New England University Transportation Center

NE University Transportation Center
77 Massachusetts Avenue, E40-279
Cambridge, MA 02139
Phone: 617-253-0753
Fax: 617-258-7570
web.mit.edu/utc

Principal Investigator:	Yossi Sheffi, Ph.D.	**Co-Principal Investigator:**	Jarrod Goentzel, Ph.D.
Title:	Elisha Gray II Professor of Engineering Systems	**Title:**	Director, Humanitarian Response Laboratory
University:	Massachusetts Institute of Technology	**University:**	Massachusetts Institute of Technology
Email:	sheffi@mit.edu	**Email:**	goentzel@mit.edu
Phone:	617-253-5316	**Phone:**	617-253-2053

Final Report

Project Title:

Big Data During Crisis: Lessons from Hurricane Irene

Project Number:	Project End Date:	Submission Date:
MITR24-10	June 1, 2014	March 2, 2015

Maximum 2 Pages

Problem

Transportation networks connect people with the goods and services that they require on a daily basis. In a disaster or emergency, they serve the same role, but often for more urgent needs. Transportation networks provide access to food, water, shelter, medical care, and other essential services for a disaster-affected population. Transportation plays a key role in enabling individuals to seek provisions in nearby communities that are not affected and/or enabling various types of response organizations to bring supplies into the affected community. If even small portions of these networks are damaged entire towns may be cut off from aid, as was the case in Hurricane Irene in the northeastern United States.

Traditional methods used by transportation agencies for gaining situational awareness after a disaster or emergency often rely on assessments made by trusted individuals, whether that be a transportation agency or other government official, partner agency, or other trusted source. This can be a time consuming process, where transportation agencies may not be aware that communities have been cut off for some time after the event has occurred. This project presents the results of an assessment of the ability of Twitter data to supplement these traditional information-gathering processes to create actionable information in an emergency.

Approach

The approach taken in this project was to conduct a case study on transportation outages in Hurricane Irene. In this case study, natural language processing (NLP) techniques were used to analyze social media data for transportation outage information. The results of that analysis are then compared with data collected by state transportation agencies using traditional methods to identify whether these outages were identified on social media. The intent of the case study was to characterize the potential of big data from sensor networks to compliment existing sensor networks to create actionable information in a disaster, and to further develop methodologies to analyze these sources of data.

Methodology

In order to create a comparison between traditional methods of collecting information on transportation outages and social media data, two types of data were collected. Transportation outage data for Hurricane Irene were collected from the four states of New York, New Jersey, Vermont and New Hampshire, and Twitter data were collected. The four states were selected based on the insured losses as reported by the Insurance Services Office (ISO) for the storm. Per capita insured losses were calculated using the US Census Bureau's estimated statewide populations in 2011, and the states selected represent a range of per capita insured losses. Transportation agencies were contacted in each of the four states being considered in this study and data on instances of transportation outages following Hurricane Irene were requested.

Relevant historical Twitter data was obtained through a third party organization with access to the Twitter fire hose through an Application Programming Interface (API). Through the API we applied a set of rules that bounded the Tweets on time and location and searched within this bounded set of Tweets for specific words and hashtags. Then, in partnership with and using tools developed by Idibon, Inc., an organization specializing in NLP techniques, the quantitative analysis was conducted on the Twitter data. The portion of the analysis using NLP involved two main activities: annotation of the data by individuals followed by the development of a model using these annotations. The goal of the model was to predict which Tweets in the remaining dataset (the Tweets not previously annotated by humans) refer to transportation issues. Analysis was conducted on the output from the model to describe the characteristics of those Tweets that, according to the model, related to Hurricane Irene.

Finally, key events from each state were identified a keyword search was conducted on the positively annotated tweets (greater than 50% confidence level from the model output) to determine whether or not these key events were discussed on Twitter. Following this keyword analysis, the Tweets were compared against the state transportation data on two key attributes: date and time, to determine whether they had been posted before transportation agencies identified the outage.

Findings

The key observations and findings of this project included:

- Individual states' collection and organization of transportation outage data differs in key ways that may negatively impact information sharing among states in a crisis.
- The volume of positively annotated Tweets per day identified by the model was too great for a human to monitor in a crisis. Thus, Tweets with a confidence level in the 99th percentile, in this case, would represent a reasonable load.
- Assessing the words used to communicate transportation failures in Tweets, the word 'http' appears in both New York and New Jersey keywords but does not appear in Vermont or New Hampshire, which may be indicative of the fact that news coverage of specific road closures in New York and New Jersey is more active than in Vermont and New Hampshire. Also, the word 'rt' is in the top 15 words in three of the four states. This word either refers to a retweet 'RT' or is the abbreviation of the word 'Route'. If a retweet, this and 'http' indicate that individuals share information over Twitter following an event by sharing webpages or other users' posts.
- In its development and refinement, this case study presents a replicable and generalizable methodology for conducting an analysis on unstructured Twitter data with the intent of identifying instances of communication regarding a certain type of information. Data such as this Twitter data is more likely to be applied for decision-making if it is in a format that can be compared with other streams of existing data. In the transportation context this existing data might be coming from DOT maintenance employees in the field or from emergency response professionals.

Conclusions and Recommendations

Key conclusions and recommendations that are derived from this study include:

- State transportation agencies should make an effort to establish a standard for collecting and organizing data in order to facilitate the sharing of these data following a disaster. This is especially important because transportation infrastructure crosses state lines, and an impact in one state may also impact the neighboring state.
- NLP is a powerful tool for taking a large dataset and identifying relevant documents. The one key challenge is that in order to build a model like the one used in this study, these machine learning techniques require training data. As such, the ideal model would be developed on training data real time following a crisis, but given the need for immediate information the recommendation would be to develop a model based on commonly observed crises (floods, hurricanes, wildfires, earthquakes) and use that on an initial basis until a more detailed model can be developed.

A more extensive final report can be obtained by contacting Dr. Jarrod Goentzel at goentzel@mit.edu or 617-253-2053.