Solving the Difficult Problem of Topic Extraction in Thai Tweets

Rungsiman Nararatwong[1, 2], Roberto Legaspi[3], Nagul Cooharojananone[4],
Hitoshi Okada[2], Hiroshi Maruyama[3]
[1]*The Graduate University for Advanced Studies, Kanagawa, Japan.*
[2]*National Institute of Informatics, Tokyo, Japan.*
[3]*Research Organization of Information and Systems, Transdisciplinary Research Integration Center, The Institute of Statistical Mathematics, Tokyo, Japan.*
[4]*Chulalongkorn University.*
rungsiman@nii.ac.jp

***Abstract*—We tackled in this study the difficult problem of topic extraction in Thai tweets on the country's historic flood in 2011. After using Latent Dirichlet Allocation (LDA) to extract the topics, the first difficulty that faced us was the inaccuracy the word segmentation task that affected our interpretation of the LDA result. To solve this, we refined the stop word list from the LDA result by removing uninformative words caused by the word segmentation, which resulted to a more relevant and comprehensible outcome. With the improved results, we then constructed a rule-based categorization model and used it to categorize all the collected tweets on a per-week scale to observe changes in tweeting trend. Not only did the categories reveal the most relevant and compelling topics that people raised at that time, they also allowed us to understand how people perceived the situations as they unfold over time.**

***Index Terms*—LDA; Topic Extraction; Thai Tweets.**

I. INTRODUCTION

The differences in languages remains to be a challenging puzzle yet to be solved by researchers working on natural language processing. While powerful and well-established approaches work imposingly for specific languages, they need further development come other languages. One fundamental problem that we present in this study, apart from semantic and structural issues, is that some languages constitute sentences that have no word delimiters, e.g., Chinese, Japanese and Thai. In Thai, existing word segmentation algorithms perform rudimentarily well. However, with more sophisticated inputs, we observed a significant decrease in segmentation accuracy. This presents a potential problem when word segmentation outputs are to be fed as inputs to machine learning algorithms. What used to be meaningful words end up with distorted or incoherent implications when broken down into their constituent roots. This inability to comprehend the true meaning of the words only makes inevitable the decline in the performance of machine learning algorithms.

The problem above was only the first problem we encountered in our effort to categorize a collection of tweets into a definite set of topics. The tweets are on a specific domain and context, specifically, on the historic flooding in Thailand in late 2011. The prime difficulty is that although we were able to track the formation and development of user communities, it became inadequate for us to portray the whole picture of what came to be an unprecedented disaster in Thailand since we lacked understanding of what people were actually talking about. What we needed was the knowledge of the most relevant and compelling topics that were raised at that time.

To address our research problem, we sought for unsupervised machine learning approaches that could identify the topics among the collected tweets. Latent Dirichlet allocation (LDA) topic modeling [1] came to our attention chiefly because of its promising efficacy on microblog datasets. We first cleaned up the tweets, performed word segmentation, and applied LDA to our dataset. LDA basically produced the lists of words related to each topic. However, a manual refinement process was needed due to the imperfection in word segmentation. After several user-guided refinements, we were able to obtain sufficiently meaningful lists of words that helped us construct manually the rules to categorize all our tweets into major topics. The resulting categories not only helped us understand semantically what people were discussing, they also helped reveal how people perceived the unfolding situations.

The challenge in our research is therefore to address the difficulties associated with word segmentation and topic extraction specific to processing Thai texts. The end goal is to understand people's perception of situations as they change over time.

II. RELATED WORK

A. Data preprocessing

Stop words appear frequently in a text, albeit they are not at all informative. In 2006, Zou et al. [2] proposed an approach that automatically constructs a list of stop words based on statistical and information models. They defined a statistical value *SAT* as a proportion of the mean of probability (*MP*) of each word in a document and the variance of the probability (*VP*). In the following equations, w_j denotes each word in the entire corpus, $P_{i,j}$ is the proportion of word w_j in document i, and N is the number of documents in the corpus.

$$MP(w_j) = \frac{\sum_{1 \le i \le N} P_{i,j}}{N} \quad (1)$$

$$VP(w_j) = \frac{\sum_{1 \le i \le N} (P_{i,j} - \bar{P}_{i,j})^2}{N} \quad (2)$$

$$SAT(w_j) = \frac{MP(w_j)}{VP(w_j)} \quad (3)$$

The logic behind Eq. 3 is that the probability of stop words should be proportional to the mean of probability, while the variance of probability interacts in the opposite way. With this assumption, we should expect stop words to be on the top of the list when ranked by the *SAT* values in descending order.

In addition to *SAT*, Zhou et al.'s study adapted the concept of entropy in information theory. The distribution of each word over a set of documents is considered as an information channel. An entropy (*H*) is defined as:

$$H(w_j) = \sum_{1 \le i \le N} P_{i,j} \times \log(\frac{1}{P_{i,j}}) \quad (4)$$

The higher the entropy means the less random the word would be in all documents. The same goes for its information values. Therefore, similar to *SAT*, a list of entropy *H* values in descending order should help us create a list of stop words. The study explained an aggregation of both lists based on Borda's Rule [3] of binary relations. The approach is, in fact, straightforward: Create the third list containing all words and assign their weights from the combination of their ranking order in both lists. For example, if a word is ranked first in *SAT* list and second in *H* list then it would have a weight of 3. The list is then sorted in ascending order.

Daowadung and Chen [4] implemented Zou's method to create stop words from 1,188 textbook articles used by Thai students between grades 1 and 6. Their result showed that the method performed well with Thai.

B. Topic extraction

LDA is a widely used unsupervised algorithm for automatic corpus summarization. In addition to the original LDA, some studies proposed supervised [5] and semi-supervised version of the algorithm [6, 7]. However, the result of LDA is often difficult to interpret. Data visualization can streamline complexity, as well as enhance the understanding and interpretation of LDA.

In 2014, Sievert and Shirley [8] presented LDAvis, a web-based interactive visualization of topics estimation on top of LDA. They defined the relevance of a word in a corpus to a particular topic within a set of topics defined in LDA. A relevance of term $w \in \{1, \ldots, V\}$ to topic $k \in \{1, \ldots, K\}$ is defined as:

$$r(w, k|\lambda) = \lambda \log(\emptyset_{kw}) + (1 - \lambda)\log(\frac{\emptyset_{kw}}{p_w}) \quad (5)$$

where $\emptyset_{kw}$ denotes the probability of term *w* in topic *k* and p_w denotes the marginal probability of term *w* in the corpus. λ, which is between 0 and 1, determines the weight given to the probability of term *w* under topic *k* relative to its lift (the term's probability within a topic divided by its probability across all documents).

III. Topic Extraction For Tweets In Thai

Figure 1 shows the various processes that constitute our study. The tweets we collected are first processed in the topic extraction module. The result is a set of topics and lists of words associated to each topic. When needed, as would probably be the case, manual refinement would be applied to the extracted results to become meaningful categories. This would entail fine-tuning the arrangement of the words per topic. Based on the refined arrangement of topics and words, categorization rules would then be constructed manually. Finally, the categorization rules could be applied in trend detection. This section details all these processes.

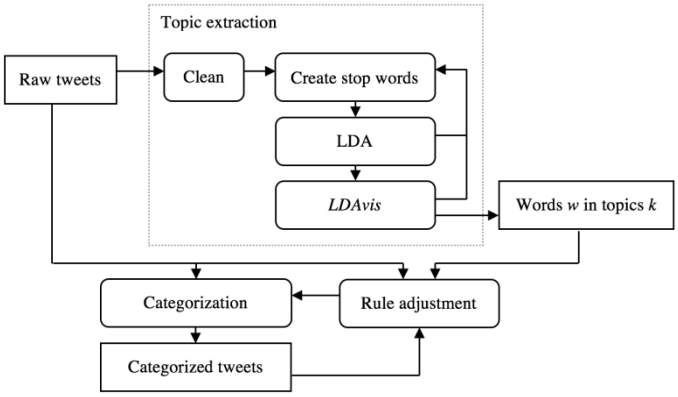

Figure 1: Topic extraction and tweet categorization

A. Data Collection

The website called *thaiflood.com* collected tweets during the 2011 Thai flood and made them publicly available afterwards. We obtained 651,183 tweets from this website that were sent between October 2010 and March 2013. All tweets contain the hashtag *#thaiflood*. By simply drawing a frequency graph on a per-day scale, we detected an unusual activity during mid-September until late December 2011. Hence, we narrowed down our interest to this particular period, which had 423,401 tweets.

In addition to the tweets, we also collected news articles related to the flooding. The articles described the intense situations in October and November and as the massive flood was approaching Bangkok. Increasing trend in tweets of up to 52% from September to October seemed consistent with the situations. The growth continued until early November.

B. Data preprocessing

Most of the tweets were written in Thai. Few were in English. Many included symbols, which were omitted since they have no meaning (e.g., a star symbol to catch attention). URLs and username tags (begin with @) were also removed. Hashtags, however, were preserved except *#thaiflood*, which is obviously in every tweet. We used LexTo to perform word segmentation. LexTo is a text tokenization tool using longest matching approach. The tool has been developed and made publicly available by Thailand's National Electronics and Computer Technology Center.

Defining stop words was complicated due to language, domain and context specific constraints. We first tried to construct a list of stop words from *SAT* and *H* values as explained in Eq. 3 and Eq. 4. This became problematic because of the very nature of microblogs: a limited number of characters. This diminished the chance of any given word to occur repeatedly. As a result, the assumption that we may extract stop words by computing the mean of probability (*MP*) and its variance (*VP*) becomes unlikely. What we would see instead is that words that are more frequent across all documents are more likely to have a higher *MP* value. The length of tweets containing the word may affect $P_{i,j}$; nevertheless, it should not be adequate to conspicuously alter the outcome. This is simply because of the limited length of tweets that moderates the effect. The same goes for *VP* and entropy (*H*) values.

To test the argument above, we conducted an experiment using the tweets we collected. We created a list of frequent words sorted in descending order. Table 1 shows a comparison of the top $T \in \{100, 300, 500, 1000\}$ words in the frequent words list against the top *T* words in *MP*, *VP*, *SAT*, *H*, and *Final* (i.e., Rank(*SAT*) + Rank(*H*)) lists in terms of how dissimilar they are, i.e., higher percentage values indicate greater difference. Stop words are supposed to be inordinate but less informative. Therefore, we expect some degree of similarity between the frequent words list and both sorted *MP* and *H* lists. If the argument is true then we should observe a high degree of similarity.

Table 1
Percentage of the top T words in the created lists that are not in the frequent words list

T	*MP*	*VP*	*SAT*	*H*	Final
100	6%	19%	100%	9%	100%
300	4%	16.67%	100%	6%	100
500	5%	15.6%	100%	7%	99.98%
1000	4.5%	16.6%	100%	6.4%	99.5%

We can see from Table 1 that, as expected, both *MP* and *H* are very similar to the frequent words list. In fact, both scores are calculated in a similar way to the frequent words list and, therefore, should not be, by themselves, an alternative list to the list. *SAT* and Final, on the other hand, are almost completely different to the list. Stop words should at least appear often. As we observed, *SAT* and Final are clearly not good indicators of a stop word. Removing words that even though uninformative but are not frequent would not effectively improve the result. We also found a considerable amount of words with an invalid *SAT* value. These words appear only once in the whole corpus, which means their *VP* value is zero. It became clear when we considered words that appeared sparsely, and yet produced a minuscule variance of probability that exaggerated the *SAT* value. Moreover, a set of retweets could contain some words that appear only within the set and the original tweet, which means these words also have an extremely high or invalid *SAT* value. All of these words obviously are detrimental to the performance of the algorithm. We finally came to a conclusion that this automatic approach is unsuitable for microblogs.

We then proceeded with a rather more conventional fashion. After cleaning tweets and performing word segmentation, we counted the frequency of each word. We manually determined the stop words from a list of frequent words. However, another problem emerged. Imperfect word segmentation broke some meaningful words into unrelated components or even completely different words. Because of the sophistication of the Thai language, we could not intuitively reassemble these words without an absolute certainty. This requires knowing the contexts that embed the words. Our choice was to include them in the stop word list. We eventually created an initial list of 100 stop words. The list was later extended to 323 words after several repetitions of LDA. Table 2 contains some examples of the stop words. Most of these stop words are ambiguous. Their meaning may change completely when combined with other words or when they are in different contexts.

C. Topic extraction

After cleaning the tweet and removing the stop words in the preprocessing stage, the tweets are finally ready for topic extraction. We performed LDA on our data using *gensim* topic modelling package version 3.4 for Python. We executed this model for a number of topics with $K = \{5, 10, 15, 20\}$, and later $K = \{4, 6\}$, as parameter. We ran the model over the corpus for up to 100 iterations. Because LDA uses topic multinomial distribution, a word may be assigned to multiple topics with different probabilities. The result was a set of words for each topic sorted by the probability of each word being in the topic. After each execution, we reviewed the top 30 words in each topic and filter out uninformative words by including them in the stop word list. Table 3 shows the top 15 words in three of five topics. From the top 30 words in each topic, we initially defined six main topics, namely, warning, news, situation report, information, volunteer and donation. Volunteer and donation are, in fact, very similar semantically and relatively (words in both groups were usually assigned by LDA to the same or homogeneous topics).

Table 3
Top 15 words in three of six topics

Topic	Words	English
1	ด่วน, ประกาศ, สถานการณ์, ฉุกเฉิน, เตือน, sm, เชียงใหม่, พื้นที่, กทม., หนัก, อาหาร, อพยพ, เตรียม, ข่าว, ระวัง	urgent, announce, situation, emergency, warn, sm, Chiang Mai [province], area, Bangkok, heavy, food, evacuate, prepare, news, warning
2	อยุธยา, ช่วย, เขื่อน, มูลนิธิ, ขอความช่วยเหลือ, ข้าวสาร, ลพบุรี, อาสาสมัคร, ตอนนี้, วันนี้, รังสิต, ระบาย, สนง., โทร	Ayuthaya [province], help, dam, foundation, ask for help, rice, Lopburi [province], volunteer, now, today, Rangsit [district], drain, office, call
3	บริจาค, อาสา, จำนวนมาก, เบอร์, เสื้อ, กรุงไทย, ชูชีพ, บัญชี, ดุสิต, สภากาชาด, scb, บริเวณ, เงิน, นนทบุรี, ร่วม	donate, volunteer, large amount, number, shirt, Krung Thai [bank], survival, account, Dusit, Red Cross, scb [bank], area, money, Nonthaburi [province], together

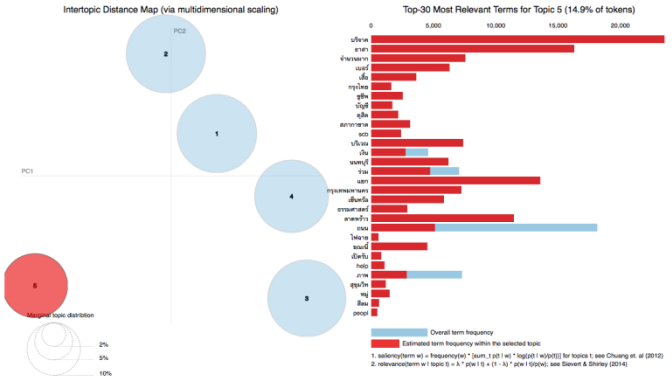

Figure 2: Intertopic Distance Map and top-30 most relevant terms (topic 5) generated by *LDAvis*

IV. TWEETING TRENDS IN THE THAI FLOOD CASE

We believe that we could learn the public's perception of the disaster from the tweets. Being able to detect tweeting trends is our key to learn how people perceived the situation. Trend, by its definition, is a topic of many posts on a social media website within some period of time. Accordingly, our trend detection process should entail two fundamental facets: topic extraction and time. The previous section explained how we achieved topic extraction. On top of that, we aimed to employ a rule-based categorization model of all the tweets. The LDA result was helpful in defining our categorization rules. It allowed us to detect tweeting trends on a temporal scale.

Although the LDA result was useful in defining the categorization rules, we should note however that the LDA result itself should not be used directly to construct a categorization model. Firstly, LDA should be told how many topics were expected. However, we had very limited and only partial knowledge of what potential topics there might be. Furthermore, LDA itself does not take into consideration the meaning of the words nor their context. Consequently, each topic in the LDA result may contain words that belong to more than one category, and some words may not at all belong to the topic they were assigned to by LDA. Secondly, some meaningful words were broken into uninformative words during word segmentation and then removed by the stop word list. Such words have to be recovered later on.

The issues above necessitated that we manually construct the rules that would improve the categorization of the words. We first studied the LDA result and then defined a set of categories. For example, words in topic 1 (see Table 1) are likely to be warnings, while topic 3 seemed related to volunteering and donation. We re-assigned some informative words from the LDA result to the categories where they are best suited. We constructed from this improved category-words associations an initial set of categorization rules. We then performed categorization on our entire dataset while adjusting the set of categories, which then necessitated adjusting the categorization rules. At the end, we were able to obtain six categories, namely, *volunteer/donation*, *warnings*, *information*, *news*, *situation reports* and *guidelines*. Tweets categorized as information provide various kinds of information, e.g., emergency telephone numbers, location of evacuation centers, maps, and volunteer opportunities. Situation reports are more specific to situations in a particular area. Guidelines involve instructions on how to help cope with flood-related issues. Some tweets can be categorized as both situation report and news, e.g., a report of flood in an area and the authorities' plan to evacuate the inhabitants.

To test the accuracy of the categorization model, we categorized 1,800 tweets and compared the result with that of our model. It is important to note here the difference between the manner we categorized the test cases and the categorization model. The full contextual meaning is evident and unambiguous when reading the test tweets, hence, we knew more how to categorize these tweets. The categorization model, however, relies on string matching between words in a tweet and words in the rules to assign a tweet into a category. Hence, the model does not consider the context in which the words are embedded – it has yet to fully capture the notion of context. This diminishes the accuracy of the model as can be seen in Table 4 where some categories have lower, but still acceptable, F-scores given the complexity of the domain.

Our next step was to categorize all tweets and observe the weekly trends, as shown in Figure 3. The news articles we collected helped us interpret the result. Tweets about volunteer/donation led the trend during the first weeks and later declined gradually. This decline can be explained by the changes in flooding situation, especially in the inundated areas. During mid-September, a large area in Thailand's northern region and north of Bangkok was severely flooded. We observed a considerable amount of tweets encouraging and providing information about donation. A week later, the flood approached Bangkok and triggered concern among people living in the area. Volunteer-related tweets became prominent. The trend was driven by an increasing need for volunteers to build temporary flood prevention systems (e.g., sandbags), produce and distribute survival kits, maintain evacuation centers, and help flood victims. Tweets providing information also diminished overtime. A simple explanation is that the concerns that motivated people to look for information were eventually relieved, and so did the need for information. Situation reports were somewhat different as the trend maintained steadily during the first four weeks, elevated during the time the flood affected Bangkok, and later declined. Guidelines, on the other hand, had not gained attention until the flood was contained. News and warning did not show any conspicuous increase nor decrease in trend. Tweets categorized as 'other' are diverse. These tweets are difficult to categorize, e.g. opinions, and can be unrelated. The 'other' category showed an increasing trend after the flood situation had been resolved.

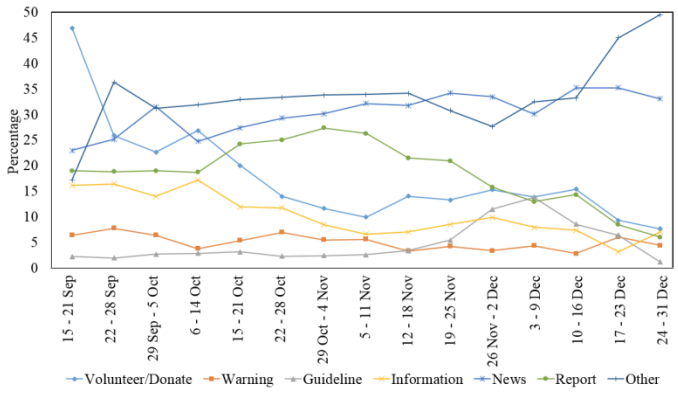

Figure 3: Tweeting trend during and after the flood

V. CONCLUSION

We presented the difficult problem of Thai language processing that we encountered while we categorized disaster-related tweets into a set of topics. Manual refinement was still an indispensable tool that allowed us to effectively improve the result of an unsupervised machine learning method, i.e., the LDA. The refinement primarily involved correcting the result of the word segmentation and making more accurate the stop word list that eventually eliminated the redundancy of words in the LDA-produced topics, as well as constructing and improving the categorization rules. The drawback of this manual refinement, however, is that in many cases it would require human expertise, which may be costly or unavailable. While the refined topic extraction and categorization results allowed us to gain deeper insights as to the activities of the tweeting communities, automating this refinement process is certainly a challenge worth exploring.

REFERENCES

[1] Bleh, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation. Journal of Machine Learning Research. 3 (2003) 993-1022

[2] Zou, F., Wang, F. L., Deng, X. Han, S., Wang, L. S.: Automatic Construction of Chinese Stop Word List. Proceedings of the 5th WSEAS International Conference on Applied Computer Science (2006) 1010-1015

[3] Myerson, R. B.: Fundamentals of Social Choice Theory. Journal of Political Science. Vol. 8. 3 (2013) 305-337

[4] Daowadung, P., Chen, Y. H.: Stop Word in Readability Assessment of Thai Text. IEEE International Conference on Advanced Learning Technologies (2012) 497-499

[5] Blei, D. M., McAuliffe, J. D.: Supervised topic models. Advances in Neural Information Processing Systems 20 (2007)

[6] Fuchs, G., Stange, H., Samiei, A., Andrienko, G., Andrienko, N.: A semi-supervised method for topic extraction from micro postings. Information Technology. Vol. 57. 1 (2015) 49-56

[7] Ramage, D., Hall, D., Nallapati, R., Manning, C. D.: Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora. Conference on Empirical Methods in Natural Language Processing. (2009) 248–256

[8] Sievert, C., Shirley, K. E.: LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces (2014) 63-70

[9] Kongthon, A., Haruechaiyasak, C., Pailai, J., Kongyoung, S.: The Role of Twitter during a Natural Disaster: Case Study of 2011 Thai Flood. Proceedings of PICMET '12: Technology Management for Emerging Technologies (2012) 2227-2232

Table 2
Example of stop words

ที่ (at), แล้ว (already), และ (and), ให้ (give), บาง (some), จาก (leave), ไม่ (no), ยัง (not yet), ของ (of), ถึง (to), เพื่อ (for), เข้า (enter), หลัง (back), แต่ (but), มาก (very), ต้อง (must)

Table 4
Accuracy test score and F-score by category

	Volunteer	Warning	Guideline	Information	News	Report	Other
Accuracy	.96	.98	.98	.91	.79	.86	.80
F-score	.87	.78	.77	.62	.66	.70	.60