Classification of Local Language Disaster Related Tweets in Micro Blogs

Randy Joy Magno Ventayen
Pangasinan State University, Philippines
dayjx@yahoo.com

Asia Pacific Journal of Multidisciplinary Research
Vol. 6 No.1, 10-14
February 2018
P-ISSN 2350-7756
E-ISSN 2350-8442
www.apjmr.com

Date Received: November 5, 2017; Date Revised: January 6, 2018

Abstract *– In Southeast Asia, Philippine is one of the disaster-prone countries which was hit by typhoon Lawin (international name: Haima), and Karen (international name "Sarika") last October 2016, the two typhoon swere named as one of the strongest typhoons that hit the country and the region 1. On some numbers of tweets in social media, there are local languages posted by the local users such as Pangasinan 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. Based on the findings, translation of selected typhoon-related words helps to filter the multilingual tweets and classified the tweets using Naïve Bayes algorithm.*

Keywords *– Natural Language Processing, Text Analysis, Data Mining.*

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 typhoon are related to disaster and emergency are written in English, but thereare 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.

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 those studies are focusing on Tagalog and English only, from these current studies, the proponent realized that there is a need of equal treatment among 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. Thus this study aims to help the government institution especially those in disaster risk management, this utility model could not be only used in the local region, but it will be a guide to those speak local language nationwide.

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.

According to the Twitter blog from @twiterindia [3], 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 with NGOs, citizens, government agencies for disaster relief operations.

Based on the findings of the paper“ Identifying and Categorizing Disaster-Related Tweets” where the tweets during Hurricane Sandy which impacted New York in 2012 was used. The researcher proposes an annotation scheme 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. [4]

Another study was conducted in Automatic Classification of Disaster-Related Tweets [5] which research about the classification of disaster-related tweets in MetroManila 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 retweet thus provide awareness to the public. This study was confirmed that re-tweet count matters that 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[6].

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. This study focusses in the Pangasinanlanguage because 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. This study could also contribute to the development of several models such as Filipino Information Extraction Tool for Twiter. [7]

OBJECTIVES OF THE STUDY

The Objective of this paper is to extract tweets from specific duration during the typhoon Karen and Lawin hit the country. Second is to filter local language tweets from Pangasinan. 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?

METHODOLOGY

Typhoon Karen and Lawin hit the Philippines last October 2016. This 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 language Pangasinan.

Tweets about disaster and emergencies will be gathered first via extracting tool. The tool will 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 based on the translated keyword. Lastly, it will manually identify disaster and non-disaster tweets as a training for classification and will use the remaining data for testing. While there is no guarantee of a number of tweets from the local language, the proponent will test an experiment from a group of the student as a fall back plan in case number of tweets is not sufficient for classification.

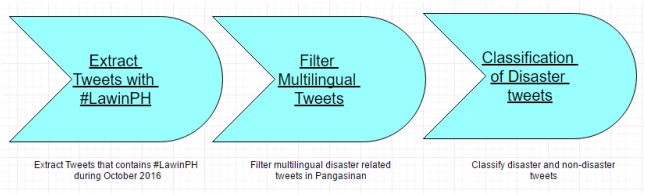

Fig. 1. Process of Classification of Multilingual disaster tweets

Data Acquisition, Extraction, and Filtering

The researcher is tested multiple analytical tools that extract twitter data. One of the tested analytical tools is rapidminer. In rapidminer, the twitter connectorallows us to access data from the social media twitter directly from the software. It can directly search phrases, tweets. The processbegins 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. The researcher used Python extraction tool to extract the data fromthe Twitter database.

After downloading the twitter data, we need to filter multilingual tweet. 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 of Pangasinan. 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.

Data Classification

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.

RESULTS AND DISCUSSION

With the use of interview from the elderly and with the translation app tool, the researcher found the translation from Pangasinan which is the language and dialect spoken in the region which could be usedto filter tweets from the extracted tweets.

Table 1. Translation of Typhoon related words

English	Tagalog	Pangasinan
Typhoon	Bagyo	Bagyo
Flood	Baha	Delap
Rain	Ulan	Uran
Bridge	Tulay	Taytay
Help	Tulong	Tulong

Translating multiple keywords from disaster and emergency-related words will help us to identify and could possibly use as a training set for future filtering and 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 with 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 since the Philippines is considered as 3rd most disaster country [9].

Extracting Tweets

Using Python extraction tool, a data was downloaded and filtered with the hashtag #LawinPH for the duration of October 17 to 22, 2017. Sample Request Code:

https://api.twitter.com/1.1/search/tweets.json?q=%23delap&since_id=24012619984051000&max_id=250126199840518145&result_type=mixed&count=4

The downloaded twitter data are stored in twitter_lawin.csvdatabase from JSON (JavaScript Object Notation) format which it human readable. Below is an example of one tweet in JSON format:

"Delap la lamet ed sikamidia! #LawinPH"

Filtering Tweets

After extracting the data from the tweeter, the proponent filtered several tweets based on the translated words. After filtering the downloaded data using Pangasinan language keywords, it shows 12 tweets for classification. Since the researcher only filtered only 12 tweets in local language which may not possible for the classification process, an experiment will be done.

Classification

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 created tweets in local language Pangasinan with hashtag #LawinPH. It generates atotal of 217 multilingual tweets. SVM was used as a

classifier because several studies concluded that SVM classifier provides better results and outperforms another classifiersuch as Naïve Bayes [10].

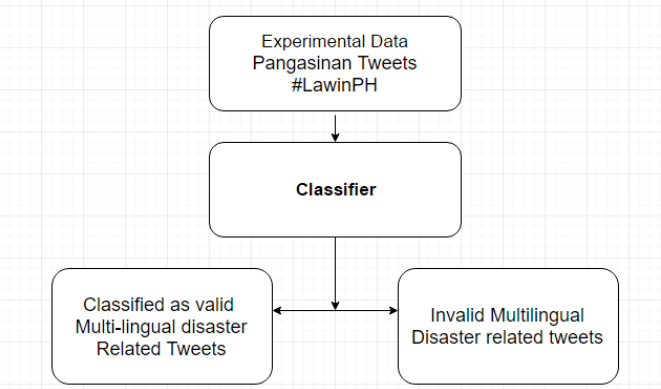

Fig. 2. Process of Classifier in Identifying disaster and non-disaster tweets

The researcher takes 50 each sample for training database and stores the frequency of disaster-related and non-disaster related tweets. Thus, for training the SVM classifier, the researcher uses 100 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.

Table 2. Classification Accuracy

Training	**# of Tweets**	
Sets	50+50	100
Samples	60+57	117
True Positive	54	90.00%
True Negative	51	89.47%
Accuracy	105	89.74%

Based on the given result, the remaining 117 tweet was used as testing samples, and out of 60 testing samples, 54 was identified by the classifier as multilingual disaster-related tweets and provides a classification accuracy of 89percent, while there are 51 identified as non-disaster related tweets.

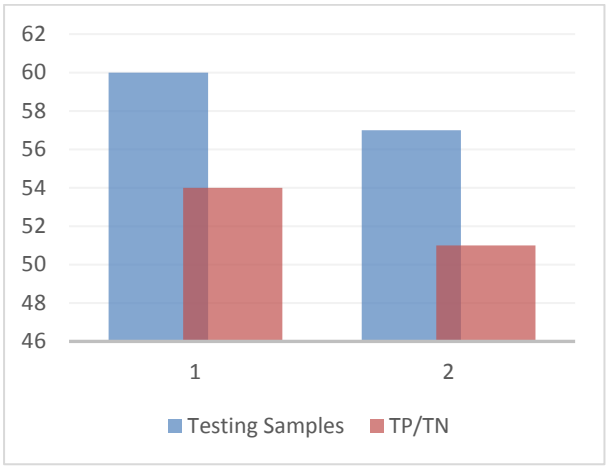

Fig. 3. Relationship between Testing Sample andTrue Positive (TP) and True Negative (TN)

CONCLUSION AND FURTHER STUDY

The extraction of multilingual disaster and emergency-related tweets is important and 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. This model and experiment could also be implemented in other local language and will help the government MIS together with another technology sector that focuses on the study of disaster management using social media analytics. Life is the most important in the universe and the goal of this study is to save a life.

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 model 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 proposes 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.

REFERENCES

[1] *Major Languages of the Philippines*. [online] Available at http://www.csun.edu/~lan56728/majorlanguages.htm [Accessed 21 Jul. 2017].

[2] Russell, J. 2011, May 15. The Philippines named social networking capital of the world. Asian Correspondent. Retrieved from https://goo.gl/hmzV4r

[3] Mahima Kaul, 2016, https://goo.gl/Pc2UV8

[4] Stowe, K., Paul, M., Palmer, M., Palen, L., & Anderson, K. (2016, November). Identifying and Categorizing Disaster-Related Tweets. In *Conference on Empirical Methods in Natural Language Processing* (p. 1).

[5] Beverly Estephany Parilla-Ferrer et. al., 2015, Automatic Classification of Disaster-Related Tweets

[6] 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.

[7] Ralph Vincent J. Regalado, 2015 FILIET: An Information Extraction System for Filipino Disaster-Related Tweets

[8] Pangasinan-English Dictionary https://play.google.com/store/apps/details?id=com.pangengdictionary&hl=en

[9] The Philippines is 3rd most disaster-prone country.n.d. Retrieved December 6, 2016, from https://goo.gl/ND7NE9

[10] Beduya, L., and Espinosa, K. 2014. Flood-Related Disaster Tweet Classification Using Support Vector Machines.Proceedings of the 10th National Natural Language Processing Research Symposium, pages 76-81, De La Salle University, Manila, February 21-22, 2014.

COPYRIGHTS