Intelligent Technology for Innovation in Urban Disasters

Joe Fan, Adam Lazrus, Anand Ravulapalli, Michael Ross

ABSTRACT

By gathering and training on data relating to hurricanes we illustrate techniques to predict which areas will be most impacted, how successful a recovery plan will be, how best to allocate resources, and the capability of social media data sources to improve the accuracy and responsiveness of damage projection models.

I. INTRODUCTION

Natural disasters in urban areas pose both a unique challenge and a unique opportunity. The high population density, potential access issues due to narrow and blocked roadways, and other issues make responding to, and planning for the response to, natural disasters in urban areas in many ways much more challenging than in rural areas where far fewer individuals and families are affected and routes of access are less constrained than in the urban landscape. Conversely, the same population density allows an economy of scale to leverage the efforts of well-planned first response resources to effect greater relief for a greater number of individuals, and the density of utilities and people make the data collected during recovery much more expansive and comprehensive. As there are many different varieties of disasters with different dynamics, our initial goal was to focus on modelling the damage and recovery from hurricanes in urban areas.

Following the landfall of hurricane Sandy in late October 2012, we quickly decided to refocus our efforts around collecting data for this event. The goal of our project was to model disaster recovery preparedness through machine learning and identify potential subject domains of data for further study. We generated systems to predict the economic, social and technological problems caused by disasters that may be used to assess how effective recovery efforts are and to make predictions for how best to prepare for and respond to a disaster recovery scenario.

We envision that with further development our models will assist in assessing how effective a particular disaster response and recovery plan is likely to be and in developing and executing response and recovery plans by deciding on efficient division of resources and manpower.

II. DATA COLLECTION

Data pertinent to the recent hurricane Sandy were collected from various federal and non-profit organization websites primarily for the date ranges from October 28 through November 1 2012. All data were provided by ZIP code or aggregated by ZIP code during data preparation. The resulting longitudinal data source by ZIP code contained primary subject domains referencing demographic information, disaster damage, utilities interruption, and twitter volume and behavior features.

Data were collected from the following sources:

- **Google Crisis Response:** Information such as storm paths, shelter locations, emergency numbers, and donation opportunities. http://google.org/crisismap/2012-sandy-nyc
- **Department Of Homeland Security - Federal Emergency Management Agency (DHS-FEMA):** Categorical flooding severity by latitude and longitude, with classifications of: Destroyed, Major, Minor, Affected and No Damage. http://goo.gl/Q7czE
- **Safe Guard Properties:** Power outages due to hurricane. http://goo.gl/7EXwL
- **United States Census Bureau:** Demographic and economic information:
  - Demographic Data by Gender and Age as of 2011: http://goo.gl/KUTt1
  - Demographic Data Pertaining to Housing as of 2011: http://goo.gl/jXK07
  - Economic Data for the Demographics: http://goo.gl/S9LnC
- **United States Geological Service (USGS):** Flood levels from data collection stations or measure by FEMA. http://water.usgs.gov/floods/events/2012/sandy/sandymapper.html

Though much of the data were sourced by ZIP code, data provided by latitude and longitude were aggregated to ZIP codes based on US ZIP code centroid reference data from the Census Bureau and Euclidian minimal distance assignment. While this method is not as accurate as geo-polygon assignment, we determined that the accuracy was sufficient given the scope and amount of time allocated for the project.

In addition to the above reference data sources, Twitter data were obtained through a combination of methods. Given the Twitter search API's limited geosearch capabilities, an HTML parser was written in Python to search Twitter via HTML and create a cohort of ~1000 twitter users who could be confidently localized to ZIP codes for which demographic and storm damage data were readily available. Due to the Twitter API's limited search history of approximately 2 weeks, a Python interface to the Topsy Otter API (Toffaletti, 2012) was written to collect all tweets for this cohort for the time period between October 1 2012 and November 15 2012, to ensure coverage of both the storm event and its aftermath and to provide a baseline period of approximately one month. From the resultant data set of approximately 200,000 Twitter status updates, we used the Python Natural Language Toolkit (Bird et al, 2011), referred to as NLTK, to tokenize the status updates and remove filler words to look for the emergence of trending words or phrases around the time of the storm landfall which may indicate the communication of storm related damage. Using these observations, we derived features representing raw update volume, the volume of occurrences of key words, and indicators of behavior changes by individual Twitter users (for example, users with over 30 updates a day who have no updates as the storm makes landfall may indicate a power outage).

The resulting longitudinal data set was comprised of 456 training examples (ZIP codes) with 175 features distributed across the following subject domains:

1. Demographic: Population, building types and density, number of automobiles
2. Disaster damage: Number of homes/units damages and their value in USD
3. Flood severity: Wave/flooding height
4. Utilities interruption: Counts of people who lost electricity
5. Social media: Keyword occurrence and indicators of changes in user volume from Twitter

The overall system design is represented in Figure 1.

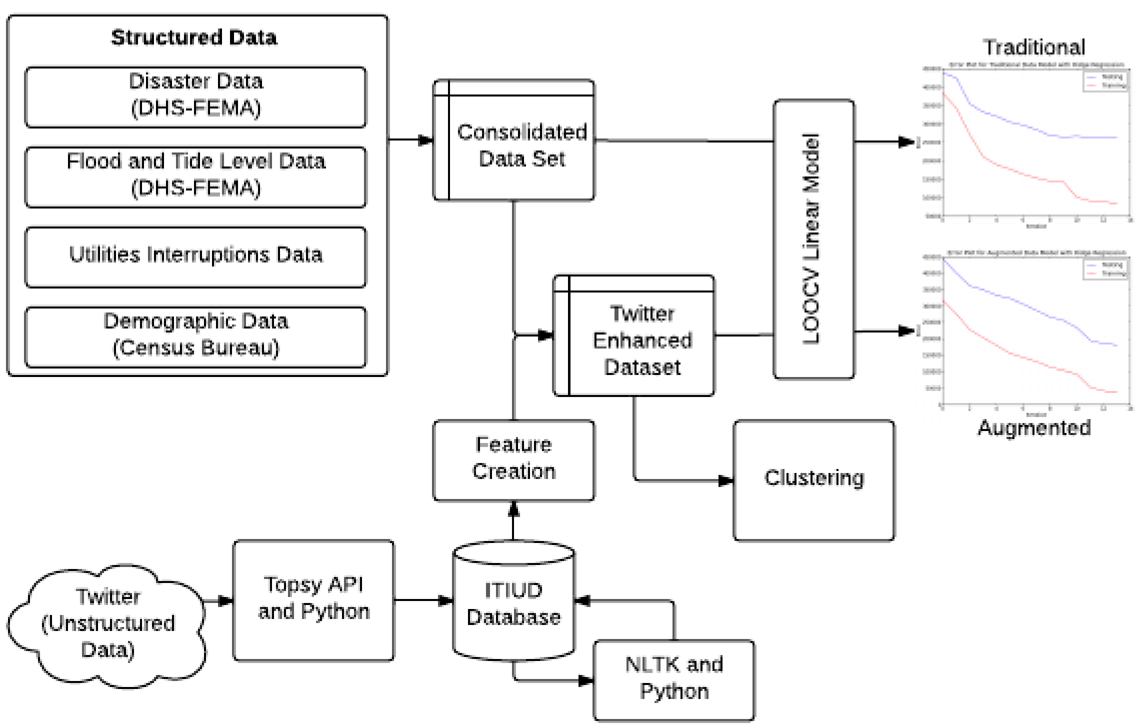

Figure 1: Overall System Design

III. METHODS

Linear Model

Our primary objective was to measure the effectiveness of using social media to enhance Machine Learning models for predicting a hurricane's impact. To this effect, we first joined the demographic and storm data with the twitter data, using ZIP code as the key. The obtained data were used as input to our models.

We first attempted to learn a Linear Regression model using regularization of parameter size. The data comprises many features, but has relatively few training examples, and the model ran the risk of overfitting. To combat this, we ran a feature selection algorithm to pick out at least 10 features to train on. The algorithm continued to pick new features beyond the first 10 as long as the model did not become too overfitted. At each iteration, a new feature was chosen and training and testing error were computed using Leave One Out Cross Validation. To determine when the model was being overfitted, we also compared the change in testing error to the change in training error in each iteration. When the training error changed drastically in comparison to the testing error, the feature selection was ended.

The regularized linear model with feature selection was implemented as a Python module integrating the Scikit-learn (Pedregosa et al, 2011) package. It was first used to train on the traditional data set, without the Twitter usage data. We then augmented the data with the dataset collected from Twitter, and ran feature selection again to retrain the model.

Clustering

As mentioned earlier, the availability and reliability of data varied. FEMA's classification of building damage was often subjective and inaccurate, the height of the storm surge could vary, and the twitter data could require additional signals. Another approach was unsupervised learning, specifically K-means because the assumption and quality of data are more amenable. The storm surge, power outage, and FEMA's multiple classifications of building damage also provided multiple response variables, so we hoped that clustering would help us identify common characteristics across damage, demography, and twitter. Clustering would work better than principal component analysis because we want to grasp the commonality across variables instead of reducing the dimensionality.

We used K-Means and top-down hierarchical K-Means in which the distance measure was Euclidean distance. We experimented with initializing various numbers of clusters and levels, and we found that the optimal number of clusters was 3, and going beyond the 1st level did not provide sufficient separation. One group member works at a company that provides in-database analytic solutions and consulting and was a key resource in implementing clustering. In-database analytics could handle massive datasets with ease, and the computational performance does not deteriorate as the data size increases. We converted the data into the appropriate format and used the company's in-database stored procedure to perform clustering. The stored procedure is implemented on Netezza and includes proprietary queries and user-defined-functions to perform the expectation maximization of the Euclidian distance. The stored procedure and user-defined-functions perform the computations in parallel across multiple computers in the server rack, which increases performance dramatically. After computing the centroids, we improved the comparability across variables by converting them to the percent rank of that variable.

IV. RESULTS

Linear Model

The regularized linear model was run through feature selection, stopping after 13 iterations to prevent overfitting. The final testing and training errors are displayed in Table 1.

	Final Training Error	Final Testing Error
Traditional Model	84,325	263,263
Augmented Model	36,762	181,330

Table 1: Linear model training and test error on traditional and augmented data

The training error and testing error were plotted against each iteration for both the traditional data and the dataset augmented with Twitter data. The plots are displayed in Figure 2.

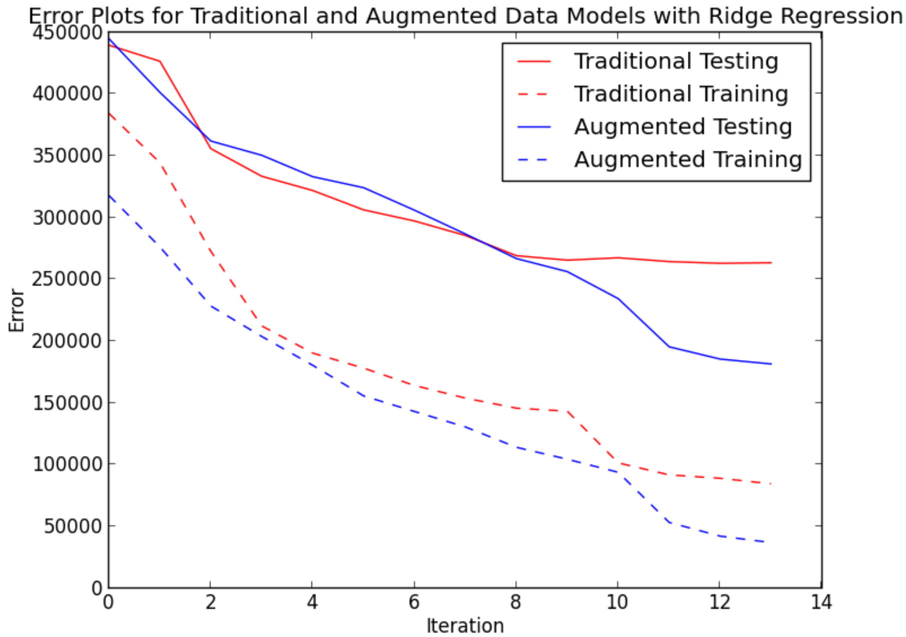

Figure 2: Error Plot for Traditional and Augmented Data Models with Ridge Regression

Clustering

The three centroids that emerged could be roughly categorized with respect to the amount of damage that Hurricane Sandy caused, and we could identify their association of other variables.

- Damage: Cluster 1 suffered significant inundation, damage, and power outage. Cluster 2 focused more on the buildings that were inundated and majorly damaged. Cluster 3 was spared of heavy damage except for minor flooding. Storm surge was highest for Cluster 2, and Cluster 1 had high amount of not inundated severely damaged building, which suggest that additional non-water response variables could be available.
- Commuting to work: There is a separation between communities that drive and do not drive to work. It is unfortunate that communities that drive to work suffered more infrastructure damage, whereas commuters who use public transportation, walked, or worked from home were spared. Also the mean travel time to work would increase even more after Hurricane Sandy for communities that already have relatively long commutes.
- Industry of employment: About 1/3 of the industries are listed here. People working in construction, transportation, warehousing, and utilities were more exposed to storm damage. On the other hand, professionals in information, finance, and management were lightly affected.
- Income: The poor were more exposed to storm damage than the rich. The centroids of Cluster 1 increases to 70% for the poor whereas the centroids of Cluster 3 increases to 100% for the rich.
- Units in structure: Buildings with higher number of units avoided significant damage. They have high centroids in Cluster 2 and Cluster 3.
- Vehicles available: Severe damage and flooding affected households with any number of vehicles. Communities that did not use vehicles had light damage, implying that urban environments fared better than suburbs.
- Twitter: There were few twitter users in high damage areas. Across the high and moderate damage areas, twitter usage decreased by 50% after Hurricane Sandy made landfall on the night of October 29. The majority of tweets about the storm were located in low damage areas whereas very little tweet usage occurred in high damage areas probably because of power outage and life-saving priorities. The keywords “storm” and “hurricane” were the most used and could be indicators of tweets related to the event.

Variables	Cluster 1	Cluster 2	Cluster 3
Damage			
Building not inundated, destroyed	94%	0%	0%
Building not inundated, major damage	91%	82%	0%
Building not inundated, minor damage	85%	52%	0%
Building not inundated, slight damage	78%	55%	0%
Building inundated, destroyed	93%	87%	0%
Building inundated, major damage	91%	81%	0%
Building inundated, minor damage	86%	79%	69%
Building inundated, slight damage	84%	72%	57%
Storm surge at nearest water station	64%	84%	35%
% Power outage in ZIP code	95%	0%	0%
Commuting to work			
Car, truck, or van (drive alone)	59%	28%	30%
Car, truck, or van (carpool)	66%	50%	41%
Public transportation (excluding taxi)	73%	88%	95%
Walked	71%	95%	100%
Other means	69%	94%	98%
Worked from home	63%	94%	99%
Mean travel time to work	51%	48%	32%
Notable industry of employment			
Construction	65%	55%	41%
Transportation, warehousing, utilities	66%	62%	50%
Information	59%	97%	99%
Finance and insurance, real estate	59%	93%	99.8%
Professional, scientific, management	59%	96%	99%
Twitter			
Total # tweets Oct 1 – Nov 4, 2012	56	1,647	7,979
# tweets within 1 week before Oct 29	26	812	3,945
# tweets within 1 week after Oct 29	13	431	3,361
Contains “flood” “rain” “water” “street”	~1	~20	~60
Contains “storm”	1	36	239
Contains “hurricane”	1	46	839

Variables	Cluster 1	Cluster 2	Cluster 3
Income (2011 inflation-adjusted)			
Less than $10,000	70%	88%	87%
$10,000 to $14,999	67%	86%	83%
$15,000 to $24,999	67%	83%	77%
$25,000 to $34,999	64%	80%	75%
$35,000 to $49,999	62%	78%	80%
$50,000 to $74,999	61%	81%	86%
$75,000 to $99,999	56%	82%	95%
$100,000 to $149,999	56%	85%	99%
$150,000 to $199,999	60%	90%	99%
$200,000 or more	65%	96%	100%
Units in structure			
1-unit detached	62%	14%	11%
1-unit attached	74%	67%	39%
2 units	72%	73%	35%
3 to 4 units	73%	87%	63%
5 to 9 units	70%	93%	95%
10 to 19 units	67%	95%	98%
20 or more units	73%	91%	100%
Vehicles available			
No vehicle	74%	91%	99%
1 vehicle	61%	74%	83%
2 vehicles	61%	24%	22%
3 or more vehicles	65%	17%	17%
Equal across all brackets			
Not explanatory			
Year structure built			
House heating fuel			
Sex and Age			

Table 2: Clustering results

V. DISCUSSION

The initial results obtained from linear regression demonstrate that social media can be effective in augmenting the model learned from the other data sources. After incorporating the twitter dataset into LOOCV, we were able to lower the testing error while not overfitting by too much, as indicated by the training error in each model. However, the plots also show that linear regression is not an overwhelmingly effective tool for modeling the data. Nevertheless, the linear model proved to be a good starting point for showing the effectiveness of social media to enhance our models.

The clustering identified similar variables across damage, demographics, and social media. The demographic variables are updated less frequently so the association between damage and demographics could serve as rules-of-thumbs for first responders and for allocating budget to the appropriate municipalities. The Twitter data shows reasonable results for associating the tweeting frequency and words associated with Hurricane Sandy. However, we should use other methods such as Naïve Bayes if we want to properly parse the tweets for richer contextual information. The clustering results provided support that this effort could be worthwhile.

VI. CONCLUSION AND FUTURE WORK

We feel further study is warranted into the near real-time identification of damage areas using social media. Though it is a ‘noisy’ data source, it can serve as an aggregator of the thousands of eyes and voices within an urban area, serving as flood and damage sensors where none are installed. Much more sophisticated techniques can be applied to derive robust features from the Twitter firehose, and more appropriate models than the linear model used here can be selected, but our experiment illustrated that social media features can be useful additions in modeling storm damage. Given that they’re available in near real-time without additional data collection efforts, social media can serve as a valuable addition in planning and responding to urban disasters.

REFERENCES

Bird, Steven, Edward Loper and Ewan Klein. “Natural Language Processing with Python.” O’Reilly Media. 2012.

Pedregosa et al. “Scikit-learn: Machine Learning in Python.” JMLR 12, pp. 2825-2830, 2011.

Toffaletti, Jason. "Topsy’s Otter API." Google Code. http://otter.topsy.com (accessed Nov 11 2012).