Extraction of Trend Keywords and Stop Words from Thai Facebook Pages Using Character n-Grams

Nattapong Ousirimaneechai and Sukree Sinthupinyo

***Abstract*—In the era of data and information, insight of user's behavior such as trend is normally used in real-time marketing for improvement of gross profit, therefore, it is beneficial to know the trend in social media. Word tokenization and stop words list are the conventional method for keyword extraction task, however for Thai language in social media platform, there are still no efficient word tokenization tools and stop words list to extract trend from platform such as Facebook. Therefore, in this research, we propose an algorithm that require no word tokenization tools and external stop words list for the purpose of *Trend Keywords* extraction. The core idea is using Character n-Grams, instead of Word n-Grams, to tokenize, process, and combine n-Grams into keyword. After that we identified *Trend Keywords* from other keywords by using our algorithm to generate stop words list for filtering out stop words. For the evaluation of result, we use human to classify the retrieved *Trend Keywords* and compare them with *Trend Keywords* from baseline method. As a result, our algorithm can identify more keyword than baseline method. Finally, the precision of generated stop words list is 97.6%, and the precision of *Trend Keywords* is 40% with the used of 1-month generated stop words list. Furthermore, by using 2-months generated stop words list, the precision can be increased to 44% by consuming more processing time for list of stop words.**

***Index Terms*—Information retrieval, keyword extraction, social media mining, stop words.**

I. Introduction

Facebook is one of the largest social media platforms. It allows users to set up *Facebook page* which is a public profile created for a specific purpose. Facebook pages can be divided into many categories, such as business, news, celebrity, brands, and organization. The amount of public information generated and provided by these pages is enormous and their contents are up-to-date; therefore, if we can extract and analyze *Trend Keywords*, which are appeared globally across multiple public pages and still appear for a duration of time, it will be possible to find global trends, events, or even behavior of the mass.

The solutions for keywords extraction in Thai language still have plenty rooms for improvement. Most of the solutions require external tools, which is word tokenization, or external database, such as training data, Thai word corpus, and stop words list. However, due to the complexity of Thai language, most of the tools are not robust enough.

The most complicacy part of Thai word tokenization is sentence segmentation. To begin with, word separation does not present between words in Thai sentence and the alphabet itself also not having their own meaning, so reliable word tokenization tool is significantly crucial. Moreover, the present official tokenization tools do have limitation to use with social media like Facebook due to occurrence of the newly adapt and slang word out of official dictionary.

For tokenization of Thai word, there is a corpus by NECTEC called BEST2010 [1], which provided data like dictionary, segmented sentences or part of speech of each word in sentences, which is the only one widely available for Thai corpus. However, this corpus is not up-to-dated.

For stop words lists, there are only small lists available. The biggest one consists of only 115 words which is not nearly enough for social media information analysis.

From reasons stated above, we decided to develop our own algorithm that require no training data or external tools. The objective of this algorithm is to extract *Trend Keywords* from Thai Facebook pages using Character n-Grams instead of word tokenization method.

To test the algorithm, we decided to collect Facebook posts from several public pages and analyzed them, using our developed algorithm, to find *Trend Keywords* during specific time periods.

II. Related Work

A. Finding Keywords

For keyword extraction, n-Grams and stop words should be stated and clarified before.

n-Grams [2] is an algorithm for slicing a message into small groups, called gram, which has constant length. n-Grams can be used to slice sentences and phrases by word or character called *Word n-Grams* and *Character n-Grams* respectively. If the message is sliced into 1-length gram called unigram. In the same way, we call bigram for 2-lengths gram and trigram for 3-lengths.

Stop words [3] are words which should be filtered out before finding other significant keywords in documents, because these stop words can appear frequently but provide no specific or important meaning. For example, common word such as "the" and "and" are considered as stop words.

Keyword Extraction [4] is described as an automatic task that can identify important terms which best represent the content of a document.

To find keywords, one of the most popular algorithm in Keyword Extraction is TF-IDF [3], TF stands for Term Frequency and IDF stands for Inverse Document Frequency. TF-IDF is TF value multiply by IDF value of the same word, so the word with high frequency and appearing in less

Manuscript received August 29, 2018; revised October 20, 2018.

The authors are with Chulalongkorn University, Thailand (e-mail: 6070188521@student.chula.ac.th, sukree.s@chula.ac.th).

doi: 10.18178/ijmlc.2018.8.6.750

documents is the keyword. On the other hand, the word with low frequency or appearing in almost documents is not the keyword.

TF-IDF with Word n-Grams can be used to find multiple-word keyword. For an example, to find keywords with 1-to-3-length, all unigram, bigram and trigram will be applied with TF-IDF, then grams with high TF-IDF value are keyword. By this method, stop words with low IDF value will be excluded from the result keyword list, but also consumed more computing power and memory. If the stop words database is provided, the calculation process will be reduced to only TF method without finding IDF value. Unfortunately, the stop word is unique and diverse across categories. [5] In some categories, ex. travel, the seems like stop word "from" and "to" will be not excluded from the keyword list according to the meaning of the location, which these words are needed, however not in others. For each specific category, stop word is not easy to come by from the accessible database, and general stop words database is not enough to categorize keyword from off-list stop words.

According from the reason above, stop word extraction task is an important task. One's solution to extract stop word is filtering by high TF or low IDF value. There are plenty room to improve such as R. T. Lo, *et al.* [6] proposed a method to build a stop words list by using Kullback-Leibler divergence on the lexicon file and select L top rank words, where L is a parameter. The result show that the new method provided higher average precision than the conventional method of using TF, term frequencies, on the corpus and perform thresholding to select stop words.

B. Finding Trend Keywords

Trend Keyword has some properties that are different from normal keyword. *Trend Keyword* should be appeared in most of all documents with high frequency like stop word, but it has a specific meaning. From this property, *Trend Keyword* must be separated apart from the extracted stop word list.

There is some similar research about this topic in Thai Language. A. Piyatumrong, *et al.* [7] want to find the keyword that can represent event in social using Twitter data. They used a word tokenization tool, LexToPlus, to tokenize Thai twitter message into a set of unigrams, bigrams with stop words removal and a set of bigrams without stop words removal, then comparing with non-required word tokenization tool algorithm that using hashtag and hashtag5, which have more than 5 characters. From 5 groups, both of only TF and TF-IDF method were applied to extract event keyword. The output result of TF and TF-IDF were similar event keyword without any statistical significant difference, so it was suggested that the IDF value could not classify event keyword from stop word. According to no different results, the only TF method were used for further analysis due to implementation, memory and processing process. At the end, across all method, hashtag5 shown the best F1 score. However, the limitation of hashtag5 was that it can only identified event keywords specified in hashtag and missed other important keywords in the content. In addition, they focused on the extracted event keywords, and founded a number of intersection between event keywords with each method is low, so it means that, the extracted event keyword cannot be relied on only one specific mentioned method.

C. Word Tokenization

Thai word tokenization tools are developed by NECTEC research organization. At first, they developed a tool named LexTo that is a dictionary-based word tokenization using longest matching, but it is not robust for words that not have in a dictionary, so the precision of this tool is very low in this case. After that, they developed a tool named TLex that is a machine learning-based word tokenization using conditional random fields and use the training data from BEST corpus that officially segmented. The precision of this tool is impressive, but some segmented words should be segmented into a smaller word, so they used dictionary-based word tokenization to help after segmented by TLex and called LearnLexTo [8]. Unfortunately, this tool still has its own limitation in some categories, ex. social media, due to intentional spelling error, usually found in social media, is generating new words and new sentences, which are not in BEST corpus, and LearnLexTo is not trained by these intentional spelling error sentences, so NECTEC decided to develop new tool from LexTo named LexToPlus [9]. With LexToPlus, user can access to add any new words to dictionary, and the method itselfs can be applied with insertion type of intention spelling error, so it will be practically used, if the up-to-date social media dictionary are provided.

D. Finding Keywords without Word Tokenization

A research of C. Haruechaiyasak, *et al.* [10] proposed an algorithm to extract keyword from categorized documents by not using word tokenization. The concept of algorithm is adding character to string and checking TF and IDF value to classify a keyword. Begin with finding TF value of all 1-character length string, then add one more character that TF value pass the threshold, the rest will be rejected. Before adding a character, each string will be check whether it is keyword or not by thresholding IDF value. The keyword string will be collected, then all of the string will be sent for the next TF value calculating. These process goes on until none of the string pass through TF value thresholding. Finally, all of the keyword list will be collected.

However, by using a data structure named suffix array to optimize the calculation process for each string contain in the documents, but the optimization is not sufficient. At the end, the process loop is terminated before completion, and getting the partially keyword, so the combination of keyword on the list is required. Moreover, it is suggested that, combining between the partial keyword under the condition that both of keyword are overlapped with the same exact TF value. Although the result is quite impress, some keywords are still lost and some in the result list are wrong.

III. Methodology

In Thai language, words are not separated by space, instead, the space is used to separate sentences. However, in social media platform, such as Facebook, the format is unrestricted, as shown in Fig. 1. It is also possible to incorporate two or more type of character, such as Thai, English, emoticon,

number, special character, and hashtag.

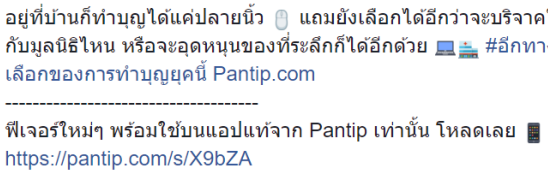


Fig. 1. A sample of Thai Facebook post.

A. Preprocessing of Original Posts

For each post collected from Facebook public pages, the post is split by space into smaller separated messages. Special characters, which are not Thai character, English Character or number, at the beginning and the end of each message are then removed by left-right message stripping. English characters are changed to lowercase, hyperlinks are removed, every split message are joined back, and finally, remove all space. The result is shown in Fig. 2.

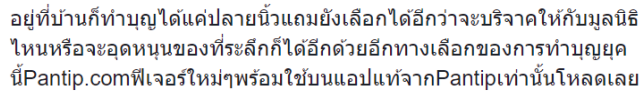


Fig. 2. After preprocessing of the example post.

B. Finding Local Keyword Grams of Each Pages

After preprocessing, posts from a specific page are selected. Each post is tokenized into multiple grams of 5 character then grouped by date of original post. In each group, we count the number of time that each gram appears on that specific date and increment by the number of time it appeared on the day before and the day after. We give an example of Gram counting in Fig. 3.

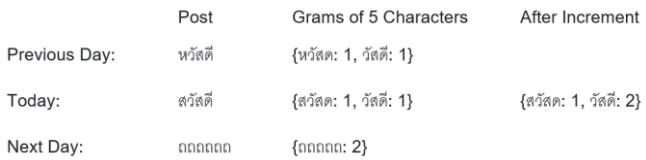


Fig. 3. Gram counting

After the process described in the above paragraph, we then have the number of frequency for each gram in a specific date. We used K-Means Clustering [11] to cluster the counted grams into k cluster, where k is the optimal number of cluster determined by Elbow Method [12] using maximum point-to-line distance function [13] as shown in Fig. 4.

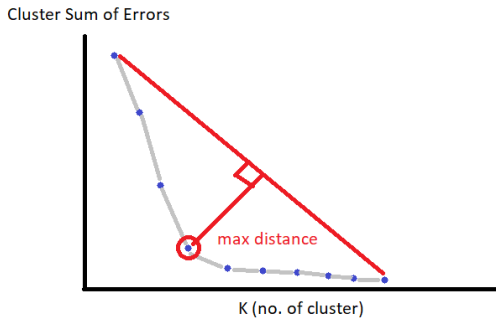


Fig. 4. Elbow Method using maximum point-to-line distance function.

Rank of each cluster is determined by mean of that cluster. The highest rank cluster is the cluster with the highest mean and the lowest rank cluster is the cluster with the lowest mean. The lowest rank cluster is discarded. The remaining clusters are kept because they have the probability to contain both keywords and stop words. We defined the remaining clusters as *Local Keyword Grams* of page.

C. Finding Global Keyword Grams

In each date, we put together the *Local Keyword Grams* from every pages and summing the frequency of the same gram. We then use again K-Means Clustering. The rank of each cluster determines by mean. We kept only the highest rank cluster and define it as *Global Keyword Grams*. The rest of the clusters are discarded as they are more likely to represent local keywords.

D. Finding Global Keywords in Each Post

To find *Global Keywords* that appear in a post, first, we need to obtain *Global Keyword Character* by evaluating every character in the post. For each character in each position, there are 5 grams which contain that character, except for the first four and the last four character of the post. Considering the set of 5 grams for a character, if there are one or more grams that match with *Global Keyword Grams*, we consider that character to be a *Global Keyword Character*. An example of *Global Keyword Character* is given in Fig. 5.

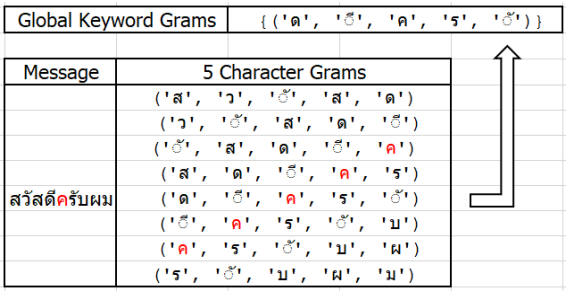


Fig. 5. Example of *Global Keyword Character*.

After that, we replace non-*Global Keyword Character* with space and split the whole post by space, so we can obtain *Global Keywords* for this post as shown in Fig. 6.

Fig. 6. Extraction of *Global Keywords* after identification of *Global Keyword Characters.*

Unfortunately, due to the nature of stop words that tend to appear in high frequency, the obtained *Global Keywords* is still containing stop words which need to be removed.

E. Building a Stop Words List from Global Keywords

After we obtain *Global Keywords* from every day of interest, it is possible to retrieve *Trend Keywords* and stop words. *Trend Keywords* appear during a specific period of days during the event, but stop words are appearing all the time. Therefore, we can set up a condition to distinguish *Trend Keywords* from stop words. Considering each *Global Keyword*, if the keyword appears during the period of 35-55

days prior to the day of interest, there's a high probability that it is a stop word. We retrieve all keywords that met with this condition and use them to build a stop words list. We can also use this process to retrieve stop words from each day and add them together.

F. Finding Trend Keywords

After we have the stop words list, we can determine if a *Global Keyword* is a *Trend Keyword*. Considering each remaining *Global Keywords* that is not stop words, if the keyword appears on 1-10 days prior to the day of interest, we can consider it as a *Trend Keyword*. The remaining keyword can be considered as a *Global Keyword* of that specific day.

IV. Parameter Discussions

A. Number of Character Grams

To choose the number of character in a gram, we must consider the number of character in Thai word. From the work of A. Piyatumrong which we discussed in related work, the average length of Thai word is 5 characters. Therefore, we decided to use 5-character n-grams in our algorithm. It is important to note that the character length of gram can be changed and is considered as a tradeoff. For example, 3 characters gram gives a high detection sensitivity of both keywords and stop words, on the contrary, 7 characters gram can only detect long words (7 characters or more) which are more likely to be keywords than stop words.

B. Range of Day Used to Weight Local Keyword Grams

For weighting of *Local Keyword Grams* in a specific day, we must consider the nature of trends and events that normally appear on social media for only a certain period. Therefore, we decided to weight each *Local Keyword Grams*, by summing the appearing time on that day, the day prior, and the day after. If we use bigger range, stop words which appear nearly all the time will gain more weight than the actual keywords.

C. Clustering Method

To cluster keywords by appearing frequency, we decided to use K-Means Clustering method which is computationally inexpensive and sufficient. To choose the optimal number if clustering (K), we used elbow method with point to line distance function to select K. With this method, we can obtain the number of clustering which is suitable for removing non-keyword grams. If the number of cluster is more than K, *Local Keyword Grams* will be further divided and this result in the vagueness between keyword grams and non-keyword grams.

D. Ranking of Global Keyword Grams

We rank each cluster by its mean. Cluster with the highest mean has the highest rank, and cluster with the lowest mean has the lowest rank. We chose keywords gram in the highest rank cluster as *Global Keyword Grams*, due to the high frequency of appearance which indicate trends and events.

E. Range of Day Used to Identify Stop Words

In our work, we chose the period of 35-55 days prior to the day of interest as a range for stop words identification. It is possible that there is a specific trend that occur on the same date of every month, so we chose the period back of more than 1-month but not reaching 2-months. The range is flexible, and it is possible that there are better range than the one we use.

F. Range of Day Used to Identify Trend Keywords

Some *Trend Keywords* appear for only a short duration of time, so we chose to consider each *Global Keywords* in the range of 1-10 days prior to the day of interest. If the *Global Keywords* appear in that range, it can be considered as *Trend Keywords*. Again, this range is flexible.

G. Number of Sample Pages

The number of chosen pages is very important because we must consider both the amount of post per page and the diversity of pages.

Concerning the amount of pages and posts per page, if we have a higher volume of sample, the keywords are easier to retrieve but the process of n-grams splitting, and counting will take a long time.

Moreover, the sampled pages must be diverse enough to make sure that the retrieved *Trend Keyword* is not keyword specific to a category of page.

In this work, we selected 10 sample pages from different categories which are news, brand, business, and organization.

V. Baseline Methodology

In social media platform such as Facebook, hashtag is a way to represent the subject of each post. Therefore, we chose hashtag as the baseline for keyword finding method

A. Finding Hashtags from Each Facebook Page

On the day of interest, hashtags of each post of the page are retrieved. These extracted hashtags are the equivalent of page *Local Keywords* of our algorithm.

B. Finding Global Keyword from Hashtags

The *Local Keywords* retrieved from hashtags can be considered as *Global Keywords* if they appeared on 2 or more pages.

C. Finding Trend Keyword from Global Keyword

We use same methodology, the *Global Keywords* that same appeared in range of 1-10 days ago are *Trend Keywords*.

VI. Result Discussions

In the experiment, we sampled 10 Facebooks pages, namely Checkbait, Brandbuffet, Investerest, Kapookdotcom, Longtunman, Pantipdotcom, Thairath, Thematterco, Themomentumco, and Underbedstar. Posts are collected from 2018/01/01 to 2018/07/07. We focus on the event and keyword trend of Thailand Cave Rescue event from 2018/06/28 to 2018/07/02.

We use human to classify the obtained result into different categories of true *Trend Keywords*, true stop words, and ambiguous words which cannot be clearly classified.

A. Stop Words

From our algorithm, we built a stop words list using posts

from 2018/06/01 to 2018/07/02. The list contains 167 words as shown in Fig. 7.

ไม่ใช่	ายที่	เกี่ยว	กับการ	งหลัง	เกาหลีเหนือ	ความเ	เรียน
ต้องการ	ความร	มากที่สุด	่าที่	ปัญหา	กำลัง	็เห็น	ระดับ
ในโลก	น่าจะ	อย่าง	ภายใน	มีความส	นนั้น	ไม่มี	เพราะ
รายได้	เหล่า	ประเทศ	ครั้งเพื่อ	เรียก	วันนี้	ความเป็น	พร้อม
บทความ	ติการ	็ความห	เตอร์	ประกาศ	ที่มี	แบบนี้	ท่าให้เ
นเรื่อง	เริ่ม	ประจำ	ติดตาม	เพิ่มเ	เจ้าของ	บริษัท	นั้น
เพื่อ	ประเทศไทย	าแล้ว	นอยู่	ท่างาน	ท่าให้	ดูเพิ่มเติม	ม่ได้
เป็นค	ปัจจุบัน	ลงทุน	เลือก	นเป็น	านที่	เข้าใจ	ตัวเอง
ต่างๆ	็ความเ	ศาสตร์	เกาหลี	หลายคน	สร้าง	กลุ่ม	เหนือ
เป็นเ	ไม่ได้	ยังไม่	กลายเป็น	ก็เป็น	นแล้ว	ันที่	อาจจะ
เกิดขึ้น	ันนี้	าเป็น	าเป็นเ	คนที่	จะเป็น	ความส	เรื่อง
องที่	่ต้อง	เพิ่มเพื่อ	ทุกคน	เป็นเรื่องที่	นอย่าง	็เป็น	เพิ่ม
อกจาก	ตอนนี้	งเป็นค	เหล่าน	นเรือ	ด้วยเ	เรื่องการ	ที่เรา
ร็อง	นอย่างไร	จากการ	ะเป็น	ที่มา	งความ	ันการ	ผู้นำ
ตั้งแ	สามารถ	เมื่อ	ได้รับ	นความ	ลงทุนศาสตร์	เป็นเรื่อง	เรื่องนี้
เท่านั้น	ยเป็น	ที่จะ	เพิ่มเติม	เหมือน	าเป็นเรื่อง	ที่เร	อย่างไร
ชีวิต	ธุรกิจ	วันที่	ตั้งแต่	็ความ	เดียว	องการ	่งที่
มีความ	สังคม	งการเ	นเข้า	เรื่อ	ที่สุด	มีการ	่านั้น
สำหรับ	ว่าจะ	ให้เรา	ต้องเ	เปลี่ยน	ล่าสุด	การณ์	ออกมา
วงการ	ในการ	เหล่านี้	เสียง	ข้อมูล	หนึ่ง	เพื่อน	ที่เพ
เขื่อ	เมือง	ครั้ง	งเป็น	งความเ	พิมพ์	็ความส	

Fig. 7. Result of stop words list. *Trend Keywords* that were misidentified as stop words are highlighted in red.

The result shown that the list has a precision of 97.6% and left 2.4% as false discovery rate such as "ประเทศไทย" (Thailand) and "เกาหลีเหนือ" (South Korea) which should be considered as *Trend Keywords*.

In addition, we gathered more stop words to improve our algorithm precision. The stop words list was gathered from 2018/05/01 to 2018/07/02 which extend to 229 words. As shown in Fig. 8, misidentified words remain the same.

ตัวเอง	ปัจจุบัน	ทุกคน	ยเป็น	นนี้เ	ภายใน	วันที่	ลูกค้า
านั้น	เหตุการณ์	เพิ่ม	ประจำ	เข้าไป	ation	ันนี้	เรื่อง
็ดตาม	ที่มา	เพิ่มเ	าเป็นเ	จากการ	จำนวน	ข้อมูล	วิทยา
นความ	งความ	ออกมา	เพื่อ	ต้องการ	ความเป็น	่าที่	นอย่าง
เป็นเรื่องที่	เกาหลี	อกจาก	็ความน	เลือก	เจ้าของ	เมื่อเ	ไม่ได้
สร้าง	ที่จะ	ายที่	บริษัท	กว่าเ	นเรื่อง	ในโลก	แล้ว
จริงๆ	หนังสือลงทุน	หรือก	เหล่าน	ในการ	เพื่อน	เพิ่มเติม	็ความห
เกี่ยว	มีความส	ากที่	เกิดขึ้น	ประกาศ	พวกเขา	พร้อม	เปลี่ยน
ับการ	ท่าให้	ใหม่ๆพร้อม	มากที่สุด	ก็เป็น	ที่เราได้	ตั้งแต่	นความเ
ต่างๆ	เริ่ม	เรื่องการ	ตอนนี้	็ะได้	คนที่	เท่านั้น	ท่าให้เ
กำลัง	นนั้น	วิทยาศาสตร์	เป็นเรื่อง	ความเ	ประเทศ	ศาสตร์	่ต้อง
มีการ	ประเทศไทย	กลายเป็น	เรื่องนี้	เที่ยว	อความ	น่าจะ	เหล่านี้
็เห็น	นเรือ	ะเป็น	ในประเทศ	ไม่ใช่	ไม่มี	นเป็น	ให้เรา
านที่	ไม่ต้อง	นหนึ่งใน	หลายค	นเข้า	อย่างไร	งหลัง	อย่าง
ความน	จะเป็น	เหนือ	เสียง	เป็นแ	โฆษณา	็ะใน	งการเ
งความเ	หนังสือ	เดียวกั	เหมือน	การณ์	น่าสนใจ	...อ่านต่อ	็ความเ
็ความส	ชีวิต	อย่างที่เ	ที่เพ	เมื่อความเ	การลงทุน	่งที่	เดียว
ดูเพิ่มเติม	ล่าสุด	สังคม	โครงการ	าแล้ว	ม่ได้	นหนึ่ง	เป็นต
เขื่อ	เพิ่มเพื่อ	ระหว่าง	ด้วยก	นประเทศ	าเป็น	ที่เรา	ติดตาม
ธุรกิจ	ท่างาน	เรียน	่านี้	ที่เร	สิ่งที่	สามารถ	งเป็น
วงการ	ระดับ	องที่	รายได้	นอยู่	สำหรับ	ความร	งเป็นค
ครั้งเพื่อ	หลังจาก	มีความ	เรียก	วันนี้	ปัญหา	ด้วยเ	เข้าใจ
แบบนี้	นอย่างไร	ลงทุน	เมือง	หนึ่ง	ลงทุนศาสตร์	ต้องเ	ความรู้
อาจจะ	หลายคน	องการ	ได้รับ	ติการ	ว่าจะ	็เป็นอย่าง	าเป็นเรื่อง
เมื่อ	เพราะ	ผู้นำ	เป็นค	ครั้ง	ตั้งแ	สร้างความเ	งที่เ
เพียง	่านั้น	ที่มี	ันที่	กับการ	เหลือ	ยังไม่	เรื่อ
็เป็นเพื่อ	ความส	มนุษย	บทความ	็เป็น	ที่สุด	ความห	เตอร์
พูดถึง	สร้างความ	นแล้ว	กลุ่ม	เกาหลีเหนือ	เหล่า	เป็นเ	็ความ
ได้ที่	ประเทศไ	นที่เ	พิมพ์	ร็อง			

Fig. 8. Result of 2-months generated stop words list.

B. Trend Keywords

Fig. 9 shows *Trend Keywords* from baseline method, Fig 10 shows *Trend Keywords* from our algorithm classified by generated stop words from 1-month data, and Fig 11 shows *Trend Keywords* from our algorithm classified by generated stop words from 2-months data. True *Trend Keywords* are highlighted in red, true stop words are highlighted in blue, and the words in black are ambiguous words.

Date	Trend Keywords				
2018-06-28	ถ้ำหลวง	พาทีมหมูป่ากลับบ้าน			
2018-06-29	ถ้ำหลวง				
2018-06-30	ถ้ำหลวง	ทีมหมูป่า			
2018-07-01	ถ้ำหลวง				
2018-07-02	ถ้ำหลวง	13ชีวิตต้องรอด	พาทีมหมูป่ากลับบ้าน	ทีมหมูป่า	

Fig. 9. The result of baseline algorithm.

Date	Trend Keywords						
2018-06-28	องประ	หลายๆ	ในถ้ำหลวง	มิถุนายน	าที่	13ชีวิตถ้ำหลวง	นที่	ต้องร
	ถ้ำหลวง	13ชีวิต	เจ้าหน้าที่	หน้าที่	หน่วย	หมูป่า	วันที่2	
2018-06-29	เข้าไป	เรื่องความ	าที่	นที่เ	ทีมหมูป่ากลับบ้าน		นางนอน	หน้าที่
	ได้ที่	หมูป่า	พิเศษ	กลายเป็นค	เป็น	หนึ่งใน	โอกาส	ประเด็น
	ลงทุนศาสตร์วันนี้		ถ้ำหลวง	ทรัพย์	การใช้	ประมาณ	นต่างประเทศ	งเพื่อน
	ที่เห	13ชีวิตติดถ้ำหลวง		าอยาก	นส่วน	จริงๆ	อื่นๆ	ามเป็น
	ศึกษา	มิถุนายน	ติดถ้ำหลวง	กลับบ้าน	ันที่	นใหญ่	สำรวจ	่แล้ว
	ได้อย	นหนึ่ง	นล้าน	นการเ	นถ้ำหลวง	เป็นอ	านการ	เป็นเ
	ในถ้ำหลวง	ทั้งห	มากกว่า	หนังสือ	ทีมหมูป่า	่เป็น	จนถึง	แล้วท
	เจ้าหน้าที่	ราคาท	เป็นส	นตอนนี้	ชีวิตค	พวกเขา	ระหว่าง	13ชีวิต
2018-06-30	หมูป่า	เรื่องหนึ่ง						
2018-07-01	ารที่	ุ่วิว่า	ชีวิตอย่าง	เป็นข	บข่าว	ทีมหมูป่ากลับบ้าน		ภารกิจ
	หน้าที่	หมูป่า	เชียงราย	ถ้ำหลวง	13ชีวิตต้องร	เรื่องการค	แต่ก็	กลับบ้าน
	ต้องร	ด้วยก	นการเ	ในถ้ำ	นถ้ำหลวง	13ชีวิตค	านการ	ในถ้ำหลวง
	หนังสือ	ทีมหมูป่า	ค้นหา	ชีวิตค	ทีมหมูป่ากลับบ้านวันนี้		13ชีวิต	
2018-07-02	ุ่วิว่า	่าที่	หน้าที่	หมูป่า	ข่าวดีถ้ำหลวง	ปลอดภัย	เชียงราย	ถ้ำหลวง
	นแล้วท	ต้องก	ในถ้ำ	นถ้ำหลวง	ในถ้ำหลวง	ทั้ง13ชีวิตในถ้ำหลวง		พยาบาล
	ทีมหมูป่า	ข่าวดี	เจ้าหน้าที่	เตรียม	13ชีวิตถ้ำหลวงเชียงราย		13ชีวิต	

Fig. 10. The result of our algorithm classified by 1-month generated stop words list.

Date	Trend Keywords						
2018-06-28	องประ	หลายๆ	ในถ้ำหลวง	มิถุนายน	าที่	13ชีวิตถ้ำหลวง	นที่	ต้องร
	ถ้ำหลวง	13ชีวิต	เจ้าหน้าที่	หน้าที่	หน่วย	หมูป่า	วันที่2	
2018-06-29		เรื่องความ	าที่		ทีมหมูป่ากลับบ้าน		นางนอน	หน้าที่
		หมูป่า	พิเศษ	กลายเป็นค		หนึ่งใน	โอกาส	ประเด็น
	ลงทุนศาสตร์วันนี้		ถ้ำหลวง	ทรัพย์	การใช้	ประมาณ	นต่างประเทศ	งเพื่อน
	ที่เห	13ชีวิตติดถ้ำหลวง		าอยาก	นส่วน		อื่นๆ	ามเป็น
	ศึกษา	มิถุนายน	ติดถ้ำหลวง	กลับบ้าน	ันที่	นใหญ่	สำรวจ	่แล้ว
	ได้อย		นล้าน	นการเ	นถ้ำหลวง	เป็นอ	านการ	เป็นเ
	ในถ้ำหลวง	ทั้งห	มากกว่า		ทีมหมูป่า	่เป็น	จนถึง	แล้วท
	เจ้าหน้าที่	ราคาท	เป็นส	นตอนนี้	ชีวิตค			13ชีวิต
2018-06-30	หมูป่า	เรื่องหนึ่ง						
2018-07-01	ารที่	ุ่วิว่า	ชีวิตอย่าง	เป็นข	บข่าว	ทีมหมูป่ากลับบ้าน		ภารกิจ
	หน้าที่	หมูป่า	เชียงราย	ถ้ำหลวง	13ชีวิตต้องร	เรื่องการค	แต่ก็	กลับบ้าน
	ต้องร		นการเ	ในถ้ำ	นถ้ำหลวง	13ชีวิตค	านการ	ในถ้ำหลวง
		ทีมหมูป่า	ค้นหา	ชีวิตค	ทีมหมูป่ากลับบ้านวันนี้		13ชีวิต	
2018-07-02	ุ่วิว่า	่าที่	หน้าที่	หมูป่า	ข่าวดีถ้ำหลวง	ปลอดภัย	เชียงราย	ถ้ำหลวง
	นแล้วท	ต้องก	ในถ้ำ	นถ้ำหลวง	ในถ้ำหลวง	ทั้ง13ชีวิตในถ้ำหลวง		พยาบาล
	ทีมหมูป่า	ข่าวดี	เจ้าหน้าที่	เตรียม	13ชีวิตถ้ำหลวงเชียงราย		13ชีวิต	

Fig. 11. The result of our algorithm classified by 2-months generated stop words list.

The precision of baseline is 100%, which mean that all keywords extracted using baseline method are *Trend Keywords*. However, a lot of keywords are missed and cannot be found by this method.

Our algorithm can extract a larger list of keywords, but it still contains stop words and some of the keywords can also be considered as just *Local Keywords*. As the result, our algorithm gave a precision of 40% with 1-month data. Consecutively, the precision is shown up to 44% with 2-months data, and the result can be more accurate depends on size of stop words list.

VII. Limitation and Future Work

A. Cleaning Noise from Character n-Grams

Algorithm suggested in this work, sometimes, produce keyword character with noise. The example of noise is given in Fig. 12. Some keywords obtained are not a complete word or contain extra character.

Message	Keyword with noise
ผมชอบบอล	
เตะขอบบอล	อบบอล
เขาปลอบบอล	

Word Tokenized Message			Keyword
ผม	ชอบ	บอล	
เตะ	ขอบ	บอล	บอล
เขา	ปลอบ	บอล	

Fig. 12. Example of noise from Character n-Grams.

The issue may be fixed with spell-correction algorithm of Thai language.

B. Getting Keyword Grams

In our work, we used K-Means Clustering as the main method to identity *Keyword Grams*, however there is still room for improvement.

The difficult point is that the frequency of keyword appearing is vary in each page, so the conventional thresholding method to determine keyword is not suitable. New solution is needed.

C. Validation Method

The problem of validation method in this work is the insufficiency of robustness, since we use human validation to classify the output of the algorithm because we cannot find a standard testing set to evaluate our algorithm.

VIII. Conclusion

In this work, we propose an algorithm which use Character n-Grams to extract keywords and stop words without using word tokenization or other training data set and corpus. However, the limitation of the proposed algorithm is the noise which is mistakenly identified as keyword. Moreover, the precision of the result depends heavily on generated stop words list, which mean if we can improve the stop words list, the precision of the algorithm can also be improved.

Acknowledgment

Nattapong Ousirimaneechai, author, would like to thank Tanawat Chansophonkul, Kidthiphutn Tejakumput for reviewing and editing the content of this paper, and Napat Simsomboonphol for support.

This research is supported by the 90th Anniversary of Chulalongkorn University, Rachadapisek Sompote Fund.

References

[1] NECTEC. (December 2009). BEST2010. *NECTEC*. [Online] Available: http://thailang.nectec.or.th/downloadcenter/

[2] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," *Ann Arbor Mi*, vol. 48113, no. 2, pp. 161-175, 1994.

[3] A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, Stanford University, California, Cambridge University Press, 2011, ch. 1, pp. 1–17.

[4] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An Overview of Graph-Based Keyword Extraction Methods and Approaches," *Journal of Information and Organizational Sciences*, vol. 39, no. 1, pp. 1–20, 2015.

[5] P. Daowadung and Y. H. Chen, "Stop word in readability assessment of Thai text," in *Proc. 2012 IEEE 12th International Conference on Advanced Learning Technologies*, July 2012, pp. 497-499.

[6] R. T. W. Lo, B. He, and I. Ounis, "Automatically building a stopword list for an information retrieval system," *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, vol. 5, pp. 17-24, 2005.

[7] A. Piyatumrong, C. Sangkeettrakarn, C. Haruechaiyasak, and A. Kongthon, "Finding Key Terms Representing Events from Thai Twitter," in *Proc. International Symposium on Natural Language Processing*, February 2016, pp. 73-87.

[8] C. Haruechaiyasak, S. Kongyoung, and C. Damrongrat, "LearnLexTo: a machine-learning based word segmentation for indexing Thai texts," in *Proc. the 2nd ACM workshop on Improving non English Web Searching*, 2008, pp. 85-88.

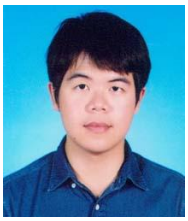
[9] C. Haruechaiyasak, and A. Kongthon, "LexToPlus: a Thai lexeme tokenization and normalization tool," in *Proc. the 4th Workshop on South and Southeast Asian Natural Language Processing*, 2013, pp. 9–16.

[10] C. Haruechaiyasak, P. Srichaivattana, S. Kongyoung, and C. Damrongrat, *Automatic Thai Keyword Extraction from Categorized Text Corpus*, 2008.

[11] J. A. Hartigan, and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[12] T. M. Kodinariya, and P. R. Makwana, "Review on determining number of cluster in K-Means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90-95, 2013.

[13] A. Perera. (October 2017). Finding the optimal number of clusters for K-Means through Elbow method using a mathematical approach compared to graphical approach. *Linkedin*. [Online]. Available: https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera/



Nattapong Ousirimaneechai was born in 1995 in Bangkok, Thailand. He received his bachelor's degree in computer engineering from Chulalongkorn University in 2017. Now he is a master student in Chulalongkorn University. His research interests are information retrieval, natural language processing, machine learning, social network analysis and social network mining

Sukree Sinthupinyo was born in 1975 in Bangkok, Thailand. He received his bachelor's, master's, and doctoral degree from the Department of Computer Engineering, Chulalongkorn University. Now he is working as a lecturer at the same department. His research interests are artificial intelligence, machine learning, innovation, social network analysis and social network mining.