Extraction of Trend Keywords from Thai Twitters using N-Gram Word Combination

Tanatorn Tanantong
Department of Computer Science,
Faculty of Science and Technology,
Thammasat University
Pathum Thani, Thailand
tanatorn@sci.tu.ac.th

Sasitorn Kreangkriwanich
Department of Computer Science,
Faculty of Science and Technology,
Thammasat University
Pathum Thani, Thailand
maysst7@gmail.com

Nasith Laosen
Department of Digital Technology,
Faculty of Science and Technology,
Phuket Rajabhat University
Phuket, Thailand
nasith.l@pkru.ac.th

***Abstract*— Extracting keywords from text on social media facilitates people to update news and trends. It reduces time spent for identifying main content from huge amount of data, and it can be used to identify situations or events that most of people mention in each period of time. This paper proposes a method for extracting keywords from Thai text on social media. A N-gram-based word-combination technique is presented to segment words that are not in dictionaries and increase the precision of word segmentation. Posts on Twitter concerning universities in Thailand are used as a case study for extracting keywords and analyzing trends. The experimental results show that the proposed method yield the highest precision of 70%.**

Keywords—Social media, twitter data, word combination, keyword extraction, Thai university

I. Introduction

Nowadays, social media plays an important role in Thailand. According to a survey of Electronic Transactions Development Agency (ETDA) Thailand [1] about the behavior of internet users in 2018, the average daily internet usage amount of Thai people is increasing. Popular social media includes YouTube, Line, Facebook, and Twitter. Twitter has revealed that it has the fastest growth rate during the past 2 years in Southeast Asia. The majority of its users (40%) are between 15-24 years old [2]. The usage of Twitter is to exchange daily-life information, which can be observed through hashtags [3]. For example, the hashtags #PM2.5, #JS100, and #TCAS are used for exchange information about air pollution, traffic routes, and education news, respectively.

Keyword extraction is a method for identifying the main content of a website. In website marketing, keywords determine prominence of a website and the chance that people will see it [4]. Keywords are also important for research articles. They describe the main content of research articles and are used for searching research article [5]. However, extracting keywords from Thai language text is challenging, especially extraction from social media websites. The reason being that Thai language writing style does not use spaces to separate words, and new words/slangs are constantly being created. These problems cause difficulties to Thai-word segmentation tools and lead to inaccurate results.

Araya Pudtal [6] conducted a research on Twitter analysis during the funeral cremation of King Bhumibol Adulyadej of Thailand. This related work introduced a method for extracting and searching events from Twitter as follows: (i) retrieve tweets that contain the words "King" and "Bhumibol", (ii) select the obtained tweets that were posted during October 1st, 2016 to December 31st, 2016, (iii) select the top ten of the most occurring hashtags, (iv) by using the search results obtained from the previous step as input, repeat steps one to three until the obtained hashtags are not changed. Then the events were analyzed from the most occurring hashtags obtained during each period of time. Moreover, the words occurring during the events were also analyzed, i.e., the words appearing in tweets that had the greatest number of retweets were analyzed and were linked to the actual events.

Natthapong Ousirimaneechai and Sukree Sinthupinyo [7] proposed a method for extracting keywords from Facebook pages using custom-made stop words and the Character n-Grams technique, which is a technique that does not rely on word segmentation tools. The most occurring words during 35-55 days before the date under consideration were identified and analyzed. Accuracy of 40% was reported for 1-month data, and 44% was reported for 2-month data. It was expected that if the number of the stop words is increased, the result will be more accurate.

Thassanee Uthaisuri [8] proposed methods for extracting keywords from English abstracts of articles using machine learning techniques, i.e., the decision tree and naïve Bayes techniques. An input abstract was preprocessed as follows: (i) segment words in the input abstract, (ii) identify part of speeches of the words, and (iii) remove stop words. Only noun words are selected and considered. This related work also proposed important features for determination of keywords, i.e., (i) the appearances of words in the title, (ii) the lengths of words, (iii) the term frequency - inverse document frequency (TF-IDF) of words, the table-term frequency (TTF) of words, and the positions of words in the abstract. The result showed that the naïve Bayes technique gave higher accuracy than the decision tree technique and the most important feature was the lengths of words.

Ekkaphum Phumiphan [9] proposed a method for extracting keywords from text in electronics word of mouth (E-WOM). A given input text was divided into several threads. A method, called the modified for thread-TFIDF (MT-TFIDF), was proposed for calculating weights of words. The experimental result showed that MT-TFIDF are better than TFIDF at 90% of confidence level.

Based on the related works mentioned above, we found that extraction of keywords from text on Twitter has some limitations. First, tweets to be analyzed are only tweets that have hashtags, and tweets without hashtags are not considered. Second, words in tweets are segmented and then TF-IDF is used for identifying keywords. However, one unit of the segmentation results equals to one word (or one incomplete word), which sometimes may not be informative enough to describe the main content and trends.

978-1-7281-6486-1/20/$31.00 ©2020 IEEE

In this paper, we propose a method for extracting keywords from Thai language text on social media. The proposed method tries to extract keywords that are compound keywords, i.e., the number of words in a keyword is more than one. Compound keywords are informative and result in higher precision of describing main content and trends on social media. An interesting word, for instance, is "งานฟุตบอลประเพณี" (the traditional football festival between Thammasat and Chulalongkorn universities). Ordinary Thai-word segmentation tools do not output this word because it is not in standard Thai language dictionaries. Manual segmentation is also difficult since main content and trends on social media are constantly changed. This paper collects posts concerning universities in Thailand from Twitter and extracts keywords from them. Twitter is considered by this work because the most of its users are students and the number of the users is increasing.

II. Methodology

An overview of an automatic trend-keyword extraction framework is demonstrated in Fig. 1. The components of the framework are detailed below.

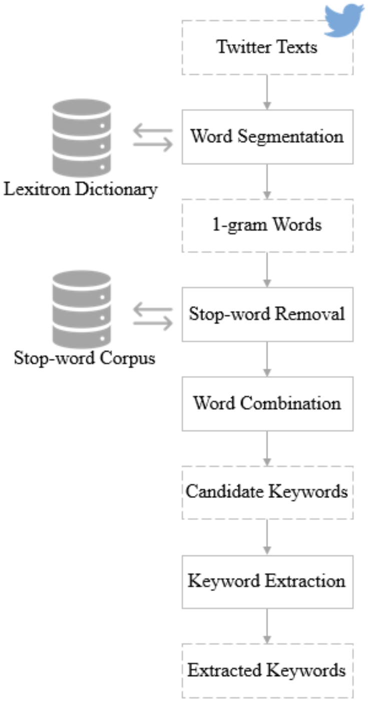

Fig. 1. Components of the proposed framework

A. Data Collection

In this research, we used Twitter API [10] to collect data related to 40 universities in Thailand, based on the ranking information from the Webometrics Ranking of World Universities website. To acquire Twitter data, we employed a university's name and its name alias [11] for being keywords in search of information about Thai universities as shown in Table I. Some Thai universities may have several name aliases. For example in Table I, a name alias for Thammasat university is "ลูกแม่โดม"(/look/mae/dome/), which refers to students that study at the university having a dome building for its signature and a name alias for Chulalongkorn university is "มหาลัยสีชมพู", (ma/haa/lai/see/chom/poo), which refers to the university that has a pink color for a symbolic color.

For experiments in this study, Twitter data was gathered for 1 week (during the 6th to 12th of February 2019).

TABLE I. A University's Name and Its Alias Name

University's Name	Name Aliases
Thammasat	"ลูกแม่โดม", "พ่อปรีดี", "#ทีมTU"
Chulalongkorn	"มหาลัยสีชมพู", "ลูกพระเกี้ยว", "#ทีมCU"
Mahidol	"ลูกพระบิดา", "มหาลัยสีเขียว", "ทีมMU"

B. Word Segmentation

For word segmentation, we used a maximal matching method [12], which is mainly based on Thai corpus (LEXiTron) developed by NECTEC [13]. The Twitter data about Thai universities was separated into the least number of words from each sentence as shown in Fig. 2. For example, "คณะ" (faculty) | "สิ่งแวดล้อม" (environment) | "และ" (and) | "ทรัพยากร" (resource) | "ศาสตร์" (study) | _ (space) | "มหาวิทยาลัยมหิดล" (Mahidol university).

คณะ | สิ่งแวดล้อม | และ | ทรัพยากร | ศาสตร์ | _ | มหาวิทยาลัยมหิดล | _ |
รายงาน | ดัชนี | คุณภาพ | อากาศ | _ | (| Air | _ | Quality | _ | Index | _ | : | _ |
AQI |) | _ |ณ | _ | จุด | ตรวจวัด | คุณภาพ | อากาศ |_| มหาวิทยาลัยมหิดล | _ |
วันที่ | _ | 7 | _ | กุมภาพันธ์ | _ |พ.ศ .| _ |2562 | _ | เวลา | _| 07 | : | 00 | _ | น. |
_ | อยู่ | ใน | ระดับ | _ | คุณภาพ | อากาศ | ดีมาก

Fig. 2. Example results of word segmentation

C. Stop-Word Removal

Stop-word removal is an important method for removing words that are irrelevant or unimportant to natural language processing tasks. Stop words are generally common words, which are used in daily life such as "เรา" (we) "ฉัน" (I) "มี" (have) and "ค่ะ" (sir). However, using only the Thai stop-word corpus, which is currently published, is not enough to remove stop words since it does not support new words occurred from social media.

For removing stop words, we first use Regular Expressions for eliminating numeric words and single characters, e.g., "ก" "k" "_" and ";". Then, we extracted stop words from segmented words from the obtained Twitter data by taking into account a number of days of segmented-word occurrence [7]. If the segmented words appear continuously for 20 days above, these words are defined as our stop words. There are four main groups of stop words in this study, i.e., three groups of stop words extracted from 1-month Twitter data from 3 months ago, 2 months ago, and 1 month ago, respectively, and a group of stop words extracted from all Twitter data during 3 months ago. Fig. 3 demonstrates remaining words from segmented words (cf. Fig. 2) after removing stop words.

สิ่งแวดล้อม | ทรัพยากร | ศาสตร์ | รายงาน | ดัชนี | อากาศ | Air | Quality |
Index | AQI | จุด | ตรวจวัด | อากาศ | กุมภาพันธ์ | พ.ศ. | ระดับ | อากาศ

Fig. 3. Example results of a stop-word removal process (comparing to the original sentence in Fig. 2)

D. Word Combination

For combining words, the obtained preprocessed words are compared with an original Twitter text. If the obtained words are located in an adjacent position, these words are considered to combine into one word and called as a candidate keyword. Fig. 4 shows a process of our proposed word combination. Examples of candidate keywords are depicted in Fig. 5.

Let W_{ij} is a position of words in each Twitter text

i is a sequence of Twitter texts

j is a sequence of words in each Twitter text

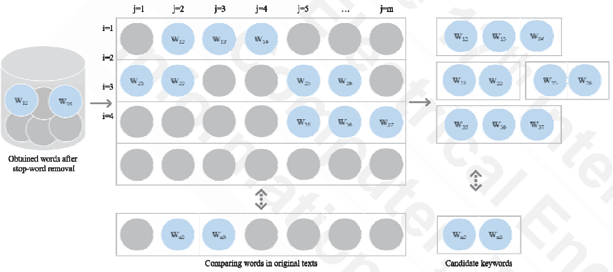

Fig. 4. Steps for combining words after stop-word removal

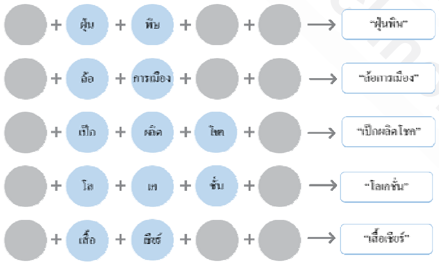

Fig. 5. Example results of word combination

E. Keyword Extraction

By taking into account candidate keywords, we extracted keywords by considering a frequency of word occurrence in each Twitter text. In this study, trend keywords are defined as extracted keywords that occur consecutively on Twitter for 3 days. And hit keywords are defined as extracted keywords that have the highest rate of occurrence on Twitter in each day.

III. Experimental Results

A. Results of Keyword Extraction

To evaluate our proposed method, four experiments are conducted as follows: (i) extracting keywords with word segmentation based on stop words extracted from 30-day Twitter data from last 3 months, (ii) extracting keywords with word segmentation based on stop words extracted from 30-day Twitter data from last 2 months, (iii) extracting keywords with word segmentation based on stop words extracted from 30-day Twitter data from 1 month ago, and (iv) extracting keywords with word segmentation based on stop words extracted from all Twitter data during 3 months ago.

For measuring performance of keyword extraction, obtained keywords were compared to annotated keywords, which were labeled by human assessors. If the obtained keywords have a similarity score greater than or equal 75% of the annotated keywords, they are determined as correct. In this study, a precision is used for a main measurement as follows:

$$Precision\ (\%) = \frac{TP}{TP + FP} \times 100$$

where *TP* (True Positive) and is the number of keywords accurately extracted and *FP* (False Positive) is the number of keywords inaccurately extracted.

Fig. 6 shows overall results of our proposed method. The obtained result of the first three experiments (Exp. #1, Exp. #2, and Exp. #3) are the results from extracting keywords using stop words from 1-month Twitter data from 3 months ago, 2 months ago, and 1 month ago, respectively. The obtained result of the fourth experiment (Exp. #4) is the result for keyword extraction, which are employed stop words extracted from Twitter data during 3 months ago. Table II demonstrated examples of keywords extracted from our proposed methods.

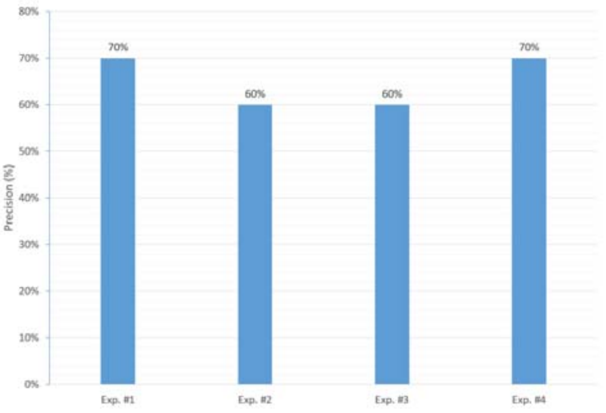

Fig. 6 Performance of the proposed method for extracting keywords

TABLE II. Examples of Keywords Extracted from Our Proposed Method

Exp. #1	Exp. #2	Exp. #3	Exp. #4
มหาวิ (university)	มหาวิ (university)	มหาวิ (university)	มหาวิ (university)
พันธ์ (February)	พันธ์ (February)	พันธ์ (February)	พันธ์ (February)
กุมภา (February)	กุมภา (February)	กุมภา (February)	กุมภา (February)
ฟุตบอลประเพณี (football festival)	บอลประเพณี (football festival)	ฟุตบอลประเพณี (football festival)	ฟุตบอลประเพณี (football festival)
เทคโน (Technology)	ฟุตบอลประเพณี (football festival)	จารย์ (teacher)	เป๊กผลิตโชค (Thai actor's name)

The 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2020)

B. *Results of Extracting Trend and Hit Keywords*

From the keywords extracted from the previous step, these obtained keywords are classified into two main groups, i.e., trend keywords and hit keywords (cf. Section II, E), as shown results in Tables III and IV. However, the obtained results from these two tables are not completed words due to limitations of the word segmentation process and a quality of writing contents on Twitter.

TABLE III. EXAMPLES OF TREND KEYWORDS

NO.	Trend Keyword	Meaning / Reffering
1	มหาวิ	university
2	พันธ์	February
3	กุมภา	February
4	ฟุตบอลประเพณี	football festival
5	เทคโน	technology

TABLE VI. EXAMPLES OF HIT KEYWORDS IN EACH DAY

Day	Hit Keywords
Day 1	มหาวิ, พันธ์, กุมภา (university, February, February)
Day 2	มหาวิ, พันธ์, กุมภา (university, February, February)
Day 3	มหาวิ, ฟุตบอลประเพณี, พันธ์ (university, football festival, February)
Day 4	ฟุตบอลประเพณี, มหาวิ, ศกเพชร. (football festival, university, Sipakorn university: Phetchaburi campus)
Day 5	มหาวิ, กุมภา, พันธ์ (university, February, February)
Day 6	มหาวิ, พันธ์, ประสานมิตร (university, February, Srinakharinwirot university)
Day 7	มหาวิ, พันธ์, กุมภา (university, February, February)

IV. CONCLUSION

We have presented an automatic method using N-gram word combination for extracting keywords from Twitter data. The experimental results show that our proposed method yielded the highest precision of 70%. Based on Twitter data about Thai universities, this method can use for extracting keywords that represent important events and activities in each time period. For example, keywords "บอลประเพณี" and "ฟุตบอลประเพณี" that can represent an occurrence of an important university event in February 2019, which is the 73rd Thammasat-Chula Traditional Football Match. In additional, our proposed word combination can partial support to solve a problem of extracting incomplete words from social media. However, further investigation is required to obtain more accurate results.

ACKNOWLEDGMENT

This work was supported by the National Innovation Agency (Public Organization), Grant No. P12-ST-61-05-043 and by the government budget of Thammasat University, Grant No. 7/2562.

REFERENCES

[1] Electronic Transactions Development Agency Thailand, *The Behavior of Thai Internet Users in 2018*, Accessed on: Mar, 30, 2020. [Online]. Available: https://www.etda.or.th/content/etda-reveals-thailand-internet-user-profile-2018.html

[2] Forbes Thailand, *Thai Twitter has the Fastest Growth in ASEAN*, Accessed on: Mar, 30, 2020. [Online]. Available: https://forbesthailand.com/news/it/twitter-ไทย-การเติบโต.html

[3] T. Tanantong and S. Thaskhwan, "Sentiment Classification on Thai University Mentions in Twitters," Proceedings of the 11th National Conference on Information Technology, pp 79-84, NCIT, 2019.

[4] OURGREENFISH, *How Keywords are Important for Digital Marketing*, Accessed on: Mar, 30, 2020. [Online]. Available: https://blog.ourgreenfish.com/th/keyword-มีความสำคัญอย่างไรต่อ/

[5] P. Hanpanich, *Academic Writing: How Keywords are Important?*, Accessed on: Mar, 30, 2020. [Online]. Available: https://www.gotoknow.org/ posts/593596

[6] A. Pudtal, "An analysis of Twitter in the passing of His Majesty KingBhumibol Adulyadej," M.S. Thesis, Dep. of Comput. Sci., Chulalongkorn Uni., Bangkok, 2017. [Online]. Available: http://cuir.car.chula.ac.th/bitstream/123456789/59629/1/5870988721.pdf

[7] N. Ousirimaneechai and S. Sinthupinyo, "Extraction of trend keywords and stop words from Thai Facebook pages using character n-Grams", *Int. J. of Machine Learn. and Comput.*, vol. 8, no. 6, pp. 589-594, Dec. 2018.

[8] T. Uthaisuri, "Keyword extraction from English abstracts," M.S. Thesis, Dep. of Comput., Silpakorn Uni., Bangkok, 2013. [Online]. Available: http://www.thapra.lib.su.ac.th/thesis/showthesis_th.asp?id=0000009680

[9] E. Phumiphan, "Extracting keywords from electronics word of mouth in webboard community," M.S. Thesis, Dep. of Comput. Sci., Chulalongkorn Uni., Bangkok, 2011. [Online]. Available: http://cuir.car.chula.ac.th/bitstream/123456789/22024/1/ekkaphumph.pdf

[10] Twitter Inc., *Filter Realtime Tweets*, Accessed on: Mar, 30, 2020. [Online]. Available: https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

[11] T. Tanantong, S. Thaskhwan and S. Kreangkriwanich, "A social media monitoring system for Thai university," *TNI J. of Eng. and Technol.*, vol. 7, no. 2, pp. 1-17, 2019.

[12] W. Phatthiyaphaibun, *User Manual PyThaiNLP 1.4*, Accessed on: Mar, 30, 2020. [Online]. Available: https://pythainlp.readthedocs.io/en/stable/pythainlp-1-4-thai/

[13] Language and Semantic Technology Laboratory, *Thai-English Electronic Dictionary LEXiTRON*, Thailand's National Electronics and Computer Technology Center (NECTEC), Accessed on: Mar, 30, 2020. [Online]. Available: https://lexitron.nectec.or.th/

The 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2020)