Twitter Analysis for Disaster Management

Varun Agarwal
Student,
IT Department,
D. J. Sanghvi College of Engineering
Mumbai.
varun.agrwl96@gmail.com

Utkarsh Dubey
Student,
IT Department,
D. J. Sanghvi College of Engineering
Mumbai.
utkarsh7dubey@gmail.com

Bhavya Shah
Student,
IT Department,
D. J. Sanghvi College of Engineering
Mumbai.
bhavyashah239@gmail.com

Stevina Correia
Professor,
IT Department,
D. J. Sanghvi College of Engineering
Mumbai.
Stevina.dias@djsce.ac.in

Abstract

Disaster always leads to response from the people and in this new technology generation, people give their feedback and raise concerns through their views on social media platform. Twitter is used for such response and information exchange. This paper reports in depth twitter analysis of Nepal Earthquake and identification of various factors in the disaster module.

Based on Twitter data collected between last week of April and first week of May 2015. 40,236 raw messages apparently related to the Nepal earthquake were retrieved, preprocessed and analysis were done on the same. This paper shows measures and analysis of the situation using geolocation feature for identifying danger zones and visual analytics of the dataset. With the help of the keyword generation algorithm different parameters were introduced. The results show that our disaster module can work on different hashtags and no need for prior defining of parameters by humans provided the data set is available for that unique problem.

Keywords- Twitter Analysis; Disaster; Tweepy; Stemming; Geo- location.

Introduction

We analyze the contents of Tweets regarding a particular Hashtag and develop a detailed analysis of the contents and mapping the co-ordinates of the users using Google maps with intent to gain a broad understanding of the nature of concerns and communication. We have used a machine Learning algorithm to automatically generate the keywords based on the data fed and then generate a detailed analysis upon those keywords while also mapping the geo-location co-ordinates of users, categorizing the users as 'In danger' and one's 'providing help' and also adding their respective tweets along the username on the map and thus creating a Danger map with twitter data

Twitter, Facebook and other social networking sites became an invaluable tool for millions of people caught up in the aftermath of the Japan earthquake which took place in 2001. Even the US State Department resorted to using Twitter to publish emergency numbers, and informing Japanese residents in America how to contact families back in Asia. Relief organizations used Twitter to post information for non-Japanese speakers to list of shelters for those left homeless and thus responding to large scale disasters. We have used the contents of Nepal earthquake which occurred on the 25th of April 2015 with a magnitude of 7.8 as a case study of the problems which appeared a week or more after the disaster occurred. A confirmed report (June, 2015) states that more than 8,702 people died. Victims suffered from shortages of critical resources such as water, energy, medicine and shelter.

Formal reports on a disaster often focus on the coordination of supplies and service while social media was instrumental in distributing resources

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

and information in the aftermath of the earthquake. Developers creating online earthquake relief resources had to devise solutions that would be easily understood and could function on limited bandwidth (NASA). Our work reports data analysis aimed at identifying the most pressing issues and tensions that arose starting about a week after the initial earthquake.

The aim is to describe how problems that appear during disaster relief can be identified based on the data gathered from Twitter and provide analysis of the same can be carried out.

Related Work

Communication means sharing or exchange of thoughts, messages, visuals or signals. It requires a sender and a receiver. Twitter plays a major role in the dissemination of information during natural disasters. However, it has several limitations, especially regarding the reliability level of the information [6].In the study carried out by Plotnick et. al. [1], it can be seen that emergency management agencies in the U.S. still do not fully exploit the capabilities of social media channels due to several barriers such as lack of policies and guidelines for its use, lack of sufficient staff, information overload and reliability of the collected information

A paper [1] mentions the trust in tweets related to disaster. Trust in tweets is explained by a paper explaining the factors like support and donations during the times of disaster.

Although Twitter may be a really helpful supply for aggregation information, it conjointly presents a giant challenge in terms of constructing sense of it due to information overload. There were a large number of duplicate tweets and tweets with missing values in the dataset [7]. Upon further investigation, we noticed that the iterative search using the Rest API of Twitter seems to cause these duplicates as Twitter API is not live stream and it provides a pool of tweets at a time window. During data cleaning, they also omitted the retweets as their aim was to specify the sentiments or opinions of individual [2].

In particular, machine learning, data mining, and natural language processing have proven very useful in extracting, processing and classifying disaster-related social media data [2-3]. For example, Ashktorab et al. [4] used a combination of classification, clustering, and extraction methods to identify and organize information from Twitter for disaster responders. Among the many different category schemes have been used in classifying disaster tweets, the categories that we used in this study were especially influenced by two sources. In coding tweets related to trust in emergency responders during Hurricane Sandy, Hughes and Chauhan [5] used categories that included donations (of time, money, or supplies to relief efforts), and support (expressions of gratitude or support).

Both of these schemes include sentiments but only positive ones. Manual examination of tweets from the Nepal data set they gathered also showed negative sentiments, even anger. Thus, in developing our categories, as explained below, we wished to capture both negative sentiments and positive sentiments.

Proposed System:

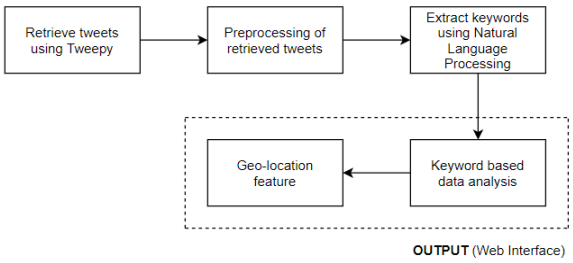

Fig 1: Proposed Architecture

Modular Description

Fig 1 describes the architectural flow of the proposed solution. The description of each block is as follows

1) Retrieve tweets using Tweepy

First step in our system is to collect the tweet data. This was accomplished by using Tweepy. The data from Tweepy was stored in a csv format. Filters used for collecting the

data were hashtags #nepal, #nepalearthquake, #nepalquakerelief, #kathmandu, “helping Nepal”, “nepal victim”, “nepal earthquake”.

We did not limit the geographical boundaries of the data collection. Final result of step 1 was a complete data set needed for our analysis in this paper.

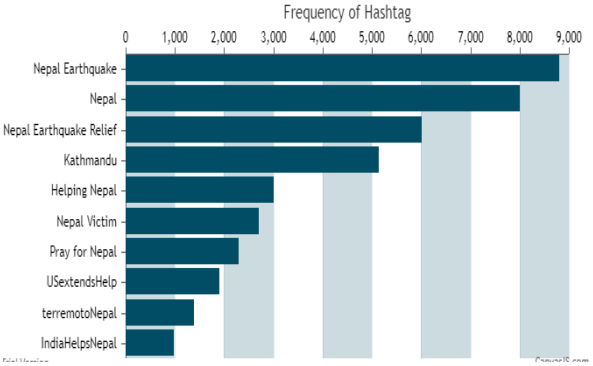

Figure 2 Frequency of Hashtag

2) Pre-processing of retrieved tweets

Figure 2 shows the frequency of occurrence of each hashtag

Next step after acquiring the required datatset was to clean the data and have only relevant information for further analysis. The dataset obtained in step 1 had a lot of issues. One major issue was removal of encoded characters and smiley faces. Most tweets had some or the other smiley depicting their emotions which when retrieved are encoded such as ‘/xe80’, ‘/xe98’ and so on. Sample data set is shown in Fig 3, removal of such encoded characters was a must. Geo coordinates of some tweets were not available so for that their home location coordinates were taken which is stored by twitter at the time of their account registration.

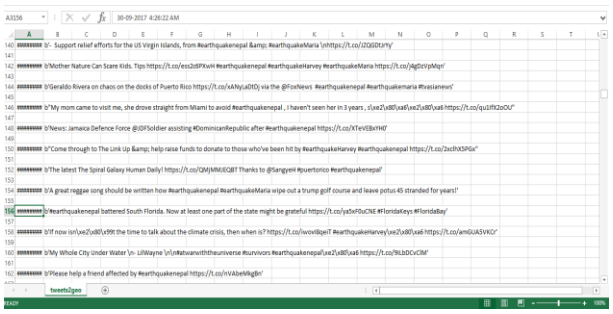

Figure 3 Raw Data Set (Not clean)

Also in dataset, if a tweet was retweeted 38 times, then it appeared 38 times in a dataset which should not be the case. So removal of duplicate tweets was required.
After all the above steps, we had our final pre-processed dataset as shown in figure 4.

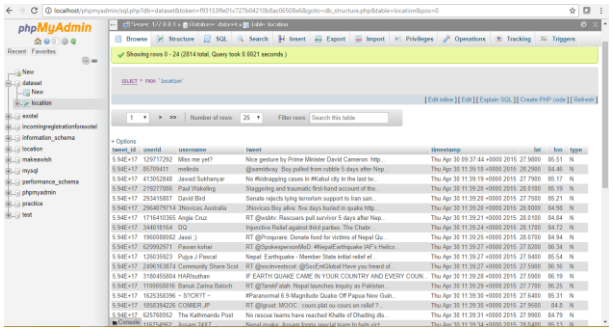

Figure 4 Standardized Data Set

3) Extract Keywords using Natural Language Processing

Next step after pre-processing was to decide on keywords. Now our main aim was to automate keyword selection based on any given dataset, and manual selection of keywords, which is usually done in all systems destroys the purpose. Thus we needed to automate the keyword extraction process from the dataset.
So firstly, we tokenized all the tweets in the pre-processed dataset. Afterwards, we removed all the stop words like “a”, “an ”, “the ” and so on along with punctuation’s to focus only on major words that would be our potential keywords. Now we also need to cluster similar meaning words to the same word root as they convey same meaning. For

example: Consider a word ‘help’ which can be written as ‘helping’, ‘helps’, ‘helped’,etc. as per the need of a sentence but should be clustered in a same root word ‘help’. This is achieved using Stemming which roots all the words to its base root word. Thus in our case, all words our rooted to base word ‘help’ which helps in our next step for keyword extraction.

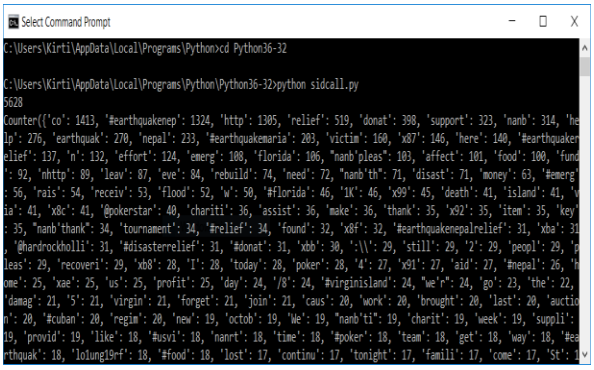

Figure 5 Keyword Generating Algorithm

Next step is to find the frequency count of each root word to find the most significant keywords. Fig 5 shows the frequency of words in our data set after applying the algorithm. Thus, we used Counter to count the occurrence of keywords and thus found the top most important keywords based on which further analysis is to be done.

4) Keyword based Data Analysis

In this phase the Output received from the algorithm used for automatically generating the keywords based on frequency of the keywords occurred in data and its significance is saved in a document which is termed as a dictionary for the dataset and is directly fed to the Canvas JS file which is used for generating various graphs, pie charts and helps in generating our desired graphical analysis results.

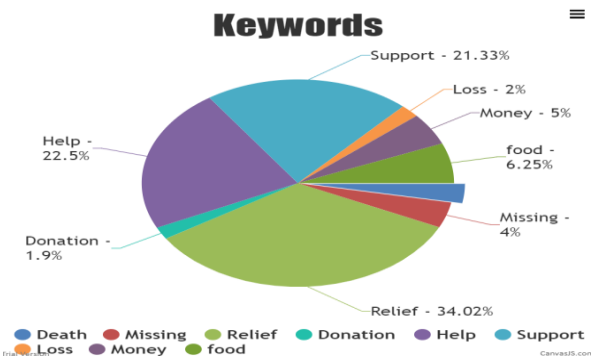

Figure 6 Distribution of Keywords

Figure 6 shows the percentage distribution of the keywords.

5) Geo-location feature

In this phase we will extract the geo-coordinates of all the tweets and classify the tweets and plot them on a map with different markers along with their tweets and username to classify the user needing help and providing help during the disaster. The Dataset is retrieved using PHP scripts and is stored in an XML file which will be store all the details in various XML tags and then this XML file will be passed to the HTML file which is used for generation and display of Maps by incorporating Google Maps API .The custom marker is then created to label the One’s who are in Danger and the one who are providing help and then this marker is plotted on the map to display all the users based on their locations and the username and Tweet posted by them is even displayed by clicking the marker which can be used for further contacting the user who is in danger or provides help. Fig 10 shows the implementation of the geolocation feature.

Benefits of proposed solution

The benefits of the proposed solution will be

- Identification of Trusted Sources
- Identifying Rumours

- Reporting Danger Areas
- Displaying Help Service Facility
- Gateway for Donation
- Facilitating the provider and the needy
- Missing and found
- Existing System Collaboration

Results

Our project is focusing on providing a detailed analysis on the contents of the twitter retrieved on basis of the keywords extracted from the dataset and generating a statistical analysis to understand the matter easily. In this project we have tried to develop our analysis in form of graphs, pie-charts and Mapping the location of users in Google Maps.

Each chart in the next two sections comprises different unique issues in each category. As Graphs are very simple to understand and easier to see the relationship between entities, We have used various types of graphs in our analysis like Bar-Graph, Line-Chart, Histogram, Customized Pie-chart, step-line charts, box and whisker charts for showing analysis using different perspective to gain a simpler and wide understanding of the matter. Our portal is an all in one stop solution where any user can visit the portal anytime and easily find all the desired results by giving the Hash tag of trending the event. The Dataset gets downloaded from Tweepy and pre-processed and a statistical-analysis is generated on basis of the keywords in very quick time and users will also be plotted to their respective location.

Figure 7 shows the number of death occurred at each day during the earthquake. Using various Line-charts, bar charts and analytical tools we have provided a statistical analysis of the loss of money, deaths, injuries and people missing, Health and Medical issues, NGO's ready to provide help for the victims, donations for the welfare etc.

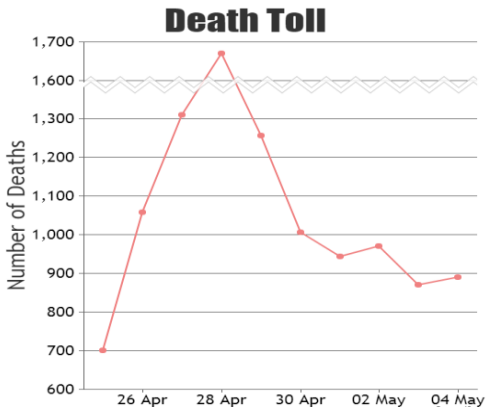

Figure 7 Death Count

In Various other charts we have explained the amount of Donation tweets posted during the Nepal earthquake, comparative analysis of the deaths and injuries during the earthquake and give an exact information about the loss of life, total damage incurred, Number of injuries during the earthquake and the exact amount being donated for victims and also provide all the necessary links of organizations who were striving to help the needy.

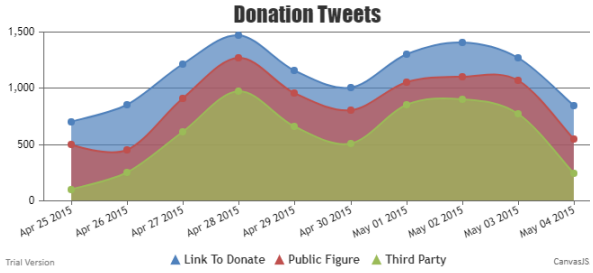

Figure 8 Donation Tweets

Fig 8 depicts donation tweets received in the time frame.

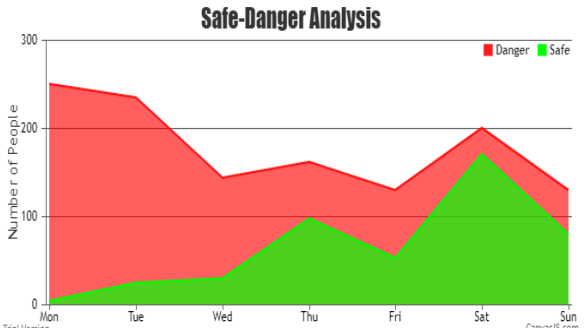

Figure 9 Safe-Danger Tweets

Fig 9 shows the analysis of number of safe people against the number of people in danger.

Our project also aims at giving a detailed list of all the users found missing during the earthquake and the one's found during the rescue process and also share the exact location of the users who are actively updating regarding the issue and plot all the users such

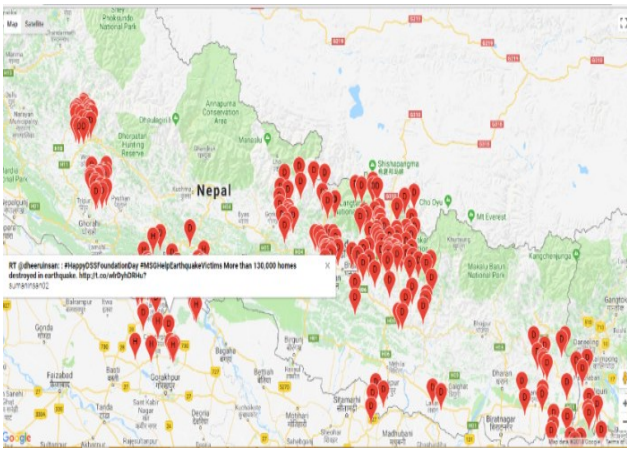

Figure10. Implementing Geo-location Feature

that any user can know the location of any user active and can contact using their details.

We have labeled users on the status of their tweets posted whether a person is in danger or is safe and plotted all the users on Google maps so they can see that whether in their surrounding people are safe or are in danger to extend their help as shown in fig 10.

Conclusion

We have collected Nepal Earthquake related tweets from tweepy and identified keywords that are to be analyzed using an algorithm. Further analysis were done based on these keywords resulting the output, an overview of disaster occurrence and its affects. Automation of keywords identification resulted in a disaster management module which will identify keywords based on any given disaster for further analysis. Future work would include creation of a generic module that would provide analysis for any hashtag.

References

[1]NazanÖztürka SerkanAyvazb ,"Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis"(Telematics and Informatics Volume 35, Issue 1, April 2018, Pages 136-147)

[2] H. Li, N. Guevara, N. Herndon, D. Caragea, K. Neppalli, C. Caragea, et al., "Twitter Mining for Disaster Response: A Domain Adaptation Approach," in the 12th International Conference on Information Systems for Crisis Response and Management, Kristiansand, Norway, 2015.

[3] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting Information Nuggets from DisasterRelated Messages in Social Media," in the 10th International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, 2013.

[4] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining Twitter to Inform Disaster Response," in the 11th International Conference on Information Systems for Crisis Response and Management, University Park, Pennsylvania, USA, 2014.

[5] A. L. Hughes and A. Chauhan, "Online Media as a Means to Affect Public Trust in Emergency Responders," in the 12th International Conference on Information Systems for Crisis Response and Management, Kristiansand, Norway, 2015.

[6]Meera R.Nair[1] G.R.Ramya[1] P. BagavathiSivakumar[1],"Usage and analysis of Twitter during 2015 Chennai flood towards disaster management",Procedia Computer Science Volume 115, 2017, Pages 350-358

[7]L. Plotnick, S. R. Hiltz, J. A. Kushma, and A. Tapia, "Red Tape: Attitudes and Issues Related to Use of Social Media by US County-Level Emergency Managers," in the 12th International.