Corpus Analysis of Earthquake Related Tweets through Topic Modelling

Lany L. Maceda, Jennifer L. Llovido, and Thelma D. Palaoag

***Abstract*—The Philippine archipelago is one of the most disaster-prone area due to its location. As Twitter grows everyday, it has now become a valuable source of people's opinion. The main purpose of the study is to use Twitter, as text corpora in the attainment of disaster risk reduction. Earthquake related tweets posted from July 1, 2017 to August 31, 2017 were programmatically collected. Data cleaning was made by removing noisy words, hence, from 90,692 collected tweets this resulted to 41,500 cleaned tweets. Topic modeling identifies patterns in a corpus. Findings revealed that there is a need to strengthen the conduct of earthquake drill and early warning notices to lessen the vulnerability of people and property, and reduce the causal damages of earthquake, thus improve its community resiliency. Topic results shall be discussed with concerned agencies for its possible consideration and inclusion to enhance the earthquake-related disaster management plan.**

***Index Terms*—Topic modeling, corpus analysis, tweets, earthquake.**

I. Introduction

A natural disaster is a major adverse event such as floods, volcanic eruptions, earthquakes, tsunamis, and other geologic processes. This can cause loss of life or property damage, and typically leaves some economic damage and the severity of which depends on the affected population's resilience, or ability to recover. An adverse event will not rise to the level of a disaster if it occurs in an area without vulnerable population.

The Philippines is a hot bed of disasters such as earthquakes, volcanic eruptions, tropical cyclones and floods as this often occurs in the country because of its geographic locations, making it one of the most disaster prone countries in the world. According to the 2016 World Risk Index, the Philippines is ranked 3rd of the most disaster-prone countries on earth [1].

Throughout the recorded history of disaster occurrences in the Philippines, earthquakes and its associated hazards has been occurring more frequently. One of this was the strong earthquake of magnitude 6.5 that shook the island of Leyte last July 6, 2017 at 4:03 pm [2]. It triggered a landslide in Ormoc City. It also caused island-wide blackouts in the provinces of Samar, Bohol, Leyte, and parts of Southern Leyte aside from damage to buildings and properties, injuries and death.

In the recent years, as the audience of social media such as Twitter grows everyday it now become a valuable source of people's opinion. Thus, the main purpose of the study is to use Twitter that contains a very large number of very short messages created by the users with messages that vary from personal thoughts to public statements, as a dataset in the attainment of Disaster Risk Reduction (DRR) in analyzing and lessening vulnerability of people and property, improve the preparedness and early warning for adverse events, and reduce the causal damages of earthquake.

Manuscript received September 30, 2017; revised December 1, 2017. This work is produced in part of the University of the Cordilleras (UC) IT research design and methodology class and was supported in part by the Commission on Higher Education (CHED).

Lany L. Maceda and Jennifer L. Llovido are with the Bicol University College of Science, Computer Science and Information Technology Department, Legazpi City, Philippines (e-mail: lanylm@yahoo.com, jllovido@gmail.com).

Dr. Thelma D. Palaoag is with the University of the Cordilleras College of Information Technology and Computer Science, Baguio City, Philippines (e-mail: tpalaoag@gmail.com).

II. Related Works

A. Social Media on Disaster

Several studies proved that social media becomes useful during natural disasters in terms of communication. According to Lam, *et al.* [3] in times of disaster, social media has been found to be a valuable source of information. It is an online communication channel that becomes popular among Internet users. It is an avenue where people can interact, share content and collaborate with other people. It is a tool that allows people to easily share information to the public.

Further, Hermida *et al.* [4] examined how Social Networking Sites (SNS) served as news and information sources through a survey of 1600 online users in Canada, concluding that while SNS have the potential to create new and personal news streams, they largely support mainstream media outlets (e.g., news organizations) in which they have retained a degree of trust.

The Philippines is named as the "social networking capital of the world". It clearly describes that the Philippines is a country that has adapted to social media. Social media continues to be an essential part in the lives of Filipinos who have access to the Internet [5]. Social media's significance to the average Filipino internet user becomes more apparent during times of natural disasters [6]. Mislos [7] cited that Filipinos use social media platforms for 53 hours per week, which is 11 hours more than the 42-hour global average.

Blanchard *et al.* [8], cited that social media and social technologies have altered communication patterns, particularly in times of disaster. Social media has been used by the public to share information during emergencies which has created an unexpected side effect where responding authorities and aid organizations are expected to be aware of and respond to emergency requests for help coming from Facebook, Twitter and text messages.

doi: 10.18178/ijmlc.2017.7.6.645

B. Twitter and Disaster

Social media serves as a platform for people to express their thoughts and opinions on various topics Social media has recently played an important role in natural disasters as an instrument of information that can be used for disaster relief and others. Social media mining is a rapidly growing field. The fast-growing interests and intensifying need to harness social media data require research and the development of tools for finding insights from big social media data. Accordingly, one (1) of the most prominent social media sites in the country is Twitter. It is a microblogging service where users can broadcast short 140-character messages called tweets, which can be personal thoughts or opinions on public statements, places, persons, disasters and others making it a useful tool for gathering information. In 2013, the Philippines ranked third worldwide in Twitter use, with 69% of Filipino Web Users having Twitter accounts, and 40% of those Web Users classified as "active users" of Twitter. It is logical, therefore, to explore Twitter as a domain within which to explore Filipino behavior regarding specific issues [9]. Twitter is increasingly being considered as a means for emergency communication during and after natural disasters due to its growing ubiquity, communication rapidity, and cross-platform accessibility [10].

Takeshi, *et al.* [11] detect earthquakes based on Twitter postings. In posting tweets, users can make use of Twitter hashtags to identify topics of posts. The hashtag is the fundamental element of Twitter's user classification system. Anatoliv, *et al.* [12], Beduya and Espinosa's [13] study took advantage on the use of tweets in Twitter in identifying the disaster-prone areas in the Philippines. This study aimed to aid the Philippine government and other concerned organizations in their disaster management plans.

The paper of Weng, *et al*., [14] focuses on the problem of identifying influential users of micro-blogging services. Twitter, one of the most notable micro-blogging services, employs a social-networking model called "following", in which each user can choose who she wants to "follow" to receive tweets from without requiring the latter to give permission first. In a dataset prepared for this study, it is observed that (1) 72.4% of the users in Twitter follow more than 80% of their followers, and (2) 80.5% of the users have 80% of users they are following follow them back.

As reported by Fugate [15] with regard to the catastrophic 2010 Haiti earthquake, even when an area's physical infrastructure was completely destroyed, the cellular tower bounced back quickly, allowing survivors to request help from local first responders and emergency managers to relay important disaster-related information via social media sites.

Similarly, the dataset used in the study of Syliongka, *et al*., [16] was obtained from, which discovered prevalent themes in Tweets about Typhoon Yolanda (Haiyan), which struck the Philippines in November 2013. Tweets containing one of the terms "Yolanda", "Haiyan", "victims", "typhoon", were collected from November 2013 to January 2014. For each month during this five month period, the researchers in performed open coding and manually identified the prevalent themes per month. These identified themes then serve as the annotations for the tweets in this study. Of course, not all tweets in a given month necessarily express the prevalent themes, so in this regard, the annotations are considered noisy labels.

C. Analyzing Data Using Topic Modeling

From the gathered data, a topic model shall be made to give meaning on it. As defined by Blei [17], topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes. This is a widely used approach for analyzing large text collections. In particular, Latent Dirichlet Allocation (LDA) is one of the most popular topic modelling approaches to aggregate vocabulary from a document corpus to form latent "topics" [18].

Ligutom *et al*. [19] puts emphasis that twitter and disaster are related to each other. It is used as a medium to convey the opinions and perspectives of a person on a particular subject or event, subjects may vary from public figures, annual events, politics and even disasters. It is through Twitter that people with Internet access spread news and information as to what is happening in the place and the people affected by the typhoon, and even how people can donate financially or materially to those in need. Help can also be tweeted by those affected where they can tweet their current location and situation. Twitter is also used by Philippine government agencies to spread news, advisories, reports and response to emergencies in times of natural disasters. Some of the agencies with Twitter accounts are PAGASA for weather disturbances; PHIVOLCS for volcanic activities, earthquakes, and tsunamis; and NDRRMC for detailed reports on the national government's disaster risk reduction efforts, to name a few.

Further, their study took advantage of the data present in tweets to get some insights, through discovered topics, on how they reflect the behavior of Filipinos during typhoons. They focused on tweets collected from February 2013 to November 2014 of Twitter users from Metro Manila. Data preprocessing was applied to remove noisy and irrelevant data, like stop word and punctuations. The tweets were set to lower case and the following were removed: Users found on the tweets based on the users found from the tweet pool; URLs; Words that contain '@' sign; Non-Printable ASCII Characters; Punctuations; Stop words; Specific word permutations that form stop words; Special character '#'; Words that are digits only; Words with lengths that are less than 4; Words with high frequency (top 15 words); The word "rt"; Tweets with less than 2 words; Duplicate tweets from the same user; Out of the 9,714 total number of tweets collected, we ended up with 6,366 tweets after cleaning. It contains 25,321 words and 2,078 unique words. Results revealed different Filipino behaviors during a typhoon such as determination to rise up after the typhoon, voicing out concerns, and using word play. Future work could experiment on selecting the appropriate number of words per topic model.

The main objective of Nassar, *et al*.'s [20] paper is to improve the quality of the Twitter Exemplar-based topic detection system. The feedback from the crowd is utilized to adjust weights of the cosine similarity function deployed in the Exemplar-based topic detection algorithm. Testing the system using the Football Association Cup (FA Cup) dataset, it is found that the crowdsourcing has achieved a constant

increase in the topic recall (by up to 15%), term precision (by up to 4%) and term recall (by up to 3%). Therefore, the new weights succeeded in increasing the three measures of topic quality significantly.

Yi and Allan [21] explore the utility of different types of topic models, both probabilistic and not, for retrieval purposes. They showed that topic models are effective for document smoothing; more elaborate topic models that capture topic dependencies provide no additional gains; smoothing documents by using their similar documents is as effective as smoothing them by using topic models; and topics discovered on the whole corpus are too coarse-grained to be useful for query expansion. Experiments to measure topic models' ability to predict held-out likelihood confirm past results on small corpora, but suggest that simple approaches to topic model are better for large corpora.

III. Methodology

The study observed the topic modeling processes as depicted in Fig. 1 below.

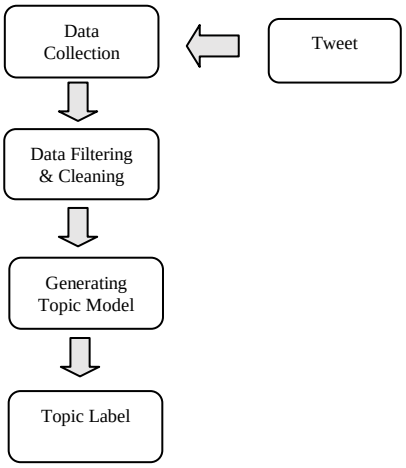

Fig. 1. Topic modeling processes.

A. Data Collection

The *GetOldTweets-python* programmatically was used to collect data from twitter. It requests tweets with keywords "earthquake leyte", "lindol leyte", "magnitude 6.5", "ormoc earthquake" and "phivolcs leyte". The tweets from July 1, 2017 to August 31, 2017 with Philippines geotagged were gathered resulting to 90,692 tweets.

B. Data Cleaning

Data cleaning is an important step in the study wherein data were cleaned to avoid producing misleading results. This was made possible by removing stop words and extra words through *bin\mallet import-file --input sample-data\leyte-all-tweets.txt --output leyte-tweets.mallet --keep-sequence --remove-stopwords --extra-stopwords sample-data/mario_extra.txt.* This removed noisy words found from the tweet pool such as URLs, word that contain @sign and #hashtag, punctuations, and the retweeted (rt) tweets. From 90,692 total number of collected tweets, after cleaning, this resulted to 41,500 number of tweets.

C. Topic Models

With the 41,500 data sets, a number of topics by 10 unique words were set using *bin\mallet train-topics --input leyte-tweets.mallet --num-topics 10 --optimize-interval 10 --num-top-words 10 --output-topic-keys leyte-tweets-keys-10.txt --output-doc-topics leyte-tweets-composition-10.txt.* Through the Dirichlet parameter with indication of the weight of each topic showed that the principal topic was Magnitude Impact with 2.57%.

D. Topic Labels

Analysis of each topic model was created based on the frequency of unique text occurrence and the interconnectivity of each word with the same and complex meaning.

IV. Discussion of Findings

The following topic models with the Dirichlet parameter indicative of the weight of each topic was presented in Table I.

TABLE I: Topic Models Results

	Topic Labels	Parameter	Topic Models
0	Social Media Post	0.00924	lindol, puro, post, nanaman, hugot, tweets, naramdaman tungkol, posts, lahat
1	After Effect	0.00684	lindol, malakas, ulan, sunod, ngayon, dulot, lalo, bagyo, pinsalang, leyte
2	Earthquake Feels	0.01701	lindol, naramdaman, manhid, kanina, ramdam, sobrang, nakaramdam, lakas, maramdaman, tulog
3	Class/work suspension	0.00656	lindol, leyte, kanina, suspended, school, klase, manila, tapos, pasok, tao
4	Unrecognized	0.00948	lindol, kanina, tapos, ulo, alam, drill, tao, earthquake, naramdaman, lumindol
5	Agency Notification Alerts	0.00951	lindol, leyte, magnitude, ormoc, niyanig, matapos, city, visayas, patay, phivolcs,_dost
6	Advisory	0.0034	earthquake, date, time, information, earthquakeph, dost, phivolcs, magnitude, lindol, niyanig
7	Earthquake Effects	0.0124	lindol, akala, kanina, kala, lakas, nahihilo, naramdaman, nahilo, tapos, upuan
8	Affected Places	0.00832	leyte, magnitude, earthquake, quake, Philippines, ormoc, hits, central city, leytequake,
9	Magnitude Impact	0.02568	lindol, lakas, safe, kanina, time, grabe, ramdam, tagal, naramdaman, floor

Findings revealed that the topic labels generated were Social Media Post, After Effect, Earthquake Feels, Class/work suspension, Unrecognized Label, Agency Notification Alerts, Advisory, Earthquake Effects, Affected Places, and Magnitude Impact. Indication on the weight of each Dirichlet parameter showed that the principal topics identified were Magnitude Impact (2.57%), Earthquake Feels (1.7%), and Earthquake Effects (1.24%).

With manual analysis, it could be inferred that most of the tweets reflect the importance in the conduct of earthquake drill and early warning notices to reduce possible damages and increase resiliency among community members. This also showed that despite of the fear on the effects of earthquake, use of social media is a great communication means for them. Further, results could serve as basis to the concerned government/non-government agency on the possible program

enhancement for disaster risk reduction as far as earthquake is concerned.

V. Conclusion

The main purpose of the study to make sense of the earthquake-related tweets sent as text corpora. This was made possible through the topic models generated such as Social Media Post, After Effect, Earthquake Feels, Class/work suspension, Unrecognized Label, Agency Notification Alerts, Advisory, Earthquake Effects, Affected Places, and Magnitude Impact. Having the Dirichlet parameter with 2.57% weight, this revealed that the principal topic from the pool of tweets is Magnitude Impact. Thus, it could be inferred that most of the tweets reflect the importance in the conduct of earthquake drill and early warning notices to reduce possible damages and increase resiliency among community members. This also showed that despite of the fear on the effects of earthquake, use of social media is a great communication means for them. Further, results could serve as basis to the concerned government/non-government agency on the possible program enhancement for disaster risk reduction as far as earthquake is concerned.

Acknowledgment

This work is produced in part of the University of the Cordilleras (UC), IT research design and methodology class.

References

[1] Philippines Now The 3rd Most Disaster-Prone Country According to Latest World Risk Index. (2016, September 03). [Online]. Available: http://www.wheninmanila.com/philippines-now-3rd-most-disaster-prone-country-according-to-latest-world-risk-index/

[2] P. Administrator. Primer on the 06 July 2017 Magnitude 6.5 Leyte Earthquake. [Online]. Available: http://www.phivolcs.dost.gov.ph/index.php?option=com_content&view=article&id=7641%3Aprimer-on-the-06-july-2017-magnitude-65-leyte-earthquake-07-july-2017&catid=60%3Alatest-news&Itemid=19

[3] L. A. Jan, N. Oco, and R. R. Edita. "Classifying Typhoon-Related Tweets using Word Embeddings and Convolutional Neural Networks," *Philippine Computing Journal*, Computing Society of the Philippines.

[4] H. Alfred, F. Fred, K. Darryl, and L. Donna, "Share, like, recommend," *Journalism Studies*, vol. 13, no. 5-6, pp. 815-824, 2012.

[5] PH is social networking capital of the world. (September 29, 2017). [Online]. Available: from http://manilastandard.net/lifestyle/101379/ph-is-social-networking-capital-of-the-world.html

[6] M. X. Y. Zia, "Aid networks to the rescue: Tweeting relief and aid during ondoy," MA Thesis. Georgetown University, 2010.

[7] M. Vincent, 2014, "Research confirms the Philippines is still the social media capital of the world," *Yahoo News!* July 3.

[8] H. Blanchard, A. Carvin *et al.*, "The case for integrating crisis response with social media," White Paper, American Red Cross, Washington, DC, p. 32, 2012.

[9] M. Marcello. (2013). "Infographic: Twitter The Fastest Growing Social Platform." GlobalWebIndex Blog, February 25.Accessed August 26, 2014.

[10] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the crowdsourcing power of social media for disaster relief," *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 10-14, 2011.

[11] S. Takeshi, O. Makoto, and M. Yutaka, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *Proc. 19th International Conference on World Wide Web*, 2010, pp. 851-860.

[12] G. Anatoliy, W. Barry, and T. Yuri, "Imagining twitter as an imagined community," American Behavioral Scientist, vol. 55, no. 10, pp. 1294-1318, 2011.

[13] L. J. Beduya and K. J. Espinosa, 2014, "Flood-related disaster tweet classification using support vector machines," in *Proc. the 10th National Natural Language Processing Research Symposium*, De La Salle University, Manila, February 21-22, 2014, p. 10.

[14] J. S. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential twitterers," in *Proc. the Third ACM International Conference on Web Search and Data Mining*, February 4-6, 2010.

[15] C. Fugate, "Five years later: An assessment of the post katrina emergency management reform act," Tech. Report US Federal Emergency Management Agency, Oct. 2011.

[16] L. R. Syliongka, N. Oco, A. J. Lam, C. R. Soriano, M. D. Roldan, F. Magno, and C. Cheng, "Combining automatic and manual approaches: Towards a framework for discovering themes in disaster-related Tweets," in *Proc. the 24th International Conference on World Wide We*, May 2015, pp. 1239-1244.

[17] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012

[18] N. Sukhija, M. Tatineni, N. Brown, M. V. Moer, P. Rodriguez, and S. Callicott, "Topic modeling and visualization for big data in social sciences," in *Proc. Intl IEEE Conferences on Ubiquitous Intelligence &Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People*, and Smart World Congress, 2016.

[19] C. Ligutom, J. V. Orio, D. A. Ramacho, C. Montenegro, R. E. Roxas, and N. Oco., "Using topic modelling to make sense of typhoon-related tweets," in *Proc.* 2016 International Conference on Asian Language Processing (IALP) , 2016.

[20] L. Nassar, R. Ibrahim, and F. Karray, "Enhancing topic detection in twitter using the crowdsourcing process," in *Proc. International Conference on Collaboration Technologies and Systems (CTS)*, 2016.

[21] X. Yi and J. Allan, "Evaluating topic models for information retrieval," in *Proc. the 17th ACM Conference on Information and Knowledge Management*, 2008.

Lany L. Maceda is a faculty member of Bicol University College of Science, Computer Science and Information Technology Department, Legazpi City, Philippines with an academic rank of Assistant Professor IV. Currently she is a Commission on Higher Education (CHED) scholar taking doctor in Information Technology (DIT) at the University of the Cordilleras, Baguio City, Philippines.

Jennifer L. Llovido is a faculty member of Bicol University College of Science, Computer Science and Information Technology Department, Legazpi City, Philippines with an academic rank of Associate Professor I. Currently she is a commission on Higher Education (CHED) scholar taking Doctor in Information Technology (DIT) at the University of the Cordilleras, Baguio City, Philippines.

Thelma D. Palaoag is a faculty member of University of the Cordilleras College of Information Technology and Computer Science, Baguio City, Philippines. Currently she is the department head of Computer Science and the College Research Coordinator. Completed her Doctor in Information Technology (DIT) at the University of the Cordilleras, Baguio City, Philippines.