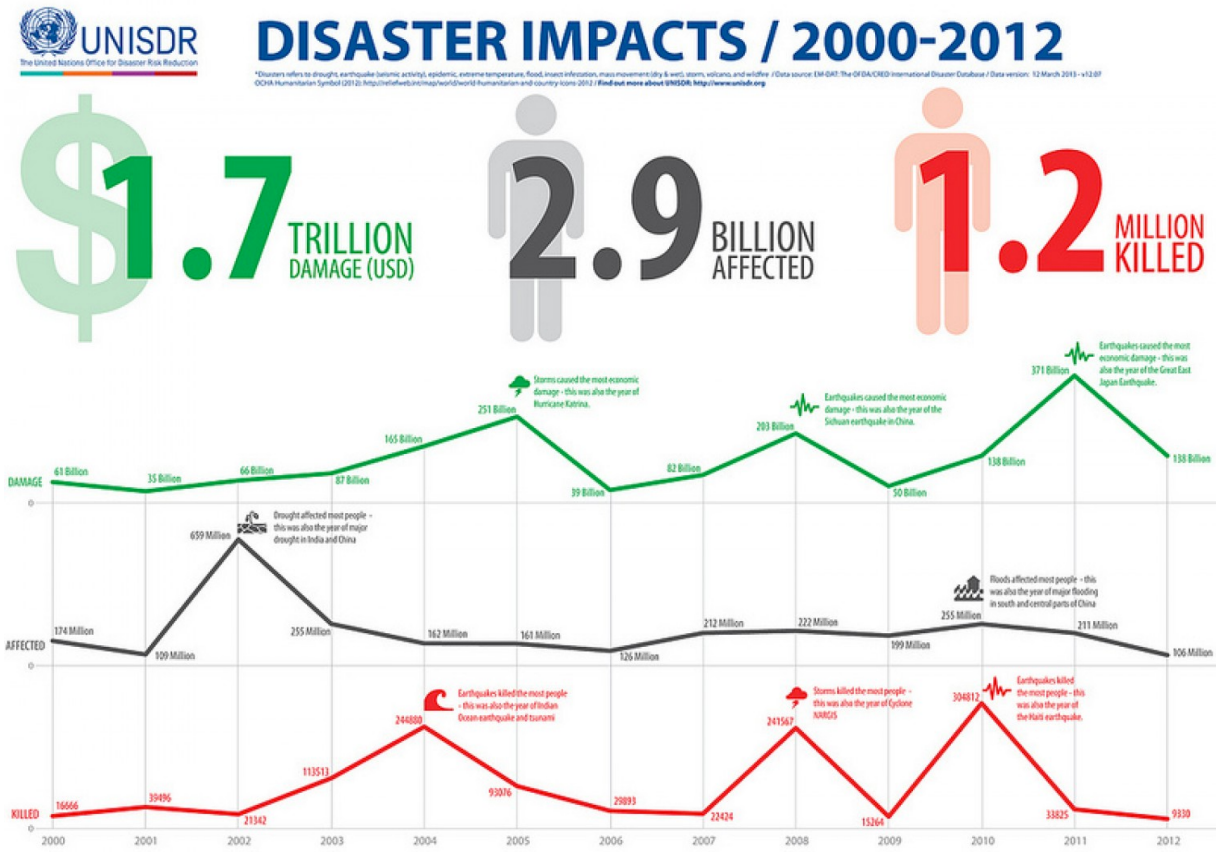
"Data For Good" - Leveraging Social-small-talks for Crisis Informatics

Soudip Roy Chowdhury

Collaborators

- Dr. Sihem Amer-Yahia, Director, CNRS, France
- Dr. Carlos Castillo, Director, Data Science, Eurecat, Spain
- Dr. Hemant Purohit, Asst. Prof. George Mason University, USA
- Dr. Muhammad Imran, Researcher, QCRI, Qatar
- Dr. Muhammad Rizwan Asghar, Lecturer, University of Auckland, New Zealand

Why do we care?



- Millions of people affected each year by disasters
- At an annual cost to the global economy that exceeds $300 billion
- In 2015
  - 26,000 persons killed
  - $85 billions lost

Reality in India is not much different?

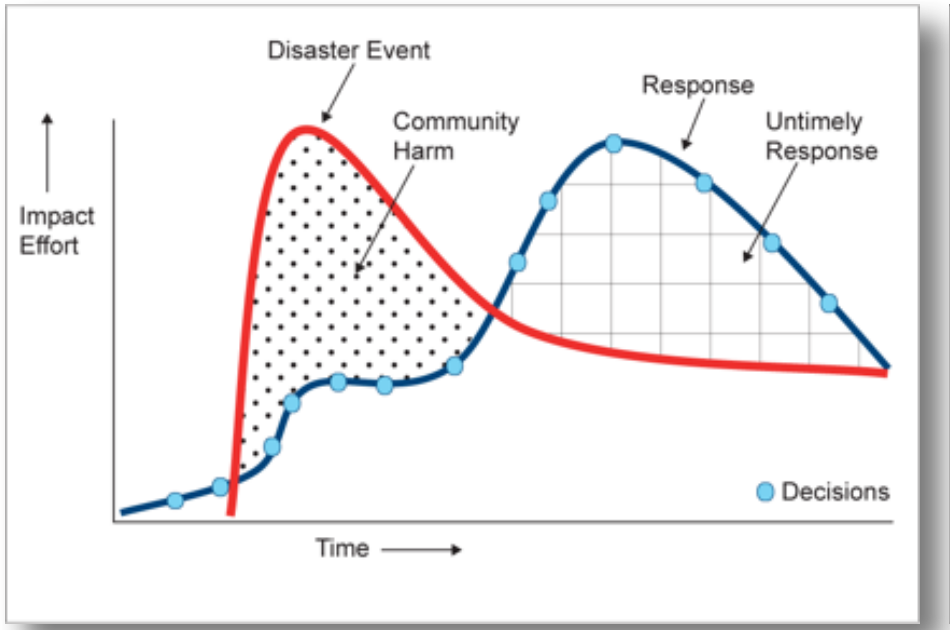
- 57% of land is vulnerable to earthquakes
- 68% of land is vulnerable to drought
- 12% land is Vulnerable to floods
- 8% land is vulnerable to cyclones

Better preparedness may not avert disaster, but can certainly save lives and properties!

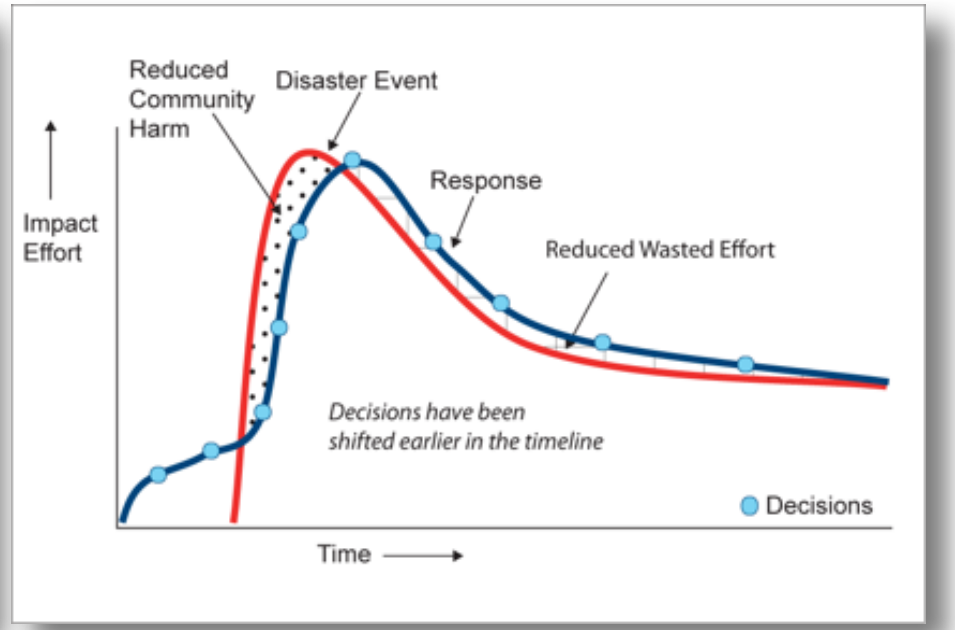
Source: National Institute of Disaster Management, Gov. India

IMPACT AND RESPONSE TIMELINE

Disaster response (today)



Disaster response (target)



Effective disaster response requires real-time coordinated actions

Department of Community Safety, Queensland Govt. & UNOCHA, 2011

Information is the key!

- People need information as much as water, food, medicine or shelter
- Information can help people-in-distress to access necessary aids
- It also helps organizations to better coordinate aiding services

Where can we find **real-time** information?

Twitter Breaks Events Faster

Hudson Plane Crash

First report

Janis Krums
@jkrums
Follow

http://twitpic.com/135xa - There's a plane in the Hudson. I'm on the ferry going to pick up the people. Crazy.

RETWEETS 244 LIKES 878

9:36 PM - 15 Jan 2009

Westgate Mall Attack

Breaks the story 33 minutes before local TV

I come from Somalia.
@geekinthejungle
Follow

gun shots in westlands? wtf??

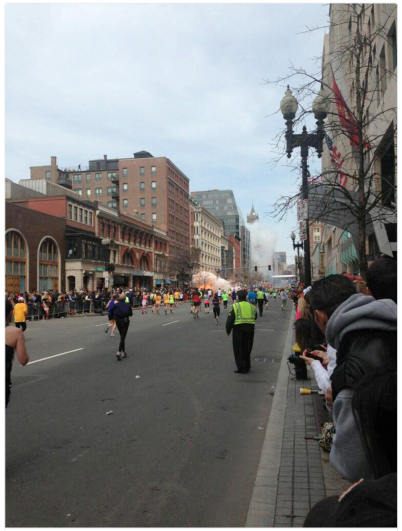
Reply Retweet Favorite Buffer Storify More HootSuite

12:38 PM - 21 Sep 13

Twitter Breaks Events Faster

Boston Bombing

First report on Twitter



Dan Lampariello
@WBBJ7Dan
Follow

Explosion at coply

9:50 PM - 15 Apr 2013

1,943 187

After 1 minute

Pure Applesauce
@stackiii
Follow

An explosion just went off in downtown Boston. Spectators fleeing the #bostonmarathon course.

9:51 PM - 15 Apr 2013

79 11

Pure Applesauce
@stackiii
Follow

I think a bomb just went off in Boston. Can't tell. Can smell smoke. Emergency vehicles everywhere. instagram.com/p/YItk1gTIUJ/

9:53 PM - 15 Apr 2013

308 21

After 2 minutes

Fox Sports 1380 KRKO
@KRKO1380
Follow

BREAKING: Per our man on the ground at the Boston Marathon, @tooblackdogs, there was an explosion. More to follow.

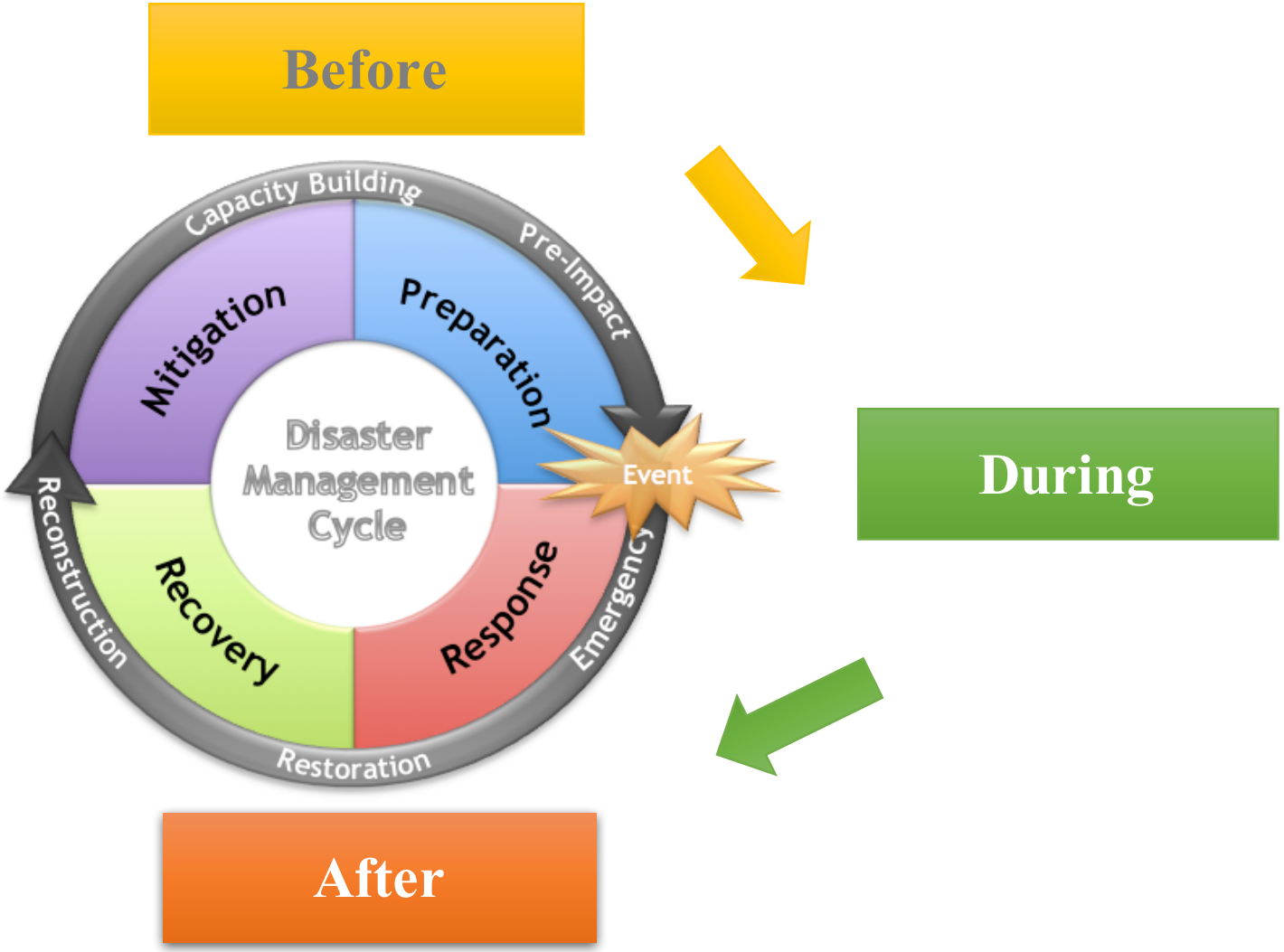
9:52 PM - 15 Apr 2013

153 3

Social media plays an important role!

- Humans is the most accurate sensor
- Human + mobile is the most sophisticated IoT device
- Social media provides the most valuable information
- Users posts in social media contribute to situational awareness
- Highly useful, if analyzed timely and effectively – *Crisis Informatics*

Social Media Response to Disaster Phases



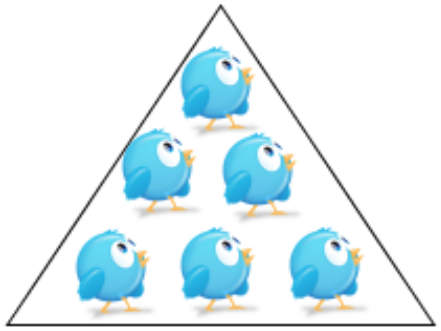
FEMA

OCHA

What are we trying to solve?

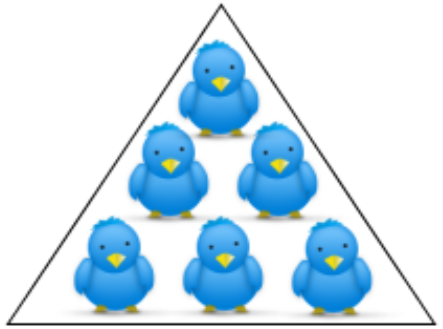
1. Identify messages related to an incident (data cleaning)
2. Classify incident-messages with the corresponding period (PRE, DURING, POST).

PRE-INCIDENT



- **Caution, warnings**
- **Alerts etc.**

DURING-INCIDENT



- **Damage**
- **Causalities etc.**

POST-INCIDENT

- **Request for help**
- **Donations etc.**

What do we mean by these periods?

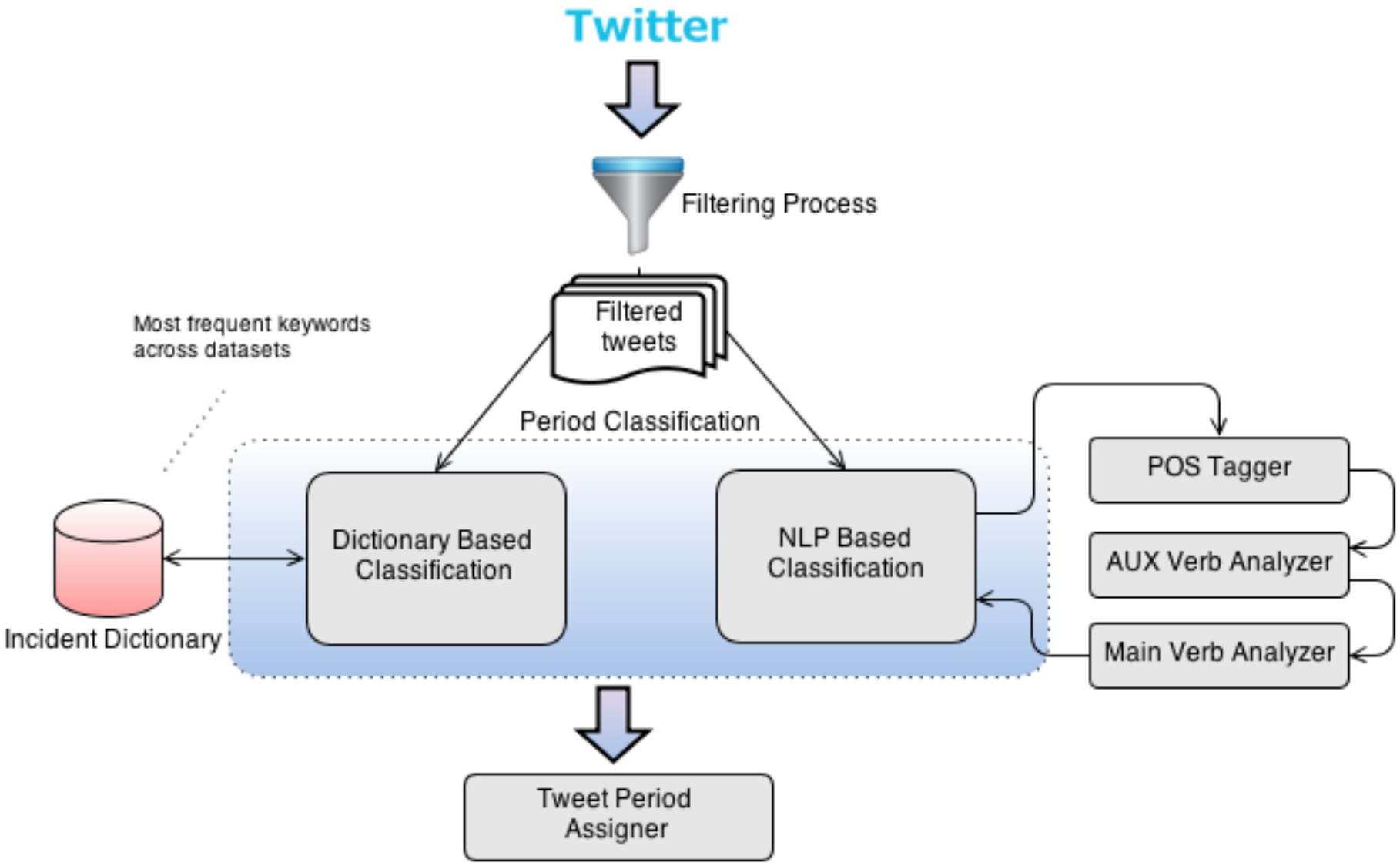
- Lets see an example - **Joplin Tornado on May 22, 2011**

[PRE] New #tropical storm forms in the West #Pacific. #Nesat may hit the #Philippines & #China as a #typhoon next week

[DURING] @Yahoo News: Powerful #typhoon with winds up to 106 mph makes landfall in #Philippines as 100,000 odered to fless homes

[POST] News5 Action center is now accepting donations for the victims of Typhoon “pedring. Drop boxes are located @ TV5 Office :)

Tweet4Act System



Data filtering / cleaning process

- Normalization: remove “RT @username” and “@username” prefixes and remove duplicate messages
- Apply the k-mediod method with the manhattan distance between medoids and messages in each cluster
- Discard all cluster having a negative number or zero as *silhouette coefficient*
- Select from each cluster the fraction *m* messages closer to the mediod

Classification Techniques

- Domain-Specific Dictionary
  - Most frequent words across datasets (tf-idf)
    - “warning” & “alert” typically found in the *Pre*
    - “now”, “sweeps” etc. typically found in *During*
    - “aftermath”, “donate” etc. typically found in *Post*

- NLP-based period classification
  - Tense of verbs can help identify period. (A. Iyengar et al., 2011)
    - Tense of aux verbs add +0.5 to pre/during/post period
    - Tense of main verb in future/present/past tense, add +0.5 to pre/during/post period, respectively.
- Weighted average of above two techniques determines the final classification label

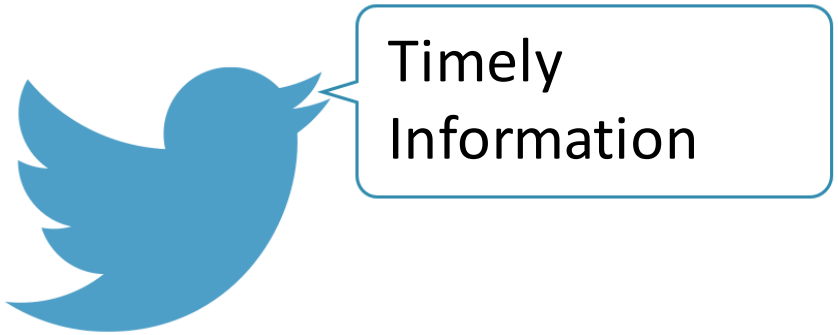
Accuracy of Tweet4Act

Period	Tweet4act			SVM			MaxEnt			Tree			RF		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
	Joplin Tornado														
PRE	0.33	**0.85**	**0.48**	0.00	0.00	0.00	**0.43**	0.21	0.28	0	0	0	0	0	0
DURING	**0.88**	0.89	**0.89**	0.32	**0.91**	0.47	0.35	0.55	0.43	0.3	0.73	0.43	0.32	0.1	0.48
POST	0.61	**0.84**	**0.71**	**0.67**	0.20	0.31	0.55	0.6	0.57	0.57	0.4	0.47	1.00	0.1	0.18
AVG	**0.61**	**0.86**	**0.69**	0.33	0.37	0.39	0.44	0.45	0.42	0.29	0.38	0.45	0.66	0.37	0.33
	Haiti Earthquake														
PRE	0.63	**1.00**	0.77	**1.00**	0.67	0.80	**1.00**	**1.00**	**1.00**	**1**	0.67	0.8	**1.00**	0.33	0.5
DURING	0.72	**0.97**	**0.83**	0.75	0.6	0.67	0.67	0.80	0.73	0.6	0.6	0.6	**1.00**	0.4	0.57
POST	0.46	0.82	0.59	0.92	0.97	0.94	**0.97**	0.95	**0.96**	0.92	0.95	0.93	0.88	**1.00**	0.94
AVG	0.60	0.87	0.71	0.89	0.74	0.80	0.88	**0.91**	**0.89**	0.84	0.74	0.78	**0.96**	0.58	0.67
	Nesat Typhoon														
PRE	0.36	**1.00**	0.53	**1.00**	0.50	**0.67**	**1.00**	0.50	**0.67**	0.33	0.25	0.28	1.00	0.5	0.67
DURING	**0.94**	0.94	**0.94**	0.79	**1.00**	0.88	0.81	1.00	0.90	0.71	0.77	0.74	0.79	**1**	0.88
POST	0.52	**0.85**	**0.65**	**1.00**	0.2	0.33	**1.00**	0.40	0.57	0	0	0	**1.00**	0.2	0.33
AVG	0.61	**0.93**	**0.71**	0.93	0.57	0.62	**0.94**	0.63	**0.71**	0.35	0.34	0.51	0.93	0.57	0.63

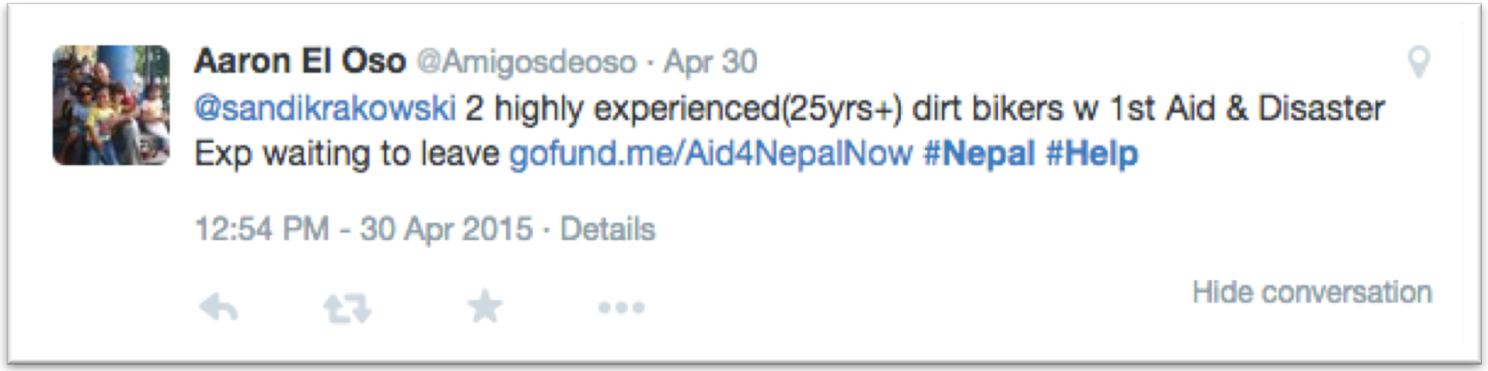
But this is not enough!

- Data curated for a specific period
  - *Must be actionable for agencies to work on it*
- Lets see an example of actionable information during disaster

Tweets *During* Nepal EQ

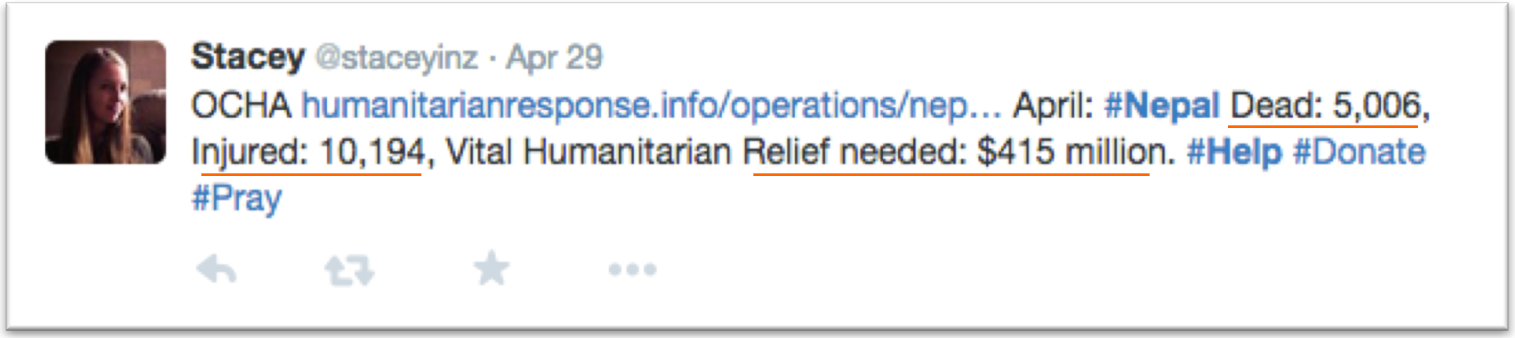


Post event Tweets…



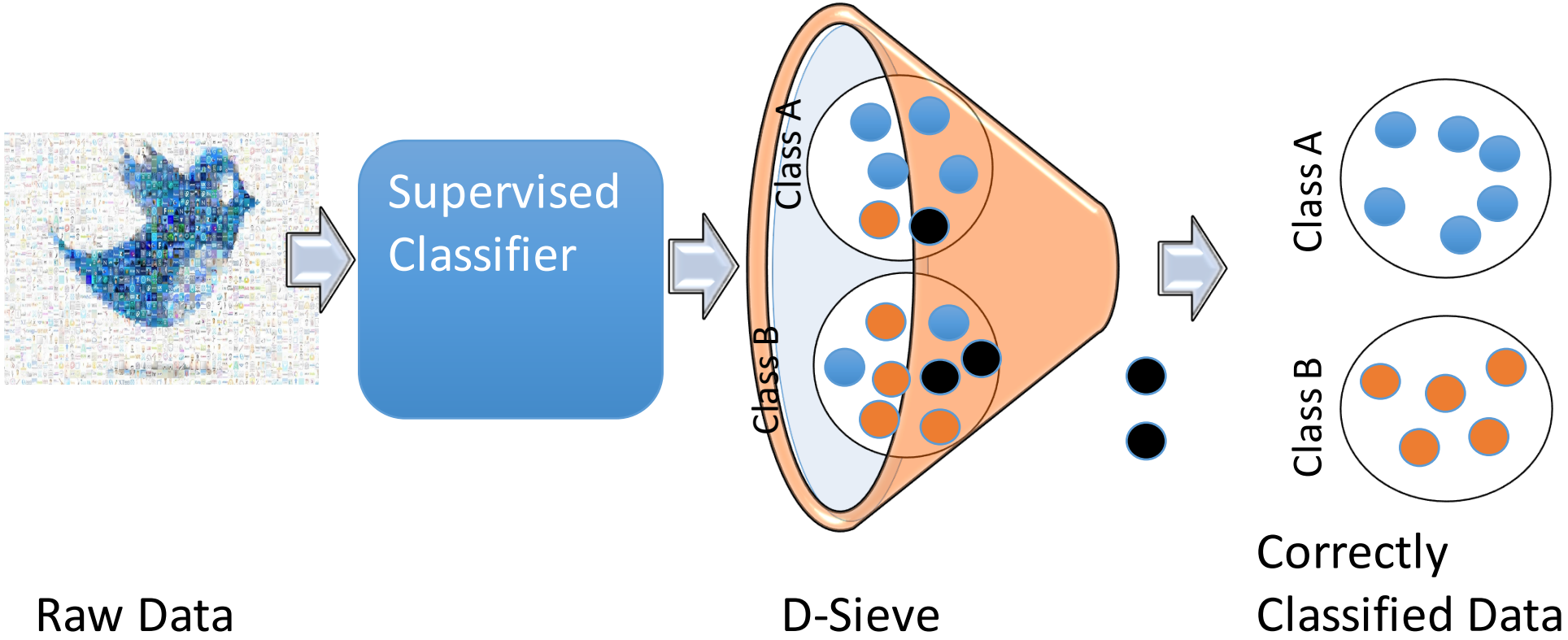
Actionable Information

And it contains…



D-Sieve

- Data processing engine in a supervised classification setting



Data Model

- A message contains

@sandikrakowski 2 highly experienced(25yrs+) dirt bikers w 1st Aid & Disaster Exp waiting to leave gofund.me/Aid4NepalNow #Nepal #Help

9:54 PM - 30 Apr 2015 · Details

TEXT

Date

Data Model

- A message contains

@sandikrakowski 2 highly experienced(25yrs+) dirt bikers w 1st Aid & Disaster Exp waiting to leave gofund.me/Aid4NepalNow #Nepal #Help

9:54 PM - 30 Apr 2015 · Details

TEXT

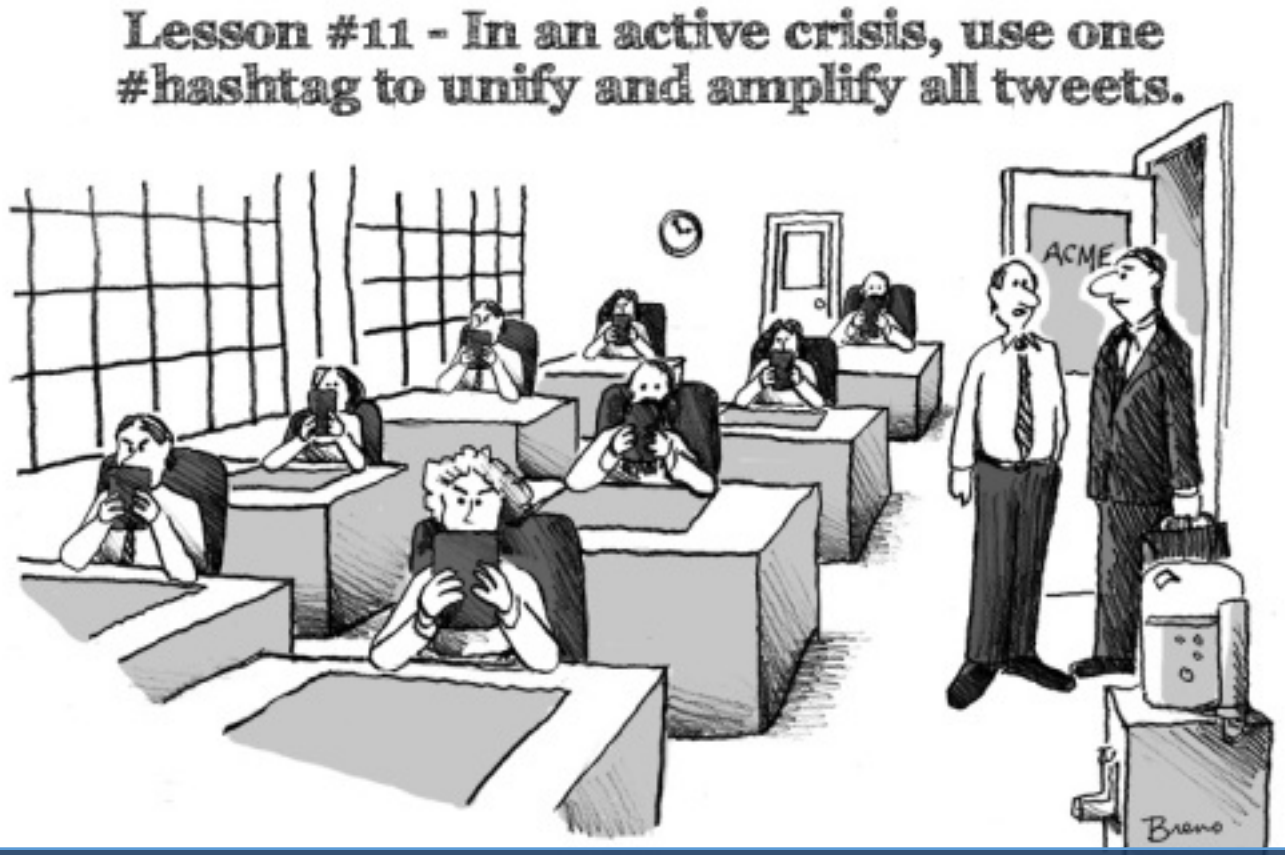
- hashtags
- content
- URL

Date

Our Approach

- D-Sieve extracts two content features to classify messages correctly
  - Stable hashtag association
  - Stable named entity association

Stable Hashtag Association



Relative frequencies of stable hashtags for an event get converged over time….

Stable Hashtag Association

- A hashtag is stable, if the moving average score of its frequency distribution (after *k* posts) becomes steady [X.S. Yang et al. 2013]
- Stable hashtag set (ϕ) contains all stable hashtags for an incident
- Stable hashtag association metric for an input message

$$V_H = \frac{|(\phi_{TP} - \phi_{TN}) \cap H_i| - |(\phi_{TN} - \phi_{TP}) \cap H_i|}{H_i}$$

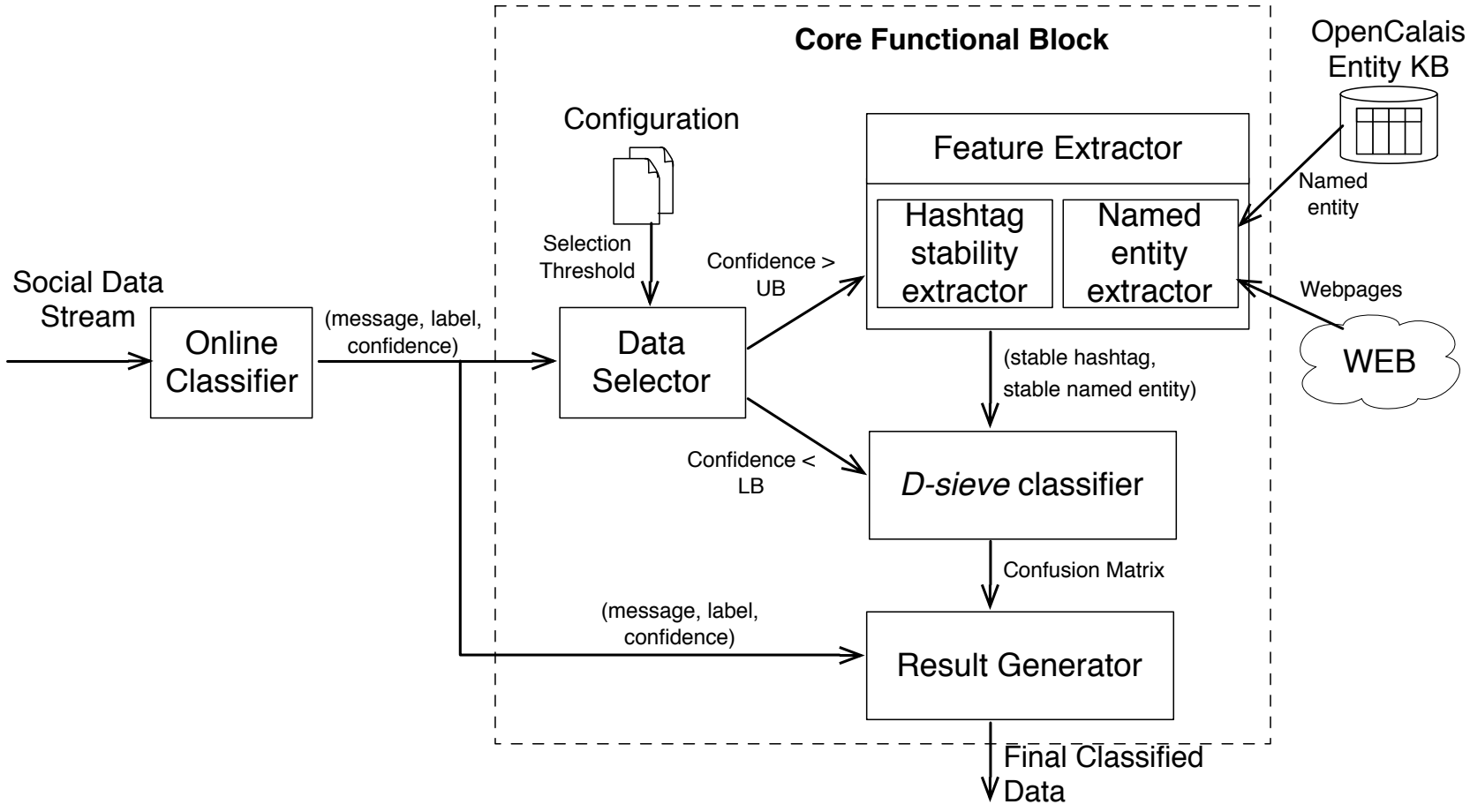
- ϕ_{TP} and ϕ_{TN} are set of stable hashtags for true positive and true negative classes

Stable Named Entity Association

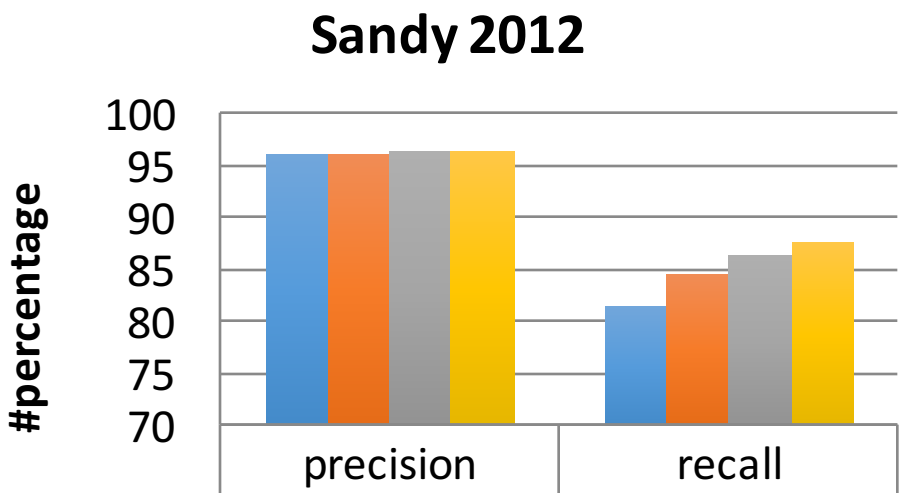
- Named entities are extracted from the message content + URL content
- Stable named entity association metric for an input message $$V_E = \frac{|(E_{TP} - E_{TN}) \cap I_E| - |(E_{TN} - E_{TP}) \cap I_E|}{I_E}$$

- We calculate the final feature association metrics $S_C = W_H . V_H + W_E . V_E$
- Feature weights (W_H, W_E) are chosen such that S_C ranges [-1,1]
- Rule of thumb weights are set as 0.5

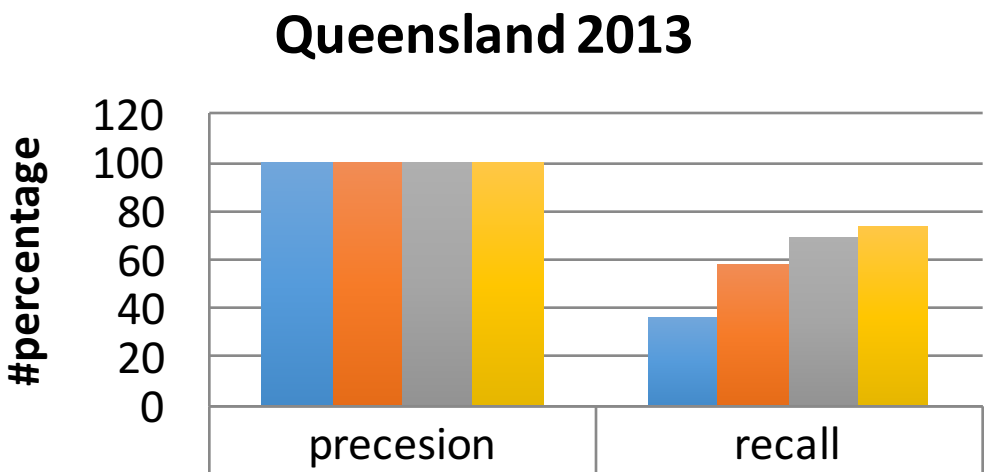
Functional Architecture



Experimental Results



	precision	recall
OC	96.09	81.42
OC+HT	96.23	84.51
OC+NE	96.31	86.49
OC++	96.35	87.55



	precesion	recall
OC	99.53	36.75
OC+HT	99.8	58.47
OC+NE	99.75	68.9
OC++	99.8	74.24

D-sieve improves recall values substantially (6% and 37% improvements for two datasets).

Take home messages

- Social-data analytics plays an important role for devising tools for effective disaster management
- Though mixed with noise, often social data contains valuable actionable/tactical information in disaster situation
- Capability to analyze social data at real-time for finding out the signal requires domain specific data engineering excellence
- Right choice of data mining and machine learning algorithms with NLP features is the recipe to devise the most powerful disaster management tool

Why your tweets could really matter during a natural disaster

References

- A. Liaw and M. Wiener. Classification and regression by randomforest. R news, 2(3):18–22, 2002.
- A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg. “Crisislex: A lexicon for collecting and filtering microblogged communications in crises”. In ICWSM, 2014.
- N. Morrow, N. Mock, A. Papendieck, and N. Kocmich. “Independent evaluation of the ushahidi haiti project”. Development Information Systems International, 2011.
- S. Roy Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo. “Tweet4act: Using incident-specific profiles for classifying crisis-related messages”. In ISCRAM 2013.
- Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. “Earthquake shakes Twitter users: real-time event detection by social sensors”. In WWW 2010.
- X. S. Yang, D. W. Cheung, L. Mo, R. Cheng, and B. Kao. “On incentive-based tagging”. In ICDE 2013.

References

- A. Iyengar, T. Finin, and A. Joshi (2011) Content-based prediction of temporal boundaries for events in Twitter. In Proceedings of the Third IEEE International Conference on Social Computing.

- K. Starbird, L. Palen, A. Hughes, and S. Vieweg (2010) Chatter on the red: what hazards threat reveals about the social life of microblogged information. In Proceedings of the 2010 ACM conference on Computer supported cooperative work, pages 241–250. ACM.

- Latonero, Mark, and Irina Shklovski. "“Respectfully Yours in Safety and Service”: Emergency Management & Social Media Evangelism.” Proceedings of the 7th International ISCRAM Conference– Seattle. Vol. 1. 2010.

thank

{you}