Privacy and Security in Online Social Media

Course on NPTEL

NOC21-CS28

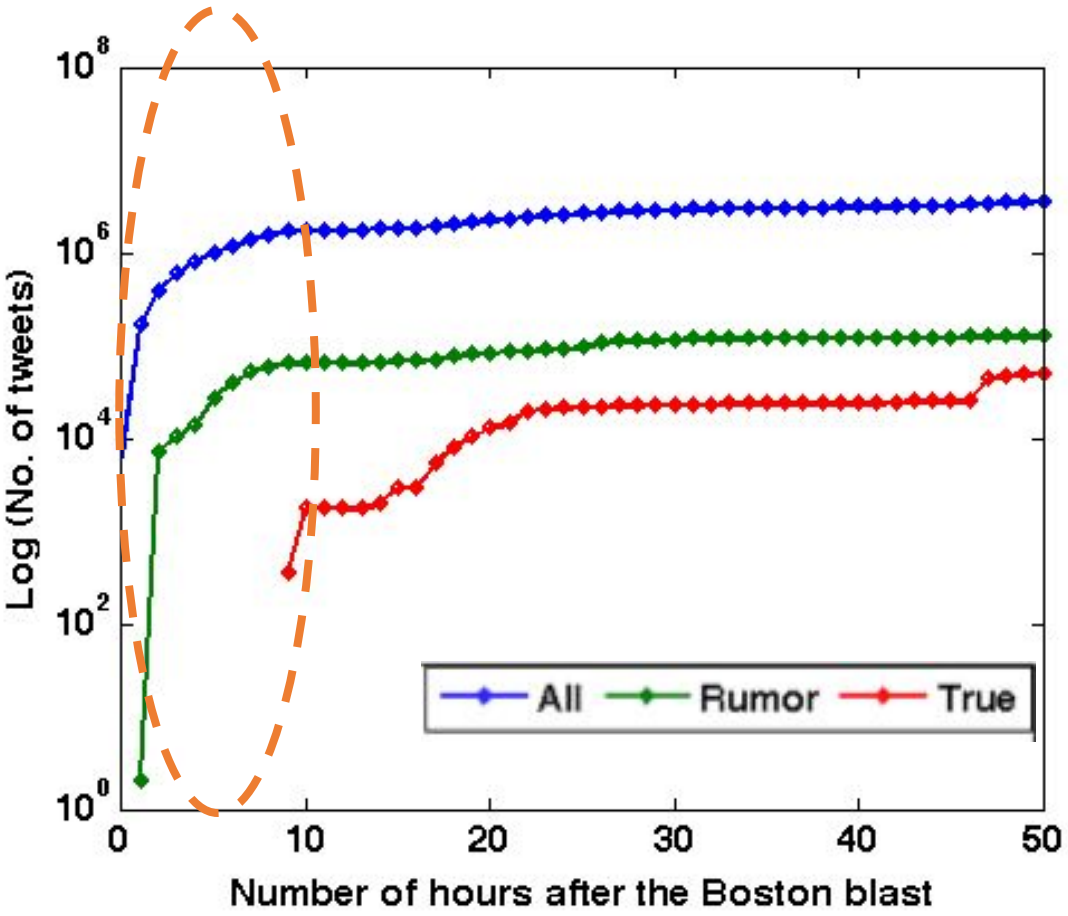
Week 2.2

Ponnurangam Kumaraguru (“PK”)
Full Professor
ACM Distinguished Speaker
fb/ponnurangam.kumaraguru, @ponguru

Topics that we will cover

- Overview of OSM
- Linux / Python / Twitter API / Mongo DB / MySQL [Hands-on]
- **Trust & Credibility**
- Privacy
- Social Network Analysis, NLTK [Hands-on]
- e-crime
- Plotly / Highcharts / Geo-location analysis [Hands-on]
- Policing
- Identity resolution
- What next – Deep learning, machine learning, NLP, Image analysis

Temporal Patterns



Fake content / rumors becomes viral in first 7-8 hours just after the event.

Misinformation on Social Media

Boston Bombing Facebook And Twitter Page 'Fakes' Set Up To Capitalise On Tragedy

Huffington Post UK | By Felicity Morse
Posted: 16/04/2013 09:34 BST | Updated: 16/04/2013 16:36 BST

Like 2,240 people like this. Be the first of your friends.

1,175 share | 266 tweet | 21 +1 | 4 email | 120 comment

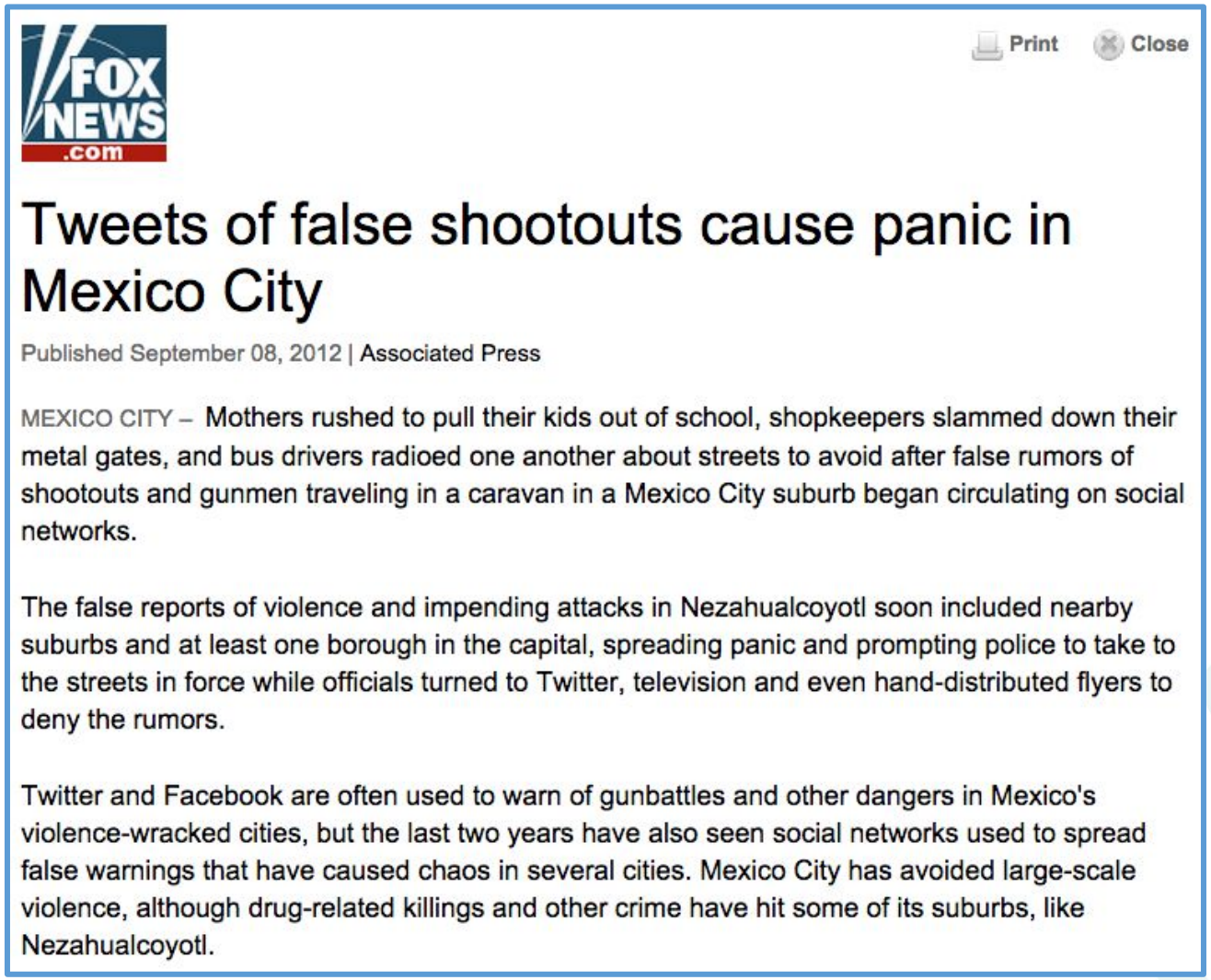
GET UK ALERTS: Enter email SIGN UP

FOLLOW: Facebook, Twitter, Video, Boston, Boston Bombing Fakes, Boston Marathon, Boston Marathon Bombing, Boston Usa, Fake Boston Marathon, Marathons, UK NEWS, UK News

A number of fake charity Twitter accounts and Facebook pages have been set up in the wake of the Boston marathon bombings in an attempt to capitalise on the tragedy.

Pictures of 'child runners' who had supposedly died in the blasts were tweeted from a 'Hope for Boston' account begging for retweets to "show respect".

Misinformation on Social Media



FOX NEWS .com

Print Close

Tweets of false shootouts cause panic in Mexico City

Published September 08, 2012 | Associated Press

MEXICO CITY – Mothers rushed to pull their kids out of school, shopkeepers slammed down their metal gates, and bus drivers radioed one another about streets to avoid after false rumors of shootouts and gunmen traveling in a caravan in a Mexico City suburb began circulating on social networks.

The false reports of violence and impending attacks in Nezahualcoyotl soon included nearby suburbs and at least one borough in the capital, spreading panic and prompting police to take to the streets in force while officials turned to Twitter, television and even hand-distributed flyers to deny the rumors.

Twitter and Facebook are often used to warn of gunbattles and other dangers in Mexico's violence-wracked cities, but the last two years have also seen social networks used to spread false warnings that have caused chaos in several cities. Mexico City has avoided large-scale violence, although drug-related killings and other crime have hit some of its suburbs, like Nezahualcoyotl.

Misinformation Tweets

DC Maryland Virginia
@DMVFollowers
Follow

McDonalds in Virginia Beach flooded.
pic.twitter.com/FZBoCydM
Reply Retweet Favorite

FAKE

AP The Associated Press
@AP
Following

Breaking: Two Explosions in the White House and Barack Obama is injured

Reply Retweet Favorite ••• More

3,063 RETWEETS 144 FAVORITES

12:07 PM - 23 Apr 13

RUMORS

#LondonRiots hearing reports that london zoo was broken into and a large amount of animals have escaped. Too far! Thats not cool :-(

@Twiggy_Garcia, 5,178 followers

Background: Hurricane Sandy

- Dates: Oct 22- 31, 2012
- Damages worth $75 billion
- Coast of NE America

Fake Image Tweets

DC Maryland Virginia
@DMVFollowers
Follow

McDonalds in Virginia Beach flooded.
pic.twitter.com/FZBoCydM
Reply Retweet Favorite

Katina
@kdekranis9
Follow

I TOLD Y'ALL! Shark on the highway in New Jersey!!!!
@maxthewanted would appreciate this. #HurricaneSandy
pic.twitter.com/kaYMjWzT
1:09 AM - 30 Oct 2012

Jamster
@jamster83
Follow

Amazing picture of hurricane #Sandy decending in New York pic.twitter.com/3mMhCbNq
4:21 PM - 29 Oct 2012

5 RETWEETS 586 FAVORITES

Motivation

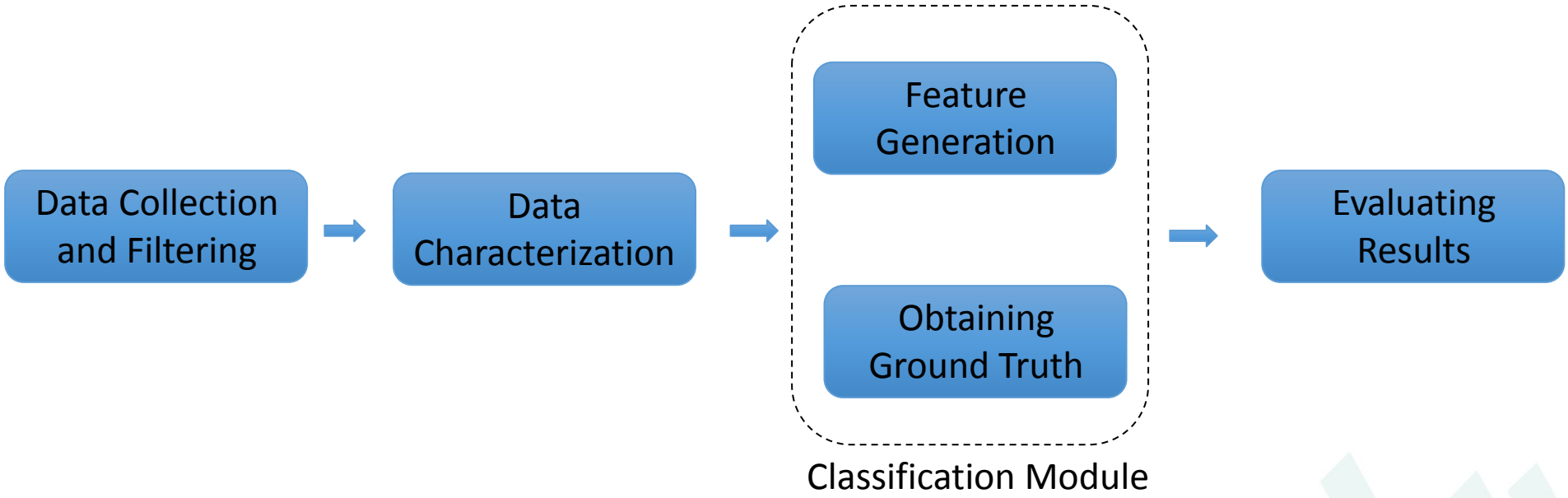
theguardian

USNEWS BLOG

Hurricane Sandy brings storm of fake news and photos to New York

Misinformation over storm spread quickly online, abetted by journalists no longer taught importance of verifying every source

Methodology



Data Description

Total tweets	1,782,526
Total unique users	1,174,266
Tweets with URLs	622,860

Data Filtering

- Reputable online resource to filter fake and real images
  - Guardian collected and publically distributed a list of fake and true images shared during Hurricane Sandy

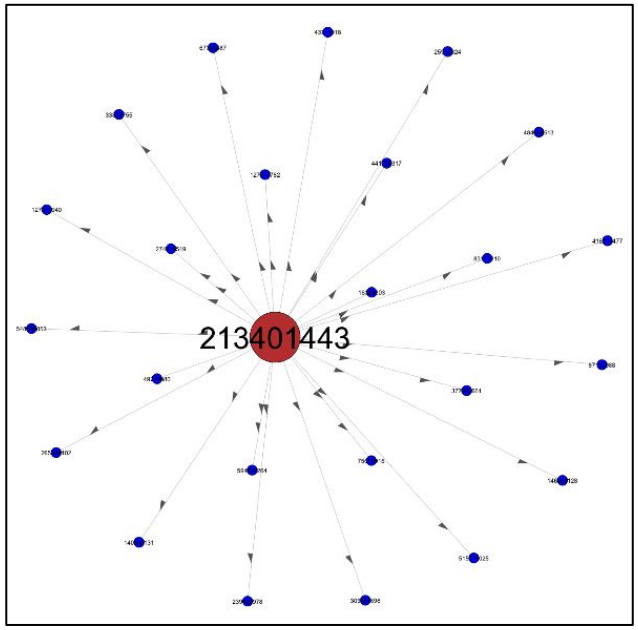
Tweets with fake images	**10,350**
Users with fake images	**10,215**
Tweets with real images	**5,767**
Users with real images	**5,678**

- One of the biggest fake content propagation datasets that have been studied by researchers

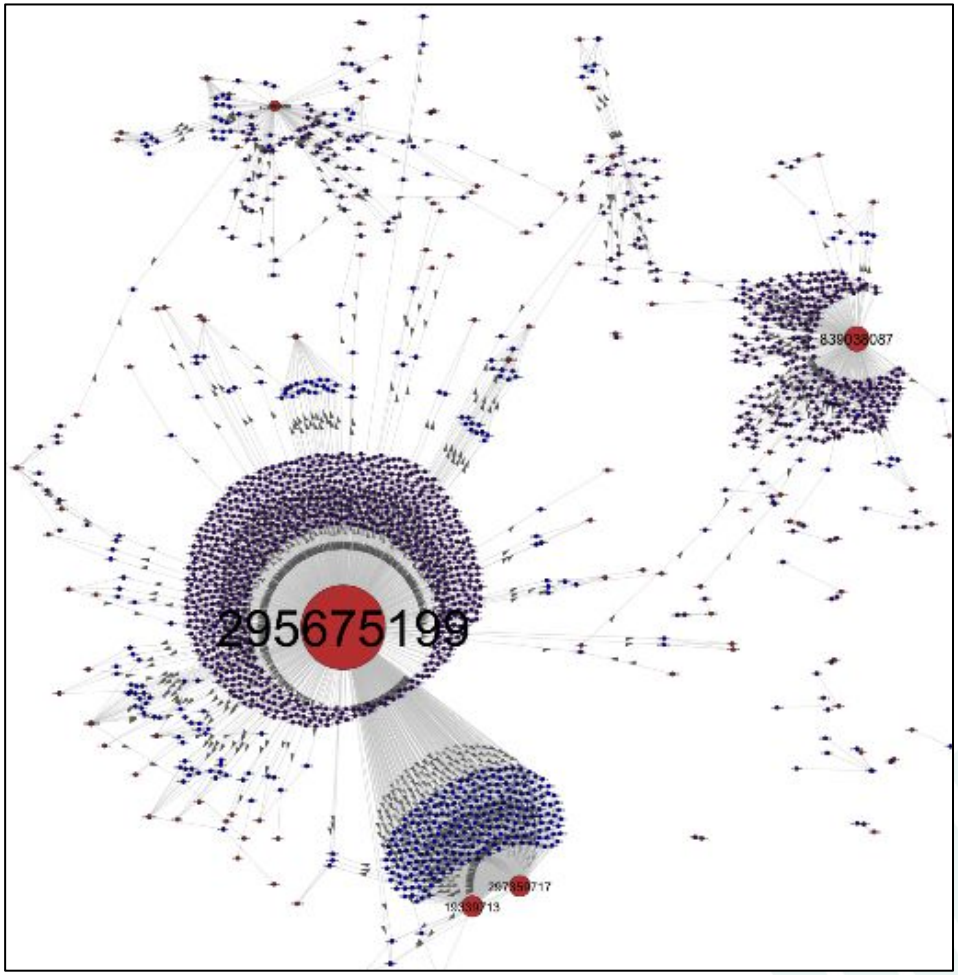
Analysis

- Who
- When
- Where
- What
- Why
- How

Network Analysis



Tweet – Retweet graph for the spread of fake images at ‘nth’ and ‘n+1th‘ hour



Classification

5 fold cross validation

User Features [F1]
Number of Friends
Number of Followers
Follower-Friend Ratio
Number of times listed
User has a URL
User is a verified user
Age of user account

Tweet Features [F2]
Length of Tweet
Number of Words
Contains Question Mark?
Contains Exclamation Mark?
Number of Question Marks
Number of Exclamation Marks
Contains Happy Emoticon
Contains Sad Emoticon
Contains First Order Pronoun
Contains Second Order Pronoun
Contains Third Order Pronoun
Number of uppercase characters
Number of negative sentiment words
Number of positive sentiment words
Number of mentions
Number of hashtags
Number of URLs
Retweet count

Classification Results

	F1 (user)	F2 (tweet)	F1+F2
Naïve Bayes	56.32%	91.97%	91.52%
Decision Tree	53.24%	97.65%	96.65%

- Best results were obtained from Decision Tree classifier. 97% accuracy in predicting fake images from real
- Tweet based features are very effective in distinguishing fake images tweets from real, while the performance of user based features was very poor.

Boston Blasts

- Twin blasts occurred during the Boston Marathon
  - April 15th, 2013 at 18:50 GMT
- 3 people were killed and 264 were injured
- First Image on Twitter (within 4 mins)

Sample Fake Tweets

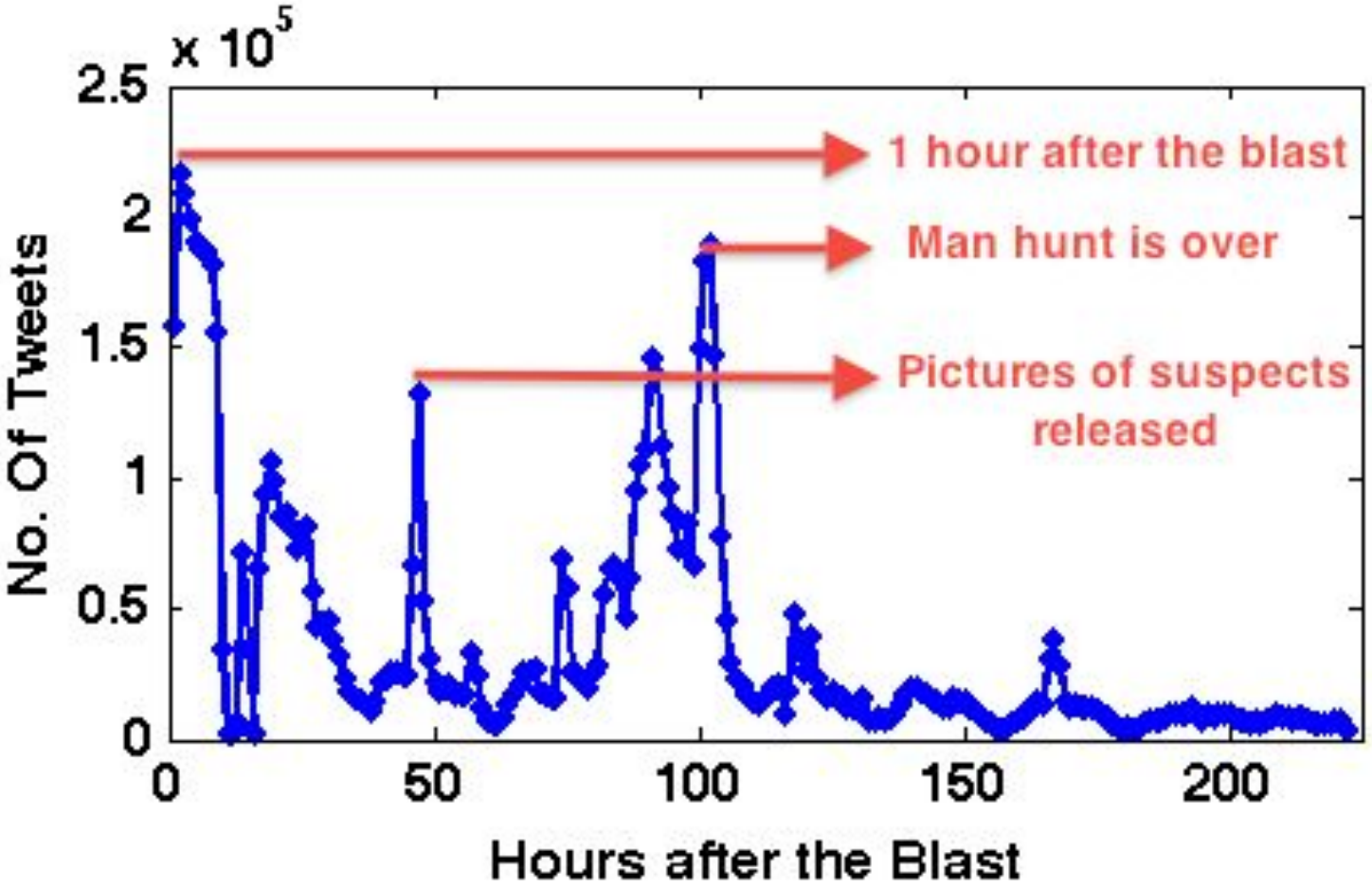
> 30,000 RTs

> 50,000 RTs

Data Description

Total tweets	7,888,374
Total users	3,677,531
Tweets with URLs	3,420,228
Tweets with Geo-tag	62,629
Retweets	4,464,201
Replies	260,627
Time of the blast	Mon Apr 15 18:50 2013
Time of first tweet	Mon Apr 15 18:53 2013
Time of first image	Mon Apr 15 18:54 2013
Time of last tweet	Thu Apr 25 01:23 2013

Data Description



Geo-Located Tweets

Identifying Rumor / True tweets

- Tagged most viral 20 tweet content
  - Rumor / Fake
  - True
  - Generic (NA)

- Six Rumors
  - 130,690 Tweets / Retweets (29%)
  - *R.I.P. to the 8 year-old boy who died in Boston's explosions, while running for the Sandy Hook kids. #prayforboston*

- Seven True news
  - 116,454 Tweets / Retweets (20%)
  - *Doctors: bombs contained pellets, shrapnel and nails that hit victims #BostonMarathon @NBC6*

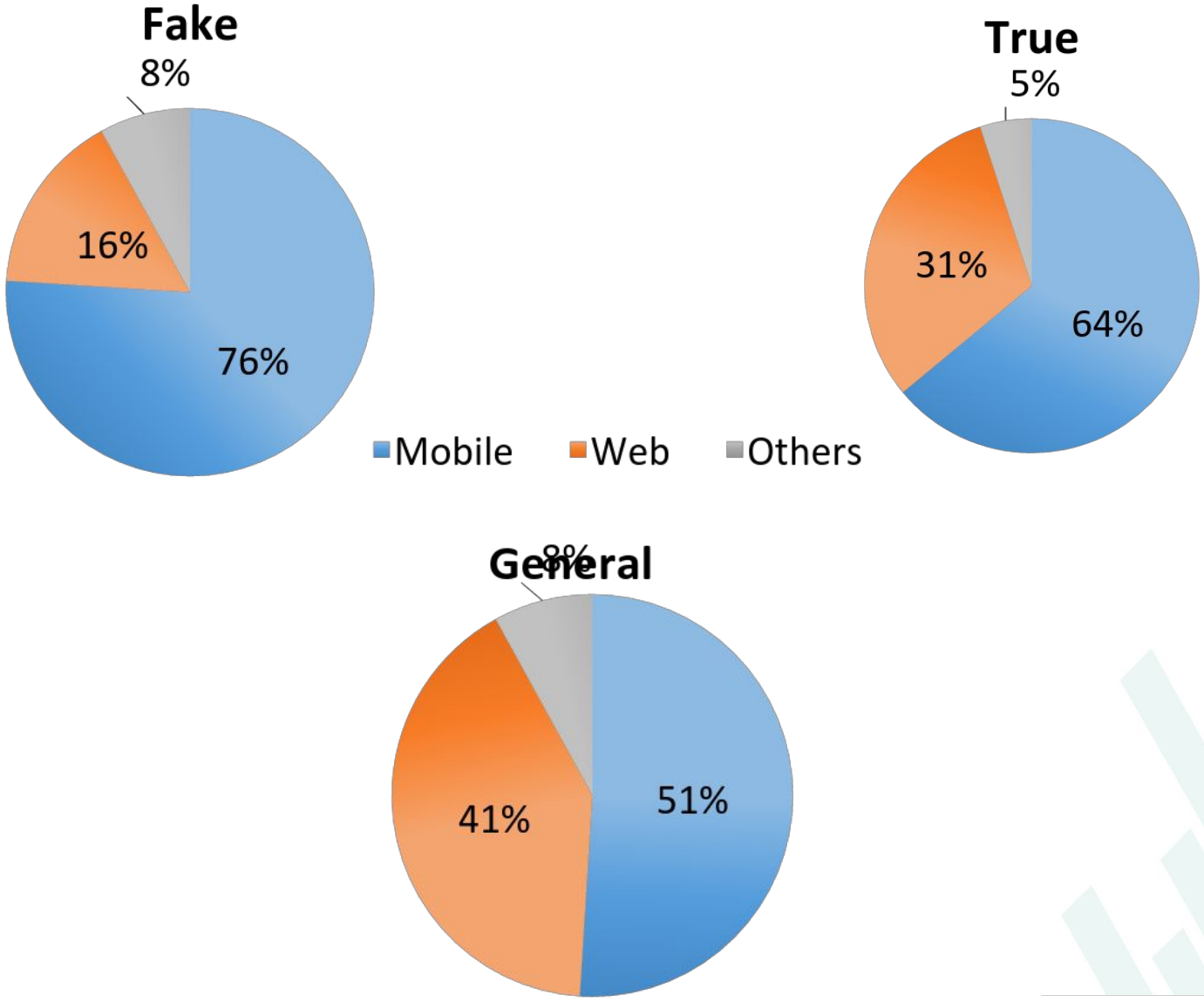
- Seven Generic
  - 206,816 Tweets / Retweets (51%)
  - *#PrayForBoston*

Fake Content User Profiles

	Account 1	Account 2	Account 3	Account 4
No. of Followers	10	297	249	73,657
Profile Creation Date	Mar 24 2013	Apr 15 2013	Feb 07 2013	Dec 04 2008
Total No. of Statuses	2	2	294	7,411
No. of Fake Tweets	2	2	1	1
Current Status	Suspended	Suspended	Suspended	Active

Username: BostonMarathons

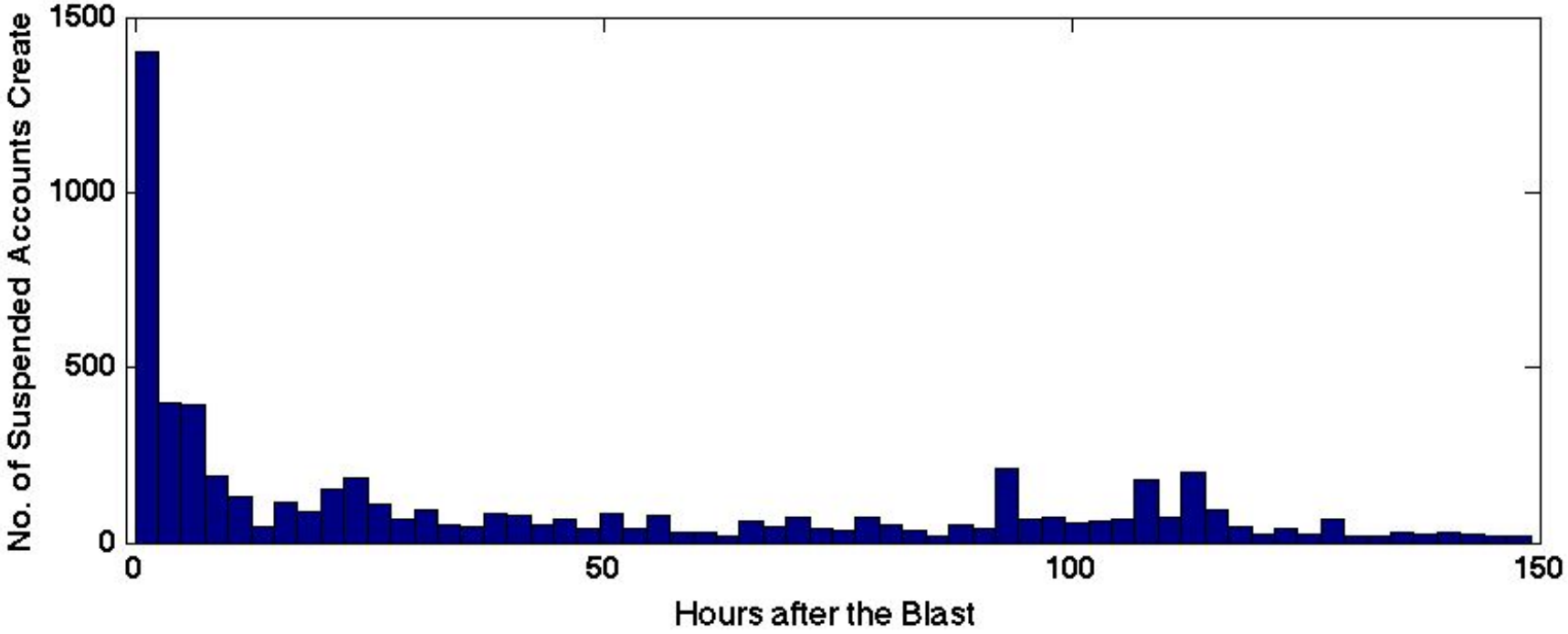
Tweet Source Analysis



Suspended Accounts

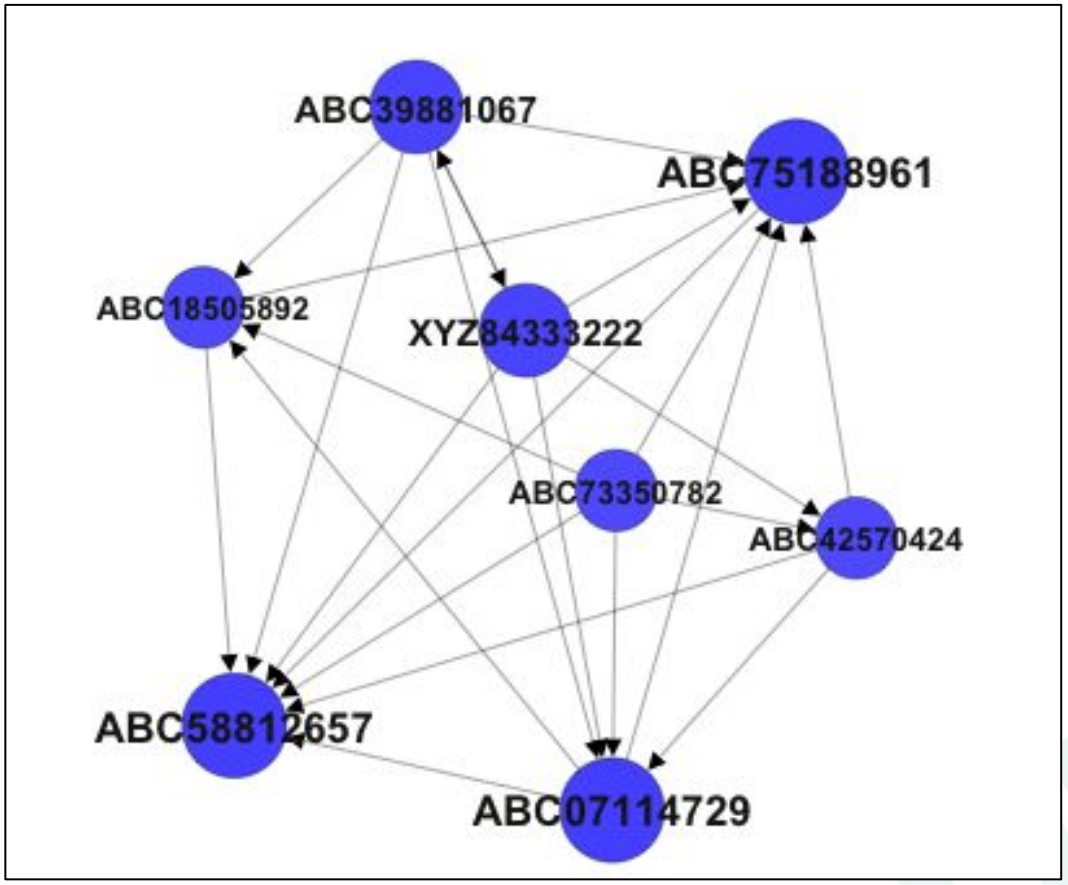
- **31,919** new Twitter accounts created during Boston blasts, that tweeted about the event
- Out of these **19%** [6,073 accounts] were deleted or suspended by Twitter

Fake / Malicious Accounts

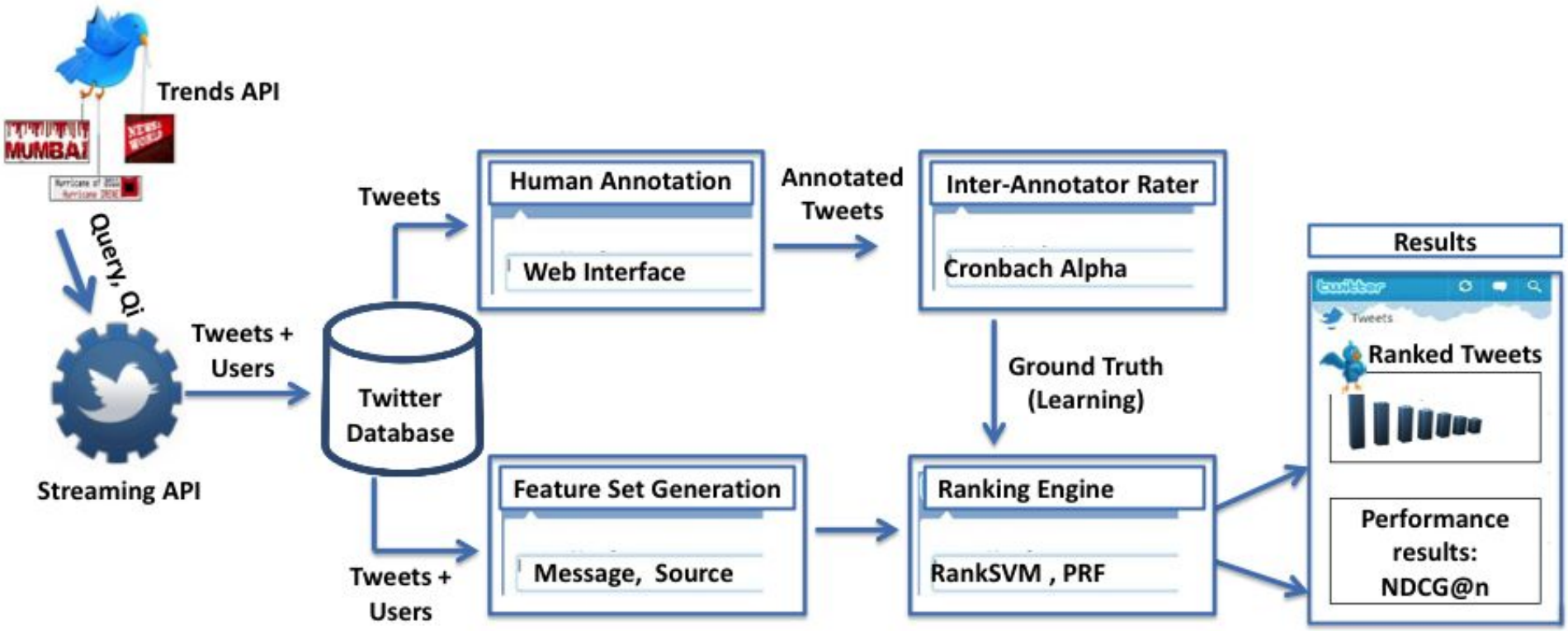


Network Analysis of Fake Accounts

Closed community



Architecture



Data Statistics

Events	Tweets	Trending Topics
UK Riots	542,685	#ukriots, #londonri- ots, #prayforlondon
Libya Crisis	389,506	libya, tripoli
Earthquake in Virginia	277,604	#earthquake, Earth- quake in SF
JanLokPal Bill Agitation	182,692	Anna Hazare, #jan- lokpal, #anna
Apple CEO Steve Jobs resigns	158,816	Steve Jobs, Tim Cook, Apple CEO
US Downgrading	148,047	S&P, AAA to AA
Hurricane Irene	90,237	Hurricane Irene, Tropical Storm Irene
Google acquires Motorola Mobility	68,527	Google, Motorola Mobility
News of the World Scandal	67,602	Rupert Murdoch, #murdoch
Abercrombie & Fitch stocks drop	54,763	Abercrombie & Fitch, A&F
Muppets Bert and Ernie were gay	52,401	Bert and Ernie
Indiana State Fair Tragedy	49,924	Indiana State Fair
Mumbai Blast, 2011	32,156	#mumbaiblast, Dadar, #needhelp
New Facebook Messenger	28,206	Facebook Messenger

Annotation

- Step 1
  - R1. Contains information about the event
  - R2. Is related to the event, but contains no information
  - R3. Not related to the event
  - R4. Skip tweet
- Step 2
  - C1. Definitely credible
  - C2. Seems credible
  - C3. Definitely incredible
  - C4. Skip tweet.

Annotation Results

- Each tweet annotated by 3 people
- Inter-annotator agreement (Cronbach Alpha) = 0.748
- 30% of tweets provide information (17% credible information) and 14% was spam

Feature Sets

Message based features
Length of the tweet
Number of words
Number of unique characters
Number of hashtags
Number of retweets
Number of swear language words
Number of positive sentiment words
Number of negative sentiment words
Tweet is a retweet
Number of special symbols [$, !]
Number of emoticons [:-), :-(]
Tweet is a reply
Number of @- mentions
Number of retweets
Time lapse since the query
Has URL
Number of URLs
Use of URL shortener service
Message based features
Length of the tweet
Number of words

Source based features
Registration age of the user
Number of statuses
Number of followers
Number of friends
Is a verified account
Length of description
Length of screen name
Has URL
Ratio of followers to followees
Source based features
Registration age of the user
Number of statuses
Number of followers

Evaluation Metric

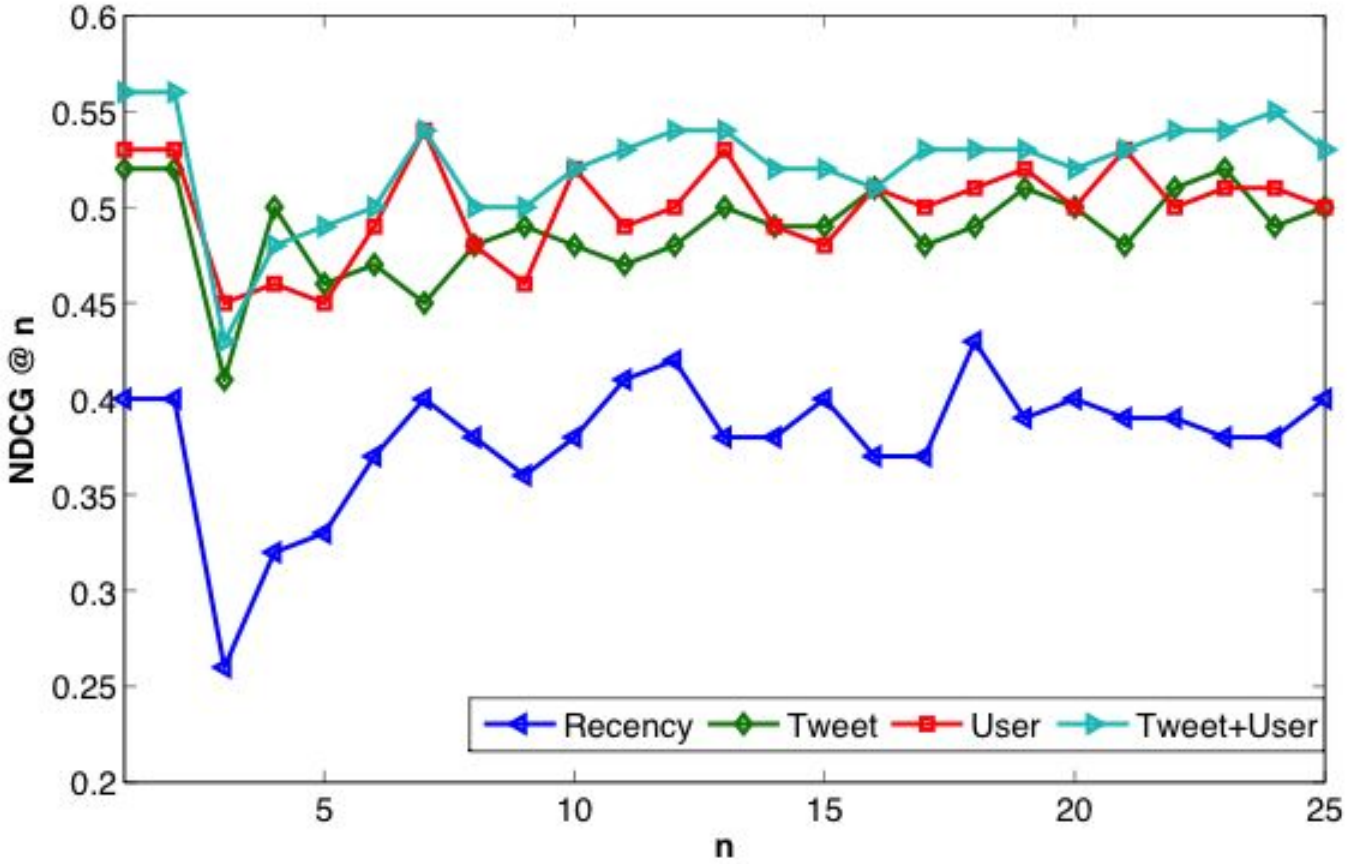
Evaluation Metric: NDCG (Normalized Discounted Cumulative Gain)

$$DCG@n = \Sigma_{i=1}^{n} \frac{1}{log_2(1+i)} (2^{label(v_i)} - 1)$$

NDCG is the standard metric used to evaluate “graded” results

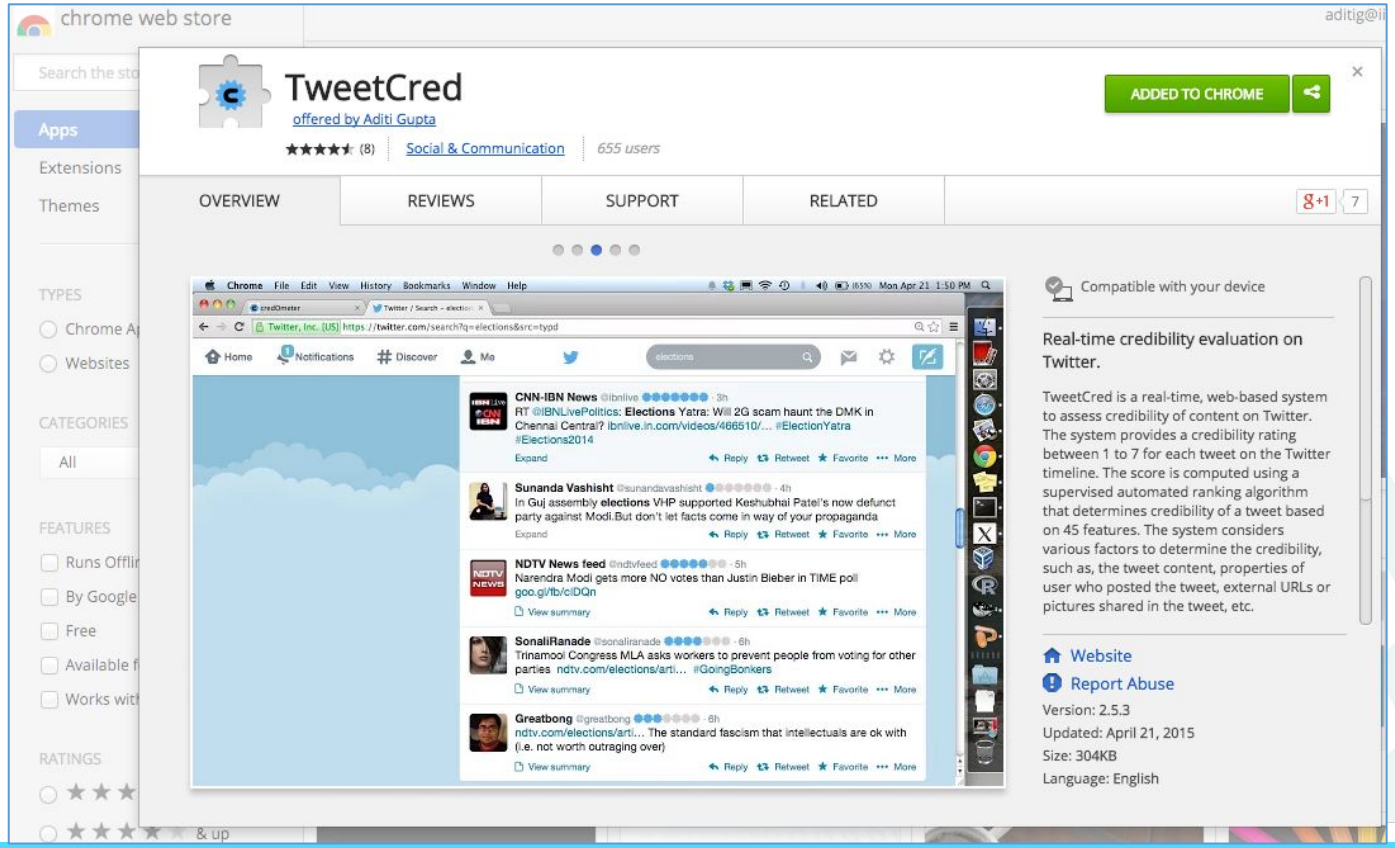
Ranking Results

- Tweet and user based features contribute in determining the credibility – it matters "*what you post and who you are*"



TweetCred

- Available as a Chrome Extension



Live Demo of TweetCred

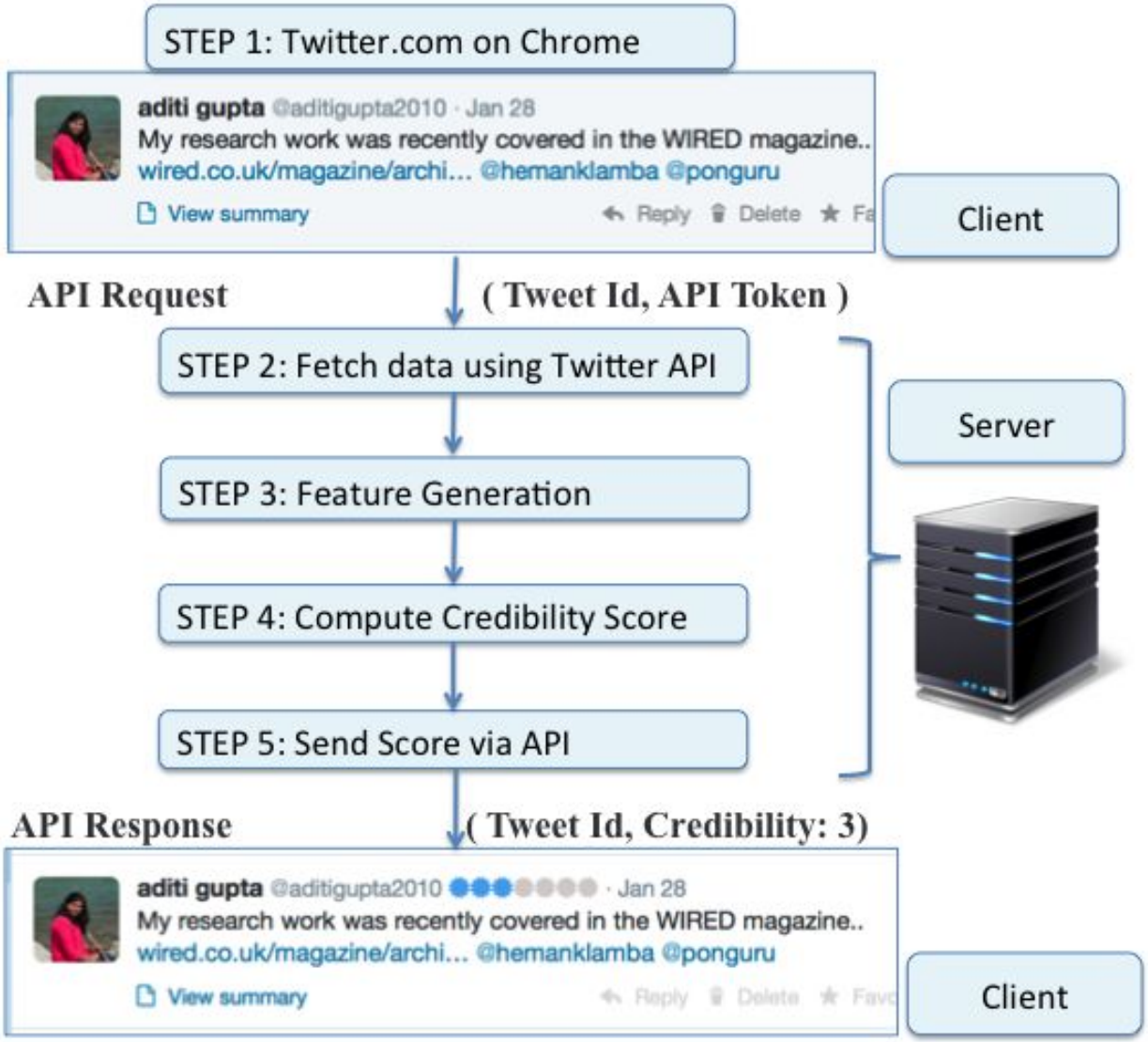
Features for Real-time Analysis

Feature set	Features (45)
Tweet meta-data	Number of seconds since the tweet; Source of tweet (mobile / web/ etc); Tweet contains geo-coordinates
Tweet content (simple)	Number of characters; Number of words; Number of URLs; Number of hashtags; Number of unique characters; Presence of stock symbol; Presence of happy smiley; Presence of sad smiley; Tweet contains `via'; Presence of colon symbol
Tweet content (linguistic)	Presence of swear words; Presence of negative emotion words; Presence of positive emotion words; Presence of pronouns; Mention of self words in tweet (I; my; mine)
Tweet author	Number of followers; friends; time since the user if on Twitter; etc.
Tweet network	Number of retweets; Number of mentions; Tweet is a reply; Tweet is a retweet
Tweet links	WOT score for the URL; Ratio of likes / dislikes for a YouTube video

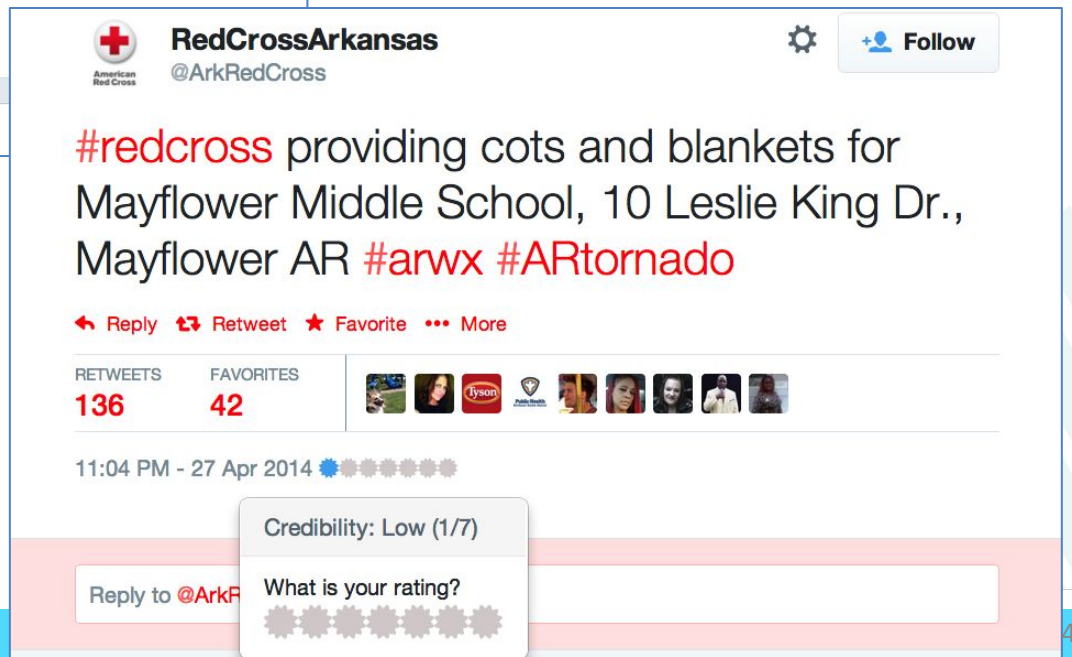
Top Ten Features

- No. of characters in tweet
- Unique characters in tweet
- No. of words in tweet
- User has location in profile
- Number of retweets
- Age of tweet
- Tweet contains URL
- Tweet contains via
- Statuses / Followers
- Friends / Followers

Implementation



Feedback by Users



Users of TweetCred

Sample users:

- Emergency responders
- Firefighters
- Journalists / news media
- General users

Quick summary for Week 2

- Frameworks / Platforms
  - APIs – Twitter & ~~Facebook~~ Reddit
  - Python
  - MySQL / MongoDB
  - PhpMyAdmin
- Rate limits
- JSON
- Graphs
- Credibility
- Data collection for an event
- Who, When, Where, What, Why, and How
- Network analysis

Takeaways / Questions?

Thank you

pk@iiitd.ac.in
precog.iiitd.edu.in
fb/ponnurangam.kumaraguru