Disaster Analysis Through Tweets

Conference Paper · August 2015

DOI: 10.1109/ICACCI.2015.7275861

2 authors:

Himanshu Shekhar
B.V. Bhoomaraddi College of Engineering and Technology (BVBCET)
3 PUBLICATIONS **63** CITATIONS
SEE PROFILE

Shankar Gangisetty
International Institute of Information Technology, Hyderabad
40 PUBLICATIONS **443** CITATIONS
SEE PROFILE

Disaster Analysis Through Tweets

Himanshu Shekhar*, Shankar Setty†
Department of Information Science and Engineering
B.V. Bhoomaraddi College of Engineering and Technology
Hubli, Karnataka, India
Email: *himanshu508@hotmail.com, †shankar@bvb.edu

***Abstract*—Social networks offer a wealth of information for capturing additional information on people's behavior, trends, opinions and emotions during any human-affecting events such as natural disasters. During disaster, social media provides a plethora of information which includes information about the nature of disaster, affected people's emotions and relief efforts. In this paper we propose a natural-disaster analysis interface that solely makes use of tweets generated by the Twitter users during the event of a natural disasters. We collect streaming tweets relating to disasters and build a sentiment classifier in order to categorize the users' emotions during disasters based on their various levels of distress. Various analysis techniques are applied on the collected tweets and the results are presented in the form of detailed graphical analysis which demonstrates users' emotions during a disaster, frequency distribution of various disasters and geographical distribution of disasters. We observe that our analysis of data from social media provides a viable, economical, uncensored and real-time alternative to traditional methods for disaster analysis and the perception of affected population towards a natural disaster.**

***Keywords*—Twitter, Natural Disaster, Sentiment Anlysis**

I. Introduction

Microblogging can been seen as a form of lightweight chat which allows users to share short messages to the internet community. There are many popular microblogging services available, which includes Twitter, Plurk, Jaiku etc. Our focus for the purpose of this research paper is on Twitter, which allows its user to share short messages called Tweets which are essentially of 160 characters or less. These messages (*tweets*) can be sent and retrieved through a variety of means and front-end clients, including text messaging, e-mail, the web, and other third-party applications, which are enabled through Twitter's public API.

Microblogs are mainly used to share information and track general public opinion during any human-affecting event such as sports matches, political elections, natural disasters etc. Particularly in recent years, during the event of a natural disaster which usually occurs without any warning, social media has gained a lot of attention as an additional medium for crisis communication. Largely Twitter, being most popular of microblogging websites, is being used by users to share news, photos and observations from the crisis site globally. Twitter being the real time data generated by the user community itself, is largely uncensored, easily accessible and is user-centric. Thus it provides an easier, comprehensive and economic analysis approach for users on fingertips. The hard problem is the sentiment analysis, specifically the behaviour and emotions that users express from natural disaster crisis site to the distributed users across the globe.

The major motivation behind the proposed work is the fact that gathering information about people's emotions during and aftermath of a disaster is a task which is not yet automated and is mainly dependent upon manual surveys and interviews. We propose to automate this task by making use of the data present on the social media websites. In this paper, we demonstrate an innovative application to analyze the effects of disaster on people and society through the use of twitter posts generated by users. The tool has ability to store the relevant (filtered) data from twitter and provides a detailed graphical analysis which encompasses user's emotions, disaster frequency and geographical distribution of several disasters such as earthquake, forest fire, floods, and droughts.

The main contributions of this work are:

- demonstration of geographical distribution of selected natural disaster within a given time period solely through the use of tweets.
- continent-wise occurrence frequency of selected natural disasters.
- analysis of people's sentiment during a disaster by applying sentiment-analysis on the tweet content.

II. Related Work

Research on social media in disaster events is growing, and ranges from examination of common photo repositories[1] to social networking sites[2]. More specifically, interest in microblogging in emergency management activities is on the rise (e.g. [3], [4] and [5]). Early research shows that critical up-to-date and on-location updates can be found in microblog messages about an unfolding crisis, precipitating an interest in robust processing and visualization tools.
Analysis of effects, frequency and distribution of natural disaster has been done in many ways by various agencies and individuals, usually with the help of dedicated satellite data, records maintained by government or by interviews taken from the affected people at disaster sites. It's imperative to find out the emotions of people who are actually affected by the disasters.

In literature we observe that tweets from twitter are used as a back-channel communication system during natural disasters[6]. Twitter has also been useful in coordinating and planning the disaster relief efforts[7]. Emotional state of people who are affected by a disaster previously were analyzed through videotape analysis[8], recordings of affected people were viewed by researchers to perceive their emotional condition. This type of analysis is particularly helpful in relief

978-1-4799-8792-4/15/$31.00

efforts in the aftermath of a disaster. Agencies such as Red Cross or WHO make exhaustive ethnographic inquiries by visiting the victims in order to provide treatment for emotional trauma suffered by the victims.

Specifically, sentiment analysis has been used to find the different levels of concerns experienced by the affected masses during Hurricane Irene[9]. The authors trained a sentiment classifier to categorize messages, which was then used to examine the level of concern and anxiety in the people before and after the Hurricane. This classification was further studied to find out the different levels of anxiety during the disaster for different genders. They also identified broad trends on the usage of words by the participants and provided a ranking of used words. Similar work has also been done to find trending topics, most favorable language used during disaster and emotions present during disaster occurrence in tweets of Filipino users[10]. This study used Latent Dirichlet Allocation and Principal Component Analysis to extract the various topics discussed in the event of a disaster and predicted the most likely topics which affected people may talk about. Specifically, this paper tried to see the difference between the direct victims and other observers of the disasters in terms of their tweet contents and analysis of topics.

Inspired by the authors work in [8], [9] and [10] we propose an automated process of analyzing users' emotions and geographic distribution of disasters using tweets from twitter.

III. Methodology and Constraints

The system works upon the basic idea that there is always a constant supply of streaming tweets in the cyber-universe containing a myriad of data about various domains. With the occurrence of any mass human-affecting event, such as a disaster, political or sporting event etc., there is usually a spike in generation of tweets from the area in question. The data from these tweets are extracted via keywords specific to disaster. The collected data is then filtered into four categories viz. Earthquakes, Forest Fires, Floods and Droughts. The analysis is concentrated around different locations within a geographical area by applying the K-Nearest Neighbour algorithm. The training dataset related to each category is created.

This dataset is a assortment of different lexicons collected from a eclectic set of twitter users from around the globe. Processing of such a diversified dataset present certain Natural Language Processing challenges which has to be taken into account for analysis and interpretation of results. These problems are noted down as follows:

- Esoteric Language and Grammar: Most of twitter post, being short in length, lack proper defined grammar and punctuation which is conventionally followed. Most of users tend to ignore the uses of articles and auxiliary verbs while posting tweets. Further, there is a copious usage of abbreviations, particularly the internet slangs such as LOL for "Laughing Out Loud". Such type of data poses major problems to traditional NLP tools such as parsers.
- Message Length: The tweets being short in length often contains minimal words which are "ideal" for sentimental analysis and have very little lexical redundancy.
- Mention of local references: Messages sometimes refer to specific location, events and other named entities, as well as implied references to locations[4].Thus, one cannot rely on pre-defined entity lists or complex named entity recognition methods.

The collected dataset is itself irregular in the aspects of size and incoming rate, and needs effective pre-processing.

Based on our analysis we found that most of the tweets are redundant, re-tweeted, misleading, unwanted, known as noisy tweets. Thus, while creating the train dataset, pre-processing was done to eliminate such tweets. The dataset is further subjected to analysis on three different cases: (i) frequency of disaster happening in different region over a given time period, (ii) distribution of disasters over different geographical locations and (iii) emotions of the people talking about the disaster on twitter.

IV. Proposed System

In this section, we discuss in detail the various processes of our system methodology. The system is composed of three different modules: Extraction of Data, Sorting of Data and Analysis of Data. Each module is further divided in sub-modules which cater to specific tasks. Figure 1 gives out the high level system diagram of the application. The different phases of the system are detailed down as follows:

A. Data Extraction:

This module deals with the extraction of tweets from the web and storing them locally for analysis.

1) Tweet Capturing: Data extraction was mainly done with the help of Twitter Streaming API. This API is capable of capturing live tweets from the web and parsing them as String objects. These String objects can be further worked upon by the program for storage and manipulation.

2) Redundancy Check: It is possible for the program to capture the same tweet more than once due to re-tweets by the user, recapturing of older tweets etc. This is avoided by regular expression series method based on set threshold wherein random section of incoming tweets are taken (minimum character length = 35) and are checked against the existing tweets. If a high match percent of pattern match is detected, the incoming tweet is discarded.

3) Data Storage: The tweets which pass the Redundancy Check sub-module are then stored systematically in files in ascending order of their time of posting. The system stores the tweet message content, the username of the person who posted the tweets and time-stamp of the tweet. The system is capable of storing even more data such as number of re-tweets and comments upon the tweet.

B. Data Sorting

The stored tweets are now analyzed and their disaster category is determined. This is done by checking each tweet

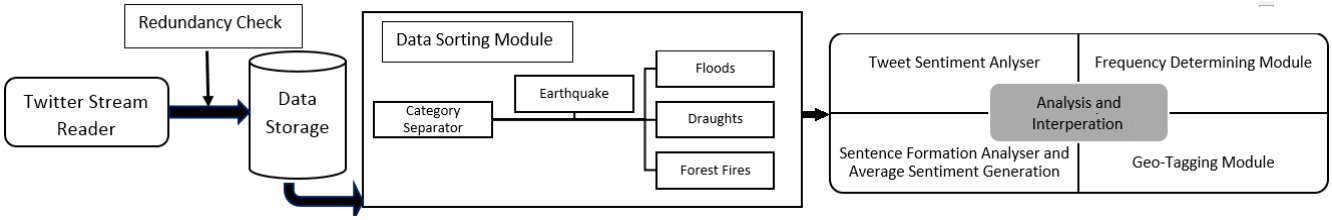

Fig. 1. System architecture for the disaster analysis using tweets

against a set of predefined weighed keywords. A tweet is said to be belonging to a particular category of disasters if it matches 40 percent of the listed keywords. Matching of higher weighed keywords provides greater relevance factor to a tweet. The tweets are then stored in then separate files pertaining to their disaster category along with their relevance factor. The tweets which do not belong to any category are readily discarded. For example, data sorting and tweet relevance for earthquakes are shown in Table I.

TABLE I. SAMPLE TWEET RELEVANCE FOR EARTHQUAKES.

Tweet	Keywords Matched	Tweet Relevance
wasnt in the middle school for 5 minutes today when there was an earthquake drill and i had to sit under a table for 10 minutes kewl	Earthquake, under	42 %
@geostuff - Earthquake rattles Midlothian - Edinburgh Evening News: Edinburgh Evening NewsEarthquake rattles MidlothianEdi... http://t.co/olRFvQFF3J	Earthquake, rattles, news	98 %
@ToranNigrelli - Today is the day people this is not a drill! 3pt Shootout in the Richter at 3. Registration at 2:30, 10 for a two person team...BE THERE!!	Richter	12 %

C. Analysis of Data

The categorized data is then passed on to the third module for analysis and interpretation. This module is divided into four distinct sub-modules which check the data for Disaster Distribution, Geo-Tagging, Occurrence Frequency and Sentiment Rating.

1) Geo-Tagging: The location of disaster is the foremost matter of interest in any type of disaster analysis. To extract the location from the tweets, a geo-filter tag is applied on the tweets to determine their point of origination. However, in many cases tweets are posted by people who are not actually tweeting from the affected region. Often in such a case the disaster location is mentioned within the tweet. The location mentioned in the tweet was determined by the help of Google Maps Javascript API wherein all the tweets are passed through a method one by one which splits the tweets into words and tries to find the co-ordinates of that word through the said API. The method returns the geographical co-ordinates if that particular word exists anywhere in the Google Maps else returns an error. Thus the possible locations mentioned in the post are then found and labeled along with tweets.

2) Disaster Distribution Analysis: This module categorizes the distribution of various types of disasters in the world. It makes use of location determined from the previous sub-module and clusters the tweets based on location co-ordinates mentioned against tweets using K-Nearest Neighbour Algorithm. The clusters are then analyzed over a time period and a distribution of different disasters in different continents over a specific time period is generated by averaging the rate of tweet flow against the location. Table II gives the result for disaster distribution by continents.

TABLE II. DISASTER DISTRIBUTION BY CONTINENTS

Continents	Earthquake	Flood	Drought	Forest Fire
Asia	42%	49%	12%	3%
Africa	12%	5%	28%	15%
Europe	3%	14%	3%	15%
North America	13%	8%	11%	4%
South America	19%	7%	10%	32%
Australia	11%	7%	36%	39%

3) Disaster Occurrence Frequency: Each type of disaster data is analyzed by monitoring the tweet incoming rate for different locations via the geo-tagging module. This live data is aggregated into clusters and different cluster sizes are then analyzed for a particular category of disasters. The number of people tweeting about a particular disaster within a given time period is taken as an indicator for the disaster frequency. This data is then normalized with reference to location-wise clusters to reduce the error due to large amount of tweets coming from a large-impact disaster. Figure 2 shows the occurrence frequency of different disasters.

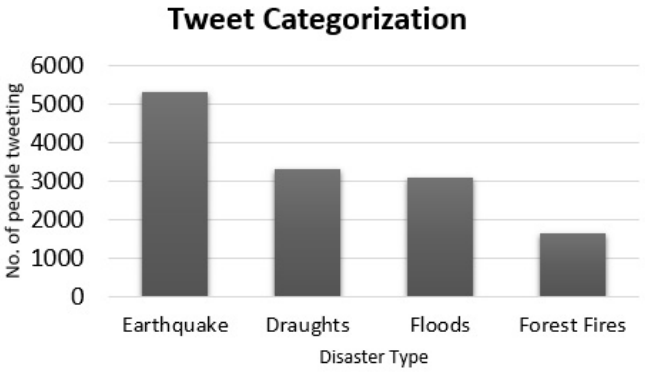

Fig. 2. Occurrence Frequency for Disasters

4) Users' Emotion Analysis: Each categorized tweet is fed to a sentiment analysis function. The function checks the tweets against a dictionary word database which contains a

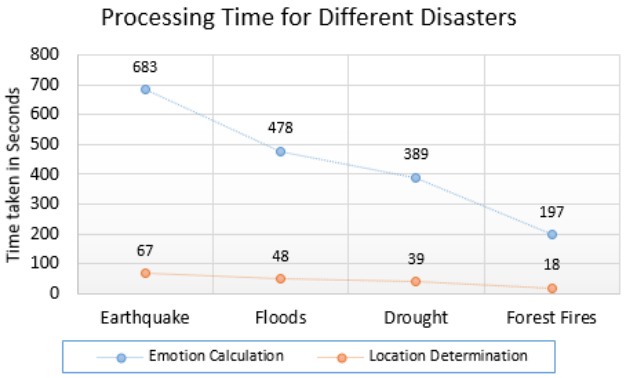

Fig. 4. Time Analysis for Various Calculations

weighted sentiment rating for each word. Each tweet is split into separate words and checked against the dictionary. The sentiment of each word is added to the overall sentiment of the tweet, with positive sentiment causing the overall sentiments to increase and vice-versa. Any abbreviations or internet slang used are checked against a slang dictionary which expands them and resend the expanded words for sentiment determination. Any smileys mentioned in tweets are similarly checked in a weighed smiley dictionary and are also given a sentiment rating. The function also takes note of sentence structure and determines the overall sentiment rating of the sentence, taking special note of any negation words used which might change the overall sentiment rating. Based on the level of negativity present in the tweets, the users' emotion for different disasters is sub-categorized as negative sentiments, unhappy, depressed and angry. Figure 3 shows the sentiment analysis of different disasters by categorization.

V. Results and Discussion

We collected dataset from Twitter microblogging service on four categories of natural disasters viz. Earthquake, Floods, Droughts and Forest Fires. These disasters are very common and usually affect a big mass of human population. The data was collected through Twitter Streaming API and was stored in separate files. The statistics of the tweet dataset from all four categories of natural disaster used in our work is given below.

- Total number of tweets for earthquakes: 21,548
- Total number of tweets for floods: 19,447
- Total number of tweets for droughts: 11,793
- Total number of tweets for forest fires: 7,731
- Average tweet length: 23 words (122 characters) with a standard deviation of 5.3

We analyzed around 60,000 tweets with our system. The time for analysis varied greatly for different type of disasters. The analysis was performed on a machine with 3.2 GHz clock speed, 8 gigabytes of primary memory and 6 & 3 megabytes of L2 and L1 cache memory respectively. Figure 4 shows the time taken for various calculations made by the system.

It can be noted that the time taken for the processing of tweets pertaining to earthquake is the largest. Supplementing this observation with Figure 2, we can observe that the frequency and the tweet streaming rate of earthquakes is the largest, followed by floods and droughts, and trailed by forest fires. Droughts and floods, though lower than earthquake, happen at almost same frequency in the world. Table II demonstrates that though earthquake is most common form of disaster in most continents, occurrence of floods, droughts and forest-fires is highly biased. However, people are more likely to tweet about earthquakes as it's usually more devastating than others. Also, the tweet streaming rate of Earthquake is very high as the disaster itself is very short lived and thus most of tweets generated for an earthquake come within a week of its happening. This isn't with the case of floods or droughts, which may last weeks or months.

From the sentiment analysis module, we can observe that during the disasters, most people who post content on the social media are usually unaffected by the disaster. They show a general negative sentiment (Figure 3), but the number of people tweeting who show high emotional distress (hence which are greatly affected by the disaster) is very low. A very small amount of people have positive sentiment during disasters. This may represent an error either in capturing of unrelated tweets or in the sentiment analysis due to eclectic variety of tweets. We observe that most data during the disaster is uploaded by people who are wishing to share information or news about the disaster, but are themselves unaffected by it. A distribution of disasters across the continents (Table II) is created solely though tweets. This can be extended to find disaster distribution in a local region, rendering this effort much more economical and easier than that of conventional methods.

VI. Conclusion & Future Work

We have presented a new technique in this paper which allows for visualizing the emotional state of masses during the event of a natural disaster occurrence through the means of analysis of tweets. Further, we have analyzed the geographical distribution and occurrence frequency of various disasters on different continents. The method is simple and effective and takes in account of almost all corner conditions and susceptibility of errors. We have demonstrated its feasibility on datasets of significant size consisting of live twitter feeds. Based on the results, we can conclude that the data from social media can be a viable complement to existing survey methodologies, and it can provide a deeper insight into real-time public perception of an event. The main contribution of this work includes proposing the usage of Social Media as an effective and uncensored source for disaster analysis on various fronts.

As a part of future work, we intend to involve the use of other popular social networking sites such as Facebook and Google+ in order to increase our database. These sites also have dedicated pages for disaster management run by various government and private agencies which can be parsed to achieve a structured data. We also intend to add a prediction model which may predict the disaster trends in various regions.

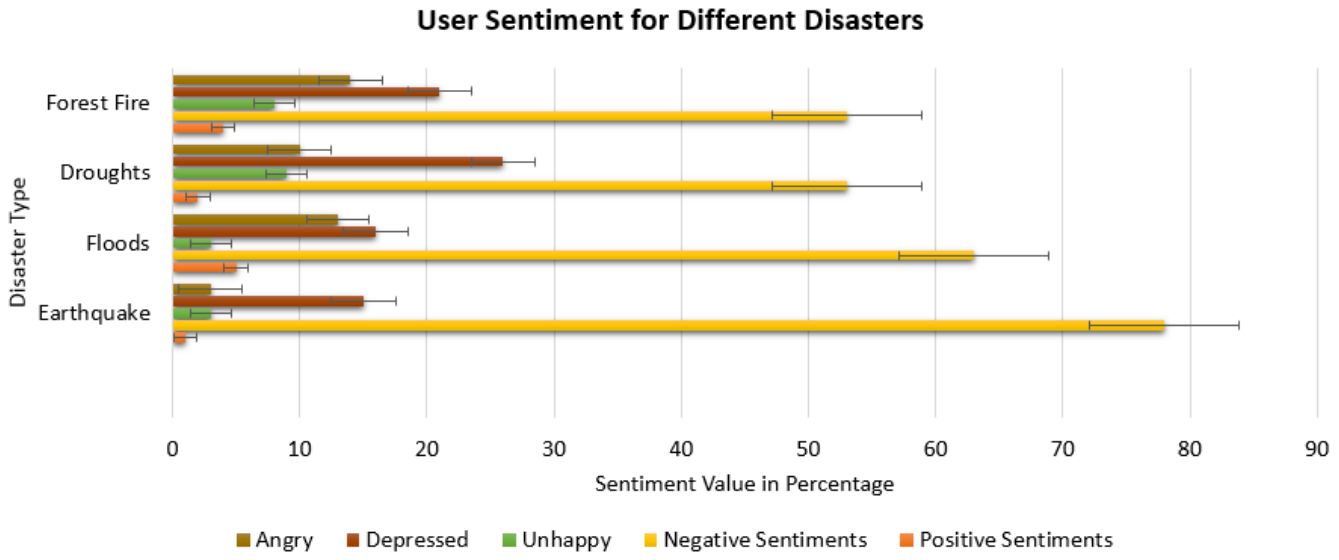

Fig. 3. Sentiment Analysis for Different Disasters

ACKNOWLEDGMENT

The authors would like to thank Ms. Ashwini Patil, Ms. Gauri Bewoor and Ms. Shruti Vasu for being an integral part of this project and working upon valuable data gathering & verification, design suggestions and system testing. We also thank our peers and anonymous reviewers for many useful comments and suggestions. Our sponsor is Department of Information Science and Engineering, BVBCET, Hubli.

REFERENCES

[1] S. B. Liu, L. Palen, J. Sutton, A. L. Hughes, and S. Vieweg, "In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster," in *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*, 2008.

[2] L. Palen and S. Vieweg, "The emergence of online widescale interaction in unexpected events: assistance, alliance & retreat," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 2008, pp. 117–126.

[3] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: what hazards threat reveals about the social life of microblogged information," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010, pp. 241–250.

[4] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1079–1088.

[5] L. Palen, S. Vieweg, S. B. Liu, and A. L. Hughes, "Crisis in a networked world features of computer-mediated communication in the april 16, 2007, virginia tech event," *Social Science Computer Review*, vol. 27, no. 4, pp. 467–480, 2009.

[6] J. Sutton, L. Palen, and I. Shklovski, "Backchannels on the front lines: Emergent uses of social media in the 2007 southern california wildfires," in *Proceedings of the 5th International ISCRAM Conference*. Washington, DC, 2008, pp. 624–632.

[7] S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu, "Tweettracker: An analysis tool for humanitarian and disaster relief." in *ICWSM*, 2011.

[8] R. E. Cohn and W. A. Wallace, *The Role of Emotion in Organizational Response to a Disaster: An Ethnographic Analysis of Videotapes of the Exxon Valdez Accident*. Natural Hazards Research and Applications Information Center, Institute of Behavioral Science, University of Colorado, 1992.

[9] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, and J. Rodrigue, "A demographic analysis of online sentiment during hurricane irene," in *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 2012, pp. 27–36.

[10] J. B. Lee, M. Ybañez, M. M. De Leon, and M. R. E. Estuar, "Understanding the behavior of filipino twitter users during disaster," *Journal on Computing (JoC)*, vol. 3, no. 2, 2014.