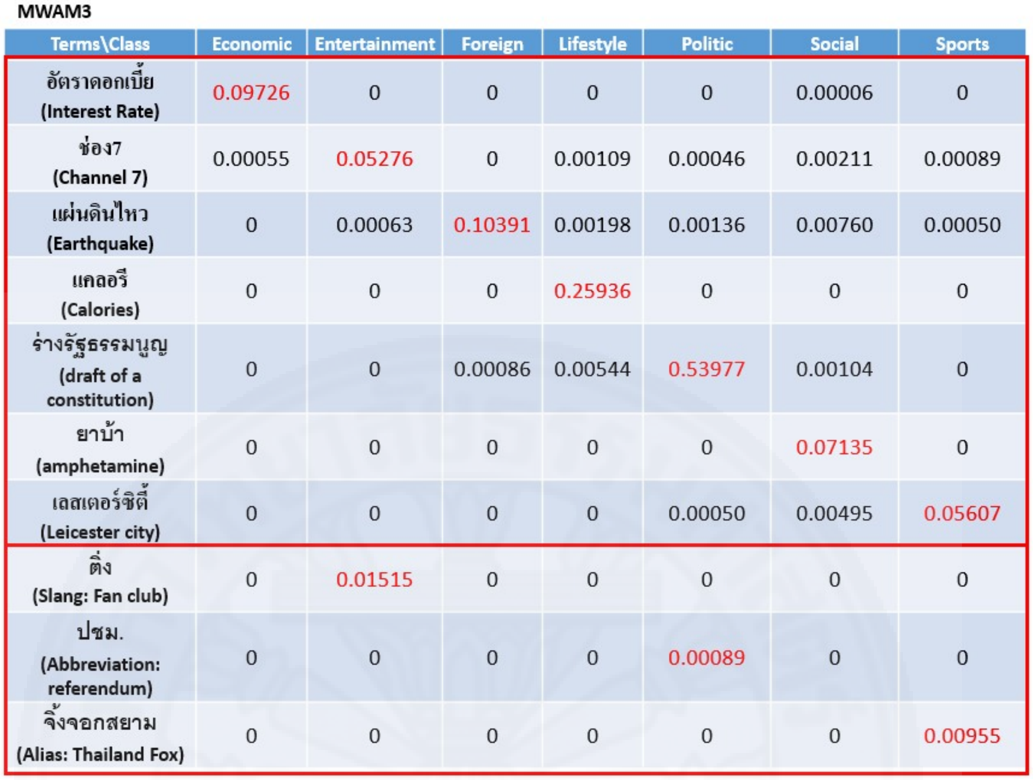
The MWAM2, and MWAM3 are created by the relevant updating with 9,516 terms, and 3,433 terms from related tweets respectively. Finally, the MWAM3 contains 39,952 terms in bag-of-words and I found that this model generate a promising result for the social media text classification as I will discuss in the next section, Evaluation Result. An example of the MWAM2 and MWAM3 are shown as Figure 21 and Figure 22.



MWAM2

Terms\Class	Economic	Entertainment	Foreign	Lifestyle	Politic	Social	Sports
อัตราดอกเบี้ย (Interest Rate)	0.04561	0	0	0	0	0.00006	0
ช่อง7 (Channel 7)	0.00022	0.07028	0	0.00044	0.00078	0.00064	0.00099
แผ่นดินไหว (Earthquake)	0	0.00009	0.46428	0.00220	0.00083	0.00688	0.00015
แคลอรี (Calories)	0	0	0	0.19399	0	0	0
ร่างรัฐธรรมนูญ (draft of a constitution)	0	0	0.00095	0.00605	0.41549	0.00116	0
ยาบ้า (amphetamine)	0	0	0	0	0	0.07766	0
เลสเตอร์ซิตี้ (Leicester city)	0	0	0	0	0.00087	0.00493	0.06609
ติ่ง (Slang: Fan club)	0	0.02415	0	0	0	0	0
ปชม. (Abbreviation: referendum)	0	0	0	0	0.00095	0	0
จิ้งจอกสยาม (Alias: Thailand Fox)	0	0	0	0	0	0	0.00998

Figure 21 An example of the MWAM2.



MWAM3

Terms\Class	Economic	Entertainment	Foreign	Lifestyle	Politic	Social	Sports
อัตราดอกเบี้ย (Interest Rate)	0.09726	0	0	0	0	0.00006	0
ช่อง7 (Channel 7)	0.00055	0.05276	0	0.00109	0.00046	0.00211	0.00089
แผ่นดินไหว (Earthquake)	0	0.00063	0.10391	0.00198	0.00136	0.00760	0.00050
แคลอรี (Calories)	0	0	0	0.25936	0	0	0
ร่างรัฐธรรมนูญ (draft of a constitution)	0	0	0.00086	0.00544	0.53977	0.00104	0
ยาบ้า (amphetamine)	0	0	0	0	0	0.07135	0
เลสเตอร์ซิตี้ (Leicester city)	0	0	0	0	0.00050	0.00495	0.05607
ติ่ง (Slang: Fan club)	0	0.01515	0	0	0	0	0
ปชม. (Abbreviation: referendum)	0	0	0	0	0.00089	0	0
จิ้งจอกสยาม (Alias: Thailand Fox)	0	0	0	0	0	0	0.00955

Figure 22 An example of the MWAM3.

4.4 Evaluation Result

All models: i-WAM, MWAM1, MWAM2, MWAM3, are evaluated with 2 types of test dataset: test dataset 1 (filtered tweets, 6,339 tweets), and test dataset 2 (raw tweets, 20,000 tweets). For the test dataset 1, I filtered all duplicated, re-tweets, and junk tweets. This test dataset is used to test the model efficiency of clean (useful) tweets. While, the test dataset 2, raw tweets, is used to perform a real-life test with tweets that contain all of useful tweets, duplicated, and junk (useless) one.

In Figure 23, the i–WAM efficiency, is very low when it was evaluated with the raw tweets, accuracy 74.64%. However, after the model was updated to be the MWAM1, the efficiency is increased significantly, accuracy 86.84%. Finally, when the MWAM2 and MWAM3 was created, the result of the evaluation seems to be more acceptable with accuracy 87.55%, and 88.75% respectively.

In addition, after the evaluation with clean tweets, test dataset 1, the efficiency of all models: the i-WAM, MWAM1, MWAM2, and MWAM3, are indicated at very high accuracy score with 98.57%, 99.38%, 99.59%, and 99.79% respectively. This result indicates an efficient solution, applying the manual filtered and tagged tweets before perform the classification process is a productive way to increase the model efficiency.

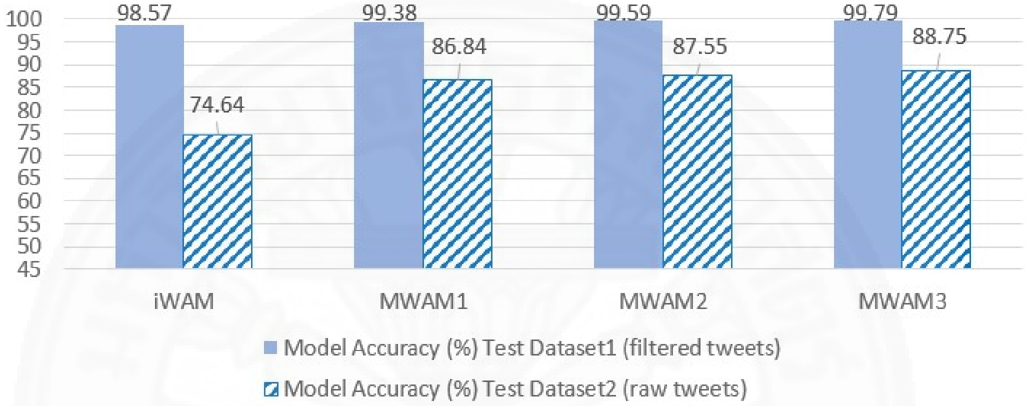


Figure 23 An accuracy (%) of all WAMs model with both test datasets.

In Table 6, Table 7. Table 8, Figure 24, Figure 25, and Figure 26, the precision, recall, and F-measure of all models which evaluated with the clean tweets, test dataset1, are shown. The result is illustrated in the same way, mostly increase when the relevant updating from related tweets is performed.

Table 6 Precision score (%) of all WAMs with test dataset1 (filtered tweets).

Class\WAM	i-WAM	MWAM1	MWAM2	MWAM3
Economic	100	100	100	100
Entertainment	90.9	100	100	100
Foreign	95	100	100	100
IT	83.33	95.23	100	100
Politics	100	100	100	100
Regionals	100	90.9	90.9	95.23
Sports	100	100	100	100

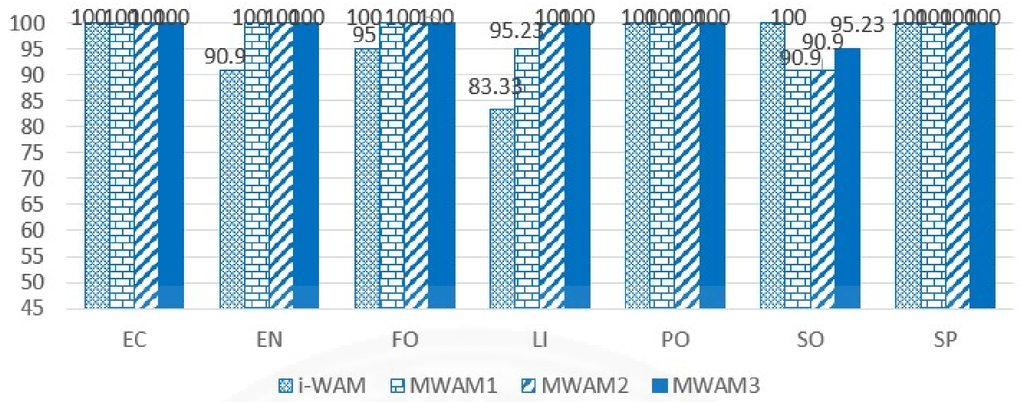


Figure 24 Precision score (%) of all WAM models with test dataset 1 (filtered tweets).

Table 7 Recall score (%) of all WAMs with test dataset1 (filtered tweets).

Class\WAM	i-WAM	MWAM1	MWAM2	MWAM3
Economic	100	100	100	100
Entertainment	100	100	100	100
Foreign	95	95	100	100
IT	100	100	100	100
Politics	80	90	90	95
Regionals	90	100	100	100
Sports	100	100	100	100

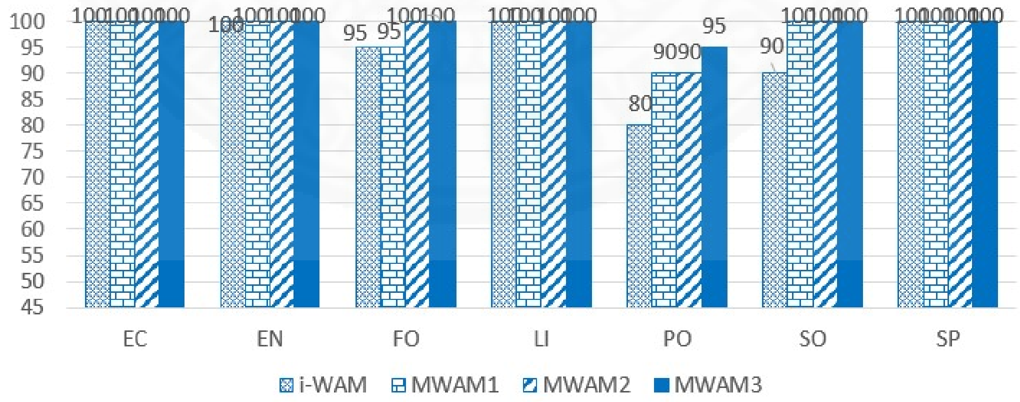


Figure 25 Recall score (%) of all WAM models with test dataset 1 (filtered tweets).

Table 8 F-measure score (%) of all WAMs with test dataset1 (filtered tweets).

Class\WAM	i-WAM	MWAM1	MWAM2	MWAM3
Economic	100	100	100	100
Entertainment	95.23	100	100	100
Foreign	95	97.43	100	100
IT	90.9	97.56	100	100
Politics	88.88	94.73	94.73	97.43
Regionals	94.73	95.23	95.23	97.56
Sports	100	100	100	100

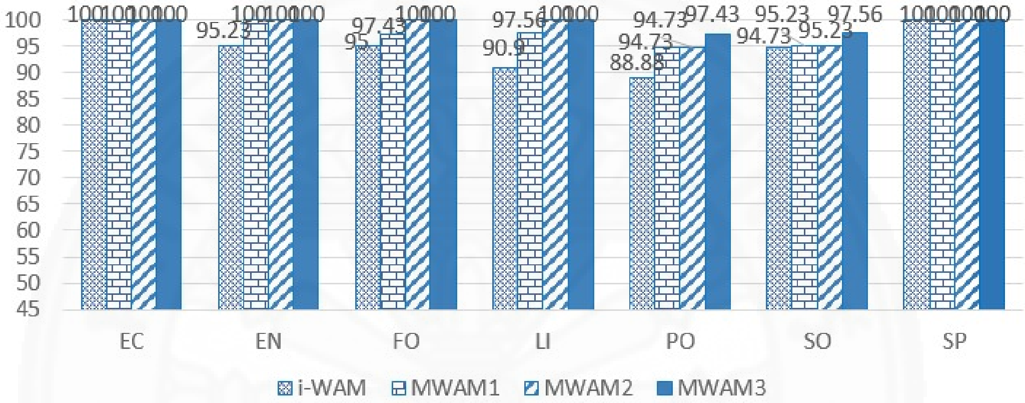


Figure 26 F-measure score (%) of all WAM models with test dataset 1 (filtered tweets).

In Table 9, Table 10. Table 11, Figure 27, Figure 28, and Figure 29, the precision, recall, and F-measure of all models which evaluated with the raw tweets, test dataset12, are shown with the same direction of the overall results.

Table 9 Precision score (%) of all WAMs with test dataset2 (raw tweets).

Class\WAM	i-WAM	MWAM1	MWAM2	MWAM3
Economic	83.16	85.25	86.67	88.14
Entertainment	63.64	81.13	81.13	81.13
Foreign	78.57	89.80	91.67	89.80
IT	68.70	91.86	92.94	92.94
Politics	95.83	78.79	81.25	81.25
Regionals	48.15	75.76	83.33	81.25
Sports	77.08	100	100	100

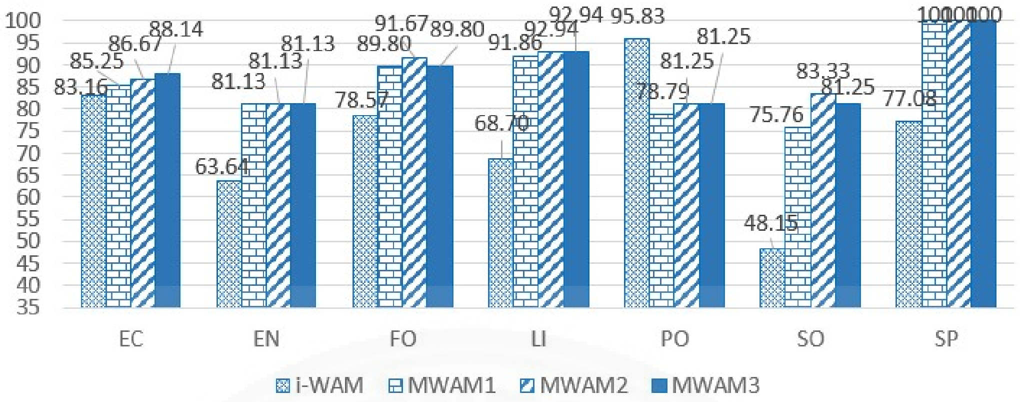


Figure 27 Precision score (%) of all WAM models with test dataset 2 (raw tweets).

Table 10 Recall score (%) of all WAMs with test dataset2 (raw tweets).

Class\WAM	i-WAM	MWAM1	MWAM2	MWAM3
Economic	71.82	94.55	94.55	94.55
Entertainment	52.83	81.13	81.13	81.13
Foreign	97.78	97.78	97.78	97.78
IT	95.18	95.18	95.18	95.18
Politics	74.19	83.87	83.87	83.87
Regionals	40.63	78.13	78.13	81.25
Sports	100	100	100	100

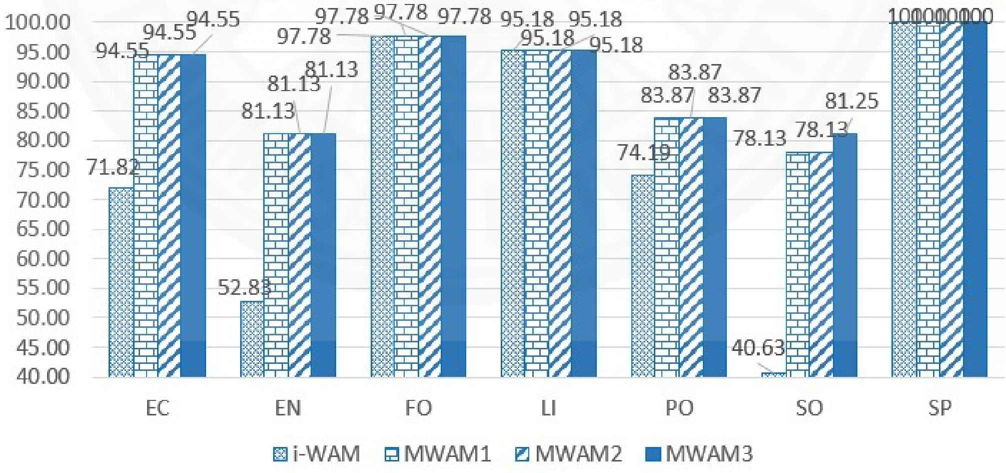


Figure 28 Recall score (%) of all WAM models with test dataset 2 (raw tweets).

Table 11 F-measure score (%) of all WAMs with test dataset2 (raw tweets).

Class\WAM	i-WAM	MWAM1	MWAM2	MWAM3
Economic	77.07	89.66	90.43	91.23
Entertainment	57.73	81.13	81.13	81.13
Foreign	87.13	93.62	94.62	93.62
IT	79.80	93.49	94.05	94.05
Politics	83.64	81.25	82.54	82.54
Regionals	44.07	76.92	80.65	81.25
Sports	87.06	100	100	100

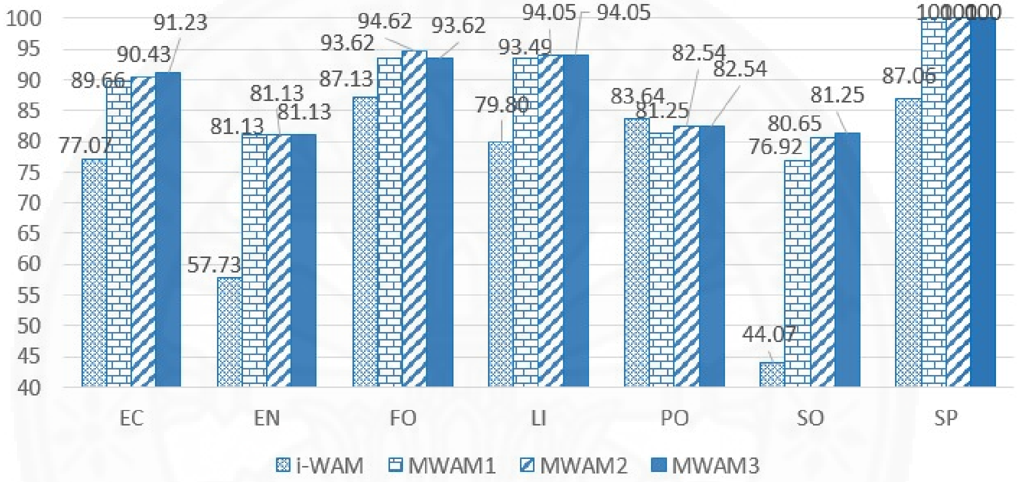


Figure 29 F-measure score (%) of all WAM models with test dataset 2 (raw tweets).

Chapter 5
Conclusions and Future Plan

"The concept of utilizing the well-formed text sources like the online news articles to build the primary model and enhanced with more specific and suitable words from the social media text itself is a productive way to build the best model for classification. While, Name Entity Recognition (NER), Deep learning, and GPU Computing be the interesting research areas to generate the most productive Word Segmentation tool."

Social Media Mining is the most interesting area of the information science field. Many useful information for decision making support could be gleaned from this novel data source. While, the classification process is a beginning step which is very important and so challenge to handle. The vector space model, Word Article Matrix (WAM), with Term Frequency-Inverse Document Frequency (TF-IDF) technique, and normalized TF-IDF Merging with Specific Terms Weighting Technique are the effective solution for the social media text classification task. Moreover, the concept of utilizing the well-formed text sources like the online news articles to build the primary model and enhanced with more specific and suitable words from the social media text itself is a productive way to build the best model for classification. The iteration of model updating concept could expand the coverage area of words in each class until the stable model is found.

As an efficiency of the vector space model depends on words quality, variety of words, and coverage area of words, Word segmentation tool is a major module which cannot be overlooked. More reliable of this module can define words boundary accurately which can increase the high quality word, correctness, generation rate. Name Entity Recognition (NER), and Deep learning be the interesting research areas to generate the most productive Word Segmentation tool. Finally, a High Performance Computing (HPC) or Graphics Processor Unit Computing (GPU Computing) is an attractive modern technique when we deal with large scale matrix and floating point calculation like the vector space model. This useful technique could reduce more computational time of the experiment dramatically.

References

Books and Book Articles

20. Luger, G.F.(2008).*Artificial Intelligence: Structure and Strategies for Complex Problem Solving*.6th edition. Addison Wesley.

28. Abel, F., Gao, Q., Houben, G.J.&Tao, K.(2011). *The Semantic Web: Research and Applications: Semantic enrichment of twitter posts for user profile construction on the social web*, 6644 of Lecture Notes in Computer Science, 375-389. Springer Berlin Heideberg.

Articles

7. Irfan, R. et al.(2015). A survey on text mining in social networks. *Cambridge Journal, The Knowledge Engineering Review*, 30(2),157-170.

8. Sorensen, L.(2009). Usermanaged trust in social networking comparing facebook, myspace, and linkdin. *Proceedings of 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic System Technology (Wireless VITAE 09)*. Denmark, 427-431.

9. Kano,Y. et al.(2009).Data Mining: Concept and Techniques. *Oxford Journal of Bioinformatics*, 25(15), 1997-1998.

10. Dai, Y., Kakkoen, T. & Sutinen, E.(2011).MinEDec: A decision-support model that combines text-mining technologies with two competitive intelligence analysis method. *International Journal of Computer Information System and Industrial Management Applications*, 3, 165-173.

11. Foreman, G.&Kirshenbaum, E.(2008).Extremely fast text feature extraction for classification and indexing. *Proceedings of 17th ACM Conference on Information and Knowledge Management*, California, USA, 26-30.

12. Yuan, L.(2010). Improvement for the automatic part-of-speech tagging based on hidden markov model. *Proceedings of 2nd International Conference on Signal Processing System IEEE (ICSPS)*, China, 744-747.

13. Strapparava, C.&Ozbal, G.(2010). The color of emotion in text. *Proceedings of 2nd Workshop on Cognitive Aspects of the Lexicon*, Beijing, 28-32.

14. Esuli, A.&Sibastiani, F.(2006). SentiWordNet: A publicly available lexical resource for opinion mining. *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Italy, 417-422.

15. Ling, H.S., Bali, R.&Salam, R.(2006). Emotion detection using keywords spoting and senmantic network. *Proceedings of International Conference on Computing and Informatics IEEE (ICOCI)*, Kuala Lumpur, 1-5.

16. Hua, J. Tembe, W.D., Dougherty, E.R.&Edward, R.D.(2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42(3), 409-424.

17. Shekar, C.B.H.&Shoba, G.(2009). Classification of documents using kohonens self organizing map. *International Journal of Computer Theory and Engineering (IACSIT)*, 1(5), 610-613.

18. Yoshida, K., Tsuruoka, Y., Miyao, Y.&Tsujii, J.(2007). Ambiguous part-of-speech tagging for improving accuracy and domain portability of syntactic parsers. *Proceedings of 2nd International Conference on Signal Processing System IEEE (ICSPS)*, China, 744-747.

19. Zhao, Y.&Dong, J.(2009). Ontology classification for semantic-web-based software engineering. *IEEE Transactions on Service Computing*, 2(4), 303-317.

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

22. Nirmala, K., Satheesh kumar, S.&Vellinggiri, J.(2013). A Survey on Text categorization in Online Social Networks. *International Journal of Engineering Technology and Advanced Engineering*, 3(9), 446-450.

23. Lee, K. et al.(2011). Twitter Trending Topic Classification. *Proceeding of the 2011 IEEE 11th International Conference on Data Mining Workshops, ICDW'11*, 251-258.

24. McCallum, A., Nigam, K., et al.(1998). A comparison of event models for naïve bayes text classification. *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, 41-48.

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

26. Chung Wong, P., Foote, H., Adams, D., Cowley, W.&Thomas, J.(2003). Dynamic visualization of transient data streams. *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium*, 97-104.

27. Weng, J.&Lee, B.S.(2011).Event detection in twitter. *The International AAAI Conference on Weblogs and Social Media (ICWSM)*, 11, 401-408.

29. Chirawichitichai, N. et al.(2010). A Comparative Study on Feature Weight in Thai Document Categorization Framework. *10th International Conference on Innovative Internet Community Services (I2CS)*, IICS, 257-266.

30. Theeramunkong,T.&Lertnattee,V. (2005). Multi-Dimension Text Classification. *Proceeding of COLING '02 Proceedings of the 19th international conference on Computational linguistics*, 1, 1-7.

31. Viriyayudhakorn, K. et al.(2011). A Comparison of Four Association Engines in Divergent Thinking Support Systems on Wikipedia. *Knowledge, Information, and Creativity Support Systems, KICSS2010*, Springer, 226-237.

32. Sornlertlamvanich, V. et al.(February 2015). *Understanding Social Movement by Tracking the Keyword in Social Media*, in MAPLEX2015, Yamagata, Japan.

33. Jotikabukkana, P., Sornlertlamvanich, V., Manabu, O.&Haruechaiyasak, C.(2015). Social Media Text Classification by Utilizing Well-Formed Text Trained Model. *Journal of ICT Research and Applications Institut Teknologi Bandung*, In Review Round 2, 23 March 2016.

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

35. Meknavin S., Charoenpornsawat P.,&Kijsirikul B.(1997). Feature-based Thai Word Segmentation. *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997(NLPRS'97).* Phuket, Thailand.

36. Ho, C.W. et al.(2008). Interpreting TF-IDF Term Weights as Making Relevance Decisions. *ACM Transactions on Information Systems*, 26(3), Article 13, 1-37.

39. Cohen J. A.(1960) coefficient of agreement for nominal scales. *Educ Psychol Meas*, 1960(20), 37–46.

Electronic Media

1. Simon, K.(2015), *Social&Mobile Worldwide in 2015*, Retrieved 9 June 2015, from http://wearesocial.net/tag/statistics/
2. Twitter Inc.(2015), *Twitter Usage/ Company Facts*, Retrieved 9 June 2015, from https://about.twitter.com/company
3. Dave C.(2015), *Global social media research summary 2015*, Retrieved 15 June 2015, from http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
4. Twitter Inc.(2015), *Entities in Objects*, Retrieved 15 June 2015, from https://dev.twitter.com/overview/api/entities-in-twitter-objects
5. Twitter Inc.(2015), *API Rate Limits*, Retrieved 15 June 2015, from https://dev.twitter.com/rest/public/rate-limiting
6. Twitter Inc.(2013), *Search API is limited to the last 7 days?*, Retrieved 15 June 2015, from https://twittercommunity.com/t/search-api-is-limited-to-the-last-7-days/11603

37. Vembunarayanan J.(2013), *Tf-Idf and Cosine similarity*, Retrieved 15 June 2015., from https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/

38. Thairath.(2016), Online News, *Thairath web*, Retrieved from 1 April 2016, from http://www.thairath.co.th/