Twitter’s Effectiveness on Blackout Detection during Hurricane Sandy

KJ Lee, Ju-young Shin & Reza Zadeh[1]

December 12, 2013

1. Introduction

Hurricane Sandy developed from the Caribbean stroke near Atlantic City, N.J., with winds of 80 mph about 8 p.m. EDT Oct. 29, 2012. When the hurricane ripped through the Northeast, thousands of buildings and facilities were catastrophically damaged. And beyond the structural damage, Sandy also left 2.5 million people without electricity mostly in New York and New Jersey[1]. It was simply the most recent time that the most populated corridor in the United States experienced massive power outages. Many victims of the power outages had to wait more than two weeks to get it back. The restoration of power outage took long time not only because the extent of damage from Hurricane Sandy far exceeded the destruction caused by previous weather-related events, but also because there was significant time lag for utilities to gather power outage information and to dispatch their limited number of field crews.

The duration of power outage restoration depends on several factors such as the main cause of power outage and the number of areas affected. Because each outage is a result of different circumstances, some may take longer to identify and restore than others. For example, if a pole near house or business goes down, it is easy to identify what caused the outage, and utilities can immediately get to work on the actual repairs. In that case, it typically takes about six hours to put up a new pole. However, during storm-related outages, restoration information may not be immediately available. Sometimes there is much damage that the utilities need to assess through a combination of remote monitoring and customer calls. Even with remote sensing and controlling system, utilities heavily rely on customer calls in the event of large blackout, resulting in time lag in their restoration efforts.

During the storm and its aftermath, social media has been critical in spreading important information and mobilizing relief efforts. Some argue that Twitter was more accessible and far-reaching than any TV network, with over 20 million Tweets posted during the height of the storm [2]. In this study, we examined and analyzed the abundant Twitter data to help minimize the damage by future catastrophic blackout events.

2. Related Work

The analysis of natural disaster using Twitter data has been done by Burks et. Al in the study titled “Early warning of ground shaking intensity using Tweets” for earthquakes and by Sakaki et al. in the study titled “Earthquake Shakes Twitter Users Real-time Event Detection by Social Sensors” [3], [4]. These paper demonstrates results for giving advanced warning to critical facilities of the magnitude of ground motion intensity by k-nearest neighbors and a generalized linear model on tweet behavior.

3. Objective

Our goal in this paper is to improve the outage detection process through the utilization of social media data. This goal consists of two sub-objectives, both of which involve machine learning using Twitter data: 1) to develop an algorithm for early detection of the blackout

[1] Reza Zadeh provided the raw Twitter data.

locations and 2) to pick the optimal locations for the blackout response crews to set up their camps.

4. Methodology

The second-costliest hurricane in United States history caused twitters to send more than 20 million tweets about the storm [5]. After filtering to collect the outage-related data, we developed an algorithm to identify tweets which correctly indicate blackout in the tweet origin location. This would significantly reduce time required to collect blackout location information for power utilities. Then, we created clusters of valid outage tweets so that utilities could use the cluster info to dispatch their field crews effectively and efficiently. The process map below (Figure 1) shows complete methodology steps.

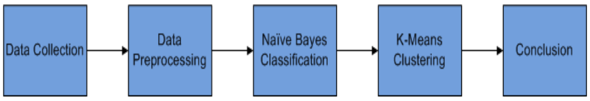


Figure 1. Methodology Process Map

4.1. Data Preprocessing

4.1.1. Filtering Twitter Data : First, we obtained the raw tweeter data containing all the tweets created in the United States during the storm from 10/30/2012 to 11/5/2012. We processed the data containing "hurricane sandy" to obtain tweets related to blackout of New Jersey. Each tweet line contains time lag between a tweet and the landfall of hurricane Sandy (EDT 01:00:00, Oct 30, 2012) and the latitude and longitude of a tweet, and the content of a tweet. To restrict target area to New Jersey, tweets written at latitude 38˚6' N to 41˚21' N, longitude 73˚ 54' W to 75˚ 34' W were chosen. Then, we matched keywords indicating blackout to each tweet line: "blackout", "power outage", "no electricity", and "no light". Also, to minimize outliers, we excluded tweets written before the landfall of hurricane. Finally, from the data filtering, overall 52000 lines of target tweets related to blackout of New Jersey by hurricane Sandy were selected. Some of the filtered Twitter data are shown in Table 1 below.

Table 1. Sample Twitter Data

ID	Time stamp	Latitude	Longitude	Text
1	2012-10-30 21:49:21	40.650	-73.950	I have no work for the rest of the week altho Ill have no electricity either #thankssandy
2	2012-11-01 04:39:03	40.768	-74.168	Shots in the dark #sandy #blackout #tequila http://t.co/5kVRs85u
3	2012-10-30 17:35:30	40.742	-74.002	Still no electricity because of hurricane Sandy Been over 24 hours now. Kitty is sad and bored!

4.1.2. Obtaining true blackout data mapping it to the filtered twitter data: First, we obtained a colored map representing blackout regions during the storm (Figure 2). Using ArcGIS software, we plotted the filtered Twitter data on NASA SPoRT map. Then, we identified whether each tweet was created in blackout region. Figure 3 shows the plotted Twitter data with outage labeling.

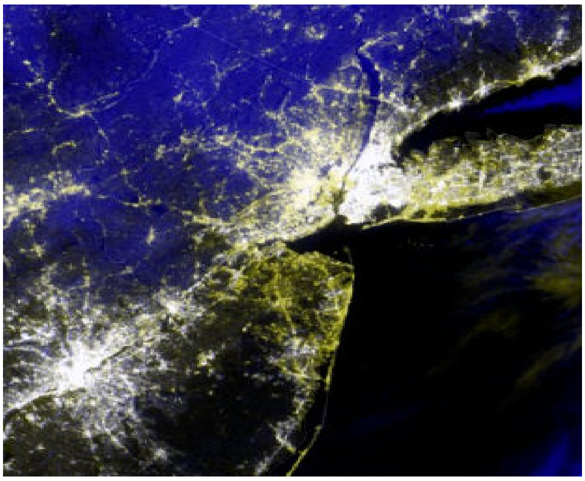
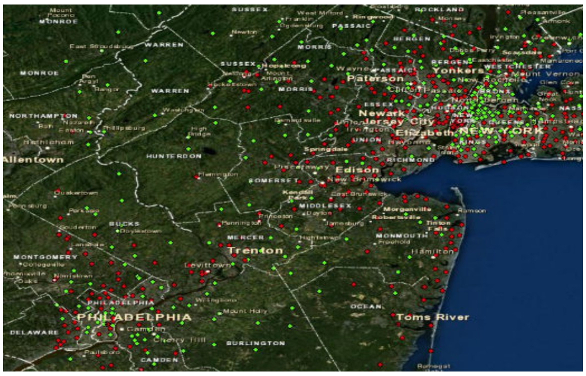


Figure 2. NASA SPoRT Map Hurricane Sandy Blackout



Red Dot: *Tweet on blackout region*
Greet Dot: *Tweet on none-blackout region (False alarm)c*

Figure 3. Plotted Twitter data with blackout labeling

4.2. Machine Learning Analysis

4.2.1. Naive Bayes classifier using the multinomial event model and the Laplace smoothing

To enhance the reliability of filtered blackout data, we built a Naïve Bayes classifier which canvalidate the true power outage tweets. By applying this machine learning, outliers such as tweet lines mentioning news or not related to blackout occurred by Sandy were excluded. The accuracy of the classification by the Naïve Bayes classifier was estimated by comparing with the labeling by a blackout satellite map.

Word x_i denotes the identity of the i-th word in a tweeter line and y indicates
whether the tweeter line informs real blackout or not.

Given a training set $\{(x^{(i)}, y^{(i)}; i = 1,...m\}$ *where* $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_{n_i}^{(i)})$
$(n_i$ *is the number of words in the i-training example),*

The likelihood of the overall probability of tweeters is given as follow:

$$L(\phi, \phi_{k|y=0}, \phi_{k|y=1}) = \prod_{i=1}^{m} \left(\prod_{j=1}^{n_i} p(x_j^{(i)} \mid y; \phi_{k|y=0}, \phi_{k|y=1} \right) p(y^{(i)}; \phi_y)$$

To maximize of the likelihood of the data, the estimates of the parameters
are as below:

$$\phi_{k|y=1} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \cap y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i + |V|}$$

$$\phi_{k|y=0} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \cap y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} n_i + |V|}$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

4.2.2. K- means clustering algorithm

For the efficient allocation of power resources to manage blackouts induced by hurricane, clustering the locations of blackouts based on spatial extent is necessary. Here the k-mean clustering algorithm was applied to clustering the distributed blackout spots and finding optimal locations of power resources as below:

```
Step 1. Initialize cluster centroids $\mu_1, \mu_2, ..., \mu_k \in \mathbb{R}^2$
        randomly within latitude 38˚6’ N to 41˚21’ N,
        longitude 73˚ 54’ W to 75˚ 34’ W
        where k is the number of power resources.
Step 2. Repeat until the centroids converge: {
            For every i, set
                $c^{(i)} := \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2$
             where x is a blackout spot.
            For every j, set
                $\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$
             where m is the total number of blackout spots.
             }
```

4. Analysis Result

Figure 4 depicts how accurately the Naïve Bayes classifier labels the true power outage tweets. As can be seen, five training sets in different sizes were used to train a model, and the error rate of classification decreased to about 30% when we employed the maximum size of a training set. Although the error rate is acceptable, it implies that there are difficulties in thoroughly distinguishing between the validate tweets and outliers because the context of tweets cannot be understood sorely by the frequency of words.

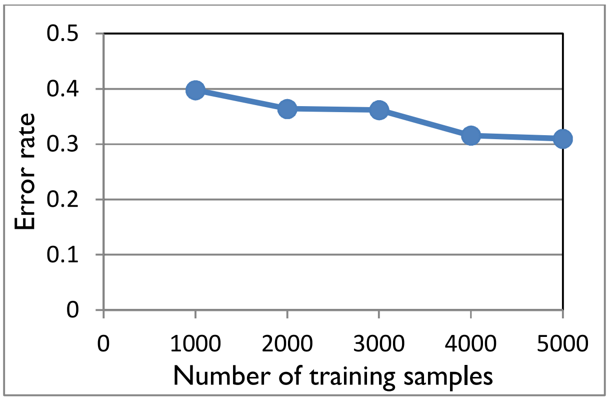


Figure 4. Error rate of classification of true blackout tweets by the Naïve Bayes classifier, used five different size of training sets.

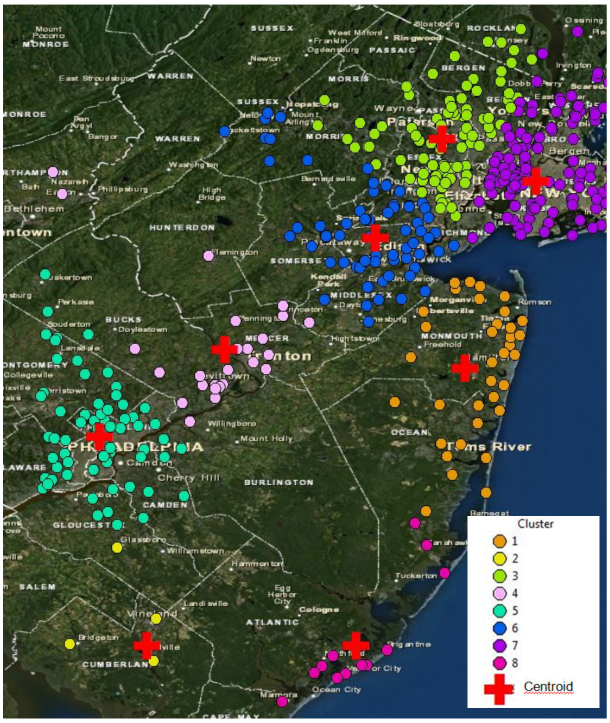


Figure 5. Allocation of power resources by k-mean clustering algorithm.

We conducted 1000 iterations of the k-mean clustering algorithm for the convergence to 8 clustering centroids as shown in Figure 5. The spatially optimized centroids described in Table 2 can be regarded as the efficient locations to supply power to blackout spots clustered by centroids. If there are a limited number of power resources and limited locations to place them, the k-mean clustering algorithm can result in very feasible clustering by adjusting the number of centroids and selecting centroids in restricted coordinates.

Table 2. Location of 8 power resources converged from k-mean clustering.

No.	Latitude	Longitude
1	40.200	-74.148
2	39.420	-75.043
3	40.845	-74.211
4	40.252	-74.822
5	40.009	-75.175
6	40.568	-74.399
7	40.726	-73.947
8	39.420	-74.455

5. Conclusion and Future Works

The Naïve Bayes classifier for identification of blackout using Twitter data seems to yield relatively high error rate (30%~40%) regardless of the size of training set. We believe that this high error rates are due to various reasons, such as small sample size, lack of feature word patterns, etc. This could be improved if there were more blackout-relevant hashtags in Twitter, in order to identify true blackout tweets. Also, there could have been errors in labeling true blackout tweets using NASA SPoRT map. Although the map is relatively accurate, the error rate could have improved if we used the actual blackout survey data from FEMA, which was unavailable to public unfortunately.

During blackout, 8 optimized location of power stations can be designated by K-mean clustering expeditiously based on 5204 twitter data. Though we selected 8 locations in our studies, the number of centroids can vary depending on the number of available emergency response teams from utility companies.

The methodology we developed in this study can also be applied to other types of disasters in densely populated areas where social media is heavily used. We hope that our study could become a foundation for further related studies in prediction of various types of mega disasters for more rapid response to life-threatening events.

6. Reference:

[1] Huffington Post, “Sandy: An Eye-Opener for the Power Grid” http://www.huffingtonpost.com/daniel-mcgahn/power-grid_b_2192554.html

[2] iRevolution, “Using Twitter to Map Blackouts During Hurricane Sandy” (July 3, 3013) http://irevolution.net/2013/07/03/using-twitter-to-map-blackouts-during-hurricane-sandy/

[3] Burks, L., Jordan, C., Miller, M., Zadeh, R., (2013). “Early warning of ground shaking intensity using Tweets”, *Association for the Advancement of Artificial Intelligence* (in review), 6.

[4] Sakaki, T., Okazaki, M., Matsuo, Y. (2010). “Earthquake Shakes Twitter Users Real-time Event Detection by Social Sensors”, *WWW2010*, April 26-30.

[5] Hurricane Sandy and Twitter http://www.journalism.org/2012/11/06/hurricane-sandy-and-twitter/