DOM: A big data analytics framework for mining Thai public opinions

Santitham Prom-on, Sirapop Na Ranong, Patcharaporn Jenviriyakul,
Thepparit Wongkaew, Nareerat Saetiew and Tiranee Achalakul
Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi
Bangkok, Thailand

***Abstract*—This paper presents the development of DOM, a mobile big data analytics engine for mining Thai public opinions. The engine takes in data from multiple well-known social network sources, and then processes them using MapReduce, a keyword-based sentiment analysis technique, and an influencer analysis algorithm to determine public opinions and sentiments of certain topics. The system was evaluated its sentiment prediction accuracy by matching the predicted result with the human sentiment and tested on various case studies. The effectiveness of the approach demonstrates the practical applications of the engine.**

***Keywords*—opinion mining; big data analytics; MapReduce; public sentiment**

I. Introduction

We, human being, have never been more connected through the emergence of social networks. Social networks, in terms of both data and users, have been exponentially growing and connect our lives together in various dimensions. We can connect with people across the planet with a touch of a finger. In every second, hundred thousands of messages are shared through social media such as Facebook, Twitter, Foursquare, Pantip, etc. They are about our life, feeling, experience and opinion. This practically represents the 21st century of our civilization, "The era of social network".

Social media networks generated huge volumes of data. They have been use in various types of applications including public health [1], emergency coordination [2], news recommendation [3], and stock market prediction [4]. The data from social media networks gathered under the catch-all term, "big data". However, as much as 90% of the data stored is "unstructured," meaning that it is spontaneously generated and not easily captured and classified. Big data is only valuable if it tells a story. The fuller the story your data tells, the better you'll be able to take advantage of that data. While recognizing a trend can help you make better decisions, understanding the cause behind that trend is even more valuable. The organizations that can use stories to make sense of big data are going to excel.

In this paper, we presents the developments of DOM (Data and Opinion Mining), a big data analytics engine that is capable of mining Thai public opinions regarding specific issues discussed on the social network sites, and its corresponding mobile solution for answering public opinions about events and locations. Software features and design will be discussed in Section II. Section III explains how the software was implemented using cloud-based technology. Section IV shows the evaluation of the DOM effectiveness in predicting the sentiment score of public opinions. Usages of DOM for different tasks are presented in Section V. Comparisons of DOM with respect to others and the future steps in the development are discussed in Section VI.

Table 1. *Sources of Social Network Data.*

Source	Data Description
Twitter	Twitter messages, also known as tweets, are short 140-character text messages. Tweets are all public.
Facebook	Facebook data can only be retrieved if the privacy is set to public. They are in forms of status posts and Facebook Page posts.
Foursquare	Foursquare provides both text comments and review score of a number of places.
Pantip	Pantip data are in forms of webboard threads. It is one of the prominent Thailand online social communities.

II. DOM

A. Data Sources

We collected data from four different data sources; Twitter, Facebook, Foursquare and Pantip, as described in Table 1. These social network data, if the locations can be specified, were collected in scope of Bangkok area. For Twitter data, we used Search API [5] provided from Twitter Inc. to collect tweets without any keywords. We collected approximated 15 million tweets or about 12GB uncompressed data each month. Each tweet contains multiple data fields, including time, username, user followers, retweet, count, location, and the textual comment. For Facebook, we used Graph API [6] developed by Facebook Inc. Unlike Twitter, we can only request and collect data from Facebook fanpage which consists of posts and comments of specific topics. We collected Facebook data about 5,000 messages, which is approximately

978-1-4799-4575-7/14/$31.00 ©2014 IEEE

4MB per fanpage each month. Graph API provides attributes including time, username, number of Like, location as well as textual comment for each message. For Foursquare, the sitation is like Graph API. Foursquare provides their API for developers to gather data named Venues and Tip search API [7]. Foursquare provides comments of places. In each month we collected approximated 500 messages or 0.4 MB per place. Foursquare data includes time, username, like count, location and the textual comment. Our last data source is Pantip.com, one of the prominent Thailand online social communities. We developed a web crawler to gather the data on this website, since they do not provide an API to gather data. The web crawler was designed to have features like Search API. First we simulated the browser by set user-agent to be Mozilla, and then assigned the keywords to the search form of web and submitted the request. We found that approximately 300 messages or about 0.2MB were collected for each topic. Each Pantip thread contains time, username, like count and the text comment.

B. System Architecture

We categorized components into two sides: server-side and client-side. The architecture design of our whole framework is illustrated in Figure 2. The components of DOM engine are classified into server-side which is cloud-based cluster. DOM engine is responsible for collecting, analyzing data and distributing the analyzed data to client-side. AskDOM components are client-side. The client-side requests the analyzed data, queries and displays them to end-users.

Workflow of our framework is as follows. Public messages are collected from social networks, blogs and forums using DOM's crawler module. All collected messages are stored in MongoDB, an unstructured database. After that each message is then processed using basic Natural Language Processing (NLP) technique to parse the text data, categorize its topic, compute its sentimental score and analyze its influences. DOM also uses MapReduce technique based on Apache Hadoop framework to reduce the processing time. DOM periodically processed the data to compute their sentimental score. Finally AskDOM, the mobile application, gets the analyzed data, queries and displays the information to users according to the inquired topics.

In this paper, we focus the usage of DOM as a Thai public opinion mining framework to track social issues and provide sentiment rating and information of point of interest (POI) based on public opinions. However, the core functions of DOM engine was designed to support dynamic data. There are several features that could be added or further developed to provide additional functionality (e.g. adding more data sources, supporting other languages). Since DOM is cloud-based engine, scalability is also available.

Furthermore DOM can be easily applied in various types of usage, either community side or commercial side. There are case studies in section IV that shows some potential usage of DOM.

The current version of DOM consists of the following: modules:

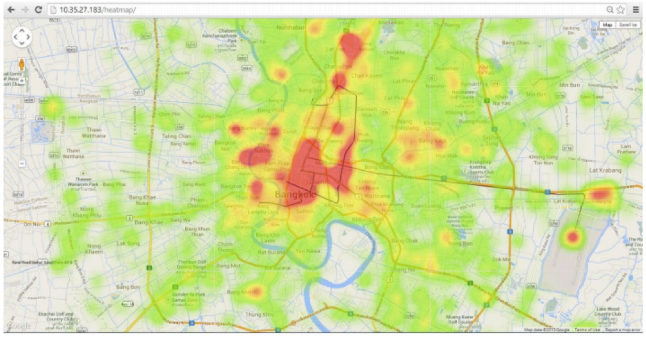

Fig. 1 Twitter activity heatmap in Bangkok area.

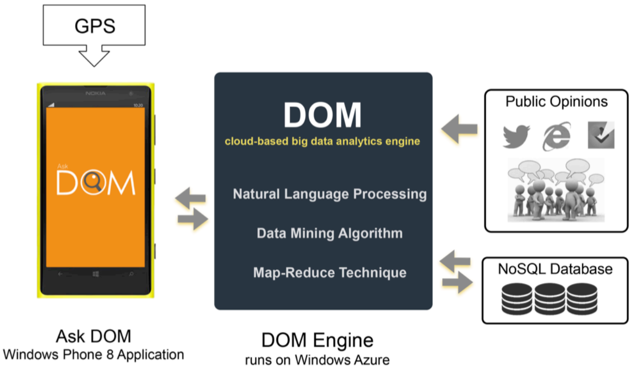

Fig. 2 Conceptual framework of DOM and its corresponding mobile application, AskDOM.

C. MapReduce Framework

Since huge data are involved in this project, MapReduce [8], the high performance computing technique, is used. This is because if the data is to be processed sequentially, the processing time would be too large for the practical application. MapReduce technique on Apache Hadoop framework is therefore the best way to accelerate the analysis speed.

In this paper, the MapReduce technique separates the mining process into two main steps; Map and Reduce. Map function takes the entire text input, breaks it into subsets to be evaluated for their sentiment scores and distributes them to worker nodes. Reduce function combines the resulting sentiment scores from each small worker nodes by grouping keywords of specific topics of interest and summarizing the sentiment scores into final results.

D. Sentiment Analysis

In this work we targeted words in which opinions are expressed in each sentence. A simple observation was that these sentences always contain sentiment words (e.g. great, good, bad, worst). To simplify the process, if the sentences do not contain any sentiment words, their sentiment values will be neutral (non-opinion). So we designed our framework to classify the sentiment of each sentence based on its sentiment words and the combination of them.

Furthermore we designed the system to be able to process Thai conditional sentences, which are sentences that describe

Table 2. *The examples of sentences in the corpora.*

#	Type of Corpus	Word	Value
1	Positive words	เทห (smart)	3
		ด (good)	3
		เยยม (best)	4
2	Negative Words	เสอมโทรม (decadent)	-3
		แย (bad)	-3
		หวยแตก (worst)	-4
3	Modifiers	ไม (not)	-1
		คอนขาง (likely)	0.5
		ทสด (best)	1.5
4	Conjunctions	แต (but)	2
		และ (and)	1
		รวมไปถง (including)	1
5	Name of places	สวนลมพน (Lumphini Park)	-
		สยาม (Siam)	-
		จตจกร (Chatuchak market)	-

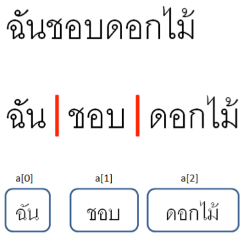

Fig. 3 Example of Thai word tokenization

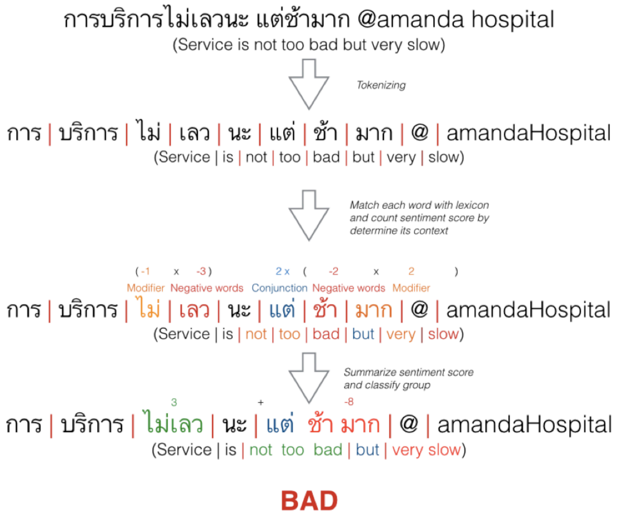
การบริการไม่เลวนะ แต่ช้ามาก @amanda hospital
(Service is not too bad but very slow)
Tokenizing
การ | บริการ | ไม่ | เลว | นะ | แต่ | ช้า | มาก | @ | amandaHospital
(Service | is | not | too | bad | but | very | slow)
Match each word with lexicon and count sentiment score by determine its context
(-1 x -3) 2 x (-2 x 2)
Modifier Negative words Conjunction Negative words Modifier
การ | บริการ | ไม่ | เลว | นะ | แต่ | ช้า | มาก | @ | amandaHospital
(Service | is | not | too | bad | but | very | slow)
Summarize sentiment score and classify group
3 + -8
การ | บริการ | ไม่เลว | นะ | แต่ ช้า มาก | @ | amandaHospital
(Service | is | not too bad | but | very slow)
BAD

Fig. 4 Example of Thai Sentiment Analysis

implications or hypothetical situations and their consequences. For example, in the sentence like 'I like the location of this company but I do not like their staffs.' The sentiment of 'location' is positive but negative on 'staffs'. We found that most conditional sentences contain modifiers and conjunctions (e.g. but, and, or).

To classify each message as positive, neutral or negative, we employed a lexicon-based algorithm to measure sentiment score of each message. We defined five corpora including positive words, negative words, modifier, conjunction as well as name of point of interest. Each word in two sentiment corpora, positive words and negative words, contains sentiment rating ranging from -5 to 5. The examples of our corpuses are shown in Table 2.

DOM detects and matches words and its sentiment polarity by using these corpora. Since the nature of Thai sentence structure is continuous without any whitespace breaks between words, we need to tokenize each sentence into group of words In this process we used 'LexTo' [9], the opensource Thai word tokenize tool, to tokenize words in each sentence and then store them as arrays using the longest word matching algorithm [10]. The example of this procedure is shown in Figure 3.

DOM generates small jobs to detect words of each sentence in parallel. First of all, DOM filters the non-related sentences out by matching words with name of POI corpus. After that only sentences that relate to specific topics of interest (in this case is point of interest) would remain. DOM then iteratively matches sentiment keywords with remaining corpuses. If there are sentiment words in array, DOM collect its sentiment score and summarize at the end of each sentence. DOM then automatically classifies each sentence into sentiment group; positive, neutral and negative, depending on its score band (the range of distributed sentiment score). DOM not only determines keyword from sentences, but also determines context of each sentence. The positions of words, modifiers, conjunctions as well as emoticons are also determined in our framework. In some cases these words can be important clues to emphasize the mood of the sentences. Especially for the modifier keywords, they can invert the sentiment score if their positions are adjacent to the sentiment words as illustrated in Figure 4.

E. *Influencer Analysis*

The rise of social media platforms such as Twitter, with their focus on user-generated content and social networks, has brought about the study of authority and influence over social networks to the forefront of current research. For companies and other public entities, identifying and engaging with influential authors in social media is critical, since any opinions they express can rapidly spread far and wide. For users, when presented with a vast amount of content relevant to a topic of interest, sorting content by the source's authority or influence can also assist in information retrieval. In the social network community, a variety of measures were designed for the measurement of importance or prominence of nodes in a network [11, 12]. In the following, we will briefly summarize the centrality measure that we have used to describe possible candidate indicators for the power of influential in message diffusion. For DOM engine, we have used "Degree centrality" to identify influential users in the Twitter's networks.

Degree centrality is the simplest centrality measure, as illustrated in Figure 5. The degree of a node i denoted by k_i, is the number of edges that are incident with it, or the number of nodes adjacent to it. For networks where the edges between nodes are directional, we have to distinguish between in-degree and out-degree. The out-degree centrality is defined as

$$C_{D_O}(i) = \sum_{j=1}^{n} a_{ij} \qquad (1)$$

where a_{ij} is 1 in the binary adjacency matrix A if an edge from node i to j exists, otherwise it is 0. Similarly, the in-degree centrality is defined as

$$C_{D_I}(i) = \sum_{j=1}^{n} a_{ji} \qquad (2)$$

where i describes the node i and a_{ji} is 1 if an edge from node j to i exists, otherwise it is 0.

F. *AskDOM: Mobile Application*

To utilize DOM to its fullest extent, we developed AskDOM, a mobile solution designed to use DOM to provide a means for general publics to help improving their own communities by providing reviews, feedbacks, and rating of service providers automatically analyzed from public opinions on social networks (Twitter, Facebook, Pantip and Foursquare). AskDOM comprises two important modules: (a) front-end interface with features designed to connect users to service providers such as I-Share (direct feedback), Map (traffic and incident map), Anomaly (ab-normal situations reports), and (b) DOM Engine, the back-end system that periodi-cally gathers and processes social network data, performs public sentiment analysis, determines relationship influencers, and conducts natural language processing for both Thai and English. The integration of both modules will increase the transparency of the service businesses, make the agencies more accountable for their service quality, and provide a means for general citizens to involve with the improvement of the public services in terms of both information availability and improvement. Such an involvement will improve not only the quality of service, but also create a sense of community to the general citizens that they have to be part of the social function.

•

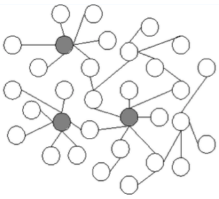

Fig. 5 Simulation of Influencer network graph in the Twitter's networks

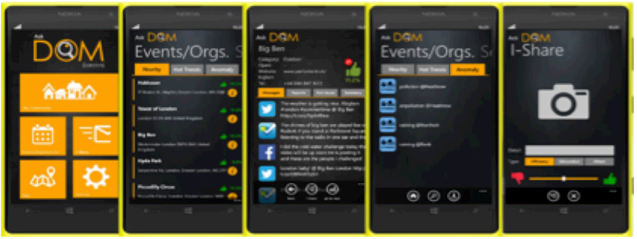

Fig. 6 AskDOM Application

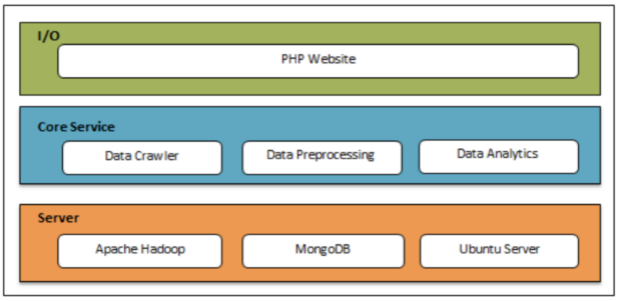

Fig. 7 DOM engine architecture

III. Implementation

Figure 7 shows the overall implementation architecture of DOM engine. The structure has three main components which are Server, Core Service and I/O.

A. *Server*

Server section consists of three components; Ubuntu server, MongoDB as well as Apache Hadoop. We implemented DOM engine based on Apache Hadoop MapReduce which on Ubuntu server. MongoDB, the famous unstructured database was also used in this framework. The unstructured database is often highly optimized key–value stores intended for simple retrieval and appending operations to improve the performance in terms of latency and throughput.

B. *Core Service*

Core Service, the main part of our framework, consists of three components.

1) Data Crawler: This module provides automatically raw data feed from social network and stores them in the database, MongoDB. Each crawler code is specific for each social network or websites.

2) Data Preprocessing: This component prepares raw data to be ready for the analysis part by tokenizing Thai and

English words from sentences and removing outliers and reformatting data. Then the cleaned data will be sent to Data Analysis part.

3) Data Analysis: There are two main analyses in this component:

a) Sentiment Analysis evaluates sentiment in twitter text and find people's mood in particular topic. For example, how people think about traffic in Bangkok.

b) Influencer Analysis determines people's positions in network, which indicate how influential they are. The influential people are more likely to acquire connections and have more connections.

C. I/O

I/O, the web-service implemented using PHP, receives the result from Core Service and then sent them to client-side to display in JSON format. Since the number of data in social network is increasing every second, using the static resources (e.g. static server) may not be practical. So we designed to run DOM on the cloud. Cloud provides ability to add a blob storage depending on the size of data. Furthermore DOM has ability to scale the number of processer. In other word, DOM can increase or decrease the number of mapper and reducer for running job.

IV. Validation

To validate the effectiveness of DOM, we conducted a subjective experiment to assess the sentiment prediction accuracy. In the following, we will describe the validation procedure and discuss on validation results.

A. Validation Parameter

- 184,184 messages from Facebook, Twitter and Foursquare (both positive and negative messages) divided into short and long messages, including 172,717 short messages (≤ 150 characters) and 11,467 long messages (> 150 characters).
- 12 subjects (6 males and 6 females) were participated in the experiment. They were students at the Computer Engineering Department, King Mongkut's University of Technology Thonburi, Thailand.

B. Validation method

1. For the human end, 184,184 messages were divided into 12 parts, each of which was assigned to each subject. They classified the messages in to positive and negative classes.
2. For DOM engine, 184,184 messages were classified by the engine into positive and negative classes.
3. The results of both human and DOM were compared and analyzed together to assess the system prediction accuracy.

Table 3. *Summary of prediction accuracy.*

Message Type	Positive Comment Accuracy (%)	Negative Comment Accuracy (%)	Total
Short	79.75	56.33	75.99
Long	86.53	38.95	81.29
Total	80.19	55.57	76.32

Table 4. *Detail analysis of the system effectiveness.*

Msg. type	TP	FP	TN	FN	Precision	Accuracy (%)
Short	115,643	12,103	15,613	29,358	0.905	75.99
Long	8,830	771	492	1,374	0.919	81.29
Total	124,473	12,874	16,105	30,732	0.906	76.32

C. Validation results

Table 3 and 4 shows the comparision results of 12 students and DOM engine. We found that DOM engine can classified messages and do sentiment analysis with accuracy over 75%. The accuracy of DOM engine is in the standard of text classification [13], so DOM engine is practical to use in social network analysis and can be applied to many dimensions in the real word.

V. Case Studies

In addition to the evaluation of the system effectiveness, we tested DOM engine further on various case studies that were of interest of Thai public during the time periods. Each case study aims to explore either specific social or political issue that people were discussed widely on the Internet, thus offers a summary of Internet public opinions of that issue.

A. Political opinion: #prayforthailand

Around the end of 2013, citizen of Bangkok faced with multiple rounds of political protests, and violent acts toward both protesters and officers. Hashtag "#prayforthailand" is one that was frequently used in social media to express the concerns over the situation. Different opinions were expressed regarding this political issue. We used DOM to mine the general public opinions that were expressed in the social network to determine the political climate at that time. We collected tweets around Bangkok area that contain the hashtag "#PrayForThailand." There were over 100K tweets collected from 29 November to 7 December 2013. We implemented Naïve Bayes and Support Vector Machine (SVM) to DOM engine to classify political opinions into six predefined categories as shown in Table 5. DOM can accurately put tweets into categories with more than 85% accuracy.

B. Bangkok traffic congestion ranking

Bangkok's traffic problem is one of the most serious problem that urban citizen have been facing in their daily life. Knowing such information on which streets the traffic jams often occur would allow citizens to prepare to encounter the problem and allows the government to find a way to solve it.

We used DOM engine to track traffic jams keywords, name of streets, intersections as well as famous places in Bangkok, Thailand that contained in public tweets, and then rank the streets that were mostly mentioned about traffic jam based on 22K tweets collected from 17 February to 8 March 2014.

The results as shown in Table 6 are consistent with what Thailand's Department of Highways hotline gathered the statistics from phone calls. However using DOM engine is much faster and cheaper.

VI. Discussion and Conclusion

This paper presents the development, evaluation, and case studies of DOM, a big data analytics framework for assessing public sentiments of specific social issues. DOM, which is an opinion mining and sentiment analysis engine, is encapsulated as a mobile application known as AskDOM, that allows users to interact and find information of places suggested by the sentiment ratings. We have demonstrated both accuracy, as discussed in Section IV, and generalizability, as shown in Section V, of the engine in the analysis of various topics that are relevant to public interests.

Further improvements are still needed to make DOM engine more adaptive and robust. First, the sentiment score associated with each keyword is currently context independent and come mainly from the manual adjustment by the administrator. A context-dependent keyword-score association study is needed for each of the task required. After obtaining these related associations from different contexts, rules can be derived so that the system can work effectively on different tasks. Second, public opinions usually contain a lot of personal messages that are irrelevant to the places under discussion. A filter that is capable of detecting the context of the message is required.

Acknowledgement

We would like to thank for the financial support the Faculty of Engineering, King Mongkut's University of Technology Thonburi through the research grant (to SP) and the Office of Higher Education Commission through the National Research University (NRU) grant, fiscal year 2011-2013 (to TA).

References

[1] M. J. Paul and M Dredze, A Model for Mining Public Health Topics from Twitter. Technical Report. Johns Hopkins University. 2011.

[2] H. Purohit, A. Hampton, V. L. Shalin, A. P. Sheth, J. Flach, and S. Bhatt, "What Kind of #Conversation is Twitter? Mining #Psycholinguistic Cues for Emergency Coordination," Computers in Human Behavior, vol. 29, pp. 2438-2447, November 2013.

[3] O. Phelan, K. McCarthy, and B. Smyth, "Using Twitter to recommend real-time topical news" Proceedings of the third ACM conference on Recommender systems, New York City, NY, USA, 22-25 October 2009.

[4] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," Journal of Computer Science, vol. 2, 1-8, March 2011.

[5] Twitter Search API, https://dev.twitter.com/rest/public/search

[6] Facebook Graph API, https://developers.facebook.com/docs/graph-api

[7] Foursqaure API, https://developer.foursquare.com

[8] Sathya, S., Jose, M.V. Application of Hadoop MapReduce technique to Virtual Database system design. Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference: IEEE, 2011

[9] Lexto, www.sansarn.com/lexto/

[10] Haruechaiyasak, C., Kongthon, A. LexToPlus: A Thai Lexeme Tokenization and Normalization Tool, The 4th Workshop on South and Southeast Asian NLP (WSSANLP) International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013.

[11] L. C. Freemann, Centrality in social networks: I. conceptual clarification, Social Networks 1 (215-239).

[12] C. Kiss and M. Bichler, "Identification of Influencers - Measuring Influence in Customer Networks," Decision Support Systems, vol. 46, pp. 233–253, December 2008.

[13] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., Deng, X. (2013) "Exploiting topic based twitter sentiment for stock prediction" The 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, August 4-9, 2013.

Table 5. *Summary of opinions with "#prayforthailand."*

Opinions	Percentage
Oppose to the government	29.45
Loyal to the king	20.91
Feeling depressed about the situation	15.61
Oppose to both government and protests	0.82
Oppose to protesters	0.01
Others	33.2

Table 6. . *Bangkok traffic congestion ranking*

Rank	**Streets / Intersections**	**Percentage**
1	Ladprao - Paholyothin	19.47
2	Vibhavadi - Rangsit	11.62
3	Petchaburi	7.76
4	Sukhumvit	4.71
5	Ramkumhaeng	4.13
6	Others	52.31