ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Incident Detection Techniques for the Thai language on Twitter

Korn Puangnak[1] and Natworapol Rachsiriwatcharabul[2]

ABSTRACT

Nowadays, the rate of road incidents is continuously increasing as a result of the elevated capability of vehicle acceleration that increases the risk of driver's mistakes. Such road incidents directly impact the flow of traffic in an area and affect the economy directly and indirectly. These incidents also have an impact on society and the environment. Incident monitoring and detection in Thailand is currently done by the responsible authority through CCTV and traffic flow data from traffic flow measurement. Both monitoring and detection have high operation costs. Online communication, on the other hand, has seen significant growth in the present day resulting in a fast growth of online social media use for various characteristic of communication, replacing telephone calls.

This article will present forms of incident detection from social media posts that have been data-mined from Twitter with an autonomous API designed to screen for messages related to incident detection. But there are an enormous number of messages on Twitter. It will be difficult to detect only content that is relevant or of interest. Therefore, this research presents a collection of search terms that are related to incidence detection for the TF-IDF, Word2Vec, and Markov Chain methods. This incident detection method consists of 4 steps. The experiment demonstrated the ability of the proposed method to detect incidents in Thai language with a detection rate of 81.71%, and a false alarm rate of 10.85%, based on the top 5 ranked keywords out of a list of the 20 first keywords.

Article information:

Keywords: Twitter, Incident Detection, Word Cloud

Article history:
Received: December 31, 2021
Revised: June 4, 2022
Accepted: June 25, 2022
Published: August 6, 2022
(Online)

DOI: 10.37936/ecti-cit.2022163.247328

1. INTRODUCTION

Previous literature reviews related to incident management by traffic information analysis to determine the location or site of incident, indicates that most of the research focuses on application of various existing traffic information formats. The practice of such research started back in 2016 due to the rise of big data to the Petabyte-level, combined with global 4G infrastructure development at the same time along with 5G pioneering projects moving toward the fully functional Smart City concept. One of the research projects showed the rising popularity of E-bike usage in China due to its accessibility, low running cost, flexibility, and mobility, making it suitable for densely populated cities. However, 30 to 60 percent of road incidents involve E-bikes. Approximately 70 E-bikes were stolen on a daily basis in Wenzhou, so the government had to install low-energy transmitters on E-bikes that work with stations to track the stolen vehicles. Xiaoxia Jia, Peng Cheng, and Jiming Chen [1] developed an analysis and visualization system that aims for E-bike monitoring, mobility analysis in a web application framework, user behavior analysis, speed monitoring, and E-bike movement using the Mapbox API. At the same time, Wei Peng et al. [2] researched platforms for traffic management using distributed storage and parallel computing to monitor drivers' behaviors and interpret traffic status for relevant information. Parallel computing and clustering helped improve responsiveness in search and analysis of the large amount of data. The parallel computing clusters utilized Flume and Hive's database system and Apache Spark, similar to some other research. Weijian Sun et al. [3] researched specific mobility characteristics of 4G cellular network users and stated that

[1,2]The authors are with Faculty of Engineering, Rajamangala University of Technology Phra Nakhon, Thailand, E-mail: Korn.p@rmutp.ac.th and Natworapol.b@rmutp.ac.th

at present, the mobility of users could be tracked from call detail records (CDR) or Wi-Fi while the users make calls or text messaging or use the Wi-Fi tracking function. However, it does not allow the researcher to access or understand the mobility of large-scale cellular network users, so the research aimed to specifically study the difference in mobility data obtained from a 4G network compared to those from CDR and 3G networks through processing of 6 Terabytes (TB) of big data from a 4G network. The result produced was more detailed than data obtained from CDR and 3G networks in similar research. Xiaoxia Wang and Zhanqiang Li [4] compiled visualization and analysis tools developed from previous research on smart traffic systems into an integrated platform and tested them with traffic data of Beijing, China, for assessing capability of big data processing. The test used the A* Algorithm for analysis. Moreover, research regarding traffic data was also developed into other uses by Zhiyng Cui et al. [5]. They improved the DRIVE Net developed by STAR Lab (US), an online digital roadway interactive visualization and evaluation network that provided data about traffic information on various roads. The improvement included "multi-sourced travel time analysis" with capability to display results in more formats. Mass transit data, motorcycle drivers' data, parking lot data, data from Car2go, shared data from various sources, and other data was collected, analyzed, and visualized through data analysis within the same year. Satyannarayana Nandury and Beneyaz Begum [6] investigated traffic system data handling in smart cities. They stated that handling big data from instruments, sensors, and various IoTs was becoming a challenge as the overall architecture of the system remained unclear. The SWIFT architecture was designed to serve as a connection between smart objects, smart devices, and smart systems. The research presented an experiment where the SWIFT architecture was tested with a large collection of traffic information data to determine various parameters such as traffic density, traffic signalling, parking management, navigation, and vehicle pollution monitoring.

On the other hand, other research about utilizing social media to navigate traffic information big data started back in 2013. Duckwon Chung et al. [7] presented an architecture used in massive real-time traffic information big data distributed complex event processing (CEP). The processing used a NoSQL database utilizing Hadoop and HBase to collect and analyze data from the California Highway Commission. The assessed data came from the performance measurement system (PeMS), collected from loop detectors, including speed, flow, occupancy, and from the TASAS (traffic accident surveillance and analysis system) which provided the coordinated date, time, characteristics, severity, and number of vehicles involved in incidents during the past 10 years. The 1 TB of data was used to summarize the probability of occurrence according to the upstream and downstream velocities, divided into intervals of 5 mph. The probability would then be tested in predicting future occurrence in I-880N highway, California. In the following year, the same process was repeated in Korea by In Jung Lee [8]. In 2015, Shen Zhan et al. [9] presented a methodology that combines Latent Dirichlet Allocation (LDA) with topic wording in text strings and document clustering to visualize the significance of each specific word of varying sizes to categorize incident reports in Twitter. Currently, there are approximately 200 million active users of Twitter and about 500 daily tweets. The first 3 words that would imply than an incident occurred were "vehicle", "driver", and "tow". The second words that would imply the situations were "call", "check", "assist", and "push". This research was tested in Seattle with DBSCAN algorithm, and the reports covered an area within a radius of approximately 600 meters from the site of incident. It was also found that the denser the area is, the closer the reports would be related to the site of incident. Theodore Georgiou et al. [10] summarized a collection of tweets related to traffic in areas of interest that contained relevant keywords such as "this traffic" or "on my way" in a time-based grouping called "Social Volume". The social volume was then compared to the actual traffic volume of California freeway I-405 and it was found that the interpretation was similar and contained a linear relationship. The ratio between traffic volume and social volume was grouped based on time period in a model called the "shift-based model" that allowed the traffic status to be measured with lower investment. Mochamad Vicky et al. [11] presented a design and development of natural language processing for Indonesian to distinguish words and interpret texts from Twitter with a tool called Twitter API. This was an important origin to the trend of design and development of incident detection from social media without installation or investment in additional instruments to obtain optimal benefit out of the significantly expanded social reporting volume.

2. DEFINITION AND CLUSTERING OF DATA

2.1 Clustering of News and Incident Reports

As a prerequisite, the data must be categorized and clustered into 2 groups: news reports and incident reports [12]. The process of data collection must also include the status of the report to indicate whether it falls into either of the clusters as described in Table 1. The test must also account for the accuracy in distinguishing between news and incident data.

Table 1: *Traffic Relevance Determination.*

Index	Interpretation	Description
0	News Report	General updates that are not relevant to traffic or are outdated such as public statements of organizations, local activity announcement, or posts about old incidents.
1	Incident Report	Reports about traffic-relevant incidents that are current or reports about resolution of such incidents.

2.2 Clustering of Occurrences

Reports from different sources usually have different formats. From past research, it was found that the researcher could categorize the reports into 3 clusters [13]. In this study, the author has defined reference "0" as undefined incidents to be excluded from the learning process. A further description is provided in Table 2.

Table 2: *Categories of Occurrences.*

Index	Interpretation	Description
0	Undefined	Occurrences unable to be categorized.
1	Traffic	Traffic volume or flow reports that affect severity clustering but do not describe the cause of the traffic volume or flow observed.
2	Incident	Undesired events resulting in temporary obstruction such as car crashes, breakdowns, or fallen objects.
3	Disaster	Natural occurrences that slow down or obstruct traffic flow such as floods, fallen timber, or mudslides.
4	Potholes	Occurrences that affect road surfaces requiring the drivers to adapt or decelerate such as holes and maintenance blockage.
5	Other	Other occurrences that could not be otherwise be categorized such as demonstrations, building collapses, and roadside fires.

2.3 Clustering of Severity

This study categorizes occurrences into 3 levels according to other past research [14-15] with an additional 0th level for an unidentifiable level of severity. A further description is given in Table 3.

Table 3: *Severity of Occurrences.*

Index	Interpretation	Description
0	Uncategorized	Occurrences without severity level or without clear event descriptions, making them unable to be categorized in the data collection process.
1	Normal	Low severity occurrences with no effect on traffic.
2	Intermediate	Occurrences with moderate severity that affect 1-2 traffic lanes or cause traffic flow slowdown.
3	Lane closure	Occurrences with high severity that result in total blockage or an area locked in a standstill.

3. INCIDENT DETECTION PROCESS FROM THAI LANGUAGE

The incident detection process from Thai language utilizes natural language processing like that which has already used in various languages around the world, but adapted for use with Thai language. There are multiple instances of NLP research for Thai language and more research being conducted continuously. Thai language is considered a continuous script without spaces or sentence terminators. Data is mined from messages on the Twitter platform. One of the ubiquitously used methods is to create a vocabulary from the frequency of words related to traffic. It was expected that there would be a large number of words involved and that a relevance checking method would be needed.

3.1 Message Collection and Recording

Data is collected from the Twitter accounts that regularly and mainly report about traffic situations such as @js100radio (an online extension from radio traffic reports with over 3 million followers), @Traffic_1197 (a twitter account under supervision of a traffic police department with over 20,000 followers), and @fm91trafficpro (a twitter account of the communications police department with over 2 million followers). The data was collected during the period from 5 February 2021 to 23 March 2021 with Twitter's support by letting the developer access tweets through tools using the Twitter API. The data was received in the form of JSON-Objects comprising 3,400 messages. The data was then clustered according to each message's nature using features such as relevance, categories, and degree of severity.

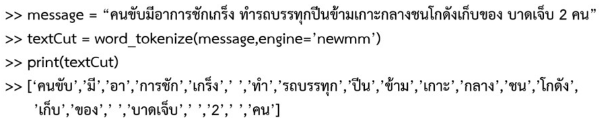
The messages for processing were analyze by experts working on traffic surveillance in Thai governance such as Inter - City Motorway Division, Department of Highways (DOH), and Expressway Authority of Thailand (EXAT). To see how messages were categorized, the message groups before starting processing are show in Table 4. Table 4 shows the example message in first column, traffic relevance determination in R column, categories of occurrences in C column, and severity of occurrences in S column, respectively.

Table 4: *Example Grouping of Messages.*

Message	R	C	S
16:52 ผู้ว่าฯ วีระศักดิ์ อาการดีขึ้นต่อเนื่อง ทุกอย่างเป็นในเชิงบวก ผลเอกซเรย์ปอดล่าสุดดีขึ้น 90% #ข่าวจริงประเทศไทย #FM91	0	0	0
13.56 "อนุสรณ์สถาน" ถ.พหลโยธิน ฝั่งขาออก ไปสี่มุมเมือง รถเคลื่อนตัวช้า สาเหตุปรับผิวจราจร "ทางขนาน" ปากทางเมืองเอก ถึงแม็คโครรังสิต #รายงานจราจร #FM91	1	1	1
18.04 ถนนบรมราชชนนีขาเข้าเยื้องๆ สน.ตลิ่งชัน รถกระบะชนกับ รถนั่งส่วนบุคคล ช่องทางด่วน เลนซ้าย มีเจ้าหน้าที่แล้ว	1	2	2
10:35 น้ำท่วมถนนลาดพร้าว ปากซอยมหาดไทย	1	3	2
22.50 ถนนรามอินทรา ขาออก ฝั่งตรงข้าม รพ.สินแพทย์ # เจ้าหน้าที่กำลังปฏิบัติงาน แนวงานก่อสร้างรถไฟฟ้าฯ #ปิดการจราจรฝั่ งขาออก แล้วเบี่ ยงให้ใช้เสนสวนในฝั่ งขาเข้าการจราจร ชะลอตัว // #FM91 #รายงานจราจร	1	4	3
01:18 หลังยุติการชุมนุมหน้า สน.ปทุมวัน สร้างความเสียหายกับรถทางราชการตำรวจ 8 คัน // #FM91	1	5	0

3.2 Word Tokenization Process

The raw data could not be used instantly because it existed in long continuous sentences. In order for the machine to be able to analyze the data, it needed to be tokenized. The word tokenization was done by a Python open-source library called PyThaiNLP which worked similarly to the natural language toolkit (NLTK), but the PyThaiNLP could be used for Thai language. Apart from tokenization, PyThaiNLP could also complete other tasks such as word correction, word categorization, and ordering. This research mainly utilizes the tokenization function or "word_tokenize(string.engine)". The parameters are "string" (the data to be tokenized), and "engine" (the method to be used in tokenization). A sample result is show in Figure 1.

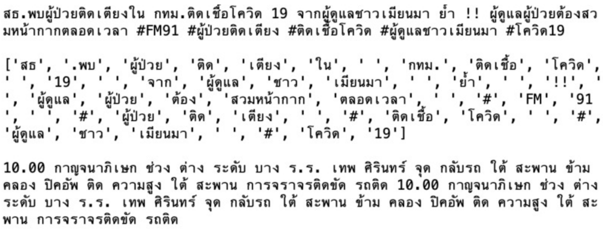


```
>> message = "คนขับมีอาการชักเกร็ง ทำรถบรรทุกปีนข้ามเกาะกลางชนโกดังเก็บของ บาดเจ็บ 2 คน"
>> textCut = word_tokenize(message,engine='newmm')
>> print(textCut)
>> ['คนขับ','มี','อา','การชัก','เกร็ง',' ','ทำ','รถบรรทุก','ปีน','ข้าม','เกาะ','กลาง','ชน','โกดัง',
   'เก็บ','ของ',' ','บาดเจ็บ',' ','2',' ','คน']
```

Fig.1: *Word Tokenization by PyThaiNLP.*

3.3 Text Filtering Process

A string consists of various letters and symbols that do not affect the processing such as '♯', "/", and various URLs. Such components are not needed in NLP and must be filtered out. This research uses regular expression search on the filtered text. It can be seen from Figure 2 that only text that included numbers 0-9 and Thai letters of ก-ฮ and ๑-๙ remained. The third section is a grouping of word cloud and term frequency.



```
สธ.พบผู้ป่วยติดเตียงใน กทม.ติดเชื้อโควิด 19 จากผู้ดูแลชาวเมียนมา ย้ำ !! ผู้ดูแลผู้ป่วยต้องสว
มหน้ากากตลอดเวลา #FM91 #ผู้ป่วยติดเตียง #ติดเชื้อโควิด #ผู้ดูแลชาวเมียนมา #โควิด19

['สธ', '.พบ', 'ผู้ป่วย', 'ติด', 'เตียง', 'ใน', ' ', 'กทม.', 'ติดเชื้อ', 'โควิด',
' ', '19', ' ', 'จาก', 'ผู้ดูแล', 'ชาว', 'เมียนมา', ' ', 'ย้ำ', ' ', '!!', '
', 'ผู้ดูแล', 'ผู้ป่วย', 'ต้อง', 'สวมหน้ากาก', 'ตลอดเวลา', ' ', '#', 'FM', '91
', ' ', '#', 'ผู้ป่วย', 'ติด', 'เตียง', ' ', '#', 'ติดเชื้อ', 'โควิด', ' ', '#',
'ผู้ดูแล', 'ชาว', 'เมียนมา', ' ', '#', 'โควิด', '19']

10.00 กาญจนาภิเษก ช่วง ต่าง ระดับ บาง ร.ร. เทพ ศิรินทร์ จุด กลับรถ ใต้ สะพาน ข้าม
คลอง ปิคอัพ ติด ความสูง ใต้ สะพาน การจราจรติดขัด รถติด 10.00 กาญจนาภิเษก ช่วง ต่าง
ระดับ บาง ร.ร. เทพ ศิรินทร์ จุด กลับรถ ใต้ สะพาน ข้าม คลอง ปิคอัพ ติด ความสูง ใต้ สะ
พาน การจราจรติดขัด รถติด
```

Fig.2: *Text Filtering for Pre-processing.*

3.4 Term Frequency Determination

Term frequency determination was used to find the most frequently used words. This research focuses on word combinations that imply road incidents. 70% of the data was sampled to determine the term frequency related to the aspects of interest. Table 3 shows the data clusters with words and their relevance. Table 4 shows the data clusters with words indicating categories of the occurrences. Table 5 shows the data clusters with words indicating degree of severity of each occurrence.

Table 5: *Word Relevance.*

No	News Report		Incident Report	
	Word	Count	Word	Count
1	และ	465	แยก	986
2	โควิด	436	รถติด	831
3	ราย	328	อุบัติเหตุ	830
4	100	255	การจราจร	646
5	การ	244	ช่วง	614
6	เขต	234	จราจร	597
7	จาก	218	รายงาน	531
8	เวลา	203	ขวาง	529
9	บาง	202	ช่อง	526
10	100	197	เคลื่อนตัว	493
11	เพลิง	194	ช้า	479
12	2564	179	กัน	477
13	วันที่	179	สะพาน	476
14	ได้	178	ขาเข้า	379
15	เพลิงไหม้	177	บาง	361
16	ซอย	177	ขาออก	355
17	ข่าว	168	คัน	321
18	แล้ว	161	จาก	289
19	ติดเชื้อ	158	ถึง	288
20	ของ	156	กับ	285

Table 6: *Word Categories (3 of 5 Categories).*

No	Traffic		Incident		Disaster	
	Word	Count	Word	Count	Word	Count
1	แยก	335	อุบัติเหตุ	830	ขัง	4
2	จราจร	229	แยก	589	กม.	4
3	รายงาน	216	รถติด	580	น้ำท่วม	3
4	รถติด	215	การจราจร	545	ช่วง	3
5	ขาเข้า	165	ช่วง	540	พายุ	2
6	สะพาน	147	ขวาง	524	ส่งผล	2
7	ขาออก	143	ช่อง	478	บริเวณ	2
8	ท้าย	140	กัน	466	การจราจร	2
9	อยู่	130	ช้า	431	สัญจร	2
10	สะสม	125	เคลื่อนตัว	430	ได้	2
11	มาก	118	คัน	316	จาก	2
12	ติด	93	สะพาน	291	สอด	2
13	ท้ายแถว	91	กับ	284	145	2
14	ข้าม	79	จราจร	274	น้ำ	2
15	บาง	78	รายงาน	267	ป้าย	2
16	ลาดพร้าว	73	บาง	251	เลน	2
17	พระราม	68	ก่อน	231	ซ้าย	2
18	จาก	66	กลาง	225	ขวา	2
19	ได้	55	จอด	220	ช่องทาง	2
20	การจราจร	54	ทางซ้าย	218	มาก	2

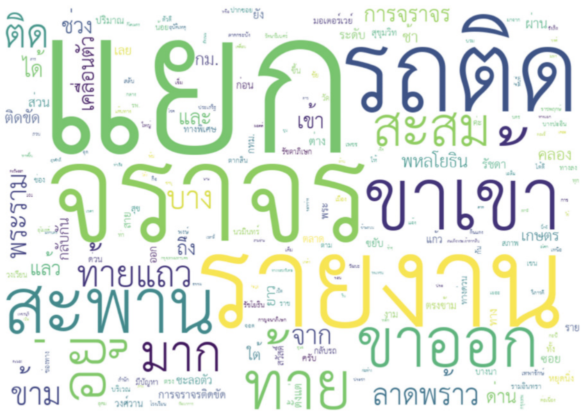


Fig.3: *Word Cloud of Traffic Categories.*

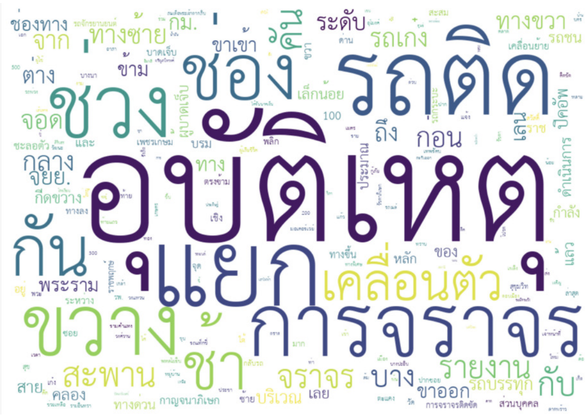


Fig.4: *Word Cloud of Incident Categories.*

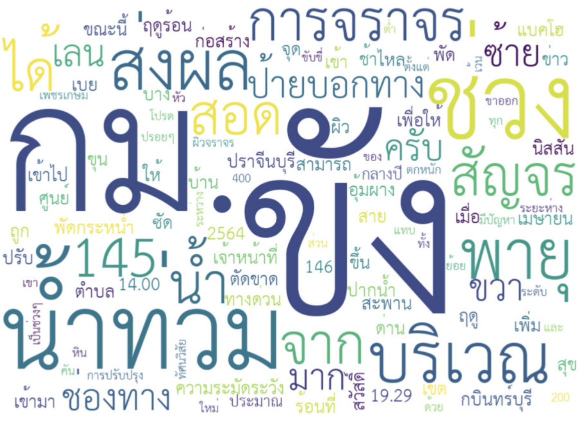


Fig.5: *Word Cloud of Disaster Categories.*

Table 7: *Severity levels of words.*

No	Traffic		Incident		Disaster	
	Word	Count	Word	Count	Word	Count
1	แยก	335	อุบัติเหตุ	830	ขัง	4
2	จราจร	229	แยก	589	กม.	4
3	รายงาน	216	รถติด	580	น้ำท่วม	3
4	รถติด	215	การจราจร	545	ช่วง	3
5	ขาเข้า	165	ช่วง	540	พายุ	2
6	สะพาน	147	ขวาง	524	ส่งผล	2
7	ขาออก	143	ช่อง	478	บริเวณ	2
8	ท้าย	140	กัน	466	การจราจร	2
9	อยู่	130	ช้า	431	สัญจร	2
10	สะสม	125	เคลื่อนตัว	430	ได้	2
11	มาก	118	คัน	316	จาก	2
12	ติด	93	สะพาน	291	สอด	2
13	ท้ายแถว	91	กับ	284	145	2
14	ข้าม	79	จราจร	274	น้ำ	2
15	บาง	78	รายงาน	267	ป้าย	2
16	ลาดพร้าว	73	บาง	251	เลน	2
17	พระราม	68	ก่อน	231	ซ้าย	2
18	จาก	66	กลาง	225	ขวา	2
19	ได้	55	จอด	220	ช่องทาง	2
20	การจราจร	54	ทางซ้าย	218	มาก	2

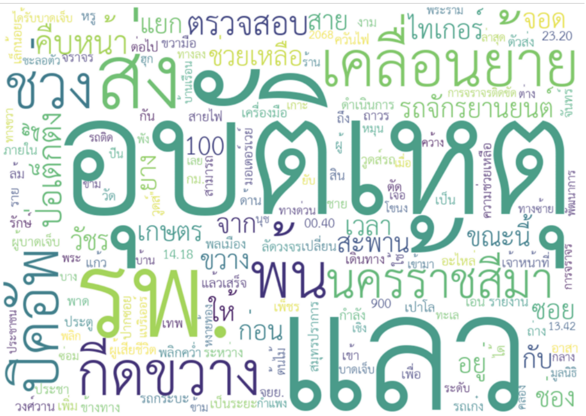


Fig.6: *Word Cloud of Normal Level.*

Fig.7: Word Cloud of Intermediate Level.

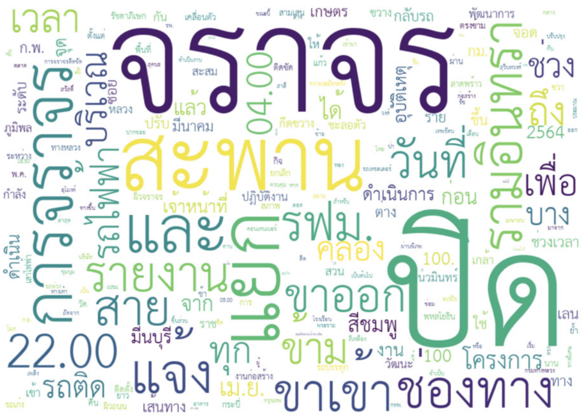


Fig.8: Word Cloud of Lane Closure Level.

3.5 Term Weight Determination

Term weight determination was used to group words by word weight to find semantic representations of text by converting the word or text of interest to a mathematical form, such as a vector of words in text. Using the word2vec techniques, similar vectors mean that the words have similar meanings. For example, accident, parking, and traffic jam have similar vector values, as shown in Figure 9.

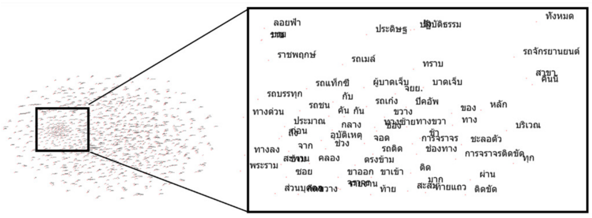


Fig.9: Word Vector Technical.

3.6 Term Probabilities Determination

Term probability determination was used to find the probabilities of link words with the Markov Chain method. The Markov Chain method finds the probability of the next word by matching the next word using the previous word. Which can be described as $S = \{S_1, S_2, S_3, \ldots, S_n\}$ where S is probability of continuous group word and S_n is sequence word as shown in Figure 10. This will find the number of most frequently used words from Table 8 showing the number of most frequently used words.

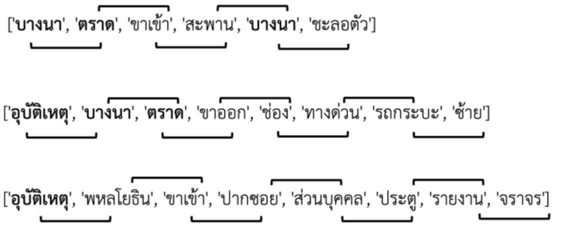

['**บางนา**', '**ตราด**', 'ขาเข้า', 'สะพาน', '**บางนา**', 'ชะลอตัว']
['**อุบัติเหตุ**', '**บางนา**', '**ตราด**', 'ขาออก', 'ช่อง', 'ทางด่วน', 'รถกระบะ', 'ซ้าย']
['**อุบัติเหตุ**', 'พหลโยธิน', 'ขาเข้า', 'ปากซอย', 'ส่วนบุคคล', 'ประตู', 'รายงาน', 'จราจร']

Fig.10: Match to Find the Probability of the Next Word.

Table 8: Severity levels of words.

S	S_1	S_2	S_3	Frequently
บางนา	ตราด	ชะลอตัว	ตราด	3
ตราด	ขาเข้า	ขาออก	-	2
สะพาน	บางนา	-	-	1
อุบัติเหตุ	บางนา	พหลโยธิน	-	2
ขาออก	ช่อง	-	-	1
ทางด่วน	รถกระบะ	-	-	1
รถกระบะ	ซ้าย	-	-	1
พหลโยธิน	ขาเข้า	-	-	1
ขาเข้า	ปากซอย	-	-	1
ปากซอย	ส่วนบุคคล	-	-	1
ประตู	รายงาน	-	-	1
รายงาน	จราจร	-	-	-

3.7 The Process for Selecting the Best Search Terms

The process of determining the best search words uses the best word test results to detect incidence and to distinguish patterns. To choose the best word, this research compares the grouping of words using three different methods. That consist of the method as below.

The TF-IDF method can generate the best possible event to detection term in 6 formats consist of traffic relevant, traffic incident, disaster, potholes, etc.,

The Word2Vec method can generate the best severity level of an incident. consists of Normal and Lane Closure.

The Markov Chain method can generate the best severity level of an incident for one pattern is Intermediate.

The best terms are selected for further use in the process of detecting and distinguishing incidence patterns as detailed in Table 9.

Table 9: *Word Detection Technique Comparison.*

Detection Group	Word Detection Technique (%)		
	TF	Word2vec	Markov Chain
News	**85.50**	85.20	80.34
Traffic	**52.80**	52.40	51.00
Incident	**72.16**	72.00	71.12
Disaster	**88.82**	87.53	54.53
Potholes	**82.10**	75.54	44.85
Etc.	**57.05**	55.96	54.67
Normal	55.80	**61.43**	48.90
Intermediate	67.00	64.21	**68.00**
Lane Closure	52.40	**53.47**	50.54

4. EXPERIMENT

4.1 Hypothesis

The term frequency is tested to determine the optimum word number to be used in Thai language for data clustering according to the data type. This experiment uses 70% of the data for term frequency determination and the remaining 30% of the data for testing. Twenty words from each cluster are then used to generate the processes for word search. The experiment is divided into 3 levels: 1. Determining whether a keyword exists in the string; 2. Keyword selection from 1 to 20 keywords per set. All keywords must be used in a completed round of test; and 3. Keyword selection from 1 to 20 keywords per set to determine condition that provide the most accuracy in each round of test. The accuracy ranges from 10 to 100 percent. The test is done in 3 clusters. Cluster 1 is used to determine the relevance of occurrences. Cluster 2 is used for categorization of occurrences. Cluster 3 is used to determine the degree of severity for occurrences.

4.2 Performance Benchmark

In this experiment, the performance of each method is evaluated by using two parameters which come from the Detection Rate (DR) that measures correction of incident detection. The calculation is shown in equation (1), and the corresponding False Alarm Rate (FAR) is shown in equation (2). D_n is number of incidents detected by this method, D_t is the number of incidents of this experiment and, N_t is the number of incidents in this experiment.

$$DR = \frac{D_n}{D_t} \times 100 \tag{1}$$

$$FAR = \frac{N_f}{D_t} \times 100 \tag{2}$$

4.3 Results

The test is done in 3 clusters where cluster 1 is the determination of the relevance of occurrences. Using the three test methods, Table 9 shows the detection group of relevant occurrences, categories of occurrences, and severity of occurrences. To find the message group that is proper for detecting an incident with a condition, the test collects groups of messages of 20 words for each detection. For the detection process, it does not need to test with all words. To detect and categorize incidents into a message group, this research tests by increasing word every testing round starting with 1 word until 20 words are used in the last testing round. Each test uses the main three methods which are the "some keywords method" (detect the message with 1 word in sentences), "all keywords method" (detect the message with all word in sentence), and "ratio keywords method" (detect the message with word ratio in sentence such as 50% detection in 10 target word in sentence that means 5 words should be detected in this sentence). The evaluation in this research considers all three methods for indicating the highest performance processing auto detection.

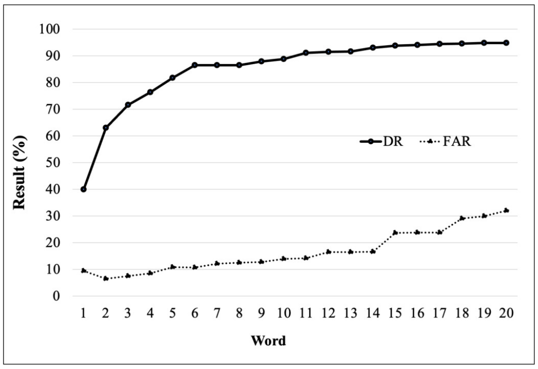


Fig.11: *Incident Detection Result (Some Keywords).*

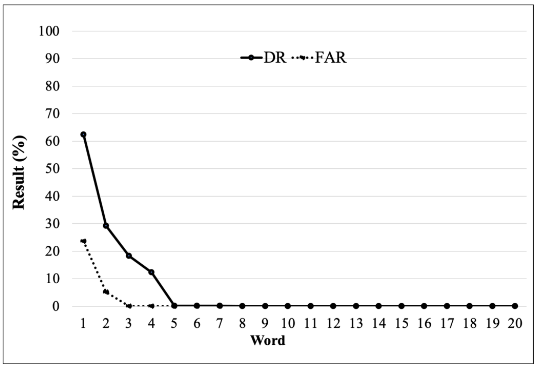


Fig.12: *Incident Detection Result (All Keywords).*

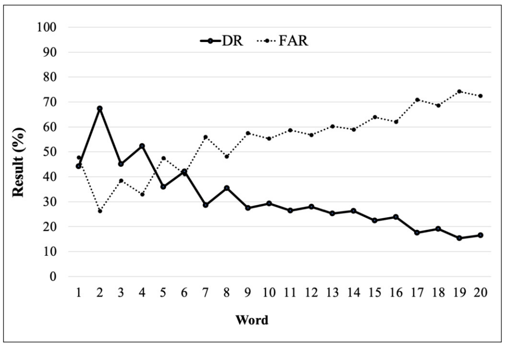


Fig.13: *Incident Detection Result (50% Ratio).*

Cluster 2 is the categorization of occurrences based on 5 main characteristics. The incident detection results from searches where at least 1 term is present are displayed in Figure 11. The number of keywords that result in the most accurate output is 5 as demonstrated in Figure 14. Figures 15-17 show that the optimum term frequency is 5. Figure 18 shows that incident detection results from searches where all terms must be present are significantly less accurate. The traffic status report accuracy is optimal at 3 keywords without the ability to detect or categorize the text string. Figure 19 displays the result of detection with ratio. The optimal result comes from 3 search terms at the ratio of 30%, as show in Figure 20.

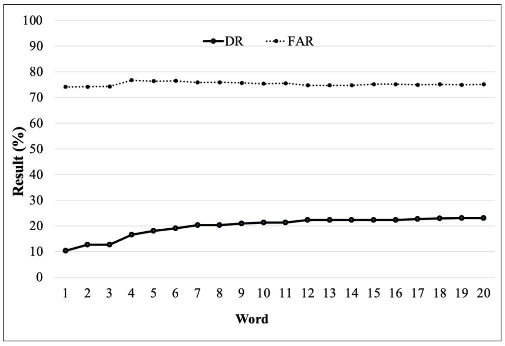


Fig.14: *Categories: Traffic Report (Some Keywords).*

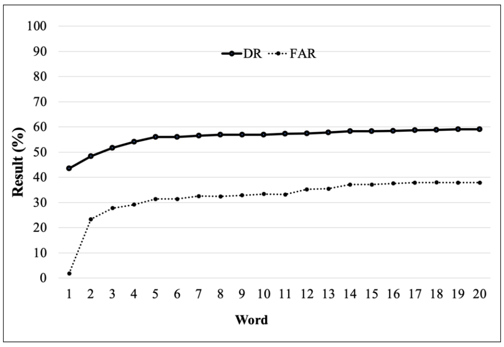


Fig.15: *Categories: Incident Report (Some Keywords).*

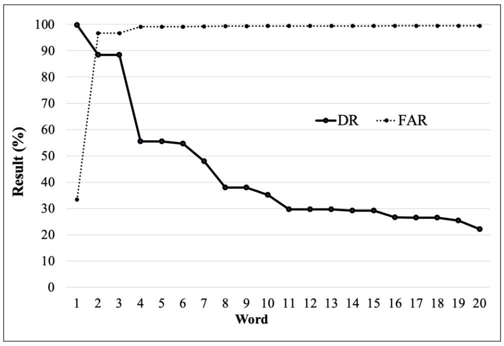


Fig.16: *Categories: Disaster Report (Some Keywords).*

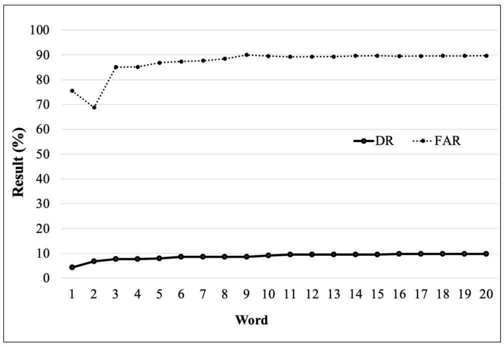


Fig.17: *Categories: Potholes Report (Some Keywords).*

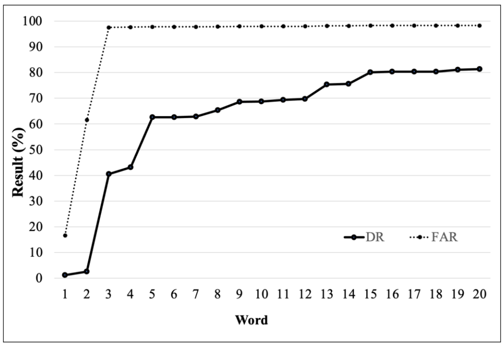


Fig.18: *Categories: Etc (Some Keywords).*

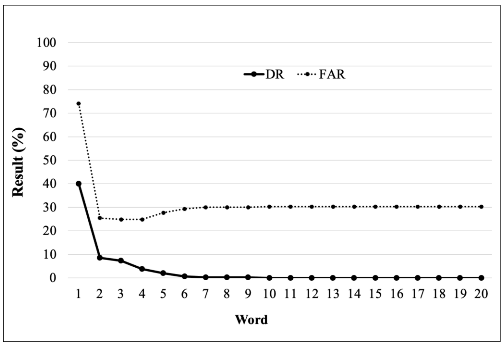


Fig.19: *Incident Categories (All Keywords).*

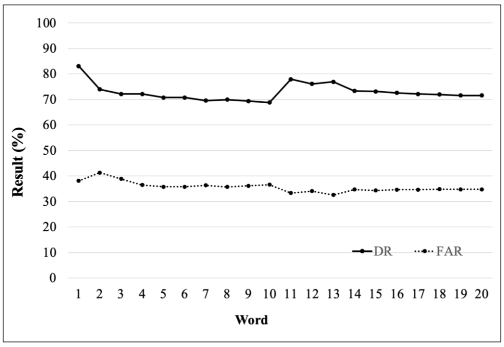


Fig.20: *Incident Categories (30% Ratio Keywords).*

Cluster 3 is the result of the detection of text strings that indicate the severity of occurrences, divided into 3 degrees. Figure 21 shows the result of detection of occurrences that do not impact the traffic with the optimum of 5 keywords. Figure 22 shows the result of detection of occurrences that moderately impact the traffic with the optimum of 5 keywords. Figure 23 shows the result of detection of occurrences that result in total blockage of traffic with the optimum of 5 keywords. Figure 24 shows that the optimum for searches where all terms must be present is also 5 keywords. From the test, the optimal result comes from 8 search terms at the ratio of 30%, as shown in Figure 25.

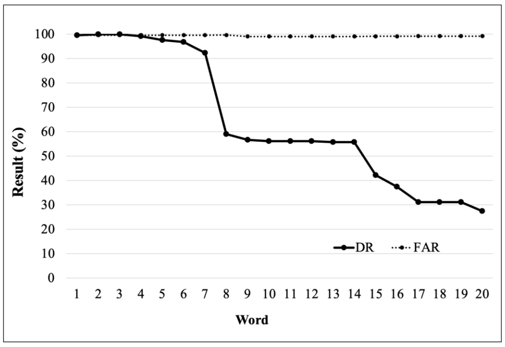


Fig.21: *Incident Severity: Normal (Some Keywords).*

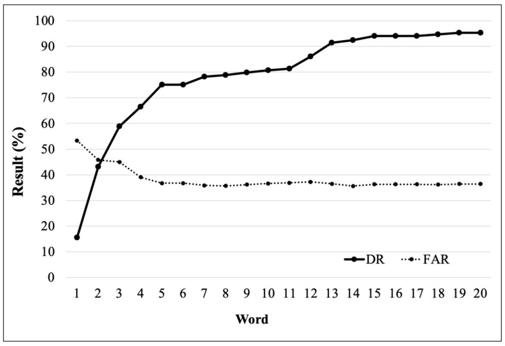


Fig.22: *Incident Severity: Intermediate (Some Keywords).*

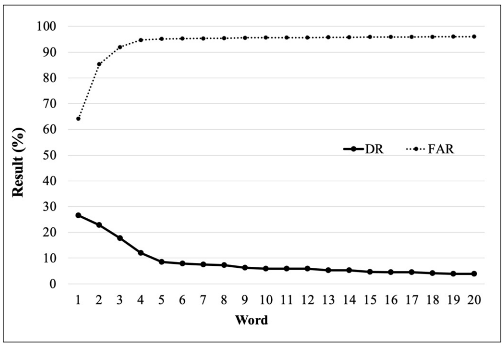


Fig.23: *Incident Severity: Lane Closure (Some Keywords).*

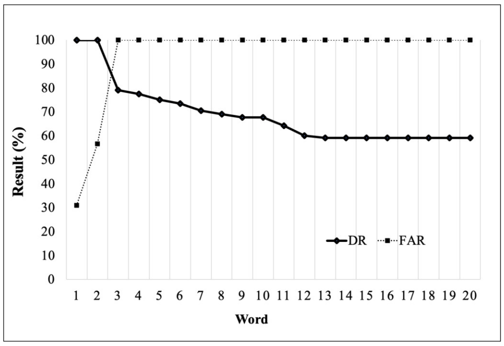


Fig.24: *Incident Severity (All Keywords).*

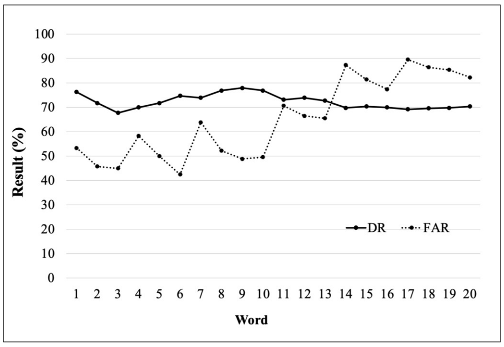


Fig.25: *Incident Severity (30% Ratio Keywords).*

Table 10: *Incident Classification Results.*

Amount of Word	Incident Classification (%)		
	Categories	DR (%)	FAR (%)
1	Traffic Relevant	10.34	74.13
	Traffic	43.54	1.79
	Incident	99.80	33.33
	Disaster	4.37	75.56
	Potholes	1.19	16.67
5	Traffic Relevant	18.09	76.36
	Traffic	56.06	31.39
	Incident	55.47	99.12
	Disaster	7.95	86.89
	Potholes	62.62	97.78

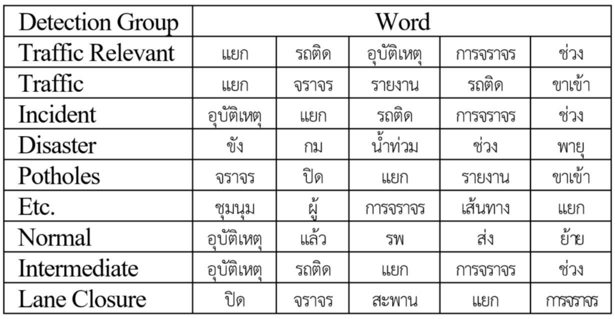
Table 11: *Incident Detection Results.*

Amount of Word	Incident Detection Event	
	DR (%)	FAR (%)
1	39.96	9.46
2	63.02	6.49
3	71.57	7.46
4	76.34	8.57
5	81.71	10.85

Table 12: *Incident Severity Results.*

Categories	Incident Severity		
	Amount of Word	DR (%)	FAR (%)
Normal	1	99.55	99.55
	2	100.00	99.59
	3	100.00	99.60
	4	99.21	99.60
	5	97.65	99.61
Intermediate	1	15.57	53.36
	2	43.10	45.81
	3	58.92	45.00
	4	66.52	39.06
	5	75.06	36.74
Lane Closure	1	26.56	64.10
	2	22.78	85.32
	3	17.78	91.96
	4	11.94	94.74
	5	8.44	95.12

Table 13: *First Five Words for Incident Detection.*



Detection Group	Word				
Traffic Relevant	แยก	รถติด	อุบัติเหตุ	การจราจร	ช่วง
Traffic	แยก	จราจร	รายงาน	รถติด	ขาเข้า
Incident	อุบัติเหตุ	แยก	รถติด	การจราจร	ช่วง
Disaster	ขัง	กม	น้ำท่วม	ช่วง	พายุ
Potholes	จราจร	ปิด	แยก	รายงาน	ขาเข้า
Etc.	ชุมนุม	ผู้	การจราจร	เส้นทาง	แยก
Normal	อุบัติเหตุ	แล้ว	รพ	ส่ง	ย้าย
Intermediate	อุบัติเหตุ	รถติด	แยก	การจราจร	ช่วง
Lane Closure	ปิด	จราจร	สะพาน	แยก	การจราจร

4.4 Analysis of the Results

In the experiment, the 20 most frequent words were taken from a grouping process using various methods. Each method was fed a specified group of words, and produced a specified pattern incidence, a classification of incidence patterns, and the severity of the incident most accurate was employed to test the efficiency of detecting and distinguishing different types of incidence patterns. This experiment confirmed that the messages received from Twitter as incident messages can be used to determine kinds of event and how severity. This is done by checking whether the message contains words that come from a given word group or not. It starts with a test on the condition that if there is only one word in the text, it is considered an event or event pattern according to that word grouping. And then gradually the words checked increases to 2 and 3 words until all words are processed. The test results are shown in Figures 11 to 25. The test results show the effectiveness of the method. Using a wide variety of different number of words. In Tables 10 to 12, from the above test results it can be concluded that using five words from a group of words ranked by frequency can best identify and distinguish the patterns and severity of the incidences, as shown in Table 13.

5. CONCLUSION

This research proposes incident detection through the method searching text messages from Twitter social media by selecting the words used in the search based on word frequencies, using TF-IDF, Word2Vec, and Markov Chain methods. This gives appropriate word groups to find and distinguish events to lead to incidence detection. Using four processes that consist of: 1. Data collection and clustering; 2. Data preparation by filters; 3. Compilation and ordering of keywords related to incidents; and 4. Determination of the appropriate number of search terms for incident detection through social media posts. The search results can be used for analysis and determination of road incidents. The method determines whether they are relevant or not, are of what category (such as traffic slowdown, incidents, disasters, and potholes), and how severe they are based on 3 degrees of severity: normal, intermediate, and lane closure. The determination could be done by determining the frequency of keywords and grouping them in Word Clouds together with comparison of the ratio of appearing keywords. The most accurate setting will result in analysis that can be used to report incidents through a single reporting platform to the public.

The results from the experiment show that the proposed method can use the collected key terms to detect and categorize the terms with a DR of 81.71% and a FAR of 10.85%. This research contributes to the development of natural language processing, especially in regard to road incidents and traffic conditions. It will serve as an important steppingstone for future research that might need to use word clouds for language processing. Fast and effective data mining about road incidents on Twitter will help reduce time for data preparation and analysis for future research.

References

[1] X. Jia, P. Cheng and J. Chen, “A data analysis and visualization system for large-scale e-bike data,” *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3998-4000, 2016.

[2] W. Peng, Y. Li, B. Li and X. Zhu, “An Analysis Platform of Road Traffic Management

System Log Data Based on Distributed Storage and Parallel Computing Techniques," *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)*, pp. 585-589, 2016.

[3] W. Sun, D. Miao, X. Qin and G. Wei, "Characterizing User Mobility from the View of 4G Cellular Network," *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, pp. 34-39, 2016.

[4] X. Wang and Z. Li, "Integrated platform for smart traffic big data," *2016 International Conference on Logistics, Informatics and Service Sciences (LISS)*, pp. 1-6, 2016.

[5] Z. Cui, S. Zhang, K. C. Henrickson and Y. Wang, "New progress of DRIVE Net: An E-science transportation platform for data sharing, visualization, modeling, and analysis," *2016 IEEE International Smart Cities Conference (ISC2)*, pp. 1-2, 2016.

[6] S. V. Nandury and B.A. Begum, "Strategies to handle big data for traffic management in smart cities," *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 356-364, 2016.

[7] D. Chung, X. Rui, D. Min and H. Yeo, "Road traffic big data collision analysis processing framework," *2013 7th International Conference on Application of Information and Communication Technologies*, pp. 1-4, 2013.

[8] I. J. Lee, "Big data processing framework of road traffic collision using distributed CEP," *The 16th Asia-Pacific Network Operations and Management Symposium*, pp. 1-4, 2014.

[9] S. Zhang, "Using Twitter to Enhance Traffic Incident Awareness," *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 2941-2946, 2015.

[10] T. Georgiou, A.E. Abbadi, X. Yan and J. George, "Mining complaints for traffic-jam estimation: A social sensor application," *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 330-335, 2015.

[11] M.V. G. Aziz, A.S. Prihatmanto, D. Henriyan and R. Wijaya, "Design and implementation of natural language processing with syntax and semantic analysis for extract traffic conditions from social media data," *2015 5th IEEE International Conference on System Engineering and Technology (ICSET)*, pp. 43-48, 2015.

[12] Y. Chen, Y. Lv, X. Wang, L. Li and F. -Y. Wang, "Detecting Traffic Information from Social Media Texts with Deep Learning Approaches," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 3049-3058, 2019.

[13] A. Agarwal and D. Toshniwal, "Face off: Travel Habits, Road Conditions and Traffic City Characteristics Bared Using Twitter," in *IEEE Access*, vol. 7, pp. 66536-66552, 2019.

[14] E. Žunić, A. Djedović and D. Donko, "Application of Big Data and text mining methods and technologies in modern business analyzing social networks data about traffic tracking," *2016 XI International Symposium on Telecommunications (BIHTEL)*, Sarajevo, pp. 1-6, 2016.

[15] A. Mulyana, H. Hindersah and A. S. Prihatmanto, "Gamification design of traffic data collection through social reporting," *2015 4th International Conference on Interactive Digital Media (ICIDM)*, pp. 1-4, 2015.

Korn Puangnak received his Master degree from Department of Computer Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. in 2011. Currently, he is an assistant professor at Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Phra Nakhon, Thailand. His research falls in the domain of machine learning, and ITS (intelligent transport system).

Natworapol Rachsiriwatcharabul received his Ph.D. degree from Engineering Management, University of Missouri, USA. in 1999. Currently, he is the President at Rajamangala University of Technology Phra Nakhon, Thailand. His research falls in the domain of energy management, transportation, smart city, and sustainable industrial management.