Detection of Damage and Failure Events of Critical Public Infrastructure using Social Sensor Big Data

Iris Tien[1], Aibek Musaev[2], David Benas[2], Ameya Ghadi[2], Seymour Goodman[3] and Calton Pu[2]
[1]*School of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Drive, 30332-0355, Atlanta, GA, U.S.A.*
[2]*School of Computer Science, Georgia Institute of Technology, Atlanta, GA, U.S.A.*
[3]*Sam Nunn School of International Affairs, Georgia Institute of Technology, Atlanta, GA, U.S.A.*

Keywords: Social Sensors, Big Data, Data Processing, Critical Infrastructure, Event Detection.

Abstract: Public infrastructure systems provide many of the services that are critical to the health, functioning, and security of society. Many of these infrastructures, however, lack continuous physical sensor monitoring to be able to detect failure events or damage that has occurred to these systems. We propose the use of social sensor big data to detect these events. We focus on two main infrastructure systems, transportation and energy, and use data from Twitter streams to detect damage to bridges, highways, gas lines, and power infrastructure. Through a three-step filtering approach and assignment to geographical cells, we are able to filter out noise in this data to produce relevant geolocated tweets identifying failure events. Applying the strategy to real-world data, we demonstrate the ability of our approach to utilize social sensor big data to detect damage and failure events in these critical public infrastructures.

1 INTRODUCTION

Public infrastructure systems provide many of the services that are critical to the continued health, functioning, and security of society. This includes energy systems that power nearly all devices, controls, and equipment, as well as transportation systems that enable the movement of people and goods across both short and long distances. Failure of or damage that has occurred to these infrastructures, whether from deterioration and aging, or from severe loads due to hazards such as natural disasters, poses significant risks to populations around the world. Detecting these damage or failure events is critical both to minimize the negative impacts of these events, e.g., by rerouting vehicles away from failed bridges, and to accelerate our ability to recover from these events, e.g., by locating the extent of power outages for deployment of repair crews.

Many of these infrastructures, however, lack continuous physical sensor monitoring to be able to detect these damage or failure events. Bridges, for example, are generally subject to only yearly inspections, and very few are instrumented with physical sensors that would be able to detect damage that may occur at any time. In addition, infrastructures that contain monitoring capabilities, such as energy systems, may have extensive networks of physical sensors at a centralized level, but less so at the distribution level. Thus, while power plants are closely monitored, maps of outages rely on individual reports.

In this paper, we propose the use of social sensors to detect damage and failure events of critical public infrastructure. Recently, there has been an exploration of the use of data from social sensors to detect events for which physical sensors are lacking. This includes the use of Twitter data streams to detect natural disasters (Sakaki et al., 2010) or the use of texts to manage emergency response (Caragea et al., 2011). In this paper, we use the LITMUS framework – a framework designed to detect landslides using a multi-service composition approach (Musaev et al., 2014a, 2014b) – to detect public infrastructure failure events. We focus on two main systems: transportation (bridges and highways) and energy (gas lines and power).

The rest of the paper is organized as follows. Section 2 provides an overview of the approach used

Tien, I., Musaev, A., Benas, D., Ghadi, A., Goodman, S. and Pu, C.
Detection of Damage and Failure Events of Critical Public Infrastructure using Social Sensor Big Data.
DOI: 10.5220/0005932104350440
In *Proceedings of the International Conference on Internet of Things and Big Data (IoTBD 2016)*, pages 435-440
ISBN: 978-989-758-183-0

to detect infrastructure failure events using social sensor data. Section 3 provides the results of the approach as applied to four infrastructures: bridges, highways, gas lines, and power. In Section 4, we provide an evaluation of the proposed approach, including filtering results for the social sensor data and visualizations of the detected events. We summarize related work in Section 5 and conclude the paper in Section 6.

2 APPROACH

An overview of the approach is shown in Figure 1. The sensor data source is Twitter. For the results presented in this paper, these are tweets pulled over the period of one month. We use October 2015 as our evaluation period. It is noted that data from any other time period can be used within this framework.

To detect infrastructure damage or failure events, all Twitter data is run through a series of filters to obtain a subset of relevant data. This filtering is done in three phases. First, we filter by search terms, which we have developed for various events of interest, e.g., “bridge collapse” to detect damage to bridge infrastructure. Second, as social sensor data is often noisy, with items containing the search terms but unrelated to the event of interest, data is filtered using stop words. Using a simple exclusion rule based on the presence of stop words, this filters out the irrelevant data. An example for detecting bridge collapses is the stop word “friendship” that refers to the collapse of a bridge or connection between two people.

Third, data is filtered based on geolocation. Although most social networks enable users to geotag their locations, e.g., when they send a tweet, studies have shown that less than 0.42% of tweets use this functionality (Cheng et al., 2010). In addition, users may purposely input incorrect location information in their Twitter profiles (Hecht et al., 2011). As geolocating tweets is an important component in being able to identify specific infrastructure damage events, including their location, the data must be additionally filtered. In this study, the Stanford coreNLP toolkit (Manning et al., 2014) is used along with geocoding (Google, 2016) to geolocate the tweet. This assigns each filtered tweet to a latitude and longitude and corresponding 2.5-minute by 2.5-minute cell as proposed in Musaev et al., 2014, based on a grid mapped to the surface of the Earth.

Once all relevant tweets are mapped to their respective cells, all tweets in a single cell are assessed to identify the infrastructure damage and failure events. In this paper, we focus on the results for tweets relating to damage detection in four infrastructures: bridge, highway, gas line, and power infrastructure.

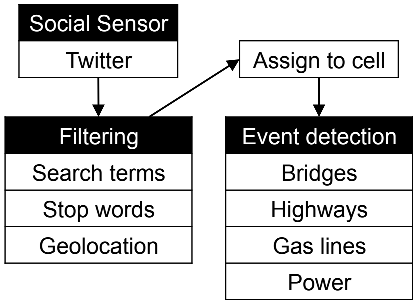

Figure 1: Overview of data, filtering, and event detection approach.

3 RESULTS

Each of the four infrastructures studied present different challenges, with particular characteristics for filtering that we discuss in this section. In addition, we present the specific search terms and stop words that we have found for use in identifying events of interest for each infrastructure. All Twitter data is filtered using these words to obtain the subset of relevant data, which is then geolocated to identify the damage or failure events.

3.1 Bridges

Bridge-related damage events tend to be major events. This includes closures of bridges that are part of major transportation arteries, or high-visibility, large-impact bridge collapses. This results in tweets that are pointing to the same incident, but are mapped to different geographical cells. Users, for example, tweet about events that are far away. A differentiation, therefore, must be made between ground users and other users. While most relevant for bridges, this difference in location proves to be applicable across infrastructures. The search terms and stop words used to detect bridge-specific damage events are listed in Table 1.

Table 1: Search terms and stop words for bridge damage events.

Search Terms	Stop Words
bridge {collapse, damaged, closure, closed, flooded, accident}	friendship, reopened, re-opened, pending, fish, bid, awe, awesome, wheelchair

3.2 Highways

Analysis of highway-related events is dependent on the severity of the event considered. For example, it was found that many Twitter users use the platform to complain about delays and increased traffic times on the highway, rather than to indicate actual infrastructure damage. Considering only major traffic or accidents that occur on the highway decreases the amount of noise in the data. As many highway damage events are due to natural disasters, future filtering of the data in conjunction with information on natural disasters may also decrease noise and enable better detection of highway damage events. The search terms and stop words used to detect highway-related damage events are listed in Table 2.

Table 2: Search terms and stop words for highway damage events.

Search Terms	Stop Words
highway {damaged, closed, blocked, accident, mud, pothole, snow, gridlock}	boating, watch, explore, delays, symbolic

3.3 Gas Lines

The social sensor data filtered to detect gas line damage events was the noisiest dataset of the infrastructures studied. While the bridge dataset includes differences in location between the tweet and event of interest, the gas line dataset also includes differences in time between the tweet and event of interest. For example, users tweet about gas leaks that have occurred in the past rather than about the current state of the infrastructure. In addition, irrelevant tweets include those complaining about the smell of gas from cars at drive-throughs, or about suspected but unsubstantiated gas leaks. Real gas leaks or damage to gas lines can result in severe health and safety consequences, so it is important to be able to detect these events. The search terms and stop words used to detect damage events related to gas lines are listed in Table 3. Note that due to the noise in this dataset and the number of stop words needed to filter out irrelevant data, a representative sample of, but not all, stop words are listed.

Table 3: Search terms and stop words for gas line damage events.

Search Terms	Stop Words
gas {leak, line}	plumbers, suspected, in-home estimate, repairs underway, drive-through, drive-thru, short line, tanker, contained, fixed

3.4 Power

In the data filtering process for power infrastructure, we are able to detect both larger-scale power outages that occur across cities and countries, e.g., the major outage in Puerto Rico on October 23, 2015, as well as smaller-scale individual outages, e.g., an outage associated with a local elementary school. For the stop words filter, we found that tweets containing any permutation of two or more of the hashtags #power, #outage, #blackout, or #grid were irrelevant. This is due to the general meanings of these words and the prevalence of these hashtags in referring to things outside the scope of events of interest. Over time, as different events occur and memes develop that utilize words associated with these critical public infrastructures but are unrelated to actual infrastructure damage, the data filtering system must be able to filter out this noise. In addition, tweets relating to news stories of past power outages, rather than the current state of power infrastructure, have to be filtered out. Future filtering in conjunction with text mining of news links in articles will facilitate this filtering. The search terms and stop words used to detect failure events of power infrastructure are listed in Table 4.

Table 4: Search terms and stop words for power infrastructure damage events.

Search Terms	Stop Words
power outage	#power, #outage, #blackout, #grid, back on, claims, resolved, files, stories, hotel

4 EVALUATION OF APPROACH

In this section, we discuss the filtering efficiency of the proposed approach, and show how results can be visualized to facilitate detection, identification, and inference about critical public infrastructure damage and failure events.

Table 5 shows the number of social media items downloaded and filtered through each step of the data filtering process. The total number of tweets remaining after each step for the four infrastructures is listed. In addition, for filter steps two and three, the percentage of data remaining after that filter step compared to the previous step is given. The relative number of tweets across the four infrastructures is an indicator of the relative prevalence of tweets related to those systems among Twitter users.

The initial filter based on search terms includes items both relevant and irrelevant to the infrastructu-

Table 5: Filtering results: number and percentage of tweets remaining after each filter step for four infrastructures of interest: bridges, highways, gas lines, and power.

Infrastructure	Filter based on search terms	Filter based on stop words		Filter based on geolocation	
	Number of tweets	Number of tweets	% remaining	Number of tweets	% remaining
Bridges	8436	8364	99.1%	1673	20.0%
Highways	5826	5817	99.8%	2368	40.7%
Gas lines	8709	8417	96.6%	2249	26.7%
Power	6648	6474	97.4%	1127	17.4%

re damage events of interest. The stop words filter out irrelevant tweets. From the first search-term filter to the second stop-word filter steps, we see that there are surprisingly low levels of noise in the social sensor data. The percentage of data remaining after the stop-word filter, however, is not 100%. This noise must be filtered out using stop words. This is important to ensure the minimization of the number of incorrect detections of infrastructure damage events.

Detections of damage or failure of critical public infrastructure have significant societal and economic impacts. If, for example, crews are dispatched to repair certain infrastructure, emergency responders are distributed to particular locations, or infrastructures are closed for safety based on this information, it is important that there is a high confidence in the inference about the infrastructure damage states before acting. This has policy implications for the accuracy of inference based on social sensor data required to transition from the data and event detections to public or community actions.

From Table 5, we see that in going from the second stop-word filter step to the third geolocation filter step, the number of results filtered out due to incorrect or insufficient geolocation information is significant. This is due to the presence of irrelevant tweets, as well as to the lack of geolocation information to confirm relevance of a tweet to an event of interest. This demonstrates the need to augment the social sensor data with other data sources, including physical sensor data, news sources, and alternate social sensor information. Doing so will reduce the loss of information and increase the resolution of the relevant information in the third filtering step. This integration across data sources will also facilitate automation in detection of infrastructure damage or failure events.

4.1 Data Visualization

In addition to the detection of an event, given the spatial distribution of public infrastructure, it is important to be able to visualize the damage or failure events. Figures 2-4 show visualizations of events of interest, including the geolocated relevant tweets and detected events.

Figure 2 shows a cluster of relevant tweets and detected events in the Johannesburg, South Africa, area related to bridge damage. The number of relevant tweets in a concentrated geographical area, i.e., the number of tweets mapped to a cell, can be used as a measure of the intensity of an event. In Figure 2, we see the relevant tweets detecting the severe bridge collapse in Johannesburg on October 14, 2015. The distribution of tweets to different cells is due to differences in identifications of geolocations. In this case, geolocations for tweets relevant to this event include Johannesburg, Sandton, and Grayston Bridge. This is because the bridge collapse event occurred in Johannesburg's Sandton district near Grayston Drive. Therefore, tweets related to the same event can be mapped to different cells due to different geolocations. Despite the distribution across cells, the number of relevant tweets in nearby cells indicates a severe event. In this case, there were two deaths and 20 injured as a result of this failure event.

Figure 2: Relevant tweets and detected bridge damage events; example for Johannesburg, South Africa.

In Figure 3, we show an example of highway damage-related relevant tweets and detected events for California, USA. The figure shows the correspondence between filtered, geolocated relevant tweets and detected events. We are able to detect damage events in both densely populated urban areas, e.g., events in the San Francisco Bay

Area, as well as in more sparsely populated rural areas, e.g., events near Lone Mountain and Death Valley. In addition, these results include a highway damage event due to a flood and subsequent mudslide, showing the ability of the approach to detect damage events due to multiple hazards.

Figure 3: Relevant tweets and detected highway damage events; example for California, USA.

For gas line damage, there were no particular events of interest, so a map is not shown here. Maps can, in general, be generated for locations or events of interest. Power infrastructure damage events are shown in Figure 4, which illustrates the widespread nature of power failures. An example of relevant tweets and detected events in the United States and Caribbean are shown. In addition to the outage events detected across the United States, we see the major power outage detected in Puerto Rico from the October 23, 2015, event.

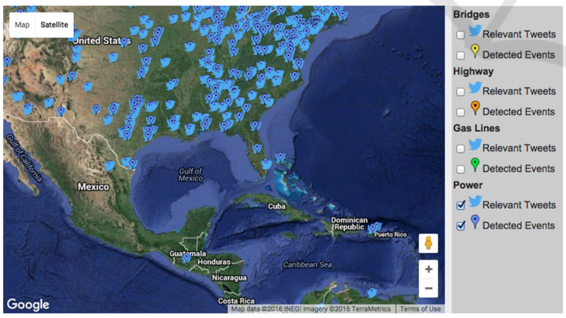

Figure 4: Relevant tweets and detected power infrastructure damage events; example for the United States and Caribbean.

In general, we are able to use the social sensor information to detect damage and failure events of public infrastructure globally. The results are not limited to any one country or region of the world, or to the type or size of a community. Of course, event detection relies on the presence of the social sensors, e.g., Twitter data streams, but as social media adoption increases around the world, the amount of relevant data available will only increase.

5 RELATED WORK

The approach for public infrastructure damage and failure event detection as described in this paper is based on the LITMUS framework for landslide detection built by Musaev et al., (2014a and 2014b). A process similar to the LITMUS filtering process was utilized to filter the noise out of infrastructure damage-related tweets. However, this work differs from the LITMUS work in that instead of detecting a single type of event, we are focusing on different infrastructures that can be damaged due to a variety of events. For example, instead of detecting a landslide, we are detecting damage to a highway that may have been caused by a landslide or other event.

There have been several studies using social sensor data to detect disaster events. This includes studies related both to man-made hazards, e.g., mass shootings (Vieweg et al., 2008; Palen et al., 2009); and to natural hazards, e.g., earthquakes (Guy et al., 2010; Sakaki et al., 2010; Caragea et al., 2011), fires (Sutton et al., 2008), floods (Vieweg et al., 2010), and tornadoes (Imran et al., 2013). Our work differs from the disaster detection literature in that rather than detection of widespread disaster events, we detect damage to specific infrastructures, which may or may not be related to or a result of a larger disaster. In addition, many studies on detecting disasters using social media data focus on the detection or description of single hazards, whereas the infrastructure damage events that we are looking at may be caused by multiple hazards.

6 CONCLUSION

Detection of damage and failure events to public infrastructure is critical to the ability of communities around the world to minimize the risks associated with both natural and man-made disasters and to recover more quickly and efficiently from the negative effects of these hazards. As many of our public infrastructure systems are not physically monitored to the degree necessary to provide relevant, detailed information about the states of these systems in real time, social sensor data is used to perform this assessment and detect damage events.

In this paper, we describe an approach to use social sensor big data to identify public infrastructure damage events. This includes a three-step filtering approach, whereby data is first filtered using search terms relevant to the event of interest. Next, noise in the data is filtered out using an exclusion rule based on the presence of stop words. Finally, data is filtered based on geolocation, resulting in each relevant filtered data item being assigned to a 2.5-minute by 2.5-minute cell in a grid mapped to the surface of the Earth.

Once all relevant data are mapped to their respective cells, all data in a single cell are assessed to identify the infrastructure damage and failure events. In this paper, we present results for detection of damage events for transportation and energy systems, and in particular for bridges, highways, gas lines, and power infrastructure. We evaluate the approach using real-world data collected from October 2015. We show the ability of our approach to use social sensor information, in this case Twitter data streams, to detect damage events. In addition, we show how results can be visualized to facilitate detection, identification, and inference about infrastructure damage.

As infrastructures are subjected to an increasing number of hazards, the ability to detect and localize damage events to these infrastructures is becoming an increasingly important task to improve the resilience of communities. In this paper, we demonstrate the ability of and value in using social sensor big data to detect damage and failure events in these critical public infrastructures.

ACKNOWLEDGEMENTS

This work was partially funded by the National Science Foundation through the CNS/CISE program (Award #1541074). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H., Mitra, P., Wu, D., Tapia, A.H., Giles, L., Jansen, B.J., Yen, J., 2011. Classifying text messages for the Haiti earthquake. In *ISCRAM '11*, Lisbon, Portugal.

Cheng, Z., Caverlee, J., Lee, K., 2010. You are where you tweet: A content-based approach to geo-locating Twitter users. In *CIKM'10*, Toronto, Canada.

Google, https://developers.google.com/maps/documentation/geocoding/intro, accessed on 2/5/2016.

Guy, M., Earle, P., Ostrum, C., Gruchalla, K., Horvath, S., 2010. Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In *IDA'10*, Tuscon, Arizona.

Hecht, B., Hong, L., Suh, B., Chi, E.H., 2011. Tweets from Justin Bieber's heart: The dynamics of the "location" field in user profiles. In *CHI '11*, Vancouver, Canada.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P., 2013. Extracting information nuggets from disaster-related messages in social media. In *ISCRAM '13*, Baden-Baden, Germany.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., and McClosky, D., 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60, Baltimore, Maryland.

Musaev, A., Wang., D., Pu, C., 2014a. LITMUS: Landslide detection by integration multiple sources. In *ISCRAM '14*, University Park, Pennsylvania.

Musaev, A., Wang., D., Cho, C.A., Pu., C., 2014b. Landslide detection service based on composition of physical and social information services. In *ICWS '14*, Anchorage, Alaska.

Palen, L., Vieweg, S., Liu, S., Hughes, A., 2009. Crisis in a networked world: Features of computer-mediated communication in the April 16, 2007 Virginia Tech event. *Social Science Computer Review Special Issue on E-Social Science*.

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10,* Raleigh, North Carolina.

Sutton, J., Palen, L., Sklovaski, I., 2008. Backchannels on the front lines: Emergent use of social media in the 2007 Southern California fires. In *ISCRAM '08,* Washington, DC.

Vieweg, S., Palen, L., Liu, S., Hughes, A., Sutton, J., 2008. Collective intelligence in disaster: Examination of the phenomenon in the aftermath of the 2007 Virginia Tech shooting. In *ISCRAM '08,* Washington, DC.

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *CHI '10*, Atlanta, Georgia.