

PAPER • OPEN ACCESS

Q-DAR: quick disaster aid and response model using Naïve Bayes and Bag-of-Words algorithm

To cite this article: R A Camama *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **482** 012051

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the collection - download the first chapter of every title for free.

Q-DAR: quick disaster aid and response model using Naïve Bayes and Bag-of-Words algorithm

R A Camama¹, J M Baldelomar², J M Claveria³, M C Espinas⁴, J M Cabardo⁵,
and L W Rabago⁶

Asia Pacific College, Magallanes, Makati City, Philippines

¹racamama@student.apc.edu.ph

²jmbaldelomar@student.apc.edu.ph

³jmclaveria@student.apc.edu.ph

⁴mcespinas@student.apc.edu.ph

⁵jayveec@apc.edu.ph

⁶lorenar@apc.edu.ph

Abstract. A real-time approach in decision-making for disaster management here in the Philippines is important, given that the country is vulnerable to disasters. Many studies show that data from social media can be of good use especially in times of disaster. This research aims to develop a model that will serve as a decision support tool for the government to respond during disasters with the use of Twitter hashtags that are based from the DRRM disaster phases; Disaster Response, Disaster Preparedness, Disaster Rehabilitation and Recovery, and Disaster Mitigation. The result of the study provides an overview of the critical level of the disaster phases with the use of Naïve Bayes and Bag-of-Words algorithm. The process of this study has four phases; the Training phase includes the labeling of the hashtags and list of Bag-of-Words, the next phase is Data Collection where Twitter data automatically updates and can be extracted every hour. The third phase is Data Pre-processing; where the tweets go through tokenization and normalization. Lastly, the dataset goes through Q-DAR with two sub-processes: The Disaster Phase Analysis using Bag-of-Words and Critical Level Analysis using Naïve Bayes in WEKA.

1. Introduction

Disasters are evident here in the Philippines. Many natural and man-made disasters have been experienced throughout the years. This is because of the current location of the country given that it is along the Ring of Fire [1], where frequent earthquakes and volcanic eruptions occur. Also, being beside the largest body of water, the Pacific Ocean, the Philippines encounters many disastrous typhoons with an average number of 20 typhoons per year [2].

The National Risk Reduction and Management Council handles various disasters with a similar framework focusing on the following disaster phases; Disaster Prevention and Mitigation, Disaster Preparedness, Disaster Response, and Disaster Rehabilitation and Recovery. It has the objective of having *safer, adaptive, and disaster-resilient Filipino communities toward sustainable development* [3].

Disaster Management in a community or area is aided by initial forecasts by the National Disaster Risk Reduction and Management Council that involves full cooperation of other government agencies



(e.g. Department of Health, Department of Science and Technology and Department of Social Welfare and Development), private agencies, as well as the community or the public. Each disaster phase is handled by a different government agency that is responsible for the overall disaster phase, along with other various implementing partners or government and local agencies. However, these local agencies are often understaffed, and a significant gap exists as the NDRRMC cannot supervise all the local councils [4]. In order to decide for the actual disaster response, it must go through all these sub-departments. Thus, this results to delayed decision-making in terms of responding and giving assistance to the people in need.

On the other hand, many studies have taken consideration of social media as an avenue for involvement of the community in times of disasters. Twitter is often noted as a source of big data that can be transformed to multiple uses such as disaster management [5] [6]. Thus, social media, on the other hand, can be an avenue for being an aid to disaster management through manipulation of data from social media sites and making use of them for decision-making. The government uses Social Networking Sites when it comes to disasters through official hashtags (e.g. #ReliefPH, #RescuePH, #PreparednessPH). People resort to these hashtags to make the government and other people aware that they are in need of help. Although the government has emergency hotlines available for the community, in the actual time of disaster, these hotlines are often not reliable.

Objectives of this research is QDAR model as a decision support tool for the government through classifying Twitter hashtags and providing an overview of the critical level of tweets per disaster phase with the use of Naïve Bayes and Bag of Words algorithm.

2. Related Works

2.1. Social Media Use during Emergency Response

Twitter, as the center of interest of the study, is a venue to find out about what is happening in the world in real-time. This type of information is important because the faster the information is available, the faster the decisions can be made. In times of extreme events such as natural disasters, real-time updates have the potential to prevent the loss of lives [7]. Based on the study conducted by C. David, J. Ong and E. F. Legara, the functions of Twitter evolved overtime showing a disaster awareness cycle on a social media platform. They examined these throughout the 20-day period before the arrival of typhoon Haiyan- the 20-foot biggest and deadliest storm surge that hit the island of Leyte in the Philippines, using Twitter users' tweets and retweets about the typhoon [2] [3]. They also found out that people are not only relying on twitter for the news, but they also engage themselves by tweeting and retweeting stories that interest them. Basically, Twitter is a site where people share information regarding the disaster relief, locally organized fundraising and finding potential volunteers and donors for the disaster [2].

2.2. S-Sense: A Sentiment Analysis Framework for Social Media Sensing

Feedback from the public cannot be utilized if not thoroughly gathered, filtered, and analyzed. This is where social media comes in. Specifically, Twitter alone has millions of users, which that there are means millions of potential sources of feedback or data. What this related literature and the current study conducted by the researchers have in common is that they both gather data and have to go through the process of filtering the data based on their relevance to the topic. To be specific, Naive Bayes is used in both studies to determine whether, for instance, a tweet, has something to do with the product or service being offered by a certain business or in the case of Q-DAR, something that pertains to the current disaster being experienced in a locality. If executed correctly, it can yield more accurate results which can then be used for wiser and faster strategic planning or decision-making.

In accordance to the purpose of this study which is to incorporate Natural Language Processing on Twitter feeds, several studies support and have taken the idea into further research. Alongside these methods, techniques, and algorithms gathered, this study also focused on the idea of establishing the disaster model of being an aid to the decision-making of the government in terms of responding to

various disaster phases. This would be with regard to how a government is disaster-prepared and how disasters are responded unto.

3. Proposed Method

The following is the methodology of the Q-DAR model as shown in Figure 1 below. It consists of the following phases: (1) Training, (2) Data Collection, (3) Data Pre-processing and lastly, and (4) Q-DAR: Quick Disaster Aid and Response.

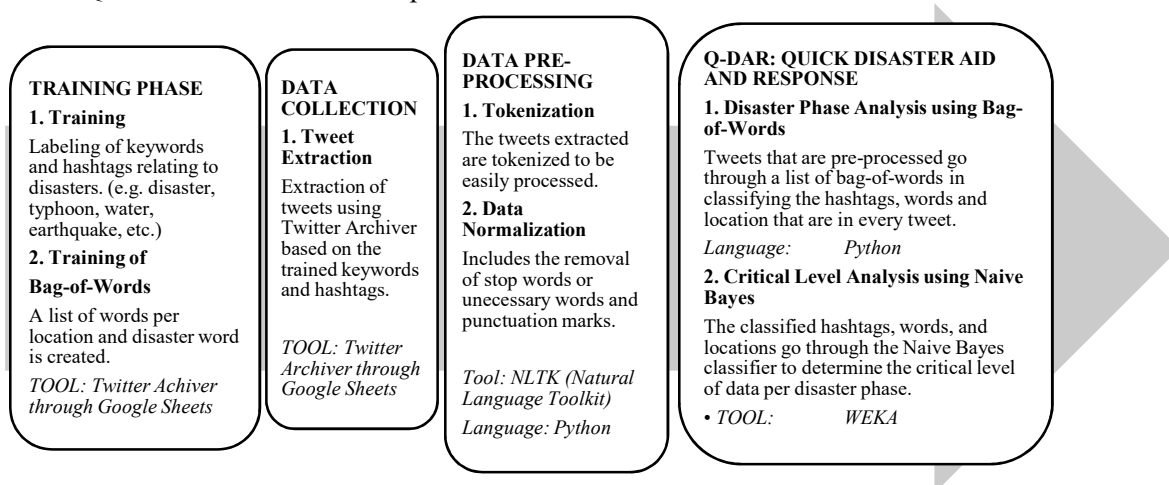


Figure 1. Q-DAR model

3.1. Training Phase

The initial step in the methodology is the training phase where the labeling of the necessary keywords, hashtags, etc. are supplied to the Twitter Archiver. These labels and hashtags are manually supplied to Twitter Archiver. The team supplied the appropriate hashtags and labels that are used to collect all the tweets.

3.2. Data Collection

Data Collection phase includes Tweet Extraction. It is where all the data that match the supplied labels and hashtags from the training phase are gathered. The tweets collected are automatically updated every hour.

This phase includes the retrieval of the tweets from Twitter using the Twitter Archiver add-on from Google Sheets. It will be achieving its real-time approach of gathering data from Twitter because Twitter Archiver is capable of being updated automatically every hour. Gathered data are collected based from the labels that are supplied in the training phase. The data is saved into a .csv file.

3.3. Data Pre-processing

Under this phase are the following processes: (1) Tokenization and (2) Data Normalization. The tool that will be used in this phase are the packages in the Natural Language Tool Kit (NLTK).

3.3.1. Tokenization. Tokenization is used in this phase to split up the texts that are gathered. For this process, the `nlk.tokenize` package, a package under NLTK, is used. The saved .csv file contains the tokenized tweets from the 520 tweets that are gathered, including the 100 dummy data that are produced for this research study.

3.3.2. Data Normalization. Data Normalization includes the removal of the stop words from the tweets. In the removal of stop words, punctuation marks are removed in the tweets, leaving out necessary words.

3.4. Q-DAR: Quick Disaster Aid and Response

3.4.1. Disaster Phase Analysis using Bag-of-Words. After data pre-processing, the tweets go through two sets of Bag-of-Words; location and disaster words. Tweets are classified by location and by words that appear in the tweet. This is also the process where the tweets are categorized into their respective disaster phases.

There are four attributes in this analysis phase, namely month, word, location, hashtag, and disaster phase. The month attribute is from the date the tweet is created. Word attribute is from the words that are used in the tweet. Location attribute is from the tweet that mentions a location that is inside the Bag-of-Words, and hashtag is also from the tweet when it mentions the specific hashtag. The tweets are then categorized into their distinctive disaster phases. The output data is then converted to an .arff file that will be processed by WEKA.

3.4.2. Critical Level Analysis using Naïve Bayes. After pre-processing, the .arff file is opened in WEKA and the following results are immediately shown in this process:

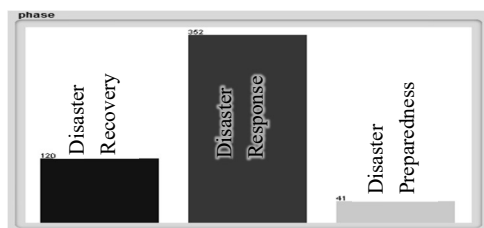


Figure 2. Critical level per disaster phase

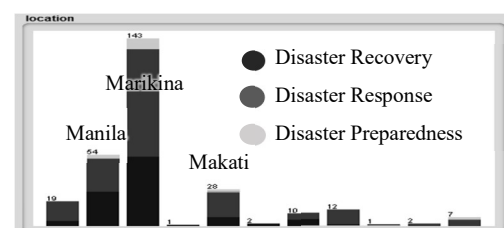


Figure 3. Critical level of disaster phases by location

In Figure 2 above, the four disaster phases (Mitigation, Response, Preparedness, and Recovery) are shown with the number of tweets per phase. Figure 3 shows the critical level of the phases per location, tested with a total of 520 tweets, including the 100 dummy tweets that are created for this study. The values show that Marikina has the highest bar in the location graph. Thus, it has the greatest number of tweets with the Disaster Response phase as its most critical phase.

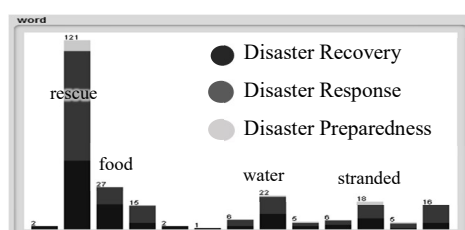


Figure 4. Critical level of words used per disaster phase



Figure 5. Critical level of disaster phases per month of tweet

Figure 4 above shows the most used word from the tweets gathered. Based on the result, the word mostly used is 'rescue', followed by 'food' and 'water'. Figure 5 shows the critical level of the disaster phases monthly. Given the date of the tweets that are processed, the bar graph shows the results in classifying the tweets in each of the disaster phases.

Figure 6 shows the hashtag attribute in determining the critical level of disaster phases. The first bar is #RescuePH and falls under the Disaster Respond phase. The second bar is #ReliefPH and is under Disaster Recovery phase. The third bar is #PreparednessPH that falls under the Disaster Preparedness phase.



Figure 6. Critical level of hashtags used per disaster phase

In order to get the best testing option for Naïve Bayes algorithm, the dataset goes through Percentage Splits and K-Cross Validation. Both testing options are evaluated based on their accuracy and kappa statistics. After getting the best testing option, the dataset is then classified using Naïve Bayes with the testing option that attained the highest accuracy for Naïve Bayes.

Results from the Naïve Bayes classification process determined the critical level of the disaster phases (Disaster Response, Disaster Preparedness, and Disaster Recovery) based on the following attributes: *disaster word used, month, location, and hashtag*.

4. Experiment and Results

The hashtags (#RescuePH, #ReliefPH, and #PreparednessPH) managed to collect disaster-related tweets that are classified by different disaster phases (*Disaster Preparedness, Disaster Response, and Disaster Recovery and Rehabilitation*). The data collection phase extracted 520 tweets from Twitter. Data Pre-processing phase managed to tokenize and remove stop words from all the tweets accordingly.

In order to analyze the best testing option for the Naïve Bayes algorithm in determining the critical level of each disaster phase, Percentage Split and K-Fold Cross Validation are compared using Naïve Bayes. The following are the results of the evaluation of two testing options.

Table 1 shows the accuracy and kappa statistic of the Percentage Splits 50-50 percentage of training data and test data. Table 2 shows Cross Validation Folds with 10 folds. Result shows that having 10 folds is more accurate between the two testing options.

Table 1. Disaster Phase percentage split

Percentage Split	Kappa Statistic	Accuracy
50-50	0.88	93.33%

Table 2. Disaster Phase Cross Validation Folds

Cross-Validation Folds	Kappa Statistic	Accuracy
10	0.90	94.44%

With a total number of 520 extracted tweets, it can be seen in Figures 3-6 that the overview of the critical level of disaster phases is given four attributes; location, word, month and hashtag. The following tables show the percentage of attributes to each disaster phase using the Naïve Bayes classifier with 10 cross validation folds.

Table 3. Tweet Location with the Highest Critical Level per Disaster Phase Using Naive Bayes

Location	Disaster Response	Disaster Preparedness	Disaster Recovery	Total
Marikina	83.0	9.0	54.0	146.0
Manila	26.0	4.0	27.0	57.0
Makati	20.0	3.0	8.0	31.0

The locations that are listed on Table 3 are the top 3 locations with the greatest number of tweets and are classified based on the disaster phase of each tweet. The results show that Marikina has the greatest number of tweets with a total of 146.0. Disaster response in that area has the highest level with the result of 83.0, followed by Manila and Makati.

Table 4 shows that the word 'rescue' is the most frequently used disaster word, with Disaster Response as its most critical disaster phase. This shows that tweets gathered and processed in that hour needs rescue.

Table 4. Words with the Highest Critical Level per Disaster Phase Using Naive Bayes

Words	Disaster Response	Disaster Preparedness	Disaster Recovery	Total
rescue	71.0	8.0	45.0	124.0
food	12.0	17.0	1.0	30.0
stranded	10.0	8.0	3.0	21.0
flood	12.0	5.0	1.0	18.0

Table 5. Hashtag with the Highest Critical Level per Disaster Phase Using Naive Bayes

Hashtag	Disaster Response	Disaster Preparedness	Disaster Recovery	Total
#RescuePH	353.0	1.0	1.0	355.0
#ReliefPH	1.0	1.0	121.0	123.0
#PreparednessPH	1.0	42.0	1.0	44.0

Table 5 shows that #RescuePH is the most used hashtag and Disaster Response as the most critical phase. This also shows that most tweets are related to rescue and the need for help.

Table 6. Result of the Highest Attributes with the Most Critical Disaster Phase

Attribute	Has the Highest Result	Disaster Response	Disaster Preparedness	Disaster Recovery
Location	Marikina	83.0	9.0	54.0
Word	rescue	71.0	8.0	45.0
Hashtag	#RescuePH	353.0	1.0	1.0

Table 6 shows all attributes with the highest results. It shows that Marikina is the location that needs rescue the most, given that Disaster Response is its highest disaster phase along with the most used hashtag #RescuePH. The most used word was rescue. All attributes have Disaster Response as its highest disaster phase as well.

Table 7. Month with the highest level of Disaster Phases using Naive Bayes

Month	Disaster Response	Disaster Preparedness	Disaster Recovery	Total
July	178.0	21.0	22.0	221.0
August	176.0	22.0	100.0	298.0

Given that the most critical disaster phase is Disaster Response as seen in Table 6, it can be concluded that it is in the month of July, as seen in Table 7, that the Philippines has encountered numerous typhoons.

5. Conclusion and Recommendations

With the Quick Disaster Aid and Response Model, the government can have a visual overview of the critical level of each of the disaster phases (Disaster Mitigation, Disaster Response, Disaster Preparedness, and Disaster Recovery) through processing and classifying the unified hashtags. It can be an aid to their decision-making, as to where and what actions are needed to prioritize, given the critical level of the disaster phases that are classified by location, word, month, and hashtags.

With the results of Q-DAR, the most critical disaster phase per location can be easily detected. The local government unit would have an idea of what location needs immediate rescue and assistance in that hour with its real-time approach of gathering and processing tweets from Twitter. The Naïve Bayes algorithm successfully classified the tweets and showed the critical level of each disaster phase as shown in Tables 3-7.

Q-DAR, on the other hand, is a study that can be further improved and capable of determining location. The concept of Geographical Information Systems can be applied to Q-DAR. The overview of the result generated can be processed and evaluated thoroughly. It can be created as a system that the government can use for a more efficient decision-making. Image processing can also be a study that can be applied to Q-DAR. In the process of collecting tweets, only texts are gathered and processed by the Q-DAR study. On the other hand, it can be further enhanced by including image

processing in collecting and processing these data. Another would be the use of other data sources aside from Twitter. Lastly, Artificial Intelligence can be supported in the study through the Bag-of-Words algorithm. Whereas, in this process, the list of bag-of-words can be trained to automatically learn new words based on the tweets that will be gathered.

References

- [1] UNICEF 2018 *Disaster Risk Reduction* Online Available: <https://www.unicef.org/philippines/risk.html#.Wy2qwu6FPIU> Accessed 23 June 2018
- [2] Wingard J and Brandlin A S 2013 *Philippines: A Country Prone to Natural Disasters* 11 October 2013 Online Available: <http://www.dw.com/en/philippines-a-country-prone-to-natural-disasters/a-17217404> Accessed 23 June 2018
- [3] NDRRMC 2011 *National Risk Reduction and Management Council* Online Available: http://www.ndrrmc.gov.ph/attachments/article/41/NDRRM_Plan_2011-2028.pdf Accessed 25 June 2018
- [4] Alcayna T, Bollettino V, Dy P and Vinck P 2016 *Resilience and Disaster Trends in the Philippines: Opportunities for National and Local Capacity Building* Online Available: <http://currents.plos.org/disasters/article/resilience-and-disaster-trends-in-the-philippines-opportunities-for-national-and-local-capacity-building/> Accessed 22 August 2018
- [5] David C C, Ong J C and Legara E F T 2016 *Tweeting Supertyphoon Haiyan: Evolving Functions of Twitter during and after a Disaster Event* Online Available: <http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0150190&type=printable> Accessed 23 June 2018
- [6] Lindsay B 2011 *Social Media and Disasters: Current Uses, Future Options, and Policy Considerations* Online Available: <https://fas.org/sgp/crs/homsec/R41987.pdf> Accessed 21 August 2018
- [7] Avci H 2016 *10 Ways Twitter Is Revolutionizing Real-Time Information* Online Available: <https://digitalculturist.com/10-ways-twitter-is-revolutionizing-real-time-information-704a894d1d0f> Accessed 21 August 2018