JEM

Extraction of actionable information from crowdsourced disaster data

Rungsun Kiatpanont, MS
Uthai Tanlamai, PhD
Prabhas Chongstitvatana, PhD

ABSTRACT

Natural disasters cause enormous damage to countries all over the world. To deal with these common problems, different activities are required for disaster management at each phase of the crisis. There are three groups of activities as follows: (1) make sense of the situation and determine how best to deal with it, (2) deploy the necessary resources, and (3) harmonize as many parties as possible, using the most effective communication channels.

Current technological improvements and developments now enable people to act as real-time information sources. As a result, inundation with crowdsourced data poses a real challenge for a disaster manager. The problem is how to extract the valuable information from a gigantic data pool in the shortest possible time so that the information is still useful and actionable. This research proposed an actionable-data-extraction process to deal with the challenge. Twitter was selected as a test case because messages posted on Twitter are publicly available. Hashtag, an easy and very efficient technique, was also used to differentiate information.

A quantitative approach to extract useful information from the tweets was supported and verified by interviews with disaster managers from many leading organizations in Thailand to understand their missions. The information classifications extracted from the collected tweets were first performed manually, and then the tweets were used to train a machine learning algorithm to classify future tweets. One particularly useful, significant, and primary section was the request for help category. The support vector machine algorithm was used to validate the results from the extraction process of 13,696 sample tweets, with over 74 percent accuracy. The results confirmed that the machine learning technique could significantly and practically assist with disaster management by dealing with crowdsourced data.

Key words: disaster management, crowdsourced data, actionable information extraction, machine learning, support vector machine

INTRODUCTION

Natural disasters cause huge losses to humanity. From 1995 to 2015, the UNISDR[1] reported 6,457 worldwide weather-related disasters. These, as recorded by Emergency Events Database (EM-DAT),[2] caused economic losses of more than US$ 1,891 billion. Specifically, there were 3,062 occurrences of disastrous floods, accounting for more than US$ 622 billion of total economic damage during the same 20 years period. More than 157,000 people lost their lives and 2.3 billion were affected.

However, the numbers above are only the tip of the iceberg, as they reflect the minimum estimation based on reports gathered from most nations. Statistics indicated many cases of underreporting, especially in low-income countries. Furthermore, in terms of economic damage, EM-DAT records cover only basic economic impacts, including homes and infrastructure damaged and destroyed. Many other economic costs are not accounted for due to quantifiable difficulties. For example, costs of repairs, rehabilitation, rebuilding, lost productivity, and increased poverty are hard to estimate.

Undoubtedly, measures to reduce disaster risk and protect people, property, and society are urgently and critically required. Traditional disaster management

DOI:10.5055/jem.2016.0302

defines the framework of activities for dealing with crisis into four phases: mitigation, preparation, response, and recovery,[3] with different kinds of activity required for each stage. For example, before a possible disaster, planning activities are required to reduce the risk of disaster occurrence (if possible) or reduce the damages from a known hazard. On the other hand, during a disaster, response activities focus on saving lives, seeking shelter, and preventing property damage. The activities following disasters concentrate on returning the situation to normal or providing financial assistance to affected people. Last but not least, activities to extract lessons learned from previous catastrophes to prevent future disasters or minimize their impacts are included in the mitigation phase.

Jennex[4] presented an interesting phenomenon regarding the different levels of activity required for each stage of disaster management. He stated that the response phase was clearly the most challenging period for a disaster manager because it required a vast number of activities performed under extreme time pressure. Also, Drabek and McEntire[5] and Helsloot and Ruitenberg[6] studied the behavioral aspects of citizens during the response phase. They looked at how people responded to disasters, including the occurrence of situational altruism.

Ashish et al.[7] defined four subphases within the response phase: (1) damage assessment, (2) needs assessment, (3) prioritization of response measures, and (4) organizational response. These four subphases work in a cycle. Situation or damage assessment is clearly a critical activity for all other activities that follow. The result of situation assessment is called "situation awareness." Endsley[8] defined situation awareness as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future." In other words, situation awareness is a process of gathering and processing information provided by various sensors in the affected area(s) in a timely manner. The latest Sendai framework,[9] published by the United Nations, also highlighted this concept by rating it as the number one priority for action. Undoubtedly, the key enablers of this activity are data science and information technologies.

In contrast with traditional disaster management, rather than focusing only on activities performed by disaster managers, the latest Sendai framework, representing a contemporary disaster management framework, also highlighted an all-of-society engagement and partnership. Hence, crowdsourcing is a potential technology to achieve situation awareness in the context of disaster by using bottom-up approach.[10-13] However, crowdsourced data contain only a relatively small portion of actionable information. Time and effort are required to process the mountain of data, and this remains the key challenge for practically adopting the crowdsourcing concept.

This article aimed, therefore, to propose a design to overcome the problem of resource requirements for processing information during a disaster. The proposed framework employed machine learning techniques to automate the actionable information extraction. This article comprises five sections. First, we refer to related research in the literature review. Cases of practical use for disaster information are then discussed, followed by an explanation of the research methodology. The results of the automated data classification are then presented, and finally, suggestions are made for future research and the results are discussed.

LITERATURE REVIEW

Disaster management

In addition to the four phases of disaster management, Lettieri et al.[14] conducted a systematic literature review. They found that there were two main approaches to manage disasters, the resistance approach and the resilience approach. Briefly, the resistance approach focuses on the period before the disaster occurs. It encourages activities to mitigate risks (if possible), identifies resistance to societal vulnerability, and also creates the preparedness of the citizens. For instance, disaster managers have to educate people to prepare food supplies and survival kits to ensure that they survive for the first 72 hours after disaster strikes. Specifically, groups who have special needs (ie, the elderly, children, people with disabilities, foreigners, or minority groups) should also be identified and taken care of separately.[15]

The resilience approach, on the other hand, focuses on the aftermath of disasters.[16] This approach entails activities that minimize the impacts of a hazard. These include issuing evacuation orders, creating evacuation centers, and the deployment of search and rescue teams. Requests for vital help such as food, water, and shelter are other good examples of actionable information in this approach. Disaster managers need to know who needs help, what kind of support is required, and where it is most needed. Limited information, time constraints, and decision load constraints are identified as key challenges in this approach.[17] Interorganizational collaboration, community involvement, and resource management are proposed as the tools required to deal with the challenges effectively.[18] The resilience approach can also be proposed as a framework to define community resilience as a set of networked adaptive capacities which are social capital, economic development, information and communication, and community competence.[16]

On the other hand, the UNDAC handbook[19] provides guidelines on how UNDAC team members should perform related activities during disasters; for instance, disaster assessment, coordination, and information management. Also, the Sendai framework represents the present disaster management paradigm in the global context.[9] It defines seven global targets and four priorities for action as the guiding direction for all nations. Besides, in addition to traditional disaster management organization, both FEMA[20] and the Sendai framework also highlight an all-of-society engagement and partnership, including both community-based organizations and nongovernmental organizations. In other words, contemporary disaster management needs to enhance collaboration among local people, especially for disseminating disaster risk information.

Regarding information and communication during disasters, Jaeger et al.[21] defined four typologies

of communication structure during disasters: (1) many-to-one, (2) one-to-many, (3) many-to-many, and (4) one-to-one. For instance, affected people calling a central hotline number is an example of many-to-one communication. One-to-many includes any announcement through mass media from the government or a centralized command center. Discussions on a Web board or wiki are considered as the many-to-many communication type, and a direct call to one's relatives or neighbors to confirm the well-being of one another is an example of the one-to-one-type.

Vieweg[22] summarized 34 types of messages contained in 37,000 tweets, from four major disasters. By following a different approach, she found that messages about the vital lines, service availability, and emergency issues were the most popular types of messages during disasters. She also discovered that the lack of appropriate communication channels could worsen the situation. For instance, many violence cases after the Haiti earthquake were caused by hunger, thirst, and those who required immediate aid. The affected people needed to communicate and escalated their problems to get attention from the public.

Crowdsourcing as an enabler for crisis informatics

Without situation awareness, disaster response is almost impossible. As a result, the steps of information acquisition and information management during disasters are included in an emerging area of study called crisis informatics.[23] In recent years, many comprehensive literature reviews have been undertaken regarding crowdsourcing as a critical role in crisis informatics tasks.[11,24-27] Crowdsourcing also provides new possibilities for people who are not physically present in the affected area, to make a real difference to the disaster situation in a country in another part of the world.

However, one of the challenges of crowdsourcing for real-world implementation is articulation work highlighted by Liu[24] as invisible coordination and negotiation activities necessary to get the work done. Consequently, Liu proposed a way to reduce this articulation work by aligning each part of crowdsourcing work based on a well-defined structure.

A well-defined structure is necessary for effective collaboration. In the early days, after the 2010 Haiti earthquake, for example, international response teams experienced difficulties in accessing first-hand information and intelligence from the local community, simply because their systems were not structured in a way to utilize the inputs from local people.[10] As a result, without a well-defined structure and supporting technologies, responding to individual messages generated by a crowd is not an easy task.

To address crowdsourcing in a crisis situation, Liu[24] defined four possible tasks of crisis crowdsourcing: (1) crowd-sensing (gathering data), (2) crowd-tagging (classifying data), (3) crowd-mapping (finding the location for each report), and (4) crowd-curating (improving data quality via filtering, verifying, synthesizing, and exhibiting). Crowd-sensing is similar to such concepts as citizen sensing or participatory sensing. The idea is for humans, equipped with mobile devices containing multiple sensors, to report observations of disaster events in near-real-time.[11] Crowd-sensing has been realized during the past few years with the use of Twitter, a popular communication tool used during disasters. Twitter is an efficient tool because it has a concise nature with a readily available public and timely information.

An example of crowd-tagging is a study by Vieweg.[22] She used an inductive approach to identify 34 message categories contributing to situational awareness, based on 37,802 tweets from four disaster events. Regarding mission 4636 in 2010 Haiti earthquake as another example, thousands of volunteers translated and categorized the short message service reports sent for free to hotline number 4636. Nevertheless, some researchers considered the hashtag of the Twitter message as a technique for crowd-tagging the data sources themselves.[28]

In addition to general message tags, crowd-mapping by tagging a particular location mean by the messages is also vital to disaster management. For example, volunteered geographic information is an approach for geotagging, based on the crowd-mapping technique.[12,29] Another approach on geotagging is the automatic location identification in the text messages, based on machine learning algorithms.[30-32]

The last crowdsourcing task is crowd-curating. According to Liu's definition, this refers to a set of

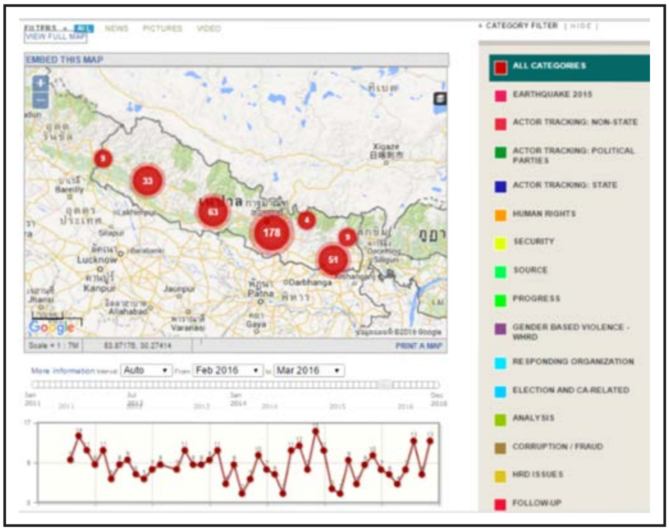


Figure 1. User interface of Ushahidi platform.[37]

activities consisting of filtering, verifying, synthesizing, and exhibiting a curated collection of data. While crowdsourcing provides new possibilities to collect extensive data within a short period, the trustworthiness of crowdsourced data has always been a concern. Weaver et al.[33] proposed three strategies to improve the reliability of crowdsourced messages. They suggested the use of group membership, vote, and machine learning algorithm. On the other hand, Ushahidi,[34] one of the most practical and proven platforms in many disaster events,[10,12,35,36] uses maps to exhibit disaster. Figure 1 shows its user interface on a map overlay with supporting data filtered via category, time, area, and information type. In summary, crowd-curating activities focus on how to improve data quality to meet the expectations of data consumers. As a result, these activities require the understanding of the different missions of each disaster responder.

ICT as an enabler for crowdsourcing

Undoubtedly, crowdsourcing provides possibilities to deal with disaster situations through the leveraging powers of the crowd. However, it also comes with some challenges. Information overload, noise, misinformation, bias, and trust are examples of major concerns for decision making based on crowdsourced data.[11] Similarly, believability, amounts of information, and relevancy are also factors affecting system architecture as stated by Hale.[38] For instance, Vieweg[22] found that using only keywords to extract information from Twitter during disasters, more than 80 percent of the extracted tweets were irrelevant or off-topic. As a result, it becomes impractical to let disaster managers to deal with this garbage by themselves.

ICT plays a crucial role regarding how to enable and implement crowdsourcing, and also deal with the challenges. The roles of ICT cover not only data visualization but also data collection and data processing. In academic fields, many technologies have been proposed to meet the challenges of crowdsourcing implementation.

Palen et al.[39] proposed a system architecture to gather and process information during disasters. Their structure added more layers to the well-known three tiers architecture: presentation layer (so-called visualization), logic tier (Web applications and service), and data layer (repository). The additional layers included the integration layer, natural language processing (NLP) layer, and trust layer. The integration layer is used to deal with limited information constraints for both quantity and quality. The NLP layer is used to partially (if not totally) automate information processing tasks and reduce information processing time requirements. Finally, the trust layer is used to improve the quality of the information.

Social media platforms (eg, Facebook or Twitter) allow new ways of communication. The potentials of using these platforms in risk and crisis communication have recently been focused from both researchers[40-44] and international organizations (eg, Organisation for Economic Co-operation and Development [OECD],[45] IRGC,[46] UNISDR,[47] and UNAPCICT[48]). In principle, as shown in Table 1, twelve good practices and 10 challenges in the use of social media in risk and crisis communication have been identified as a result of a joint meeting on June 2012 between The International Risk Governance Council and 12 OECD countries.[45]

Findings from researchers confirmed that these items being valid. For instance, Vieweg et al.[49] found that social media significantly contribute to enhancing situational awareness of the whole community. Neubig[50] illustrated how social media has been practically used to identify survivors and victims during

Table 1. Good practices and challenges in the use of social media in risk and crisis communication

Good practices	
1	Raising public awareness
2	Monitoring situation awareness
3	Improving preparedness
4	Providing information and warning
5	Mobilizing volunteers
6	Identifying survivors and victims
7	Managing reputational effects
8	Collecting funding and support
9	Learning from the crisis ex post
10	Improving partnerships and cooperation
11	Building trust
12	Enhancing recovery management
Challenges	
1	Multiple players and communication channels
2	Transparency and reliability
3	Image damage
4	Keeping in touch with all population segments.
5	Avoiding the information overload
6	Promoting open data while protecting privacy and confidentiality
7	The question of liability
8	Managing public expectations
9	Addressing security issues in a globalized context
10	Assessing the impact of the social media

2011 Japan Earthquake. Wukich[26] used content analysis of more than 80 research articles and report to reconfirm three strategic usages of social media for disaster management: information dissemination, monitoring situation awareness, and engaging the public in a conversation and/or crowdsourcing. For example, the hashtag mechanism (eg, #thaiflood) spontaneously built communities of people sharing similar concerns, regardless of their physical location. On the other hand, these social networks also allow disaster managers to collect the wisdom of the crowd for disasters occurring in their community.

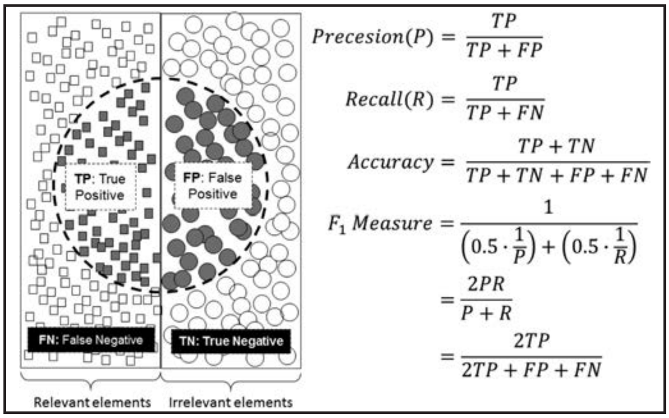
In addition to the power of the crowd provided by crowdsourcing, computational power and machine learning algorithms also play vital roles for disaster data processing. They provide the possibilities for automated data processing. For example, support vector machine (SVM), conditional random fields (CRF), and Naïve Bayesian classifiers have been used to classify tweets into predefined categories.[50-52] Also, the named entity recognition (NER) technique has been used to recognize names of people and places from Twitter.[50,53] Moreover, N-gram approximate matching, NER, and CRF have also been used for automated location detection.

Text classification

As mentioned earlier, one of the most critical steps for disaster data processing is to extract useful information from the multitude of crowdsourced data. Another important process is to discriminate data into predefined categories that may require different data treatments. In theory, these steps could be referred as document classification problems. Sebastiani[54] defined document classification as a process to label documents with some thematic categories as representative of their contents. Single-label tags only one category to a particular document. On the other hand, multilabel (overlapping categories) refers to the case where more than one category can be assigned to the document.

Many machine learning approaches have been used to implement text classifiers. Supervised techniques use a group of algorithms to let machines learn from training data to build the automated classifiers as probabilistic classifiers (eg, Naïve Bayes), Decision Tree (eg, ID3, C4.5, and C5), Artificial Neural Network, SVM (eg, LibLinear and sequential minimal optimization [SMO]), and K-nearest neighbors (K-NN).

There is no absolute answer as to which classifier is the best in all cases, as their performance comparisons are reliable only when based on experiments with the same setting, and under carefully controlled conditions. However, Kotsiantis[55] discussed their highlighted



$$Precesion(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F_1\ Measure = \frac{1}{\left(0.5 \cdot \frac{1}{P}\right) + \left(0.5 \cdot \frac{1}{R}\right)} = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}$$

Figure 2. Performance measurement of classifiers.

points as follows. Briefly, SVM has the best general accuracy through using large training data, while Naïve Bayes needs a relatively smaller training set. On the other hand, while Artificial Neural Network and SVM are the best for nonlinear problems, they are poor in terms of learning speed and interpretability of the classifier models compared to Naïve Bayes and Decision Tree. K-NN has a fast learning speed, but it is also very sensitive to noise in the data set.

The effectiveness of these classifiers can be formally measured by calculating the precision (P) and recall (R).[56] Precision (or confidence) tells how many selected items are relevant, while recall (or sensitivity) deals with how many relevant items are selected. On the other hand, accuracy is another simple measurement by summing up the correct predictions for both positive and negative sides and dividing by the number of all samples. Even though this calculation is simple and easy, it is not recognized as a reliable performance measure because of a known problem called accuracy paradox.[57] The F_1-measure, as the equally weighted harmonic mean between P and R, has gained more popularity for measuring the overall performance of predictions. Figure 2 illustrates their definitions.

INFORMATION USE CASES OF 2011 THAILAND FLOOD

Even though Table 1 could imply future direction regarding how social media would be used for disaster management, it does not reflect the current status how it has been used so far, especially in those non-OECD countries, like Thailand. With hardly any restrictions on social media deployment, Thailand is heaven for all social media platforms. For instance, on 2015, Thailand had more than 3.4 million Twitter user and 35 million[58] Facebook users (50.58 percent of Thailand population). As a result, people from all walks of life spend their waking hours on social media sites. Businesses and nonprofit organizations also use social media in every possible application. More than 350,000 tweets with hashtag #thaiflood during 2011 Thailand flood was a good example how much social medias are used for disaster management in Thailand. We therefore conducted a series of in-depth interviews with five leading disaster management organizations in Thailand to understand how disaster managers in developing countries gathered, used, and managed relevant data during disasters. As the largest flood in term of economic damage,[59] the 2011 Thailand flood has been chosen as a case study during interviews. Those five leading disaster management organizations are (1) Department of Disaster Prevention and Mitigation (DPM), (2) Pohtecktung Foundation, (3) Chaipattana Foundation, (4) #thaiflood working group, and (5) Ruam Duay Chuay Kan community radio station.

From the interviews, we learned that different organizations have different missions and focus on different things during disasters. For example, as a national government agency, responsible for all disaster-related topics at the national level, the critical tasks of DPM are to manage related information and produce disaster-related reports for other government agencies, so that they can make decisions on how to deploy resources into affected areas. These agencies are responsible for responding to an immediate disaster situation and the aftermath recovery. The other two organizations Pohtecktung and Ruam Duay Chuay Kan, on the other hand, put their effort into requests for help from affected areas because their primary mission is to distribute donations to the needy. Unlike the previous cases, the Chaipattana Foundation approach is to set up evacuation centers within the affected area. The organization believes that people prefer to live close to their hometown. The #thaiflood working group concentrates their workforces to manage information from the cloud, establishing communication channels, and manually consolidating information from many sources.

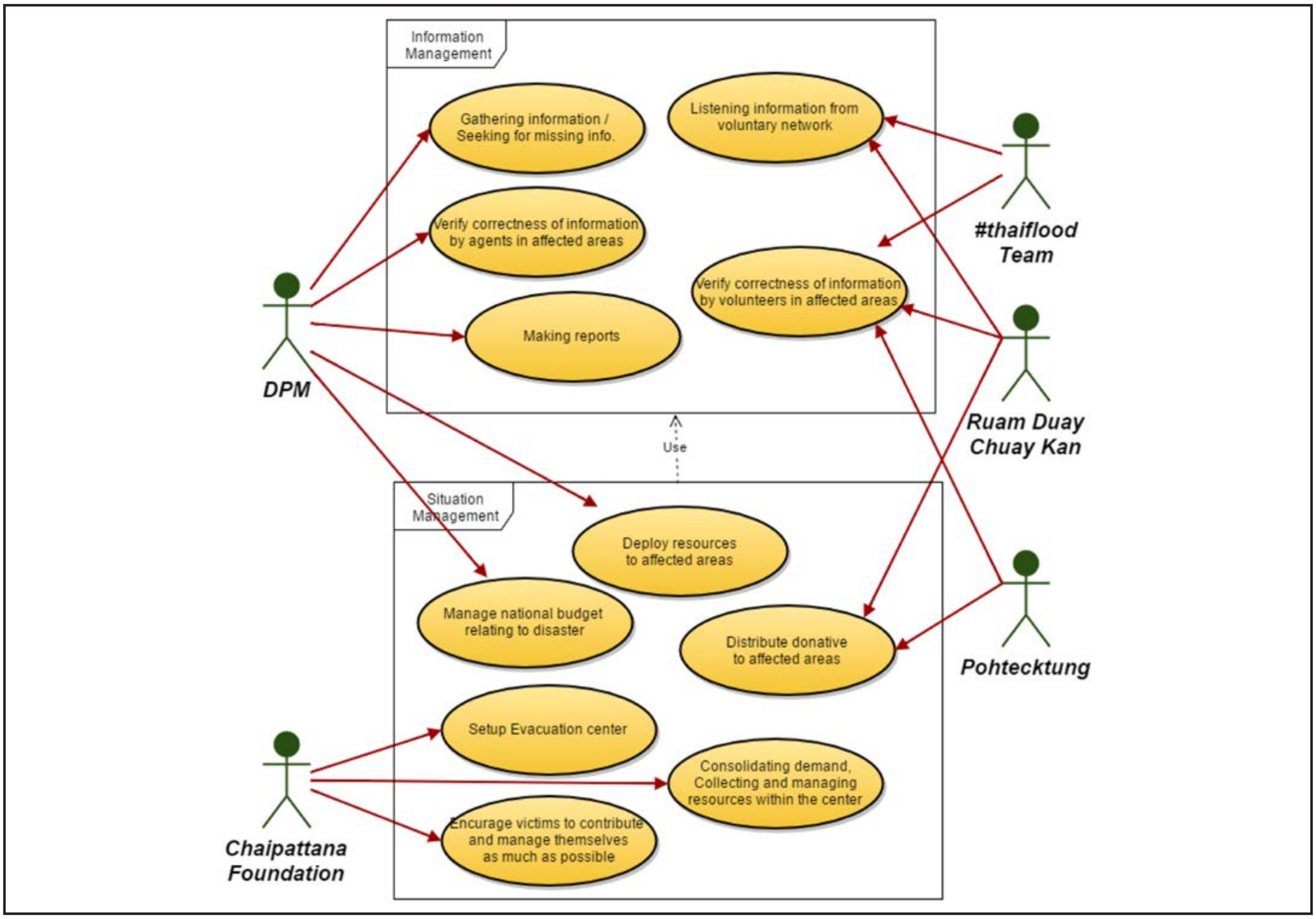

Information Management
Gathering information / Seeking for missing info.
Listening information from voluntary network
Verify correctness of information by agents in affected areas
Verify correctness of information by volunteers in affected areas
Making reports
DPM
#thaiflood Team
Ruam Duay Chuay Kan
Use
Situation Management
Deploy resources to affected areas
Manage national budget relating to disaster
Distribute donative to affected areas
Pohtecktung
Setup Evacuation center
Consolidating demand, Collecting and managing resources within the center
Encurage victims to contribute and manage themselves as much as possible
Chaipattana Foundation

Figure 3. Missions of disaster managers represented in use case diagram.

Despite the fact that different disaster management organizations have different missions, we found some commonality among their requirements. A use case diagram (Figure 3) demonstrates how these five organizations share their missions. Briefly, there are two groups of tasks, ie, information management and situation management. Information management is required to understand situation awareness. Situation management reflects how disaster managers deal with disasters in the affected area. In other words, their common mission is to deploy resources and services to the people with the most needs.

However, excepting thaiflood team who heavily used social media for their missions, usage of other organizations is limited only to providing information and warning. We realized from interviews that other social media usages were not the case because they required active engagement which those managers were not ready to afford. For example, we have been confirmed that it was practically impossible to let disaster managers deal manually with the large number of messages from social media, as very little actionable information was found in the large amount of garbage data.

In general, the term actionable should be identified by responders as to whether the messages could lead to any actions or not. Next, we developed 11 message categories from the seven message groups based on previous work by Vieweg,[22] and newly added four more categories to align with the real sampling from collected tweets. Table 2 illustrates their definition and example in detail. Moreover, as shown in Figure 3, deployment of resources and services to affected areas is one of the most critical missions for most responders.

Table 2. Message categories with examples

Category	Description	Example (translated from original tweets in Thai)
AI	Advice for information space	(recommend) list of all important numbers for help during flood bit.ly/pgJps0 #Thaiflood
D	Damage + status of hazard + general hazard information + weather	[1 NOV 18:52] 1 meter water level at Settakit village bit.ly/uDE43r #thaiflood
E	Evacuation	RT @Mayuree_NT: TU Rangsit evacuation center currently help 1031 victims. Remaining capacity is for 669 victims. They also plan to open gymnasium tower to increase their capacity of 1500 people. #thaiflood
O	Offer of help	(19:40) Who need 400 units of the packed meal? Please contact Tanthai 081-116-6899 #thaiflood #siamarsa
Q	Request for help + feeding/hydration + medical attention	Water level at Laharn temple, Bangbuatong, Nontaburi Province is at head height. Monks in the temple have nothing to eat. Please help to donate meals to the temple. #ThaiFlood
RF	Response—Formal	The Prime Minister commands relevant units to build another floodway by using Latpo cannal as the reference model. bit.ly/ov0kct #thaiflood
RC	Response—Community + personal	[13 Oct 18:48] Lardkrabang industrial park builds water barrier surround the area goo.gl/x1yTy #Thaiflood
Additional categories		
AH	Advice—How to	[1 Dec 11:04] Ideas to check electricity in the water on.fb.me/uV0Vqe #thaiflood
M	Volunteer mobilization + resource requests by relief operations centers	★ Urgent!!! Water level at Siriraj Hospital is critical ★ (tonight) Need more volunteers to build sandbags and cement wall. For more detail, please see ow.ly/77Tix #ThaiFlood #SiamArsa
C	Contextual messages (not directly relevance from disaster respond perspective)	[29 Oct 22:48] Floods in Thailand cause significant impact to the computer industry. Global hard drive supply is a shortage. bit.ly/v9Gura #thaiflood
X	Excluded messages for non-Thai tweets	(translated by google from a Japanese tweet) It feels a great flood prediction of severe of Bangkok. Thailand's flood-related hashtags # Thailand flood English #thaiflood

We, therefore, considered message category Q (request for help) as an actionable information in this study.

METHODOLOGY

Sampling frame of crowdsourced data

To understand the nature of information during disasters, the event of severe flooding that occurred during 2011 in Thailand was selected as a case study. Twitter was chosen as the data source because it represents near-real-time information during disasters, and Twitter messages are pubic to all people by default. Next, all tweets during the flooding period July 25, 2011 to January 16, 2012 were collected. To differentiate relevant messages from irrelevant ones, we used the hashtag technique. The hashtag #thaiflood was selected since it had been heavily used during the 2011 Thailand flood disaster. We programmatically acquired all the old related tweets from the Web site Topsy.com using a customized iMacros script, a web testing automation tool. As a result, more than 350,000 relevant tweets were collected with the processing time of almost 2 months.

Table 3. Results of manual classification

Category	Description	Number of tweets*	Percent
D	Damage + status of hazard + general hazard information + weather	3,617	26.09
M	Volunteer mobilization + Resource requests by relief operations centers	2,494	17.99
RC	Response – community + personal	2,477	17.86
AH	Advice—How to	984	7.10
AI	Advice—Information space	944	6.81
C	Contextual messages (not directly relevant from disaster response perspective)	913	6.58
RF	Response—Formal	791	5.70
O	Offer of help	667	4.81
E	Evacuation	442	3.19
Q	Request for Help + feeding/hydration + medical attention	367	2.65
X	Excluded messages for non-Thai tweets	170	1.23
	Total	13,866	100

*Descending sorted.

Data processing process

The machine learning approach for supervised text classification follows three main steps: (1) training data preparation, (2) classifier learning, and (3) classifier evaluation and optimization. In this research, the data source was solely crowdsourced from Twitter and contained a lot of duplicated messages. As a result, before formatting training data for learners, data preparation also included previous steps of data deduplication and manual classification of collected tweets. Next, the training data were tested against four off-the-shelf classifiers in Weka,[60] a machine learning toolbox, ie, Naïve Bayes, Decision Tree, SVM, and k-NN. The most accurate classifier was selected to be validated and optimized in the final step.

RESULTS

Data deduplication

The number of collected tweets with hashtag #thaiflood was 353,714. These were posted by 48,646 Twitter users. Unsurprisingly, as further evidence of the Pareto principle, 80 percent of all tweets were posted by around 20 percent of Twitter users. Furthermore, 99.4 percent of collected tweets were in Thai, 0.5 percent were in English, and 0.1 percent were in Japanese and Korean.

We found a lot of duplicated messages in the collected data. This occurred mainly because Twitter allows users to retweet messages very easily. By using a simple exact match technique in Microsoft Excel, the original 352,714 tweets were reduced to 131,493. In other words, the number of messages was reduced to 37 percent of all tweets.

Next, to further decrease the number of messages to be processed, we used the fuzzy lookup plug-in to find all similar texts. Empirical data indicated that a 90 percent similarity was a good threshold to use in this setting, and any messages with more than 90 percent similarity were semantically matched. As a result, the number of unique tweets was reduced from 37 to 30 percent.

Manual classification

Based on the duplicated tweets from the previous step, we selected 13,866 messages as samplings from the 131,493 tweets based on the higher number of the retweets. Next, we manually classified these samplings as single-label classifications, into the 11 categories defined in the previous step. Table 3 shows the results of manual classification, sorted by the number of messages.

Table 4. Classifier performance measurement				
Algorithm	Accuracy, percent	P	R	F_1-measure
Naïve Bayes	60.94	0.628	0.609	0.61
SVM	70.68	0.706	0.707*	0.705*
Decision tree	60.24	0.593	0.602	0.596
kNN	47.85	0.634	0.478	0.51

*Best classifier.

Best classifier identification

By excluding category X, non-Thai messages, we used all the remaining results of 13,696 tweets marked in 10 categories from manual classification as training data for all four classification algorithms. Next, we tested their performance, using a 10-fold cross-validation technique. At a glance, SVM showed the best result of all four classifiers with the highest F_1-measure value. However, the extracted messages could be very critical, especially in the case of vital requests. It was therefore important to minimize the case of false negatives, where the important messages might not be detected. From the recall definition, minimizing the FN implied maximizing the TP, which effectively meant maximizing the recall (R). By considering the recall, Table 4 reconfirmed that SVM was the best classifier in this case.

SVM classifier optimization

As there was no particular technique to guarantee optimal solutions, our best attempt identified a relative optimal classifier, based on the trial and error approach. We found the relative optimization by varying each relevant parameter one at a time. For example, while keeping other settings the same, we found that using SMO[61] gave a better result than LibSVM.[62] By changing kernels, we found that a normalized polynomial kernel[63] provided the best outcome. Using a histogram technique to filter these noise features, we found that the optimal feature vector contained those words with at least seven occurrences out of the 13,677 tweets. Next, we combined all conditions above to get the best possible classifier. The performance of the optimized SVM classifier is shown in Table 5.

Table 5. Precision, recall, F-measure, and accuracy of the model						
Category	Precision	Recall	F_1-measure	Number of tweets on train set		Number of tweets on test set
				Preclassified	Classified results	Classified results
AH	0.825	0.673	0.741	984	802	6,029
AI	0.74	0.636	0.684	944	811	6,485
C	0.659	0.47	0.549	913	651	8,878
D	0.785	0.891	0.835	3,617	4,107	46,586
E	0.668	0.706	0.686	442	467	3,874
M	0.892	0.911	0.902	2,494	2,548	16,314
O	0.766	0.607	0.677	667	529	4,465
Q*	0.849*	0.706*	0.771	367*	305*	2,828*
RC	0.564	0.679	0.616	2,477	2,984	4,674
RF	0.785	0.488	0.602	791	492	31,360
Weighted Avg.	0.753	0.747	0.744	Total = 13,696		Total = 131,493

*Focused category.

Moreover, we also validated the classifier by supplying the whole 131,493 tweets as the test data. The classified result of test data is also shown in Table 5. Next, these 2,828 messages with predicted result as category "Q" were manually reviewed as to whether their predicted result was reliable or not. In consequence, we found that 91.62 percent of these messages were correctly classified.

DISCUSSION

As mentioned before, disaster response is the most challenging phase for a disaster manager, because of the vast number of activities to be performed under extreme time pressure. Crowdsourcing provides a new possibility to collect real-time fact-on-the-ground for situation awareness. However, information overload is a key challenge in employing this new concept.

The results showed that the proposed process could significantly help disaster managers to deal with the challenges. By considering the message category Q—request for help—in Table 5 as an example, without this automated classifier, manually identifying the 2,828 tweets of category Q out of 131,493 tweets is a tedious task, and impossible during disaster response. Undoubtedly, this could significantly help disaster managers in dealing with only 2,828 extracted tweets of category Q, rather than the 131,493 semiprocessed tweets or the entire 352,714 original tweets.

Also, regarding the trained classifier, the precision values implied that 84.9 percent of those 2,828 tweets were correctly classified as category Q. This was reconfirmed by 91.62 percent correctness as mentioned earlier. On the other hand, a recall value of 70.6 percent of category Q implied that the correctly classified tweets within 2,828 tweets represented 70.6 percent of all actual tweets with category Q, and there was therefore 29.4 percent of category Q which was undetected. This number implied that the system was still not perfect and there was room for future improvement.

This could be done in many ways. First, other sets of parameters could be tested to determine whether there is any better relative optimization. Second, the NER technique could be used to label the location mean by the messages, significantly helping the crowd-mapping task in the next step. Next, the trustworthiness of the messages could be improved, using two-way communication technologies such as Twitter reply or Twitter polls. Finally, the visualization of these messages should allow leveraging of the synergy of the message categories. For example, to respond to the requests for help (Q), the system should be able to suggest who is offering related resources and services.

CONCLUSIONS

Statistical data show that humanity still has a tendency to experience big losses from natural disasters. Traditional disaster management focuses on the role of the disaster managers dealing with the situation. This contemporary concept of disaster management highlights more on community engagement for dealing with disasters.

The technological developments in recent years have provided new possibilities to allow people on the ground to contribute their knowledge and resources for improving the situation. Particularly, ICT technologies are key enablers to realize crowdsourcing concepts. Also, machine learning technologies are playing key roles here, because they provide the possibility of automated crowdsourced data processing.

By focusing on the real problems of the disaster to get useful crowdsourced data, we proposed an actionable-data-extraction process to overcome the challenge of information overload. Next, based on indepth interviews, reviews of the literature, and actual tweets, 11 message categories were defined. We used the SVM technique to realize the automated classifier. As a result, the classifier produced an impressive result with average precision and recall at over 74 percent. This new methodology can significantly help disaster management personnel to deal with crowdsourced data.

Rungsun Kiatpanont, MS, Technopreneurship and Innovation Management Program, Chulalongkorn University, Bangkok, Thailand.

Uthai Tanlamai, PhD, Department of Accountancy, Chulalongkorn Business School, Chulalongkorn University, Bangkok, Thailand.

Prabhas Chongstitvatana, PhD, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand.

REFERENCES

1. UNISDR: *The Human Cost of Weather-Related Disasters 1995-2015*. Vol 53. Geneva, Switzerland: UNISDR, 2015.

2. CRED: EM-DAT | The international disasters database website, 2016. Available at *http://www.emdat.be/*. Accessed March 2, 2016.

3. FEMA-EMI: The four phases of emergency management, 2012. Available at *training.fema.gov/emiweb/downloads/is10_unit3.doc*. Accessed March 5, 2016.

4. Jennex ME: Modeling emergency response systems. In *Hawaii International Conference on System Sciences*. Piscataway, NJ: IEEE, 2007: 22.

5. Drabek TE, McEntire DA: Emergent phenomena and the sociology of disaster: Lessons, trends and opportunities from the research literature. *Disaster Prev Manag Int J*. 2003; 12(2): 97-112.

6. Helsloot I, Ruitenberg A: Citizen response to disasters: A survey of literature and some practical implications. *J Contingencies Cris Manag*. 2004; 12(3): 98-111.

7. Ashish N, Eguchi R, Hegde R, et al.: Situational awareness technologies for disaster response. In Chen H, Reid E, Sinai J, et al. (eds.): *Terrorism Informatics*. New York, NY: Springer, 2008: 517-544.

8. Endsley MR: Theoretical underpinnings of situation awareness: A critical review. In *Situation Awareness Analysis and Measure ment*. Boca Raton, FL: CRC Press, 2000: 3-32.

9. UNISDR: *Sendai Framework for Disaster Risk Reduction 2015-2030*. Geneva, Switzerland: UNISDR, 2015.

10. Heinzelman J, Waters C: *Crowdsourcing Crisis Information in Disaster-Affected Haiti*. Washington, DC: US Institute of Peace, 2010.

11. Kamel Boulos MN, Resch B, Crowley DN, et al.: Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: Trends, OGC standards and application examples. *Int J Health Geogr*. 2011; 10(1): 67.

12. Zook M, Graham M, Shelton T, et al.: Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Med Health Policy*. 2010; 2(2): 6-32.

13. Schulz A, Ortmann J, Probst F: Getting user-generated content structured: Overcoming information overload in emergency management. In *2012 IEEE Global Humanitarian Technology Conference*. Piscataway, NJ: IEEE, 2012: 143-148.

14. Lettieri E, Masella C, Radaelli G: Disaster management: Findings from a systematic review. *Disaster Prev Manag Int J*. 2009; 18(2): 117-136.

15. Kailes JI, Enders A: Moving beyond "special needs" A function-based framework for emergency management and planning. *J Disabil Policy Stud*. 2007; 17(4): 230-237.

16. Norris HF, Stevens PS, Pfefferbaum B, et al.: Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *Am J Community Psychol*. 2008; 41(1): 127-150.

17. Cosgrave J: Decision making in emergencies. *Disaster Prev Manag Int J*. 1996; 5(4): 28-35.

18. Kim M, Sharman R, Cook-Cottone PC, et al.: Assessing roles of people, technology and structure in emergency management systems: A public sector perspective. *Behav Inf Technol*. FEMA, 2012; 31(12): 1147-1160.

19. UNOCHA: *UNDAC Handbook*. 6th ed. New York: UNDAC, 2013.

20. FEMA: *A Whole Community Approach to Emergency Management: Principles, Themes, and Pathways for Action*. Washington, DC: FEMA, 2011.

21. Jaeger PT, Shneiderman B, Fleischmann KR, et al.: Community response grids: E-government, social networks, and effective emergency management. *Telecomm Policy*. 2007; 31: 592-604.

22. Vieweg SE: *Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications* [PhD thesis]. Boulder: University of Colorado at Boulder, 2012.

23. Palen L, Vieweg S, Sutton J, et al.: Crisis informatics: Studying crisis in a networked world. Paper presented at Proceedings of the Third International Conference on E-Social Science. Ann Arbor, MI. October 7-9, 2007.

24. Liu SB: Crisis crowdsourcing framework: Designing strategic configurations of crowdsourcing for the emergency management domain. *Comput Support Coop Work*. 2014; 23(4-6): 389-443.

25. Poblet M, García-cuesta E, Casanovas P: Crowdsourcing tools for disaster management: A review of platforms and methods. In *AI Approaches to the Complexity of Legal Systems*. Heidelberg, Germany: Springer, 2014: 261-274.

26. Wukich C: Social media use in emergency management. *J Emerg Manag*. 2015; 13(4): 281-294.

27. Bennett DM: How do emergency managers use social media platforms? *J Emerg Manag*. 2014; 12(3): 251-256.

28. Potts L, Seitzinger J, Jones D, et al.: Tweeting disaster: Hashtag constructions and collisions. In *Proceedings of the 29th ACM International Conference on Design of Communication*. New York, NY: ACM, 2011: 235-240.

29. Horita FEA, Degrossi LC, Assis LFFG, et al.: The use of volunteered geographic information and crowdsourcing in disaster management: A systematic literature review. paper presented at Proceedings of the Nineteenth Americas Conference on Information Systems. Chicago, IL, August 15-17, 2013.

30. Dhavase N: Location identification for crime & disaster events by geoparsing Twitter. In *International Conference for Convergence of Technology (I2CT)*. Piscataway, NJ: IEEE, 2014.

31. Malmasi S, Dras M: Location mention detection in tweets and microblogs. In *PACLING 2015, Pacific Association for Computational Linguistics*. Singapore: Springer, 2015: 88-93.

32. Morstatter F, Gao H, Liu H: Discovering location information in social media. *IEEE Data Eng Bull*. 2015; 38(2): 4-13.

33. Weaver AC, Boyle JP, Besaleva LI: Applications and trust issues when crowdsourcing a crisis. *In 2012 21st International Conference on Computer Communications and Networks, ICCCN 2012*. Munich, Germany: IEEE, 2012: 1-5.

34. Ushahidi: Ushahidi website. 2016. Available at *https://www.ushahidi.com/*. Accessed March 2, 2016.

35. Meier PP: Verifying crowdsourced social media reports for live crisis mapping: An introduction to information forensics. 2011. Available at *http://ceulearning.ceu.edu/pluginfile.php/106991/mod_resource/content/1/M4_Meier_Crowdsourced_data-case-studies.pdf*. Accessed March 5, 2016.

36. Morrow N, Mock N, Papendieck A, et al.: Independent evaluation of the ushahidi haiti project. *Development Information Systems International* 8. 2011.

37. Ushahidi: Examples of deployments website. Available at *https://www.ushahidi.com/support/examples-of-deployments*. Accessed December 15, 2015.

38. Hale J: A layered communication architecture for the support of crisis response. *J Manag Inf Syst*. 1997; 14(1): 235-255.

39. Palen L, Anderson KM, Mark G, et al.: A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In *Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference*. Swinton, UK: British Computer Society, 2010: 8.

40. Simon T, Goldberg A, Adini B: Socializing in emergencies—A review of the use of social media in emergency situations. *Int J Inf Manag*. 2015; 35(5): 609-619.

41. Houston JB, Hawthorne J, Perreault MF, et al.: Social media and disasters: A functional framework for social media use in disaster planning, response, and research. *Disasters*. 2015; 39(1): 1-22.

42. Yin J, Lampert A, Cameron M, et al.: Using social media to enhance emergency situation awareness. *IEEE Intell Syst*. 2012; 27(6): 52-59.

43. Alexander DE: Social media in disaster risk reduction and crisis management. *Sci Eng Ethics*. 2014; 20(3): 717-733.

44. Lang G, Benbunan-Fich R: *The Use of Social Media in Disaster Situations: Framework and Cases*. Hershey, PA: IGI Global, 2010: 11.

45. Wendling C, Radisch J, Jacobzone S: *The Use of Social Media in Risk and Crisis Communication*. Vol 30. Paris, France: OECD, 2013.

46. IRGC: *Social Media for Crisis Communication*. Lausanne, Switzerland: IRGC, 2012.

47. UNISDR: *Disaster Risk Reduction and Resilience in the 2030 Agenda for Sustainable Development*. Geneva, Switzerland: UNISDR, 2015: 1-21.

48. UNAPCICT: ICT for Disaster Risk Reduction. Incheon: UNAPCICT, 2010.

49. Vieweg S, Hughes AL, Starbird K, et al.: Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *CHI 2010*. New York, NY: ACM Press, 2010: 1079-1088.

50. Neubig G, Matsubayashi Y, Hagiwara M, et al.: Safety information mining—What can NLP do in a disaster. In *IJCNLP, 2011, Proceedings of the 5th International Joint Conference on Natural Language Processing*. Vol. 11. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 2011: 965-973.

51. Imran M, Elbassuoni S, Castillo C, et al.: Extracting information nuggets from disaster-related messages in social media. In *Proceedings of ISCRAM*. Baden-Baden, Germany: ISCRAM Association, 2013: 1-10.

52. Imran M, Elbassuoni S, Castillo C, et al.: Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*. New York, NY: ACM Press, 2013: 1021-1024.

53. Lingad J, Karimi S, Yin J: Location extraction from disaster-related microblogs. In *Proceedings of the 22nd International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee. New York, NY: ACM, 2013: 1017-1020.

54. Sebastiani F: Machine learning in automated text categorization. *ACM Comput Surv*. 2002; 34(1): 1-47.

55. Kotsiantis SB: Supervised machine learning: A review of classification techniques. *Informatica*. 2007; 31: 249-268.

56. Powers DM: Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol*. 2011; 2: 37-63.

57. Valverde-Albacete FJ, Peláez-Moreno C: 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS One*. 2014; 9(1): e84217.

58. Zocial inc.: Insightful information of Thai people on social media. 2015.

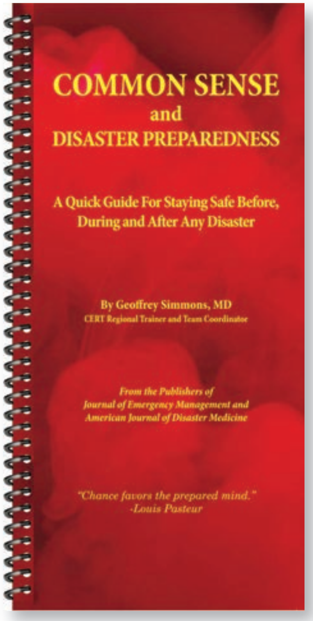
59. EM-DAT: Disaster profiles website. 2016. Available at *http://www.emdat.be/disaster_profiles/index.html*. Accessed June 14, 2016.

60. Machine Learning Group at the University of Waikato: Weka 3 website. Available at *http://www.cs.waikato.ac.nz/ml/weka/*. Accessed March 2, 2016.

61. Platt JC: Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*. 1st ed. Cambridge, MA: The MIT Press, 1998: 185-208.

62. Chang CC, Lin CJ: LIBSVM website. Available at *http://www.csie.ntu.edu.tw/ cjlin/libsvm/*. Accessed March 5, 2016.

63. Debnath R, Takahashi H: Kernel selection for the support vector machine. *IEICE Trans Inf Syst*. 2004; E87-D(12): 2903-2904.


COMMON SENSE
and
DISASTER PREPAREDNESS
A Quick Guide For Staying Safe Before, During and After Any Disaster
By Geoffrey Simmons, MD
CERT Regional Trainer and Team Coordinator
From the Publishers of Journal of Emergency Management and American Journal of Disaster Medicine
"Chance favors the prepared mind." -Louis Pasteur