Utilization Model using Support Vector Machine Classification for Disaster-Related Post on Twitter

Randy Joy Ventayen

Abstract: Philippine is known as part of the ring of fire, it is one of the disaster-prone countries which was hit by typhoon Haima (with a local name of Lawin), and followed by Sakira (with a local name of Karen) which was hit the country last October 2016, the two typhoons were named as one of the strongest typhoons that hit the country. In today's generation, Twitter served as a communication outlet and a source of information. On some numbers of tweets in social media, there is local language posted by the local users, and this study will bridge the gap since for those speaking another major language in the Philippines. The study will be sought to answer on how to download twitter data from a specific disaster duration in the region, how to extract and identify multilingual disaster-related tweets and finally how to classify disaster and non-disaster tweets in the local language. The study of classification and extraction of disaster and emergency-related tweets is important is interesting study because the life of a person which speaks a very rare dialect is important as the same as the person speaking a major language. The model presented in this study is useful as a tool to extract tweets, filter, and classify that could utilize for a faster response that could save a life.

Keywords: Natural Language Processing; Social Media; Twitter

I. INTRODUCTION

Social networking site such as Twitter is one of the most widely used as a source of news and information, and sometimes it is ahead than other media, because of its information feeds from known and unknown which is sent by users.

Typhoon Karen and Lawin hit the Philippines last October 2016 in the northern part of the Philippines specifically the Ilocos Region, it is one of the strongest typhoons with Category 5 status same as the typhoon Yolanda which hits the country last 2013. The researcher observed that not all the tweets during the during the typhoon are related to disaster and emergency are written in English, but their area also posted in the local language. Twitter is one of the most used social media platforms in the country and could be utilized to determine the location and to minimize the injury during a disaster.

Dr. Ventayen is currently designated as the University Web Administrator of Pangasinan State University, Lingayen, Pangasinan, Philippines.

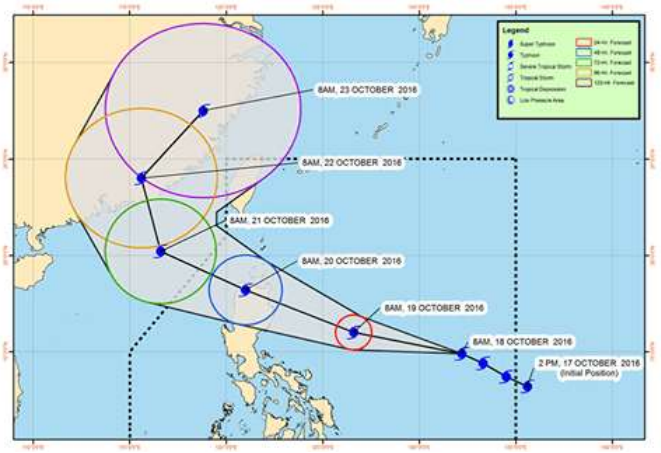

Fig. 1 shows the track of Typhoon Lawin in Region 1 (source: PAGASA)

There are 9 major languages in the Philippines and 2 of the major language and 1 minor dialect are in Region 1[1]. The region is one of the biggest regions in the Philippines and not all constituents can speak English and some have little knowledge in speaking Tagalog despite that they are Filipinos. There are many studies related to detection and analysis of tweets while those study is focusing in Tagalog and English only, from these current studies, the proponent realized that there is a need of equal treatment among those who speak those speaking local language and dialect. Since we are talking about life, the life of those speaking in national language is as important as those speaking in the local language.

Moreover, the Philippines is a country that is very attuned to social media, and it is even named as the Networking Capital of the world [2]. Some government agency in the country has social media accounts for faster information dissemination.

A. Research Objectives

The Objective of this paper is to extract tweets from specific duration during the typhoon Karen and Lawin hit the country. Second is to extract local language tweets in Pangasinan, Iloko or Bolinao. Lastly, is to classify disaster-related tweets. This paper will answer the how to extract tweets from Twitter. How to identify and extract multilingual tweets from the data? and how to classify disaster-related tweets?

B. Scope and Limitation

The scope of identification of disaster and emergency location is the four province of Region 1. While this study focuses on multilingual tweets initially in Region 1, it also gathers tweets from Tagalog and English language in the same region. The duration of the extracted data from Twitter is during the typhoon Karen and typhoon Lawin which was October 2016.

C. Significance of the Study

Life is valuable whatever language you speak. This study will help us to identify and provide a closer look at the region which is disaster-prone areas by using the tweets extracted from users. The proponent aims that this study will also help the organization and government agency such as the NDRRMC in its disaster management plans and expand future research.

II. RELATED LITERATURE

There are numerous researchers have used social media as a source of data to understand various disasters, with applications such as situational awareness and understanding the public sentiment.

In the study of Stowe, et. al [3], the tweets during Hurricane Sandy which impacted New York in 2012 was used. The researcher proposes an annotation schema for identifying tweets, it uses a system for classifying disaster-related twitter tweets. Categories were used to identify disaster-related tweets such as Sentiment, Action, Preparation, Reporting, Information, and movement. Based on its preliminary result, it shows the relevant information that can be extracted automatically via batch processing after the events, and the researcher is exploring possibilities to extend the approach to real-time processing.

According to the Twitter blog from @twiterindia, they used Twitter and worked with NGOs and another private sector with the participation of the citizen towards a strategy of disaster relief operations. They realize the usefulness of the social media during disaster relief during the Kashmir floods of 2014 and the work was replicated in 2015 when Chennai was hit with a flood. The outcome a team up and collaboration by NGOs, citizens, government agencies for disaster relief operations.

Another study was conducted by Parilla-Ferrer, et. al. [4], which research about the classification of disaster-related tweets in metro manila last 2012. The tweets were labeled as information and uninformative to check the reliability of the information posted. A machine learning algorithm was used which is the Naïve Bayes and Support Vector Machine (SVM). Based on the result of the study, SVM has a better result than the other, and it revealed that there are more uninformative tweets than the informative tweets, while the informative tweets were more likely to retweeted thus provide awareness to the public. Another study concerns RT counts on how a re-tweet count [5], the result of the human-subject experiment was when the re-tweet count of the tweet increased, the likelihood that people would share the tweet increased that the findings extend the understanding of how disaster-related information spread on Twitter.

Another study conducted tested FILIET: An Information Extraction System For Filipino Disaster-Related Tweets [6]. The study acknowledges the problem in the extraction of Filipino language, therefore, it creates a system that could extract relevant information from Filipino disaster-related tweets which they call it FILIET: Filipino Information Extraction Tool for Twiter. While the goal of this study is not just about the local Filipino language, the proponent also concerns the location [7] acknowledge the importance and impact of understanding the location information in tweets, including where the damage is.

Unlike the previous studies, this study will focus on multilingual tweets and identify locations in the region which directly provide where the disaster is happening.

III. METHODOLOGY

Typhoon Karen and Lawin hit the Philippines last October 2016. The typhoon brought weeks of torrential rain which caused flooding, landslide, damages that cause another emergency in several areas. During the disaster and its aftermath, subscribers of Twitter used this social medium to tweet information about the disaster in local languages such as Pangasinan, Iloko, and bolinao.

Tweets about disaster and emergencies will be gathered first via extracting tool and extract tweets related to disaster and emergency in the region during October 2016 with hashtag #LawinPH and #KarenPH. The translation of disaster and emergency-related tweets will be used as a filter set to identify and extract the local language. Lastly, it will manually identify disaster and non-disaster tweets as a training for classification.

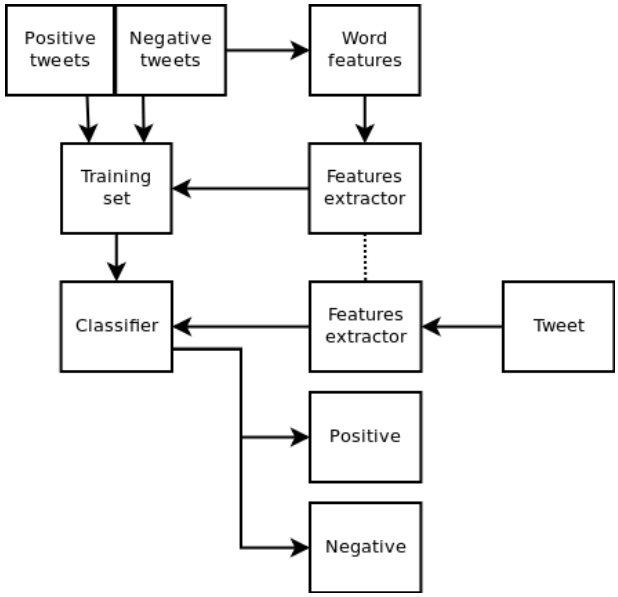

Fig. 2 shows the Process of Classification of Multilingual disaster tweets

A. Data Acquisition, Extraction, and Filtering

The researcher is testing multiple analytical tools that extract twitter data. One of the tested analytical tools is rapidminer. In rapidminer, the twitter connector allows us to access data from the social media twitter directly from the software. It can directly search phrases, tweets. The process begins with Connecting a twitter account, and a twitter connector that uses an authentication mechanism called OAuth 2.0. While the analytical software rapid miner is a good tool, it only provides limited data. Due to its limit, the researcher will use rapidminer as a database storage only.

Python is one of the growing programming languages as of this time, and there are many available python clients code to use. So the researcher will use python to extract tweets. Twitter is generous to download its data, it provides REST APIs so that we can interact with their service. Tweepy is one of the most useful to use. But to authorize our app, we need to use OAuth interface

After downloading the twitter data, we need to filter multilingual tweets using rapidminer. In order to identify and extract multilingual tweets, translation for the different disaster and emergency-related tweets will be done manually by research, interview, and local dictionary. The native and elderly are one of the targets to interview since most of them are well-versed in the local language. One application available as a translation tool for Pangasinan is the app Pangasinan-English Dictionary [8]. The translated keyword will be used as filtering word in the acquired database from disaster-related tweets. To start the filtering process, the acquired translated keyword related to disaster and emergency will be used to search specified keyword in downloaded twitter database and will be saved into another database.

B. Classification of Disaster Tweets

The classification of tweets begins when the tweets are filtered and save to rapidminer database. While extracting from multilingual tweets is done by filtering, the classification of tweets will be done automatically thru SVM. The automatic classification of tweets begins with the manual classification of a dataset which serves as the ground truth for evaluating the performance of the used machine classifying algorithms. To classify multilingual disaster tweets, we need first to manual identify which is disaster-related and non-disaster related tweets. In rapidminer, training set (data table) should be used as an input.

IV. DISCUSSION AND RESULTS

A. Overview

With the use of interview from the elderly and with the translation app tool, the researcher found the translation from Iloko, Pangasinan, and Bolinao which is the language and dialect spoken in the region which could be used later on classifying languages from the extracted tweets.

Translating multiple keywords from disaster and emergency-related words will help us to identify and could possibly use as a training set for future classification because the keywords are related to each other. But before we can use the translated keywords, we need to download the data first from Twitter for the duration of October 2016 with the hashtags keyword #LawinPH and #KarenPH.

The reason why the researcher will use the hashtag #LawinPH and #KarenPH to filter tweets because it is the most popular hashtag during the disaster, and it was even advertised media such as television and government websites such as PAGASA.

B. Extracting and Filtering Tweets

Using python extraction tool, a data was downloaded and filtered with the hashtag #LawinPH for the duration of October 17 to 22, 2017. The downloaded twitter data are stored in twitter_lawin.csv databased from JSON (JavaScript Object Notation) format which it human readable. Below is an example of one tweet in JSON format. "Delap la lamet ed sikami dia! #LawinPH# " After filtering the downloaded data using Pangasinan **language** keywords, it only shows limited results that could not be used as a training set for classification.

C. Classification and Results

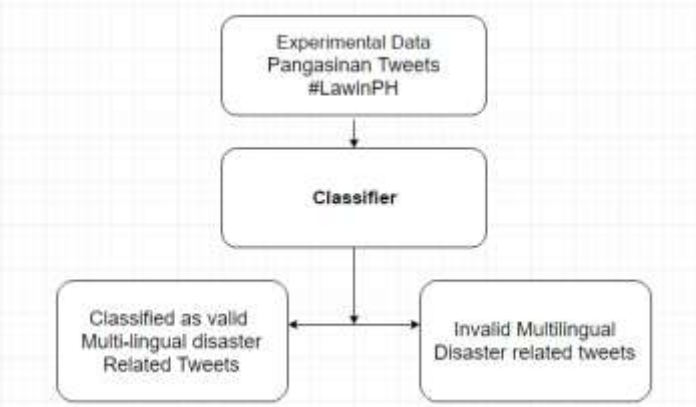

Fig. 3 Experimental Process

Due to the limited number of tweets that can be used for classification, an experiment was conducted in a group of students. 46 students participate in the experiment and assumingly created tweets in local language Pangasinan with hashtag #LawinPH. It generates a total of 435 multilingual tweets. SVM was used as a classifier because several studies concluded that SVM classifier provides better results and outperforms another classifier such as Naïve Bayes.

The researcher takes 100 samples for training database and stores the frequency of disaster-related and non-disaster related tweets. Thus, for training the SVM classifier, the researcher use 200 tweets as a training set for all the category. After training the classifier, it can now be used to classify the remaining tweets that weren't used as a training set. For testing, the researcher follows the same procedure of calculating the frequency of disaster-related and non-disaster

related and pass them as features to the classifier. The classifier classifies the tweet as disaster-related and non-disaster related tweet with the following results.

	Multilingual Disaster Related Tweets	**Non-Related**	**Total Tweets**
Training Sets	100	100	**200**
Testing Samples	120	115	**235**
			435
True Positive	**109**	90.8%	
True Negative	**102**	88.7%	
Classification Accuracy		**89.8%**	

Fig. 4 Initial Experimental Results

Based on the given result, the remaining 235 tweet was used as testing samples, and out of 120 testing samples, 109 was identified by the classifier as Multilingual disaster-related tweets and provides an overall classification accuracy of 89 percent. This study is limited only to classification, while characterization is important, a future study on characterization of multilingual tweets [9] is an interesting study to apply in the local language.

V. CONCLUSION

A. Conclusion

The extraction of multilingual disaster and emergency-related tweets is important is interesting study because the life of a person which speaks a very rare dialect is important as the same as the person speaking a major language. In the future, further study should be conducted not limiting in Region 1 but also in other language and dialect. Further study in the investigation on how to extract location information that is hidden in hashtags and from all other languages not just in the region could be an important study to determine.

B. Further Study

In the further study, the researcher plan to create a mapping system for the disaster-related tweets by identifying the tweets coordinates. After obtaining the coordinates, the location will be mapped using integrated maps with Google Maps API integration. Maps shall be used as a crowdsourcing to identify the disaster and emergency-related tweets in Ilocos Region which gives bigger changes of possible projecting the exact location. The system will also serve as an application to validate the data based on the number of tweets, The number of tweets in the location, will determine the validity of the tweet. Information from the government agencies which areas are disaster and emergency prone areas will help us also to identify the validity of the tweets. The future proposed system will be open to government agencies, LGU and to the public for possible determining the exact location of the emergency. When the system is complete, it could help the Red Cross and other organization working in disaster relief.

C. Ongoing Study

This research study is interesting to the researcher itself, while the paper and experiment are ongoing, the initial result shows that the model could be utilized for life-saving detection process of disaster tweets and could contribute for greater and faster response.

ACKNOWLEDGMENTS

The researcher would like to say thank you to the administrators of Pangasinan State University. Special thanks to our College Dean Dr. Julie Lomibao, our Research Coordinator, Dr. Nova Arquillano, and our Business Administration faculty and friends.

REFERENCES

[1] (n.d.). Retrieved July 07, 2017, from http://www.csun.edu/~lan56728/majorlanguages.htm

[2] Russell, J. (2011). Philippines named social networking capital of the world. Asian Correspondent.

[3] Stowe, K., Paul, M. J., Palmer, M., Palen, L., & Anderson, K. (2016). Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media* (pp. 1-6).

[4] Parilla-Ferrer, B. E., Fernandez Jr, P. L., & Ballena IV, J. T. (2014, December). Automatic Classification of Disaster-Related Tweets. In *Proc. International conference on Innovative Engineering Technologies (ICIET)* (p. 62).

[5] Li, H., & Sakamoto, Y. (2015). Re-Tweet Count Matters: Social Influences on Sharing of Disaster-Related Tweets. *Journal of Homeland Security and Emergency Management*, *12*(3), 737-761.

[6] Regalado, R. V. J., Kyle Mc Hale, B., Garcia, J. P. F., Kalaw, K. M. D. F., & Lu, V. E. (2015). FILIET: An Information Extraction System For Filipino Disaster-Related Tweets.

[7] Lingad, J., Karimi, S., & Yin, J. (2013, May). Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on world wide web* (pp. 1017-1020). ACM.

[8] Pangasinan-English Dictionary - Android Apps on Google Play. (n.d.). Retrieved July 07, 2017, from https://play.google.com/store/apps/details?id=com.pangengdictionary

[9] Kupsch, S. (2017). CHARACTERIZATION OF VALUABLE INFORMATION FROM SOCIAL MEDIA NETWORKS DURING NATURAL DISASTERS. *MATTER: International Journal of Science and Technology*, *3*(2).

Dr. Randy Joy M. Ventayen is currently the University Web Administrator of Pangasinan State University. He is currently taking up his Doctor in Information Technology. A native speaker of Pangasinan language

He handled different designation in the institution such as Coordinator of Web Services, and College Research Coordinator. He initiated the development of eLearning platform in the PSU Open University Systems and pioneered in blended learning approach in the Business Administration Program of PSU Lingayen Campus.

His research presentation experience of this year includes a presentation at Nanyang Technological University in Singapore for international presentation and University of the Philippines Open University – Ncodel 2017 for the national presentation. Dr. Ventayen also received the award as best paper during the 5th International Conference on Business, Law, and Education held at Hotel Benilde Maison last September 17-18, 2017.