Social Media Data Analytics Applied to Hurricane Sandy

Han Dong, Milton Halem, and Shujia Zhou
Computer Science and Electrical Engineering Department
University of Maryland Baltimore County
Email: han6@umbc.edu

***Abstract*—Social media websites are an integral part of many people's lives in delivering news and other emergency information. This is especially true during natural disasters. Furthermore, the role of social media websites is becoming more important due to the cost of recent natural disasters. These online platforms are usually the first to deliver emergency news to a wide variety of people due to the significantly large number of users registered. During disasters, extracting useful information from this pool of social media data can be useful in understanding the sentiment of the public; this information can then be used to improve decision making. In this paper, we developed a prototype that automates the process of collecting and analyzing social media data from Twitter. Furthermore, we explore a variety of visualizations that can be generated by the tool in order to understand the public sentiment. We demonstrate an example of utilizing this tool on the Hurricane Sandy disaster between October 26, 2012 to October 30, 2012. Finally, we perform a statistical analysis to explore the causality correlation between an approaching hurricane and the sentiment of the public.**

I. Introduction

Social media websites can serve as a breaking news role for natural disasters. According to recent study surveyed by the American Red Cross Society, websites such as Facebook, YouTube, MySpace, Flicker and Twitter were the most popular source of news during natural disaster events [1] [2] [3]. They are usually among the first to deliver news to a large mass of people. Natural disasters can cause a significant increase in online activity among users in order to contact family and friends in disaster zones, and seek information such as food and shelter. Therefore, social media data collected during the events of a natural disaster can be important in understanding the publics reactions and feelings.

There have been some work in the area of analyzing social media data during emergencies and disasters. Cameron et. al [4] demonstrated the Emergency Situation Awareness Automated Web Text Mining system to identify relevant Twitter messages that inform the situation awareness of emergency incident as it unfolds. Vieweg et. al [5] examined Twitter data with respect to geo-location, location referencing and situational update information in two natural hazards-based data sets: spring 2009 Red River Floods and Oklahoma grass fires. They developed a working framework to inform the design and implementation of software systems that can employ information extraction strategies. Some researchers have also used visual based methods to perform event detection during emergencies [3] [6]. Liu et. al [6] investigated if and how disaster-related Flickr activity evolved for six notable disasters between December 2004 and October 2007.

Recently, Zin et al. [14] proposed a hybrid system of extracting key information during disasters by combining visual information from YouTube and semantic information from Twitter. They developed a knowledge-based framework for detecting and analyzing key events in disasters by using rich information from social network platforms. Sakaki et al. [15] built a system for identifying earthquake zones based on Twitter data; they investigated the real-time interaction of events such as earthquakes and proposed an algorithm to monitor tweets and to detect a target event.

This paper explores the use of Twitter data in analyzing sentiment and evacuation information during the Hurricane Sandy disaster. Different from [3]–[6], our motivation is to understand how the general public on Twitter reacts to an incoming natural disaster and the impact of disaster warnings from various government agencies. Furthermore, it is to obtain insights into what factors in evacuation message encouraged certain individuals to respond to the evacuation request and why others would not evacuate. Moreover, we also identify and formulate a corpus of keywords and a broader vocabulary utilized by the public during the disaster. While our work has similarities in investigating the role of social media in disasters to [14], [15], the novelty in this paper is its application to a major hurricane disaster event. Moreover, we solve the problem from a data analytics perspective: multi-faceted visualizations with in-depth analysis.

Lastly, we propose a system that can automate the process of extracting, analyzing and visualizing social media data. Specifically, it consists of two key functions: (1) extracting live Twitter data, pre-processing the data and storing it in a Lucene index store; and (2) displaying multi-faceted visualizations comprised of word clouds, spatial maps using geolocation data, along with dynamic online activity graphs. The rest of the paper is organized as follows. An overview of the proposed system is described in Section II. Section III presents experimental results on Hurricane Sandy data along with correlation analysis by using Granger causality tests. It is then followed by the conclusion in Section IV.

II. Data Processing and Extraction

The system illustrated in Figure 1 demonstrates the tool that was built to analyze Twitter data during Hurricane Sandy. The data was collected during the Hurricane Sandy disaster from October 26, 2012 to October 30, 2012. It affected much of the east coast and was ranked as the second-costliest hurricane disaster in the U.S. at $75 billion dollars [10]. Figure 2 shows its path. A general timeline of its events is listed below:

978-0-7695-5137-1/13 $26.00 © 2013 IEEE
DOI 10.1109/SocialCom.2013.152

- On October 26, Sandy moved through the Bahamas.
- On October 27, Sandy weakened to a tropical storm and then became a Category 1 hurricane.
- On October 29, Sandy moved near Atlantic City, New Jersey, as a post-tropical cyclone with hurricane-force winds.
- On late October 29, the storm hit New York City, flooding streets, tunnels and subway lines and cutting power in and around the city.
- In early October 30, residents in several counties of New Jersey were evacuated due to floods.

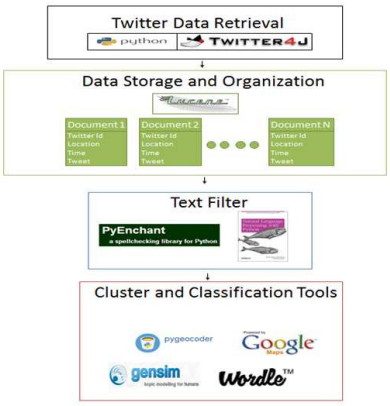

Fig. 1. Illustrates the system that was built to analyze Twitter data during Hurricane Sandy.

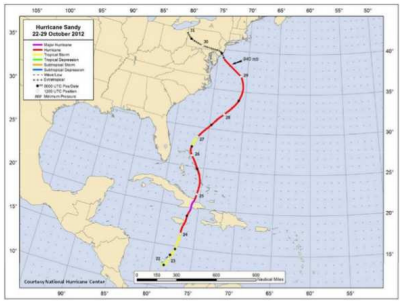

Fig. 2. Path of Hurricane Sandy approaching the east coast [10].

Twitter Search API allows a developer to specify keywords and geo-locational data to extract tweets within varying distances at a specific location. The data collected is listed in Table I and was produced by utilizing the following set of keywords:

- Evacuate, Prepare, Flood, Hurricane, Storm, Tornado, Sandy, Rain, Water, Food, Emergency, Battery, Safe, Leave

For each unique Twitter user from the dataset in Table I; the Username, Location, Time, and Comment data was first saved in an initial text file; it was then processed using the following steps:

- Special characters are removed from each tweet to avoid text processing issues.

TABLE I. TWEETS COLLECTED FROM DIFFERENT REGIONS ON EAST COAST AFTER CLEANING AND REMOVING DUPLICATES.

Location	Size of Data (MB))
Florida	100
South/North Carolina	200
Georgia	80
Virginia	100
Maryland/Washington D.C.	60
New York City/New Jersey	40
New York	100
Massachusetts/Rhode Island	50

- Due to geo-locational overlaps from retrieving tweets using the Twitter Search API, there are occasional duplicates. The duplicate tweets are removed in order to prevent putting additional weight on any specific tweet.

Finally, the data files from the different regions are merged together into a single corpus. This processed corpus is stored into a Lucene database, which currently contains 378,773 tweets.

III. RESULTS

A. Topic Models

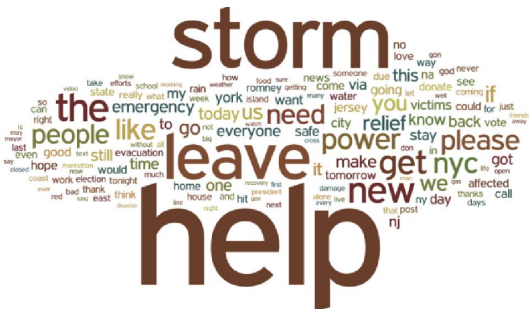

Fig. 3. Word cloud generated by bag-of-words model of entire Hurricane Sandy corpus.

Figure 3 illustrates a word cloud generated using the bag-of-words model involving the top 2000 words in the whole corpus for Hurricane Sandy. The goal of this word cloud is to assist in the intent classification of Hurricane Sandy data. For our use case, intent classification is described as a way of determining if a person will evacuate or stay given his/her tweet. Currently, there are no pre-trained corpus available for a domain that deals with disasters and evacuation warnings. Furthermore, given the sparse amount of information within a tweet, it may be insufficient to determine if he or she will actually commit to what they state on Twitter. We utilize an initial implementation below in order to tackle this problem:

- Corpus data is cleaned.
- Retrieve data that only contain pronouns in order to obtain tweets from a first person perspective.
- Extract tweets that contain keywords such as, *evacuate*, *leave*, and *go* as an indicator that a user intends to evacuate.

- Extract tweets that contain keywords such as, *stay*, *prepare*, *power*, *battery*, *water*, and *precaution* as an indicator that a user intends to stay.

The steps described above help to eliminate noisy data for the machine learning algorithms and will produce better textual visualization information as shown in the figures below. We

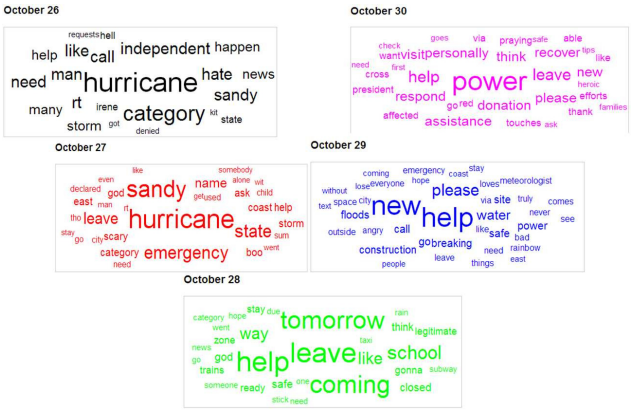

Fig. 4. Word cloud visualization of topic models from October 26, 2012 to October 30, 2012. Each color-coded text represents a single day.

used the Latent Semantic Indexing (LSI) algorithm in *gensim* [11] to produce a topic model of the Hurricane Sandy data in Figure 4. *gensim* was chosen for this task due to its ability to scale the computation to multiple nodes so as to expedite text processing. The algorithm runs for a number of iterations and returns a range of topics discovered for each day. The size of the word corresponds to its importance to the topic model. From the visual topic model, it is possible to notice the change in main topics regarding Hurricane Sandy. For example, most of the topics were regarding the *hurricane* and *category* topics on Oct. 26, 2012. This changes between Oct. 27 and Oct. 29 as Hurricane Sandy moved up the east coast with topics such as *emergency*, *scary*, *leave*, *help*, indicating a change in the community reactions regarding the incoming hurricane. Towards the end of Sandy, which is around Oct. 29 or 30, new topics such as *construction*, *power*, *help*, *assistance*, and *donation* appear. This suggests that the communities focus on rebuilding and fixing the damages from the hurricane. Therefore, this topic modeling visualization appears to be helpful in obtaining public reactions on a daily basis.

B. Intensity Maps

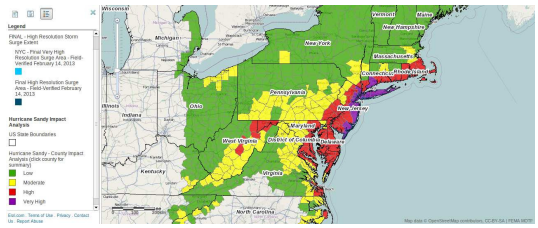

Fig. 5. Hurricane Sandy impact map by FEMA [13].

By using geo-location data from Twitter users, we are able to generate visualizations that characterize specific affected

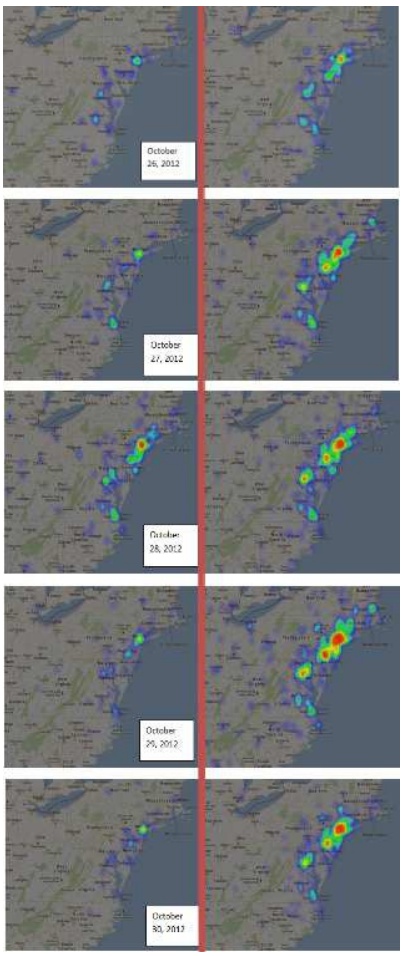

Fig. 6. Intensity map of Twitter users signaling evacuation intention. The images on the left column represent Twitter users with intention to evacuate while images on the right column indicates those who prefer to stay.

zones. Figure 5 is an impact analysis map from FEMA that was created after Hurricane Sandy had passed in order to identify regions that were severely impacted by the hurricane. We generated the intensity map in Figure 6 by combining geo-location and evacuation intention data; it lists the major Twitter activity for evacuation and non-evacuation in the scope of the five days that Hurricane Sandy impacted the east coast. Comparing Figure 5 to Figure 6, it is possible to see the intensity maps are able to indicate similar regions that have been most severely impacted. The benefit of this intensity map is that it can be produced dynamically by the system using social media data; this is different from the FEMA disaster impact maps that are usually produced after the hurricane has passed. Moreover, Figure 6 also illustrates the increase in social media activity on Twitter through the 5 days; more notably, we are able to see the increased activity zones appear around the coasts and also identify certain inlands communities that are also heavily impacted.

C. Bivariate Granger Causality Analysis

In addition, we are also interested in uncovering relationships between the distance of Hurricane Sandy and evacuation

TABLE II. GRANGER CAUSALITY ANALYSIS ON CORRELATING HURRICANE SANDY WITH EVACUATION OR SENTIMENT DATA

Lag (6-hour interval)	Evacuate	Not Evacuate	Positive	Negative
1	0.000041	0.497411	0.006910	0.005348
2	0.000506	0.274279	0.020155	0.010839
3	0.003008	0.152484	0.009751	0.005737
4	0.005761	0.220379	0.046324	0.025636
5	0.035183	0.158647	0.732366	0.406944
6	0.08065	0.533765	0.936770	0.762819

intention or general mood of Twitter users. We explore a technique utilized by recent social media research on stock markets [7] [8]. The authors in [7] [8] utilized the Granger causality analysis (GCA) to correlate Twitter sentiment mood with daily stock market closing prices. While GCA is often utilized in the field of economics, we intend to apply it in social media disaster analytics. For our use case, we utilize two sets of data collected from New York City during Hurricane Sandy. The first set consists of the number of Twitter users indicating their evacuation intention as well as the number of positive and negative comments classified with the Linguistic Inquiry and Word Count [12] tool. The second dataset consists of the distance of Hurricane Sandy to New York City. This analysis will test if the distance of Hurricane Sandy approaching the east coast contains correlations with the evacuation intentions or positive/negative sentiments of users on Twitter.

$$D_t = \sum_{i=1}^{p} \beta_i D_{t-i} + u_t \tag{1}$$

$$D_t = \sum_{i=1}^{p} \beta_i D_{t-i} + \sum_{i=1}^{p} \gamma_i X_{t-i} + u_t \tag{2}$$

We use a simple bivariate vector model with lag length p by the least squares equations indicated above. D_t represents the Twitter evacuation intention data and X_{t-i} represents the Hurricane Sandy distance data at 6-hour intervals from October 26 to October 30, 2012. In GCA, a p-value less than 0.05 indicate a significant causality relation between the two datasets. We performed the Granger causality analysis according to Equation 1 and Equation 2 using a web based tool [9] and utilized time lags of 6-hour intervals.

In Table II, we observe that the p-values for the evacuate data points were significant in that we can reject the null hypothesis that the Hurricane Sandy distance data *does not* have a granger-causality correlation on the evacuation intention data. For someone that is signaling their intention to stay, this could be explained by the fact that an approaching hurricane may not have a significant impact on their decision making. Furthermore, it might be beneficial to further explore the positive and negative sentiment Granger tests by using machine learning algorithms.

IV. CONCLUSION

In this paper, we present a system capable of processing, analyzing, and extracting useful information from Twitter during the Hurricane Sandy disaster in 2012. This system combines text processing and machine learning algorithms to uncover the sentiment of the public with real-world data collected between October 26, 2012 and October 30, 2012. Furthermore, our results show the tool can enable better situation awareness during a disaster event and provide visualizations in better understanding the general sentiment of the public. Lastly, we uncover a Granger causality between the approaching hurricane and the evacuation intentions of Twitter users.

ACKNOWLEDGMENT

This work was supported in part by a grant from NOAA NCEP through NSF CHMPR. We would like to thank Ben Kyger for helpful discussions.

REFERENCES

[1] D. Velev and P. Zlateva, "Use of social media in natural disaster management", Intl. Proc. of Economic Development and Research, Vol. 39, pp. 4145, Jun. 2012.

[2] L. Palen, "Online social media in crisis events," EDUCAUSE Quarterly (EQ), vol. 31, no. 3, pp. 7678, 2008.

[3] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D.S. Ebert, and T. Ertl, "Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition", IEEE Conf. on Visual Analytics Science and Technology (VAST2012), pp. 143152, Oct. 14-19, 2012.

[4] Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. 2012. Emergency situation awareness from twitter for crisis management. In Proceedings of the 21st international conference companion on World Wide Web (WWW '12 Companion). ACM, New York, NY, USA, 695-698. DOI=10.1145/2187980.2188183 http://doi.acm.org/10.1145/2187980.2188183

[5] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. 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 (CHI '10). ACM, New York, NY, USA, 1079-1088. DOI=10.1145/1753326.1753486 http://doi.acm.org/10.1145/1753326.1753486

[6] Liu, Sophia B., Leysia Palen, Jeannette Sutton, Amanda L. Hughes, and Sarah Vieweg. "In search of the bigger picture: The emergent role of online photo sharing in times of disaster." In Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM). 2008.

[7] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market."Journal of Computational Science2.1 (2011): 1-8.

[8] Rao, Tushar, and Saket Srivastava. "Twitter Sentiment Analysis: How To Hedge Your Bets In The Stock Markets."arXiv preprint arXiv:1212.1107(2012).

[9] Wessa P., (2013), Bivariate Granger Causality (v1.0.3) in Free Statistics Software (v1.1.23-r7), Office for Research Development and Education

[10] Hurricane Post-Tropical Cyclone Sandy, October 2229, 2012 (Service Assessment). United States National Oceanic and Atmospheric Administration's National Weather Service. May 2013. p. 10. Archived from the original on June 2, 2013. Retrieved June 2, 2013.

[11] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora." Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50, May 22, 2010

[12] Pennebaker, J. W., Francis, M. E., Booth, R. J."Linguistic inquiry and word count: LIWC 2001". Mahway: Lawrence Erlbaum Associates. 2001.

[13] FEMA Modeling Task Force (MOTF). https://fema.maps.arcgis.com/home/user.html?user=FEMA.MOTF

[14] Zin, Thi Thi, et al. "Knowledge based Social Network Applications to Disaster Event Analysis." Proceedings of the International MultiConference of Engineers and Computer Scientists. Vol. 1. 2013.

[15] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." Proceedings of the 19th international conference on World wide web. ACM, 2010.