Thanks to all my collaborators.

Machine Learning and Applications on Social Media Data

Janani Kalyanam

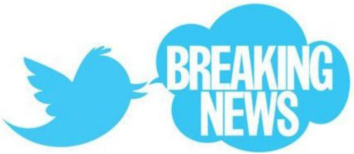
May 19, 2017

Background

- Advances in digital communication (smart phones);
- Shift from informational web (web 1.0) to interactional web (web 2.0)
- Statistics
  - ➢ 78% of all Americans have some form of social media presence
  - ➢ Teenagers spend 6 to 8 hours per day on social media
  - ➢ On Twitter, more than 6000 messages are published every second

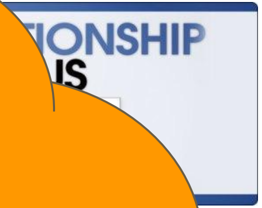
Background

Changed the way we operate as a society.



- news
- formation of support groups aı communities
- “social” aspect to everything
- societal norms

Background



Changed the

society.

- electronic snapshot of life
- use to answer important social questions

● r

● fo

co

● “so

● societal nor

Unique Challenges

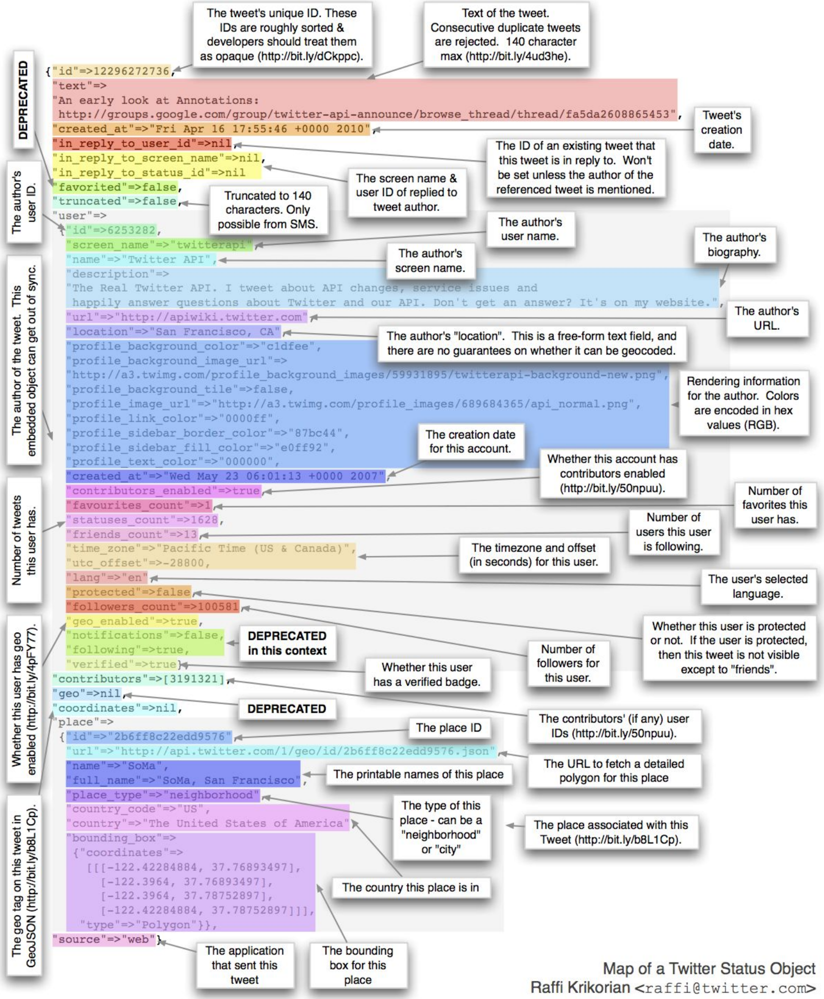
Background

- Content generated by a large demographic
  - GOOD: rich data

Background: Unstructured Multimodal Content

Lots of metadata

- Tweet text, its timestamp
- Whether it’s a “reply”
- Geolocation of the tweet
- Information about user
- Whether it’s a retweet
    - Information about original tweet
    - Information about original user



Map of a Twitter Status Object
Raffi Krikorian <raffi@twitter.com>
18 April 2010

Background

- Content generated by a large demographic
  - ➢ GOOD: rich data
  - ➢ BAD: missing information

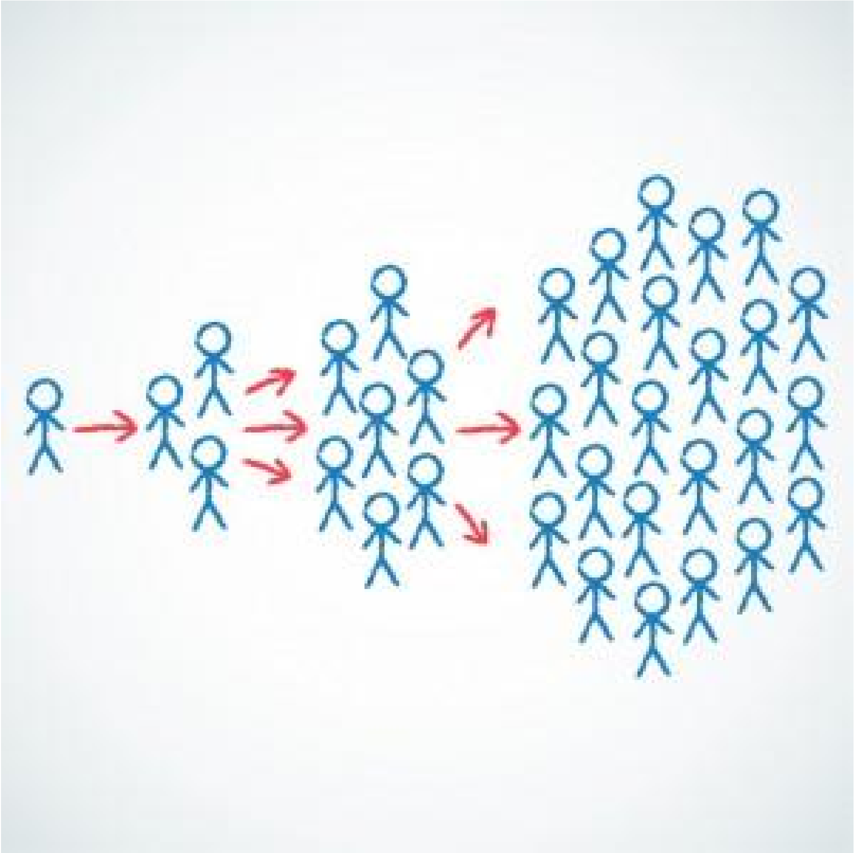
Background

- Content generated by a large demographic
    - GOOD: rich data
    - BAD: missing information
- Uncurated, uncensored, unedited
    - GOOD: no bias
    - BAD: increases the “unstructured” nature of the data

Background

- Content generated by a large demographic
  - GOOD: rich data
  - BAD: missing information
- Uncurated, uncensored, unedited
  - GOOD: no bias
  - BAD: increases the “unstructured” nature of the data
- Propagation

Background: Propagation



- Can we study the impact of events through the intensity of reactions it creates on social media?
- Can we predict this impact early enough in the event life cycle?

Background

- Content gener
  - GOOD: rich
  - BAD: m
- Uncurated
  - GOOD: n
  - BAD: inc
- Propagation

Apophenia

Contributions

design efficient and robust computational methods to analyze the data generated from the collective online footprints and the digitized archival of human communication and provide answers to some important questions and help improve quality of life.

Contributions

design efficient and robust computational methods to analyze the data generated from the collective online footprints and the digitized archival of human communication and provide answers to some important questions and help improve quality of life.

1. Methods to effectively use metadata
2. Analyze social media reactions to events
3. Infoveillence

Contributions

design efficient and robust computational methods to analyze the data generated from the collective online footprints and the digitized archival of human communication and provide answers to some important questions and help improve quality of life.

1. Methods to effectively use metadata
2. Analyze social media reactions to events
3. Infoveillence

Part 2

Studying events through the lens of social media reactions

Motivation

Event

Reaction

Motivation

For every action, there is an equal and opposite reaction, plus a social media overreaction.

Event

Reaction

Studying Events through Social Media Reactions

- Characterize events through the reactions it creates on social media
- How to quantify the impact of an event?
- Soon after outbreak, can early signals predict impact?

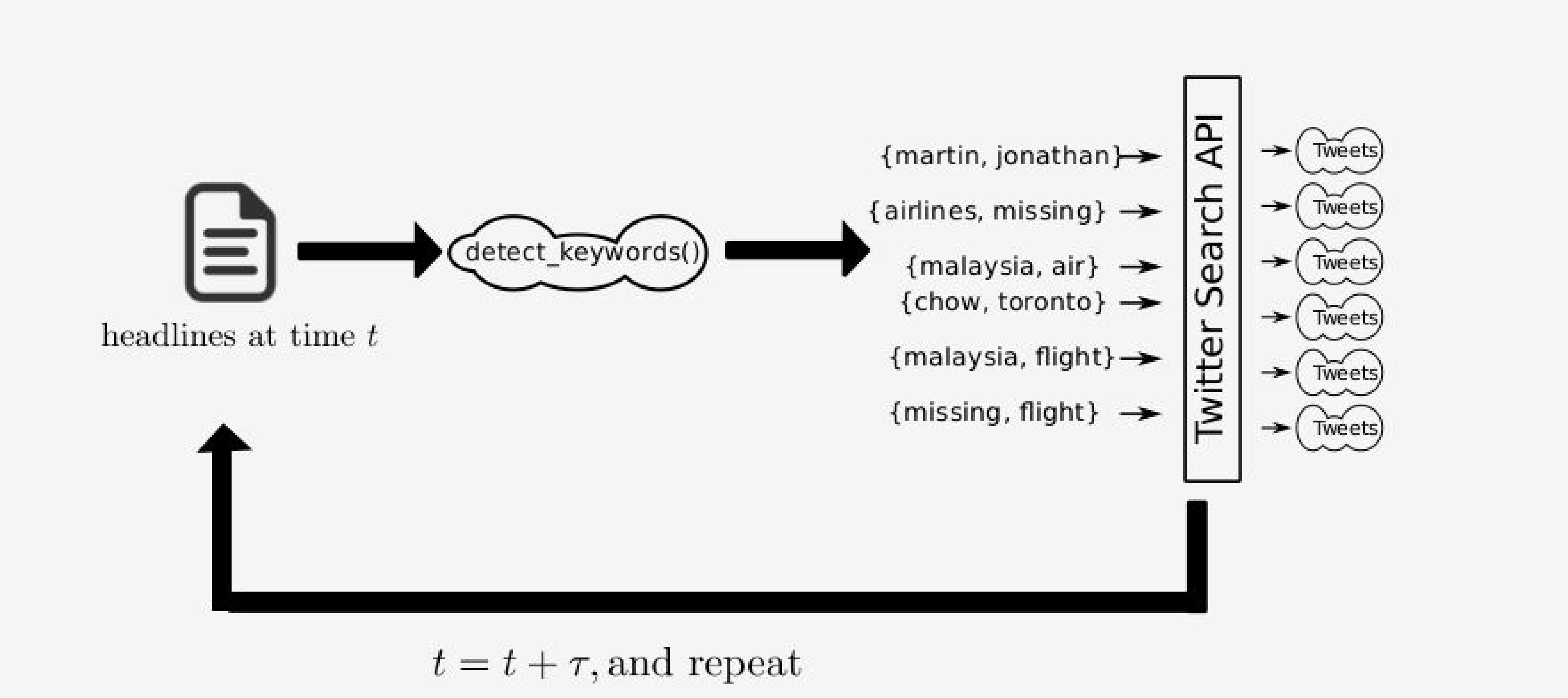
Background

- Identify social media messages about an event
- Need to obtain all the early posts about the event

The data collection methodology should, in real time, collect social media posts about events.

Data Collection

Data Collection



Data Collection

- extract common keywords across headlines
- form itemsets
- pick top-2 keywords from each itemset, and search Twitter API

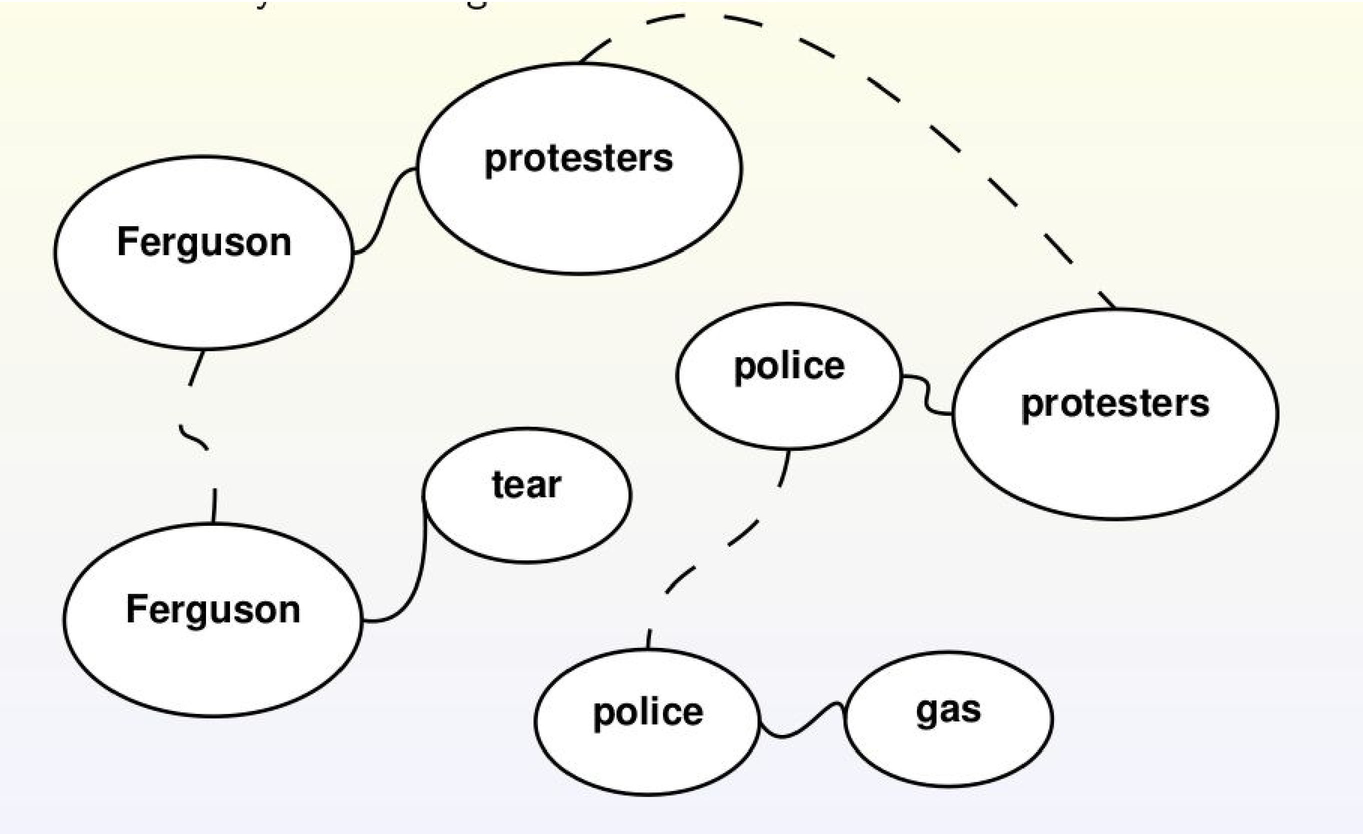
Why some **protesters** in **Ferguson** have been forced to choose between speaking out or keeping their jobs

Things in **Ferguson** have gotten so unruly that the National Guard has been called in #MikeBrown

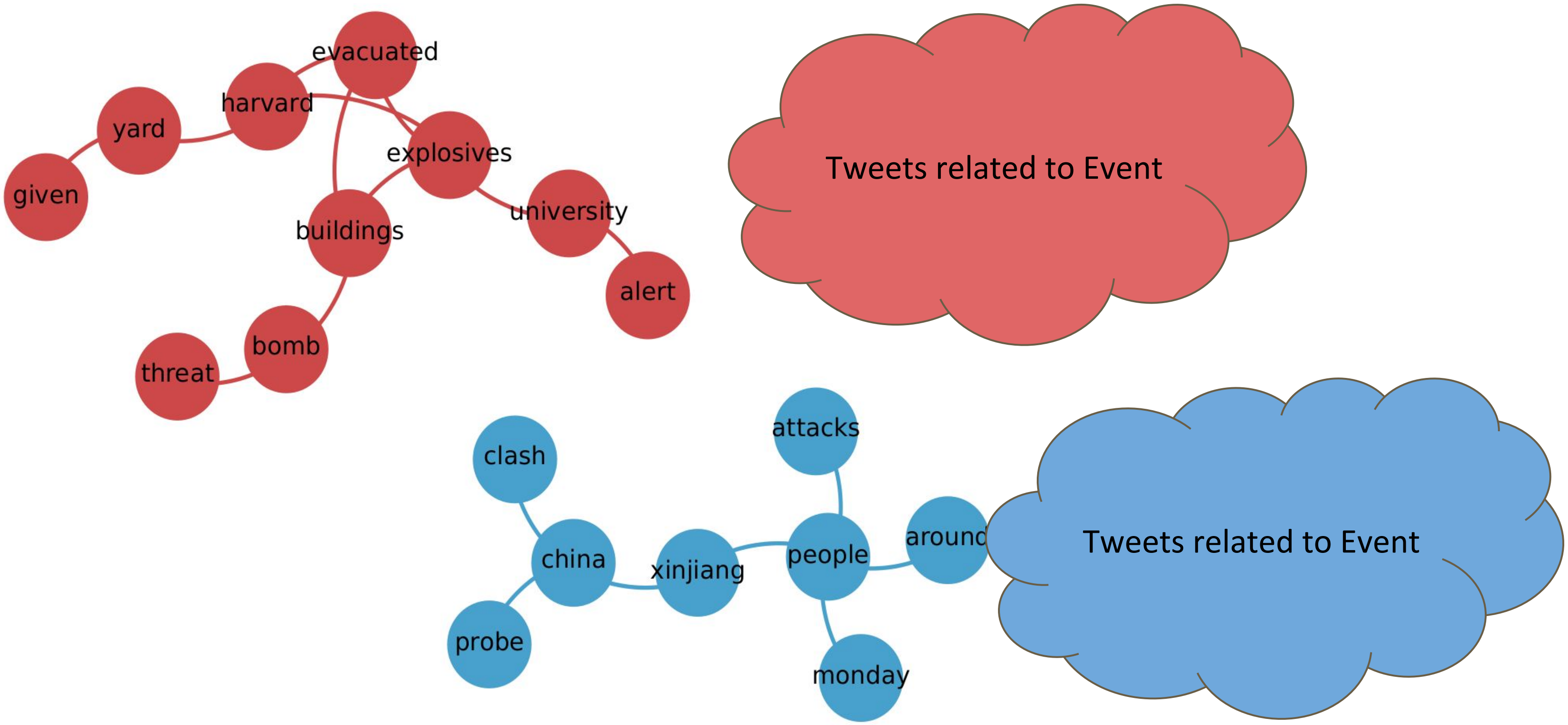
Police launch **tear gas**, flash grenades at **protesters** in #**Ferguson**

Tear gas is illegal in war under treaties signed by the U.S. Yet, the US uses it against its own people in #**Ferguson**

Example Event Components



Events


evacuated
harvard
yard
given
explosives
buildings
university
alert
bomb
threat
Tweets related to Event
attacks
clash
around
china
xinjiang
people
Tweets related to Event
probe
monday

Impact of Events

Quantify the impact an event has through social media reactions

- Number of tweets
  - it is subject to normalization biases. Size alone is not always important.
  - Does not encompass impactful, but local events
  - Example: recent shooting in La Jolla.
- Average tweet arrival rate
  - Does not depend on the size (good thing)
  - However, single number can be too restrictive.

Impact of Events

Quantify the impact an event has through social media reactions

- Number of tweets
  - it is subject to normalization biases. Size alone is not always important.
  - Does not encompass impactful, but local events
  - Example: recent shooting in La Jolla.
- Average tweet arrival rate
  - Does not depend on the size (good thing)
  - However, single number can be too restrictive.

Impact of Events

Quantify th impact an even s throu cial media reactions

- Number of t

  t.

  - Exa

- Av

  - Does not d c

  - However gle num be too ve.

Want a rich descriptor quantifying the "buzz" or "chatter" surrounding an event

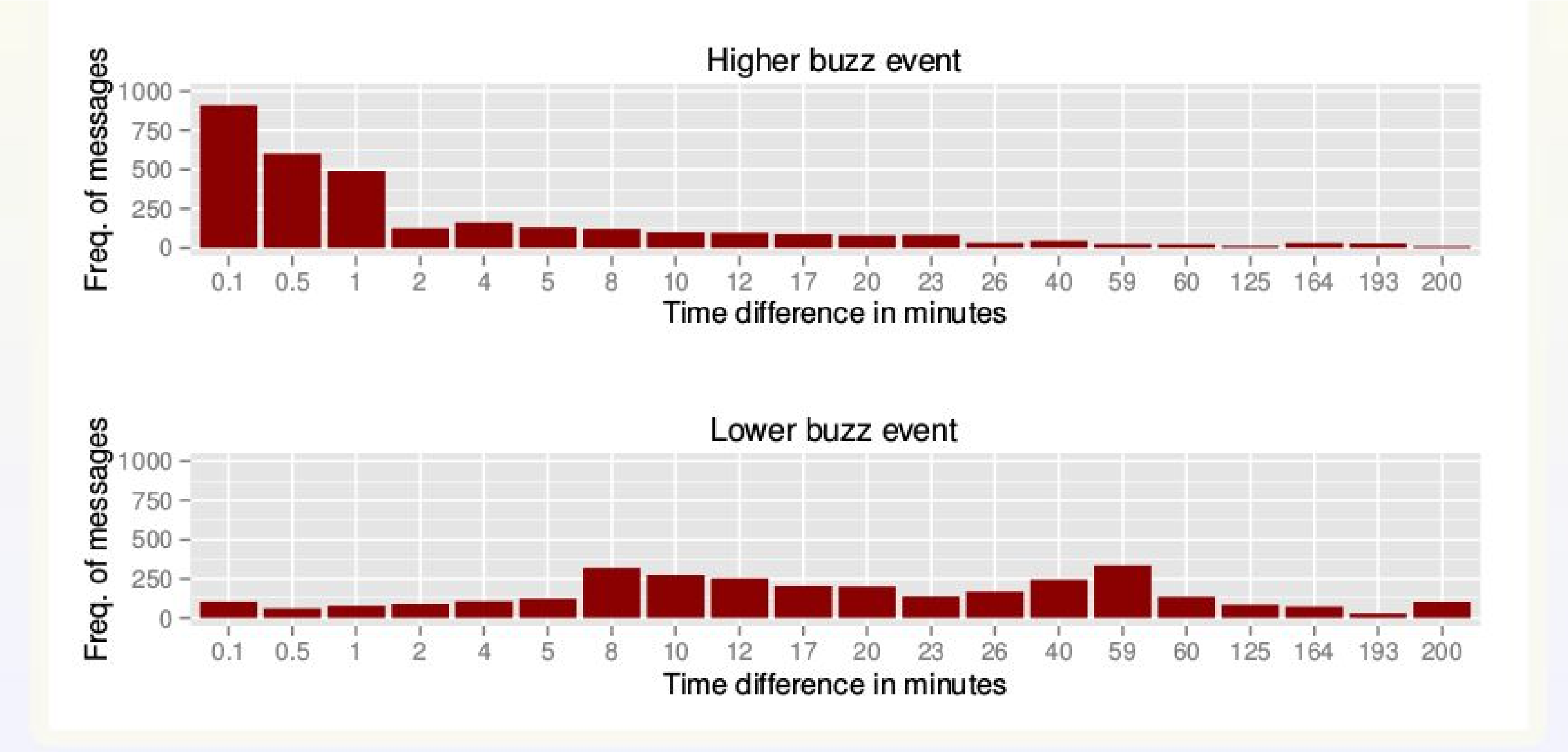
Impact of Events

- interarrival times between consecutive social media messages within an event
- Time in between 2 messages $d_i = t_{i+1} - t_i$
- for a given event, how are the differences distributed?
  - Event with most {di}s very very small
  - Event with most {di}s quite large

VQ Event Model for Measuring Impact

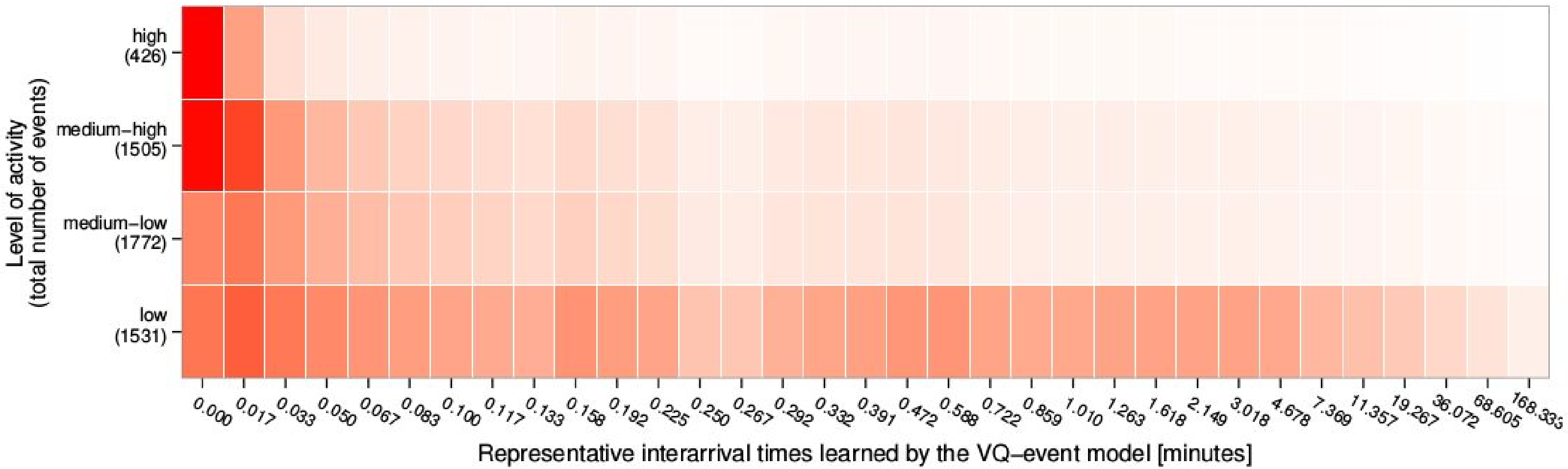
- Learn some “representative inter arrival times” from a large corpus of events (clustering)
- Vector quantization of the inter arrival times
- For each event, quantize the inter arrival times to the closest “representative”
- Thus construct a histogram.

VQ Event Model for Measuring Impact



VQ Event Model

- Cluster events



Results

- 82 world wide news sources
- 5,234 events and 43 million tweets over 5 months (October 2013 to Feb 2013)

Examples

Event	Sample Tweets
Description: Death of South African politician Nelson Mandela. **Keywords:** [nelson, mandela] **Date:** 2013-12-05 **Size:** 134,637 tweets	@DaniellePeazer: RIP Nelson Mandela..... what a truly phenomenal and inspirational man xx @iansomerhalder: Im in tears.The world has lost one of its greatest shepherds of peace. Thank you Mr.Mandela for the love you radiated. http://t.co/u39MVVEKe8 @FootballFunnys: This is so true. RIP Nelson Mandela. http://t.co/vF9xri8LdP @David_Cameron: I've spoken to the Speaker and there will be statements and tributes to Nelson Mandela in the House on Monday.
Description: 2013 Mumbai Gang Rape **Keywords:** [rape, mumbai] **Date:** 2013-08-24 **Size:** 1,705 tweets	@TheNewsRoundup: Mumbai gang-rape: Second accused confesses to crime: Mumbai Police - Daily News Analysis http://t.co/KnabwhqH66 @vijayarumugam: An interesting take on the Mumbai rape: http://t.co/ylBmW4l8sA @LondonStephanie: Two arrested over gang rape of Mumbai photojournalist that sparked renewed protests in India http://t.co/McYfLNDvaE @GanapathyI: Most brutal rapist of Delhi gang-rape was 17. Most brutal rapist of Mumbai gang-rape is 18. Worst Young generation I have seen in my life.

Examples

Event	Sample Tweets
Description: Teen survives hiding in a plane wheel. **Keywords:** [teen, survives, old, well, skydivers, plane, wheel, flight] **Date:** 2014-04-21 **Size:** 18,519	@ToniWoemmel: 16-year-old somehow survives flight from California to Hawaii stowed away in planes wheel well: http://t.co/IGiJa60SiK @iOver_think: 38,000 feet at -80F: Teen stowaway survives five-hour California-to-Hawaii flight in wheel well http://t.co/ejXQH9VZyT @TruEntModels: GOD IS GOOD...runaway TEEN hid in plane's wheel for 5 HOUR flight during FREEZING temps and survived http://t.co/6g6Cqhs9Ib @DvdVill: A 16-year-old kid, who was mad at his parents, hid inside a jet wheel and survived flight to Hawaii. http://t.co/c82GbjrfUH
Description: Surveying the damages of recent tornado in Canada. **Keywords:** [canada, tornado] **Date:** 2014-06-21 **Size:** 1,033	@Kathleen_Wynne: Visited #Angus today to survey the damage. Thankfully no fatalities or major injuries from recent tornado. http://t.co/xRQyRWg5Vw @SunNewsNetwork: PHOTOS & VIDEO: Hundreds displaced after tornado hits Ontario town, destroying homes http://t.co/L38rG6N1a6 @CBCToronto: Kathleen Wynne is speaking at site of tornado damage in Angus, Ont. now. Watch live here: http://t.co/EDKNUiZo0X #cbcto @InsuranceBureau: @CTVBarrieNews: Insurance Bureau of Canada is setting up a mobile unit in #Angus today to help residents affected by #Tornado

High Impact vs Low Impact

- # retweets 2.4 times higher
- Most-retweeted tweet ⟶ propagates 7 times more in the network
- Number of tweets which are retweeted is lesser
- Initial messages get retweeted very quickly, extensively through forwarding.

High Impact vs Low Impact

- Sparks more conversation, 33.3% more
- Number of unique users who engage with posts is also 33% more

High Impact vs Low Impact

- Diversity in hashtags is 7 times more for low impact events
- High impact events seldom loses focus

Early Prediction

- Early 5% of tweets
- Average time of 1.5 minutes

FP-Rate	0.009
Precision	0.819
Recall	0.555
ROC-Area	0.900

Part 3

Infoveillence

Infoveillence

- Wide demographic on social media
- Surrogates for qualitative and quantitative research designs such as surveys, in-depth interviews.
    - Surveys take time to compile
    - Survey bias
- Useful to identify
    - Macro level trends
    - Micro level behavior

Trends of Drug Abuse

- Opioid abuse is a grave threat to national public health
- Record number of deaths from drug overdose in 2014
- 4 fold increase in opioid related mortality since 1999
- Heroin abuse along with NMUPD

Trends of Drug Abuse

- Complementing surveys based instruments, can also use surveillance on social media
- Twitter Streaming API to download tweets for three drugs “oxycodone”, “oxycontin” and “percocet”
- AIM: focus on individual user tweets about abuse of NMUPD behavior.

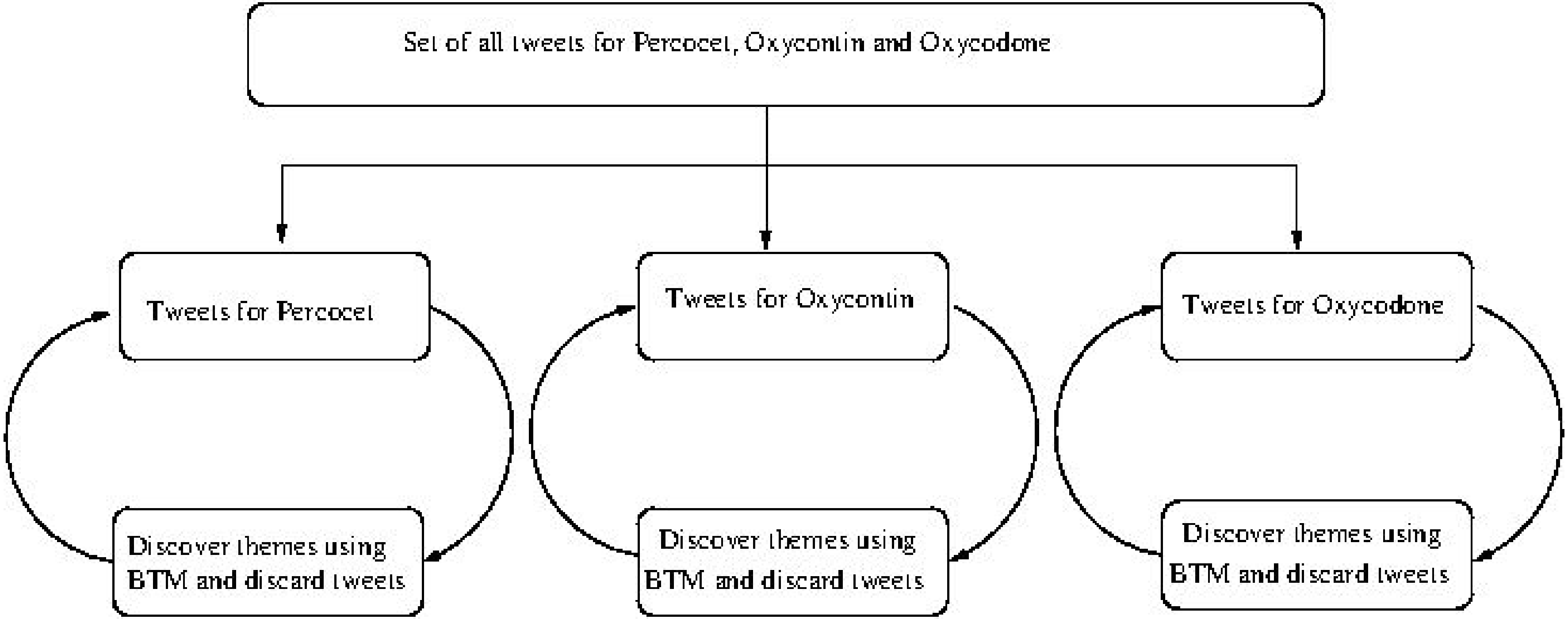
Trends of Drug Abuse

Initial glance at the data: predominantly contained news. But we need risk behavior related tweets.

Inclusion/Exclusion criteria

- Contain INN or slang names identified by NIDA
- Mention other illicit drugs (heroin, cocaine, marijuana)
- Mentions substance abuse risk behaviors (overdose, injection, withdrawal)
- Contains adjectives related to substance abuse behavior (popping, high)

Overview of Methodology



Results

Total of 11M tweets were collected.

Drug	%-of-tweets retained between round 1 and round 2	%-of-tweets retained between round 2 and round 3
Percocet	24%	84%
Oxycontin	36%	72%
Oxycodone	29%	74%

We can infer that the signal to noise ratio is much better between rounds 2 and 3, suggesting that we have eliminated a lot of the noise.

Results

Some examples of topics which were included and excluded

EXCLUDE	INCLUDE
Percocet, super, high, best, buy, online, place, offer compare, quality	Percocet, liquor, pour, dose, weed
Oxycodone, drug, approval, fda, media, reports	Oxycontin, bottle, cocaine, drug, love, wrong

Results

“i got xanex percocet promethazine with codeine”

“i fell in love with a trap mami she be snortin cocaine and molly sometimes she be poppin oxycontin blue pill she be smokin them roxis”

“my mom is ritalin my dad is oxycontin”

“i need the zans and oxycontin christ every 2 hours”

“i sure wish i had a few beers and maybe an oxycodone to make this afterglow even more pleasurable”

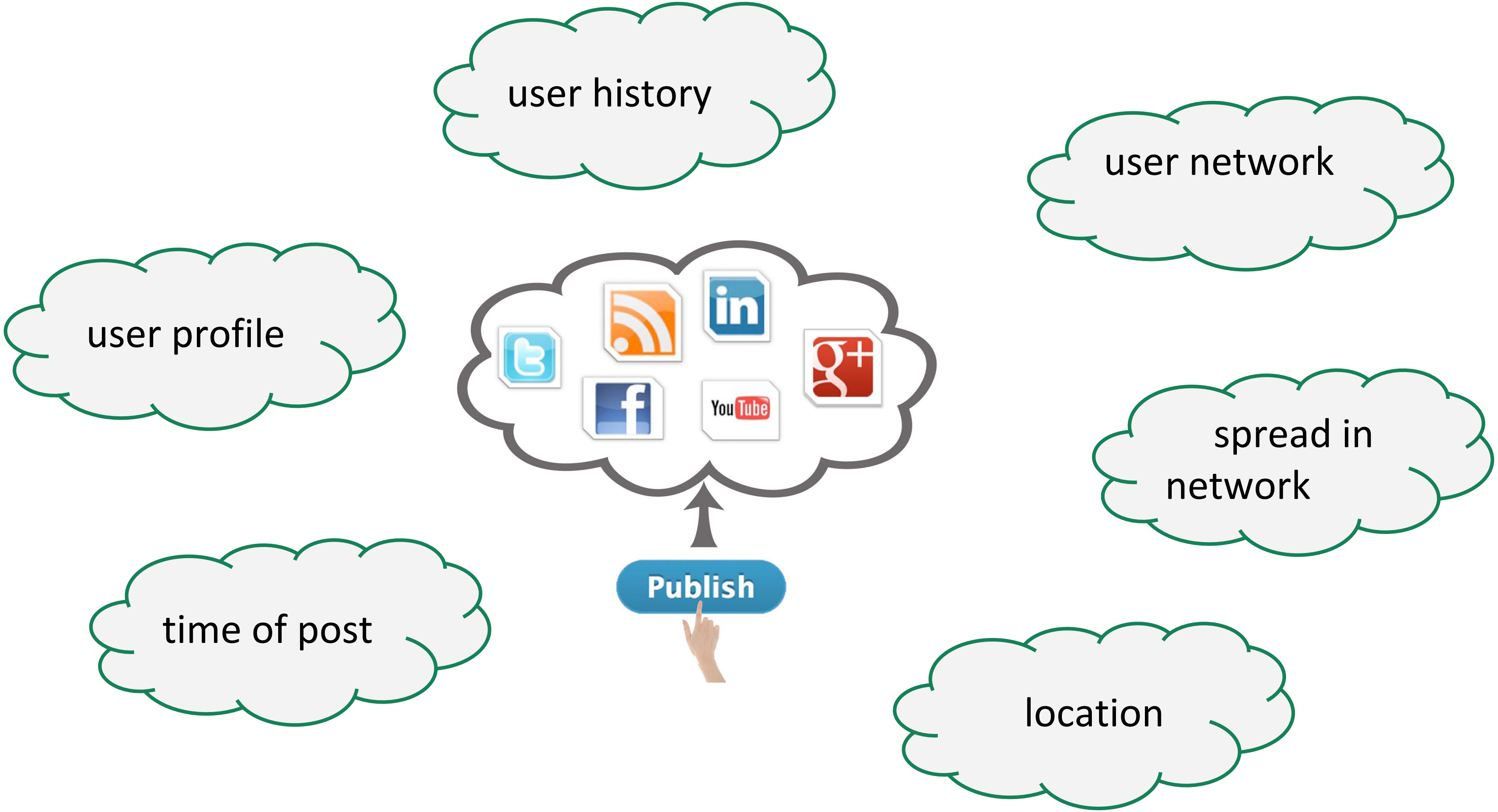
“w00t oxycodone and morphine i feel like lindsay lohan”

Conclusion

- Methodology to identify individual user tweets
- Central trend of drug abuse Polydrug abuse
    - Informed caregivers can provide appropriate advice about particularly dangerous combinations of drugs
- Presence of illicit online pharmacies
    - Paper under submission

Part 1

Framework to Combine Metadata


user history
user network
user profile
spread in network
time of post
location
Publish

Identify Underlying Topics

- Non-negative Matrix Factorization
- Probabilistic Topic Models (like Latent Dirichlet Allocation)

focus on content

Some topics are difficult..

Bird flu outbreak; everything you need to know about protocols goo.gl/F1dnfk #birdflu

Selina Gomez and Justin Bieber: “just friends” goo.gl/M9dlhj #celebritygossip

U.S poultry devastated by birdflu outbreak goo.gl/1gX8FC #birdflu

Kim Kardashian: pregnant again! goo.gl/Ir1knd #celebritygossip

USDA to review protocols following birdflu outbreak goo.gl/X88iSe #birdflu

Lindsay Lohan messed up contract with Oprah goo.gl/Ir1knd #celebritygossip

Metadata on Twitter

- Each tweet comes with numerous side-information

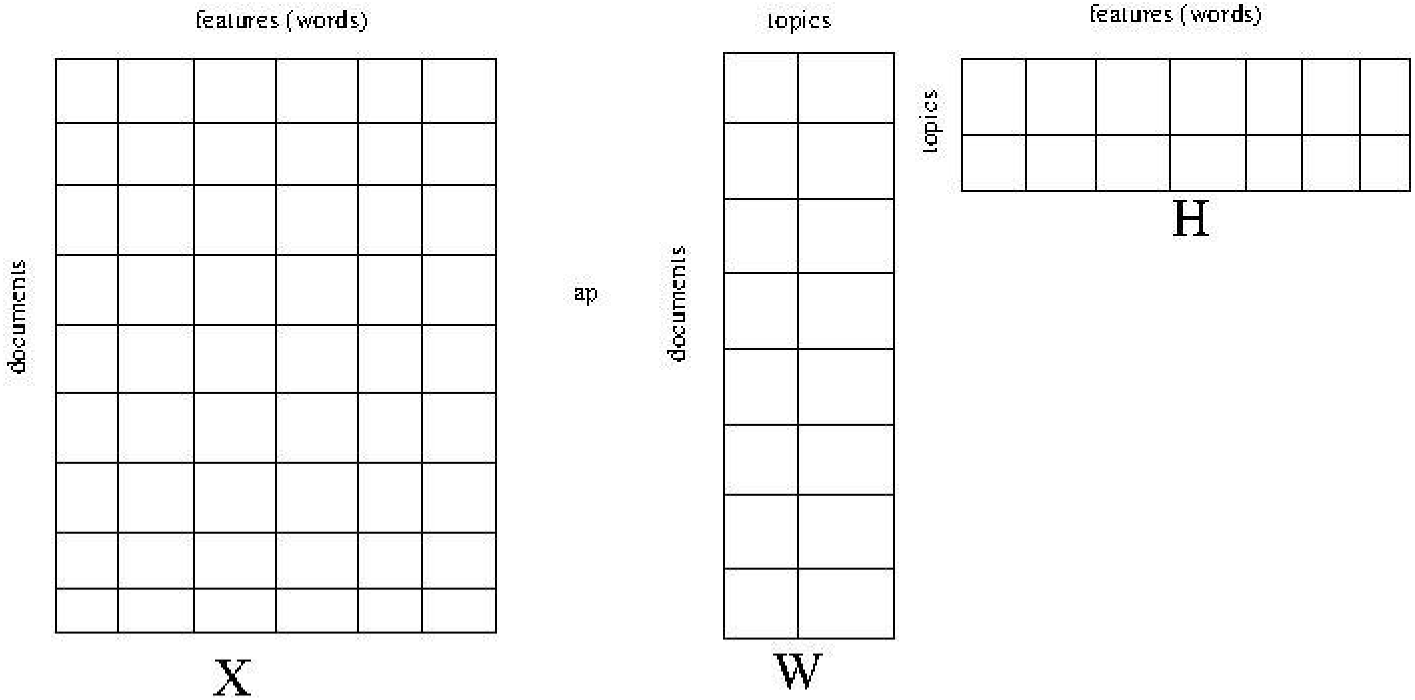
- Many of these can be very useful in detecting topics
  - ➢ A specific community of people, perhaps in the teenage demographics, talk about celebrity gossip

Method

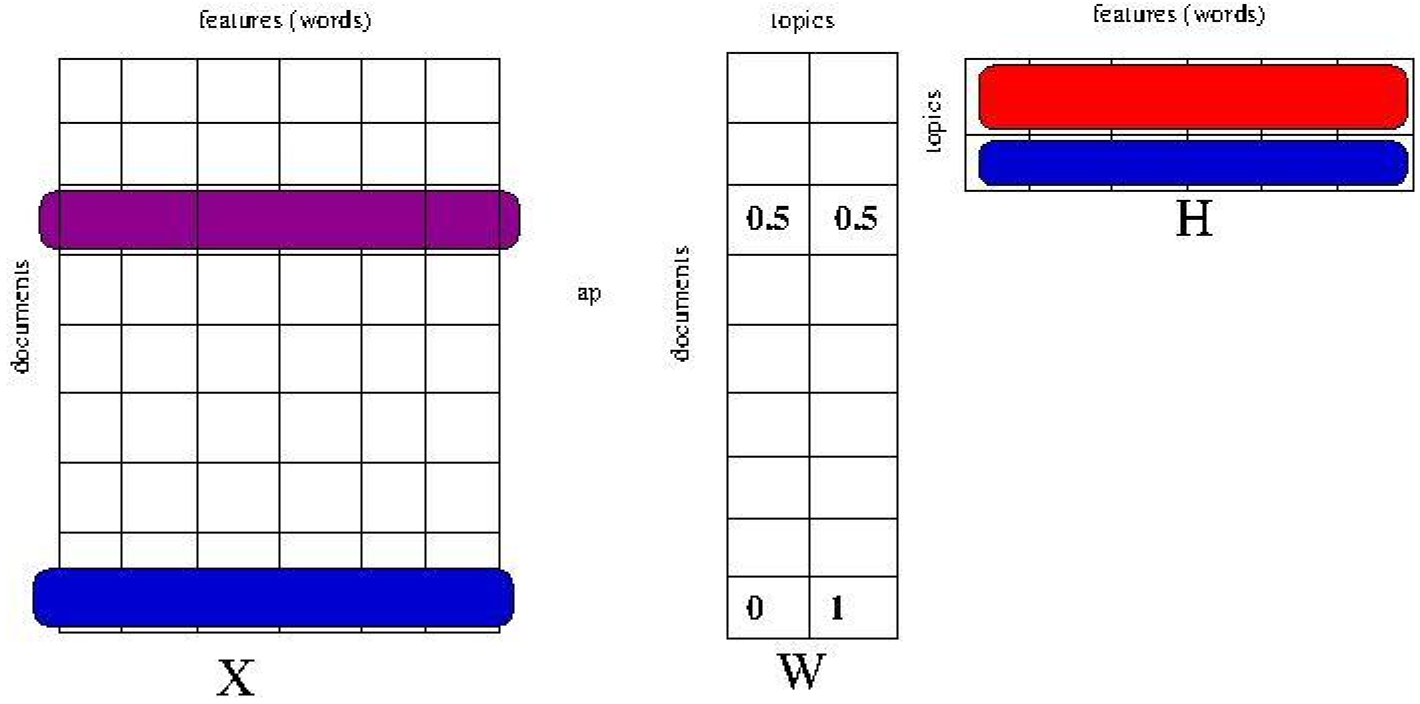
NMF-based method.

$$\mathbf{X} \approx \mathbf{W}\mathbf{H}$$

$\mathbf{X}$	doc-by-words
$\mathbf{W}$	doc-by-top
$\mathbf{H}$	top-by-words


features (words)
documents
X
ap
topics
documents
W
features (words)
topics
H

Method


features (words)
documents
X
ap
topics
documents
0.5
0.5
0
1
W
features (words)
topics
H

Method

NMF-based method.

$$\mathbf{U} \approx \mathbf{W}\mathbf{G}$$

$\mathbf{U}$	doc-by-users
$\mathbf{W}$	doc-by-comm
$\mathbf{G}$	comm-by-users

Method

$\mathbf{X}$ Data Matrix

$\mathbf{U}$ Data Matrix

$\mathbf{W}$ Decomposition Matrix

$\mathbf{H}$ Topic Matrix

$\mathbf{G}$ Community Matrix

Method

$$\mathbf{X} \approx \mathbf{W}\mathbf{H} \qquad\qquad \mathbf{U} \approx \mathbf{W}\mathbf{G}$$

Common Decomposition Matrix

- Every topic has a corresponding dedicated community of users
- Hence, the decomposition of the document into its topics or communities is same.
- Hence common “W”

Method

$$L = ||\mathbf{X} - \mathbf{WH}||_F^2 + ||\mathbf{U} - \mathbf{WG}||_F^2$$

$$L = \mu L_T + (1 - \mu)L_C + \alpha(R)$$

μ Importance parameter

Optimization

- Collective Matrix Factorization (Singh and Gordon 2008)
- Multiplicative Updates to find local minimum (Lee and Seung 2000)

Experiments

Dataset

- Publicly available dataset.
  - ➢ News articles;
  - ➢ Tweets which linked to them until 12 hours after publication;
  - ➢ Considered only verified news sources;
  - ➢ Extracted the text;
- From the tweets, also extracted usernames
  - ➢ Hashtags (for evaluations)
  - ➢ Usernames

Experiments

- $\mathbf{X}$
  - ➢ TF-IDF features
- $\mathbf{U}$
  - ➢ 0/1 matrix of which users tweeted about that document
  - ➢ Usernames

Experiments

- Identified the “difficult” articles, and created three categories
  - “Content” stable
  - “Community” stable
  - “Mixed” stable

Evaluation

- Centroid of all the documents of that hashtag

- Obtain the top-10 words and compare the rankings on NDCG and MAP

Results - Community Stable

Our Method

	K = 5	K = 10	K = 15	K = 20
NDCG	0.4081	0.4800	0.5029	0.5129
MAP	0.2653	0.3637	0.4007	0.4173
	μ = 0.01	μ = 0.5	μ = 0.5	μ = 0.5

Baseline Approach; NO CONTEXT

	K = 5	K = 10	K = 15	K = 20
NDCG	0.3699	0.4496	0.4608	0.4138
MAP	0.2191	0.3596	0.3462	0.3420

Results - Content Stable

Our Method

	K = 5	K = 10	K = 15	K = 20
NDCG	0.6888	0.6055	0.6317	0.6623
MAP	0.5655	0.4784	0.5115	0.5559
	$\mu = 1$	$\mu = 1$	$\mu = 0.75$	$\mu = 0.75$

Baseline Approach; NO CONTEXT

	K = 5	K = 10	K = 15	K = 20
NDCG	0.6888	0.6055	0.4885	0.6504
MAP	0.5655	0.4784	0.3089	0.5411

Results - Mixed Stable

Our Method

	K = 5	K = 10	K = 15	K = 20
NDCG	0.9005	0.8868	0.9249	0.9089
MAP	0.7783	0.7965	0.8964	0.8845
	$\mu = 0.25$	$\mu = 0.75$	$\mu = 0.25$	$\mu = 0.25$

Baseline Approach; NO CONTEXT

	K = 5	K = 10	K = 15	K = 20
NDCG	0.8771	0.8762	0.4251	0.4580
MAP	0.7762	0.7783	0.3232	0.3644

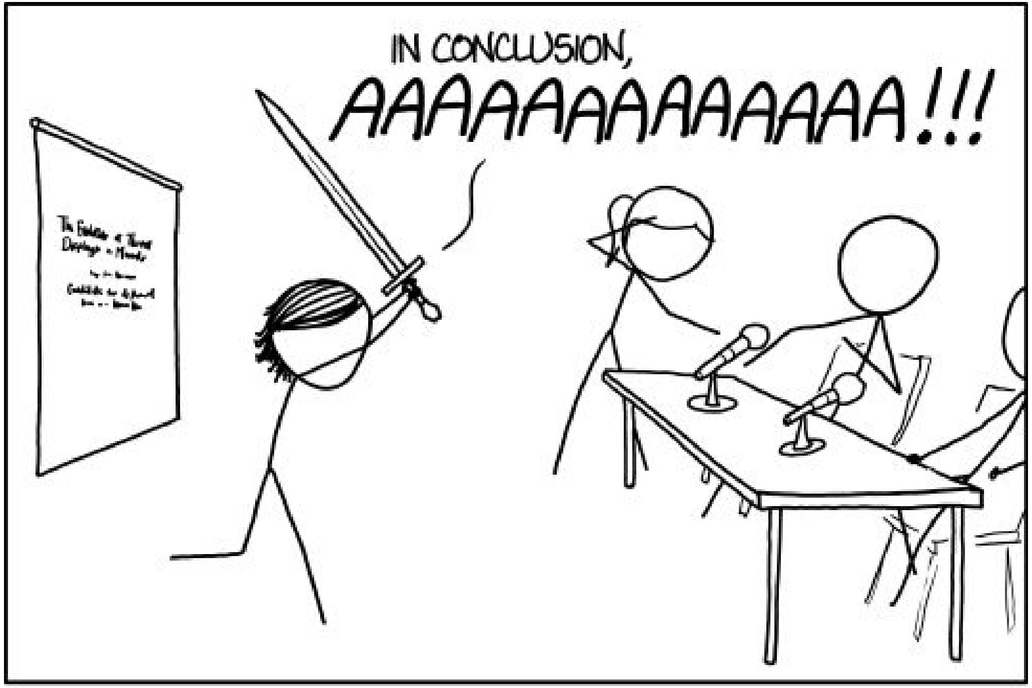
Conclusion

design robust and efficient computational methods to analyze social media data

1. Methods to effectively use metadata
   - ➢ Used text and user interactions to learn better topics (NMF-based)
2. Analyze social media reactions to events
   - ➢ Quantified the “buzz” of an event.
   - ➢ Independent to the size and scope of the event.
   - ➢ Early prediction of impact
3. Infoveillence
   - ➢ Iterative use of topic modeling to prune social media
   - ➢ Detect trends of prescription drug abuse

Thank you!

Questions?



THE BEST THESIS DEFENSE IS A GOOD THESIS OFFENSE.

Publications

PUBLICATIONS

Janani Kalyanam and Gert Lanckriet, "Learning from Unstructured Multimedia Data", *Proceedings of the 23rd International Conference on World Wide Web*, 2014.

Janani Kalyanam, Amin Mantrach, Diego Saez Trumper, Hossein Vahabi and Gert Lanckriet, "Leveraging Social Context for Topic Evolution", *Proceedings of the 21st International Conference on Knowledge Discovery and Data Mining*, 2015.

Janani Kalyanam, Sumithra Velupillai, Son Doan, Mike Conway and Gert Lanckriet, "Facts and Fabrications about Ebola: A Twitter Based Study", *Proceedings of the 21st International Conference on Knowledge Discovery and Data Mining Workshop on Connected Health in Big Data Era*, 2015.

Janani Kalyanam, Sumithra Velupillai, Mike Conway and Gert Lanckriet, "From Event Detection to Story Telling on Microblogs", *Proceedings of the ACM/IEEE Conference on Advances in Social Network Analysis and Mining*, 2016.

Janani Kalyanam, Takeo Katsuki, Gert Lanckriet and Timothy Mackey, "Exploring Trends of Nonmedical use of Prescription Drugs and Polydrug Abuse in the Twittersphere Using Unsupervised Machine Learning", *Addictive Behaviors*, 2016.

Janani Kalyanam, Mauricio Quezada, Barbara Poblete and Gert Lanckriet, "Prediction and Characterization of High-Activity Events in Social Media Triggered by Real-World News", *PLOS ONE*, 2016.