Crowdsourcing Incident Information for Emergency Response Using Open Data Sources in Smart Cities

Fan Zuo [a], Ph.D. Candidate, Abdullah Kurkcu [a, b], Ph.D. Candidate, Kaan Ozbay [a, b], Ph.D., Jingqin Gao [a], Ph.D. Candidate,
[a] Department of Civil and Urban Engineering, Tandon School of Engineering, New York University
[b] Center for Urban Science and Progress (CUSP)

NYU
NYU Center for Urban Science + Progress
C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Abstract

Emergency events affect the human security and safety as well as the integrity of the local infrastructure. Emergency response officials are required to make decisions using limited information and time. During emergency events, people post updates to social media net-works such as Twitter containing information about their status, help requests, incident re-ports and other useful information. In this research, the Latent Dirichlet Allocation (LDA) model is used to automatically classify incident related tweets and incident types using Twitter data. The LDA is an unsupervised learning model which can be utilized directly without prior knowledge and preparation for data in order to save time during emergencies. Twitter data including messages and geolocation information during the recent Chelsea explosion and Hurricane Sandy both in New York City are used as two case studies to test the accuracy of the LDA model for extracting incident-related tweets and labeling them by incident type.

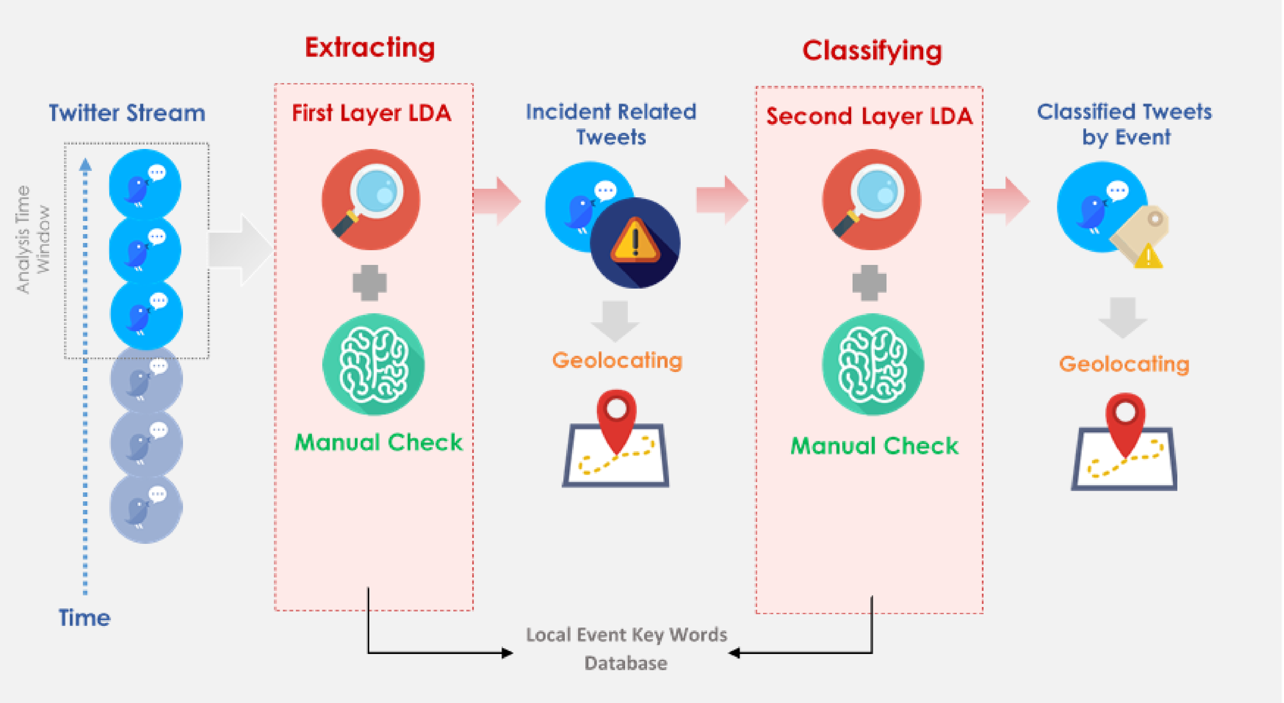

Figure 1. Proposed Tweets-Based Emergency Response System Architecture

Conclusions

- Results showed that the model could extract emergency events and classify them for both small and large-scale events, and the model's hyper-parameters can be shared in a similar language environment to save model training time.
- Furthermore, the list of keywords generated by the model can be used as prior knowledge for emergency event classification and training of supervised classification models such as SVM and Recurrent Neural Network.

LDA Model Pseudocode

Topic Layer
for *all topics* $k \in [1, K]$ **do**
sample mixture components $\beta_k \sim Dirichlet_V(\eta)$

Document Layer
for *all documents* $d \in [1, D]$ **do**
sample mixture proportion $\theta_d \sim Dirichlet_K(\alpha)$
Word Layer
for *each word* $n \in [1, N]$ *in document d* **do**
sample topic index $Z_{d,n} \sim Multinomial_K(\theta_d)$
sample term for word $W_{d,n} \sim Multinomial_V(\beta_{Z_{d,n}})$

Joint Distribution of All Variables in LDA Model

$$p(\vec{\beta},\vec{\theta},\vec{Z},\vec{W}|\vec{\alpha},\vec{\eta}) = \left(\prod_{k=1}^{K} p(\beta_k|\vec{\eta})\right)\left(\prod_{d=1}^{D} p(\theta_d|\vec{\alpha})\prod_{n=1}^{N} p(Z_{d,n}|\vec{\theta_d})\,p(W_{d,n}|\vec{\beta_k},\vec{Z_{d,n}})\right)$$

Objective Function of Model Training

$$\ell(\vec{\alpha},\vec{\eta}) = \log p(D|\vec{\alpha},\vec{\eta}) = \log\prod_{d=1}^{D} p(\vec{w_d}|\vec{\alpha},\vec{\eta}) = \sum_{d=1}^{D}\log p(\vec{w_d}|\vec{\alpha},\vec{\eta})$$

Figure 2. Additional Incident type generated from Twitter data

Important Hyper-parameters

K : The number of topics. This is the most important parameter directly affecting the training result. This will be further discussed in the case study section;
α : The prior of topic Dirichlet distribution $\beta_k \sim Dirichlet_V(\eta)$. It indicates how many topics a document may have, $\alpha = 0.01$ was set as default;
β : The prior of word Dirichlet distribution, $\beta = 0.01$ was set as default.

Table 1: The Number of Tweets by Incident Type

Incidents type	Number	Percentage	Incidents type	Number	Percentage
Communication	76	4.04%	Traffic	176	9.36%
Debris	62	3.30%	Transit	236	12.55%
Flooding	393	20.90%	Tree	345	18.35%
Gasoline	472	25.11%	Wind	120	6.38%

TABLE 2a. Performances of Classification Training by the LDA Model with $K = 20$

Incident Type	Percentage	Accuracy	Precision	Recall	F1
Community	4.04%	0.975531	0.720588	0.644737	0.680556
Debris	3.30%	0.954255	0.363636	0.516129	0.426667
Flood	20.90%	0.830851	0.587822	0.638677	0.612195
Gasoline	25.11%	0.895213	0.793177	0.788136	0.790648
Traffic	9.36%	0.896809	0.434783	0.340909	0.382166
Transit	12.55%	0.912766	0.627660	0.750000	0.683398
Tree	18.35%	0.924468	0.883019	0.678261	0.767213
Wind	6.38%	0.945213	0.559441	0.666667	0.608365
Total	100%	0.916888	0.621266	0.627939	0.618901

TABLE 2b. Performances of Classification Testing by the LDA Model with $K = 20$

Incident Type	Percentage	Accuracy	Precision	Recall	F1
Communication	3.39%	0.967797	0.529412	0.450000	0.486486
Debris*	3.56%	0.000000	0.000000	0.000000	0.000000
Flood	20.34%	0.791525	0.487603	0.491667	0.489627
Gas	24.24%	0.761017	0.505952	0.594406	0.546624
Traffic	9.49%	0.896610	0.456140	0.464286	0.460177
Transit	12.88%	0.854237	0.443182	0.513158	0.475610
Tree	19.49%	0.833898	0.613333	0.400000	0.484211
Wind	6.61%	0.923729	0.442308	0.589744	0.505495
Total	100.00%	0.861259	0.496847	0.500466	0.492604

*: Event "Debris" was not detected

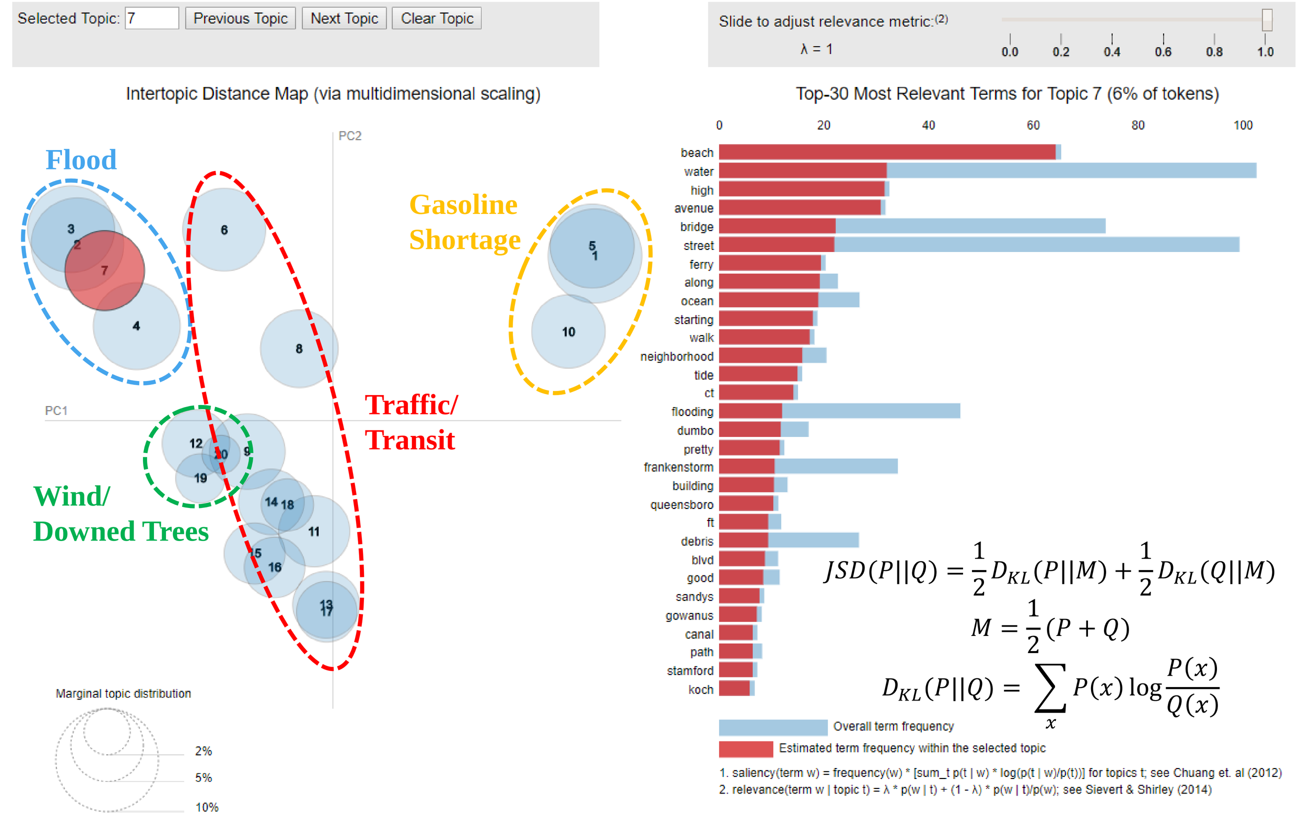

Figure 5. Inter-Topic Distance Map via Multidimensional Scaling and Manual Clusters
(Python Package: LDAvis, Sievert, C., and K. E. Shirley)

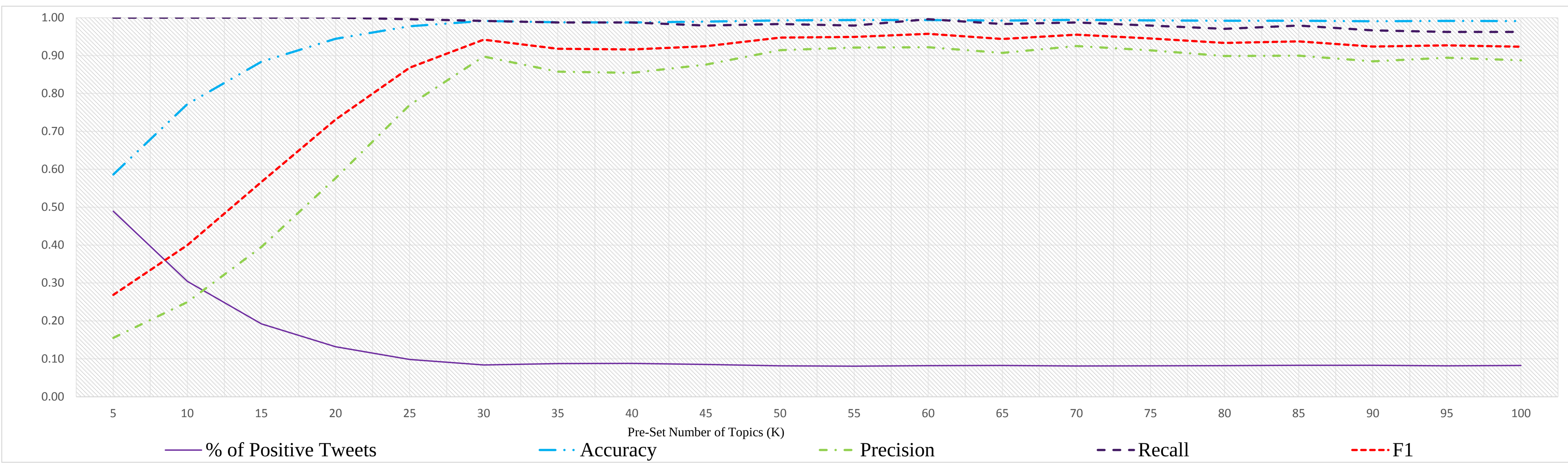

Figure 3. Event Extraction Performance of the LDA Model for Hurricane Sandy

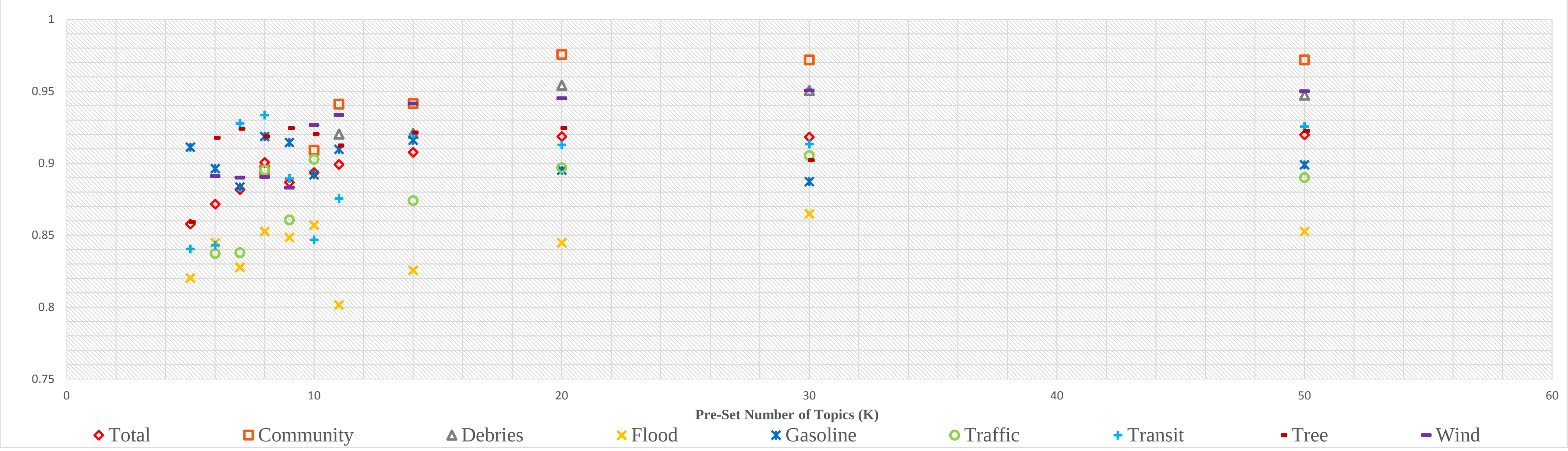

Figure 4. The Accuracy Distribution of LDA Model for Events Classification Training

Case Study 1: Hurricane Sandy

- **258** manually labeled incident-related tweets are selected out of **3,131**.
- The dataset was randomly separated into **2087** for **training** and **1044** for **testing**.

Case Study 2: Chelsea Explosion

- **61,089** tweets after the explosion, the vocabulary of $V = 10213$ unique words with total document size $N = 321478$ words.
- The dataset was randomly separated into **40306** for **training** and **20783** for **testing**.

Table 3: The Ratio and Top 10 Keywords of Chelsea Explosion Related Topics

K	Topic selected from generated topics correlate with Chelsea explosion (Presented by top 10 keywords)	% of Total Tweets
6	chelsea game day one today get night giants explosion go	27.81%
8	chelsea get shit know go really fuck im time explosion	18.27%
10	chelsea one explosion time know never us trump live get	13.67%
13	chelsea trump explosion safe news one bomb dvd cases know	8.34%
15	chelsea explosion safe one news everyone bomb last stay today	7.75%
20	chelsea explosion safe stay bomb everyone news night morning manhattan	5.48%
25	chelsea explosion safe stay bomb manhattan news united bombing police	4.08%
30	chelsea explosion safe stay everyone news bomb cases manhattan nypd	2.53%
	chelsea everyone safe stay know hope live going explosion real	1.12%
40	chelsea explosion bomb news bombing police safe manhattan alert nypd	2.14%
50	chelsea explosion safe bomb stay news manhattan everyone bombing police	2.03%
60	chelsea explosion safe stay manhattan everyone hope ok bomb away	1.74%
	chelsea bombing alert police nj rahami suspect ahmad khan act	1.27%
80	chelsea news explosion police alert bombing nypd bomb rahami suspect	1.11%
	chelsea safe stay everyone explosion hope manhattan away tonight heard	0.95%

Contact:
fz380@nyu.edu

Connected Cities for Smart Mobility toward Accessible and Resilient Transportation Center (C2SMART)
http://c2smart.engineering.nyu.edu/