Windstorm Hazard and Vulnerability Characterization using "Human-Sensor" Data

Franklin T. Lombardo, University of Illinois at Urbana-Champaign

Summary of DPRI Research Visit – October 10-14, 2016

I. Introduction

Windstorm losses are the greatest of any natural hazard in the U.S. and Japan (and a significant number of other countries) annually. A significant amount of these losses are caused by localized events (e.g., thunderstorms, tornadoes) [1]. In order to mitigate these losses we need to improve our understanding of both the wind hazard itself and the impacts it causes. However, due to the small-scale nature of these events current measuring networks do not capture most occurrences. These networks currently have poor resolution – typically tied to aviation or other transportation. Poor sensor resolution is especially true for developing countries, which may lack any sophisticated observation techniques and warning procedures. The resolution of sensors causes under-sampling of windstorm events, which by their very nature are of high intensity and relatively low proximity. This makes assessment of these events and their associated risk from traditional sensor data difficult. For some widespread and extreme events, damage-based assessments can be used to gauge intensity. These damage (i.e., impact) assessments routinely suffer from a lack of available manpower, inability to arrive at the scene rapidly and difficulty accessing a damaged site [2]. For real-time or near real-time projections, the poor resolution of the data and associated uncertainty can have impacts on decision-makers that weigh the cost benefit of a performing a certain action given a hazard and its potential impacts.

During natural disasters, when the impact is significant and proximity is low, it is imperative to rapidly estimate the overall consequences such as the hazard characteristics, physical damage and human injury in areas where traditional, dedicated sensors may not be present or limited. As population rapidly increases on a global scale [7], this increase is expected to lead to a large increase in the amount of data produced by "human-sensors" such as smartphones, computers and tablets. For example, over 3% of the global population has sent 170 billion Twitter messages in only seven years of its existence which produces rapidly disseminated information on a global scale [8]. Human-sensor data has the potential to radically improve fundamental understanding of hazard and disaster processes. However, very little research on these topics has been performed. There are some 'citizen-science' applications in existence that allow for input of certain environmental conditions or the intensity thereof. An example, is "Did You Feel It?" [9], a United States Geological Survey (USGS) application that asks users, once an earthquake occurs, the level of 'shaking' that was felt. This information on shaking is used, in part, to generate estimated intensity maps (i.e. ShakeMaps) for a given earthquake. Another application is MPING [10], developed by the National Oceanic and Atmospheric Administration (NOAA) in the United States. As weather hazards cannot be detected near the ground by radar, NOAA uses information from observers on meteorological conditions such as 'rain' at their location to verify (or not verify) the condition shown by radar at some distance above. Information from users on what meteorological condition is occurring is geo-tagged at their location. These programs have shown that data from the public can be used to understand both the hazard and its impacts. However, no applications have been specifically developed to map windstorm hazards, vulnerabilities and impacts. Doing detailed scientific research on windstorms and other hazards using citizen science and 'crowdsourcing' has been called a 'grand challenge' with significant potential [11].

Given this background information, it is time to formally include public participation in wind hazard and damage assessment. Ubiquitous cell phone and internet availability, including the use and rapid dissemination of social media, the power of crowds engaged in scientific endeavors [3], and the public's awareness of vulnerabilities [4] all point to a paradigm shift in how events and hazards in particular can be sensed in a rapid manner. We propose to better represent, extract and track events that could lead to disasters by using 'human-sensor' data. Information extracted from these sources can feed into the following areas: Numerical Weather Prediction (NWP), advances in hazard and disaster characterization and modeling, social-cognitive theories of human behavior and decision-making for hazard mitigation. The objectives of this research project, with a focus on windstorms are listed below.

This white paper will discuss the following topics. In Section 2 the justification and description of specific datasets and methods will be carried out with respect to a single human-sensor – Twitter. Section 3 will describe a preliminary analysis of Twitter data from a tornado event in the United States in a case study format as well as a 'proof-of-concept' showing that similar analysis can be applied to Japanese Twitter data and events. The final section (Section 4) will outline continuing and future work on the topic.

II. Data and Methods

The initial data collection for this project which spanned August-October 2016 was done in an ad-hoc manner. Twitter data was extracted from the Twitter streaming API (https://dev.twitter.com/streaming/overview), which allows access to all global Tweets, through a MATLAB code 'twitty' which is available through the link (http://blogs.mathworks.com/loren/2014/06/04/analyzing-twitter-with-matlab/). Once the 'twitty' code is initialized, the program is able to access any Tweets beginning with the time of initialization. For the purposes of this work, a keyword (or keywords) were chosen to filter the large stream of Twitter data to only extract Tweets containing those keywords. The initialization of the 'twitty' script was done manually by the author when, based on following local or global weather conditions, that an extreme windstorm may occur or was occurring in some location (in the U.S. only for this portion of the work). When initialized, it is possible to set a sample size (N) of Tweets to collect before ending the extraction process. For this preliminary work, N = 1000 or 10000 depending on the event and discretion of the author. After completing, the script was initialized additional times to increase the size of the sample. After the running 'twitty' has completed, the data output is contained in a 'structure' format in MATLAB. Within this structure, N 'statues' are stored. These statues contain relevant information on the time of tweet, text of the tweet, any user information such as handle, geolocation (if available), the presence of any media information such as photos and videos and whether the tweet was a 're-tweet' of an original message. This relevant information can also be extracted for the purposes of the research. As of October 14, 2016, sampled information on five (5) 'events' was collected through Twitter. These events are shown in Table 1.

Table 1. Events on which Twitter samples were collected.

Event Description	Date(s) of Data Collection	Number of Samples (N)
North Dakota/Manitoba Tornado	August 3, 2016	1000
Indiana Tornado Outbreak #1	August 15, 2016	1000
Indiana Tornado Outbreak #2	August 24, 2016	3000
Typhoon Meranti (Pacific)	September 14, 2016	10000
Hurricane Matthew (Atlantic)	October 6, 2016	2000

III. Case Study and Proof-of-Concept: Kokomo, IN, US Tornado and Japan Access

a. Kokomo, IN, U.S. Tornado

1. Overview

For the purposes of this white paper to highlight some of what the Twitter data can provide, the focus will be on a single event that occurred in the 'Indiana Tornado Outbreak #2' on August 24, 2016. Specifically, the tornado that affected the community of Kokomo, Indiana. This tornado was rated EF-3 on the Enhanced Fujita (EF) Scale (http://www.weather.gov/ind/august242016severe) with estimated peak winds of 152 mph (68 m/s). This tornado, which was on the ground for 14 minutes (3:20P-3:34P EDT) injured 20 people and produced damage to numerous homes in Kokomo as well as completely collapsing a Starbucks. An example of the damage is shown in Figure 1. Additional damage photos, obtained from Twitter, focusing on the Starbucks will be analyzed later this section.

Figure 1. Picture of residential damage in the Kokomo, IN tornado. (Source: NWS Indianapolis)

2. Twitter Analysis

Text Analysis

Specific to the Kokomo tornado, a total of 2000 tweets were archived. The inclusive keyword filter for all the tweets was 'tornado' neither of which were case-sensitive. The first 1000 tweets were collected from 3:20P-3:27P EDT, the first seven (7) minutes of the tornado lifespan. Based on the damage map presented in [NWS link], there was some significant damage around this time, however the tornado had not entered more populated areas of the city. In these 1000 tweets that had already been pre-filtered, there was only one mention of the word 'Starbucks' and it was unrelated tweet to the event of interest. As the first set of 1000 tweets did not mention the focal structure of interest (i.e. Starbucks) as it had not likely been damaged yet the focus will be on the second set of Tweets collected from 4:06P-4:10 EDT.

It should be noted that the earliest mention of the words 'tornado' and 'starbucks' in the same tweet occurred at around 3:30P EDT and the first image was sent at 3:33P EDT, three minutes later. This information was obtained via the Advanced Twitter Search page (https://twitter.com/search-advanced?lang=en) which operates under similar queries as the streaming API. It is safe to assume that the tornado destroyed the Starbucks at around or shortly before 3:30P EDT.

The second set of 1000 tweets, as stated previously was collected between 4:06P-4:10 EDT. At this point, it was widely circulated that the Starbucks has been destroyed and this was evident in a text analysis of the Twitter data. Figure 2 illustrates a 'word count', sorted by alphabetical order from the 1000 tweets, where select words that were mentioned more than 50 times are listed. The frequency of words that appear in the entire sample of tweets are tabulated by removing special characters (e.g., $,*) , 'stopwords' (e.g., a, at, or), delimiters and numbers to preserve whole words which are then counted. As may be expected variations in the word tornado appear in all 1000 tweets (first filter on the data). Other frequently occurring words include 'kokomo', 'warning', 'starbucks', and 'indiana', which gives some idea about the location and event taking place around the time of the data collection. Variations on the word 'damage' (e.g., damage, damages, damaging) were mentioned 54 times in the 1000 tweets. More intensive text analysis (e.g., investigating outliers in frequency of occurrence) will be a part of future work for the project. Another interesting observation to be looked at is the joint occurrence of some words just in case the single occurrence of words are unrelated. For example the words 'starbucks' and 'damage' occur 15 times while 'starbucks' and 'tornado' occur 97 times in the 1000 tweets much more likely than a random sample of 1000 tweets is likely to include.

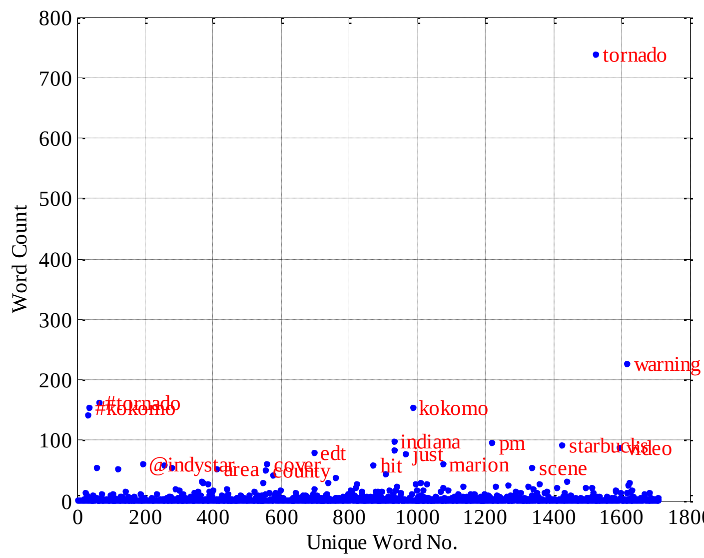

Figure 2. Word count from Twitter data around the time of the Kokomo tornado

Image Analysis

Of the 1000 tweets, 426 of them contained some sort of still image. Each of these 426 images were saved separately from the text (coupled versions still remain) and stored. The first tweet to contain an image of the Starbucks damage (in this set of 1000) is shown on the left hand side of Figure 3. This tweet is a re-tweet (RT) of the original and first photo taken of the Starbucks damage. Other photos showing the damage from different vantage points also were present in the tweet sample. These are also shown in Figure 3 separate from the text accompanying them. Many duplicates of these images existed due to RT. Many pictures of tornado itself also existed in the image set. Videos also existed of the tornado and of the Starbucks collapse itself. Future work will aim to extract and archive videos in the same manner as is done for images. This image and video information could be very valuable to engineers and scientists from a forensic point of view to estimated damage (or tornado) intensity in relatively short period of time after the tornado or for a reanalysis of any tornado event with archived Twitter data (see Future work). For example, images could be rapidly sent to damage assessment experts to estimate tornado intensity and wind speed using the EF Scale. Images of the same structure can be coupled with image analysis software [ref] to create 3-dimensional models of the damage to "recreate" the event. If images of the tornado are available estimates could be made of the damage width which can then be correlated to intensity. All of this information can then be geolocated if a location can be determined. Discussion of geolocation will take place in the next sub-section.

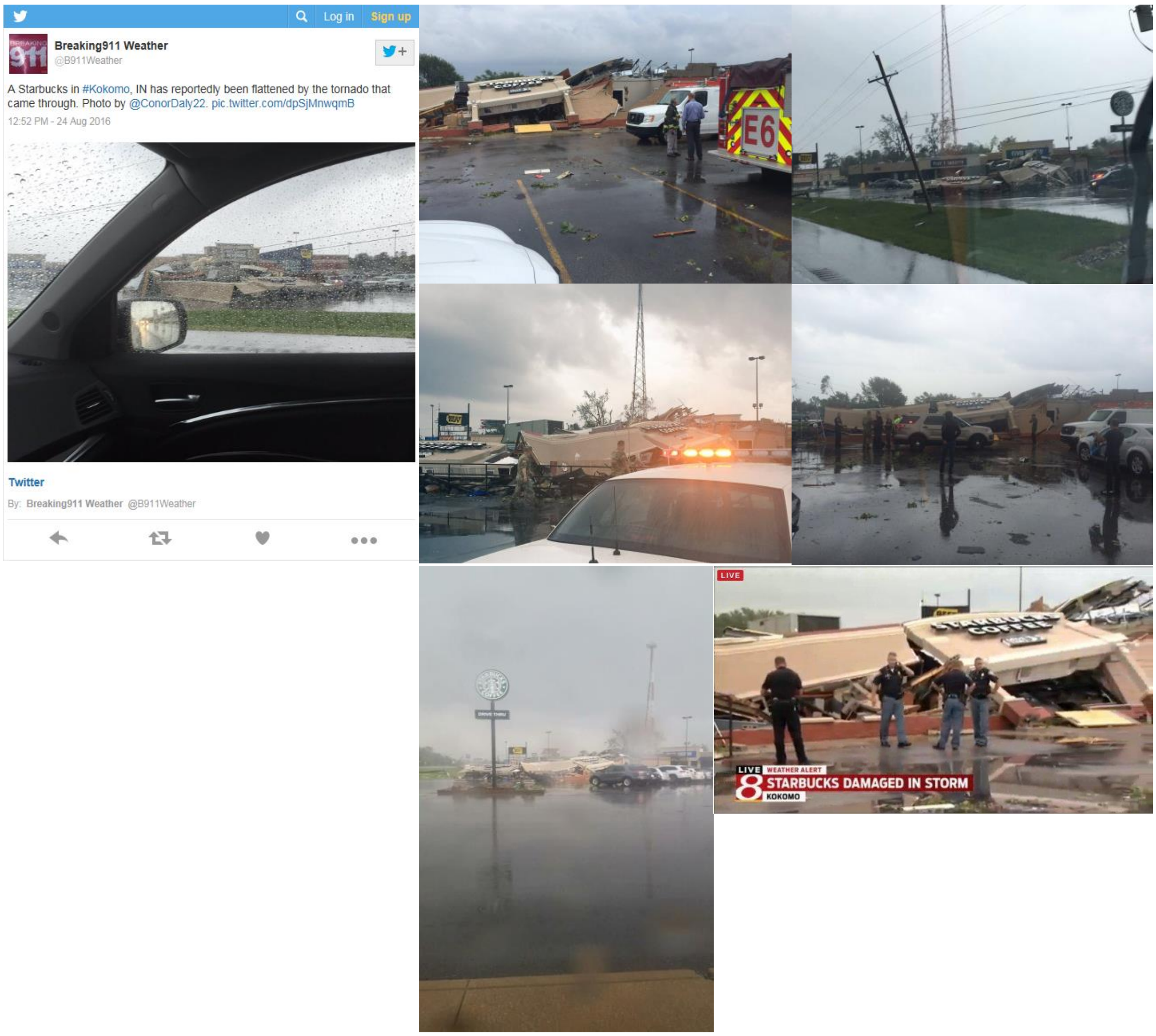

Figure 3. All images of the Starbucks damage extracted from Twitter data around the time of the Kokomo tornado.

Geolocation

For the second set of tweets, only five (5) of the 1000 tweets (0.5%) had encoded coordinates in the Twitter metadata. None of these coordinates were in the vicinity of the Kokomo tornado, however 4 of the 5 were sent by National Weather Service (NWS) accounts referencing other ongoing tornadoes (time was after Kokomo tornado). A small percentage of geolocated tweets is very common and has been noted for other hazard events [ref]. A possible way to ameliorate this problem is to use the text information 'kokomo', 'indiana', 'starbucks' to be able to pinpoint a particular location (or possible locations) of an ongoing event [ref].

b. Japan Access

The 'twitty' code was shared with our counterparts part at Kyoto University's Disaster Prevention Research Institute (DPRI). DPRI attempted to extract tweets that contained keywords in the form of Japanese characters sent from Japan. They were successful in doing so. However the stored data did not retain the Japanese characters for text analysis. The researchers at DPRI believe they have a solution to the problem through an additional MATLAB script and will continue to work on accessing Twitter data regarding Japanese windstorm events as part of ongoing and future work.

IV. Conclusions and Future Work

Windstorms are the cause of significant losses throughout the world, including the U.S. and Japan. Due to the small-scale of some events, measurements often fail to capture relevant details of the event due to poor-resolution or outright lack of sophisticated measurement devices in developing countries. As 'human-sensors' (e.g., iPhones, computers) become ubiquitous they become a promising source of supplementary data for hazard and damage analysis as well as other potential research avenues. In this white paper, samples of Twitter data, including times, text and images relating to a number messages were extracted and archived, using MATLAB, for particular windstorm events in the United States and abroad. A case study focused on a damaging EF-3 tornado that occurred in Kokomo, IN on August 24, 2016. Text analysis from the Tweets revealed the approximate time when a Starbucks suffered catastrophic damage (i.e., collapse) as well as highlighted frequent occurrences of keywords (e.g., starbucks, tornado, kokomo) that suggest location and damage properties. Image analysis revealed 426 of 1000 tweets contained a still image, many of which showed different angles of the Starbucks damage.

Information collected from Twitter data such as the large volume of text and images offer rich avenues for future research. These interrelated research avenues include the following:

- Correlation of text and image properties with actual measurements and intensity of windstorm events. This research could include relating mentions of a specific keyword increasing abnormally compared to 'ambient' tweets coupled with actual wind speed and/or damage measurements.
- Geolocation and prediction of events for decision-making and public awareness including sending pertinent information back to social media sources
- Accessing the full archive of Twitter data in addition to continuous 'real-time' Twitter analysis
- Improved text and image analysis of Twitter (e.g., sentiment analysis, point-cloud model creation)
- Estimation, validation and mapping of windstorm hazard and damage characteristics similar to DYFI using social media data as catalyst
- Extracting and archiving videos on Twitter for further analysis (e.g., photogrammetry)

An example roadmap of these (and other) research avenues planned for windstorms and human-sensor data are included in Figure 4.

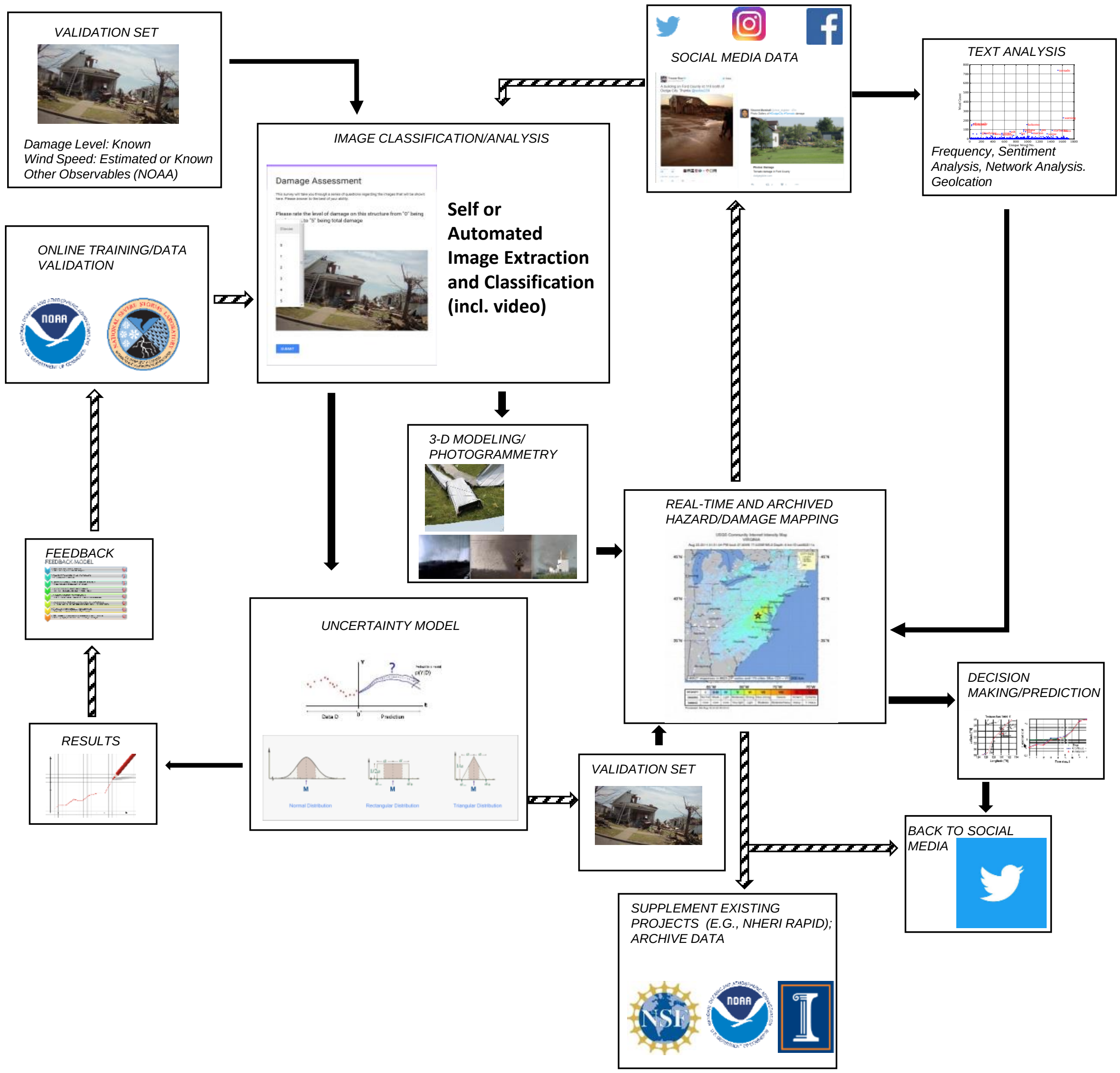

Figure 4. Roadmap of 'human-sensor' research for windstorms.

V. References