Studying Anonymous Health Issues and Substance Use on College Campuses with Yik Yak

Michael Paul, University of Colorado

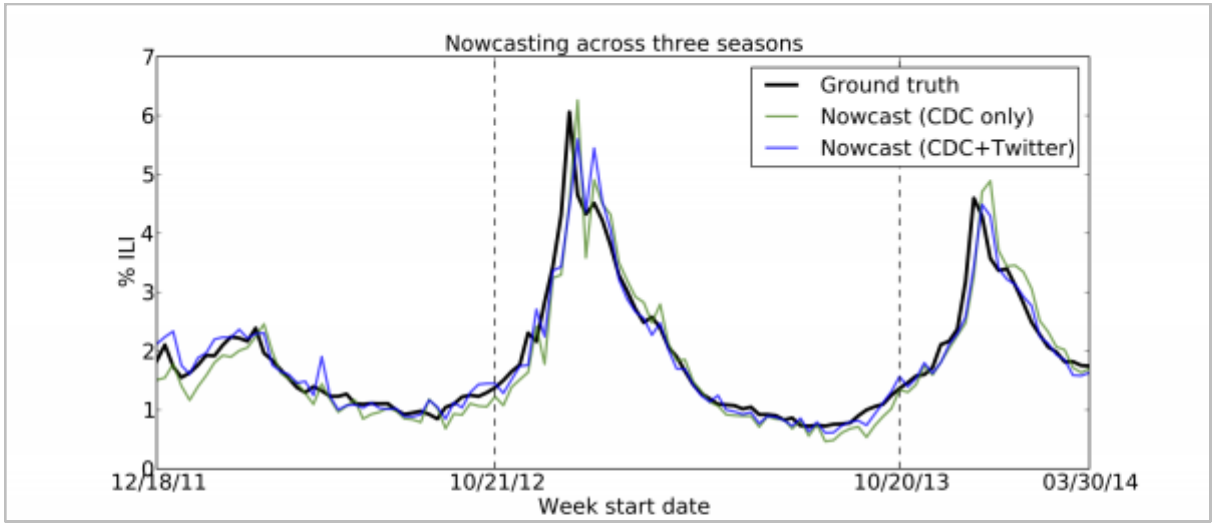
W3PHI | Feb 12, 2016

with:
Animesh Koratana, Mark Dredze, Margaret Chisolm, Matthew Johnson

Health in Social Media

People publicly share a variety of self-reported health information on social media

- Medication adverse reactions
- Healthy behaviors
- Illness
- Smoking
- Pain
- Mood



Health in Social Media

Typical social media platform:

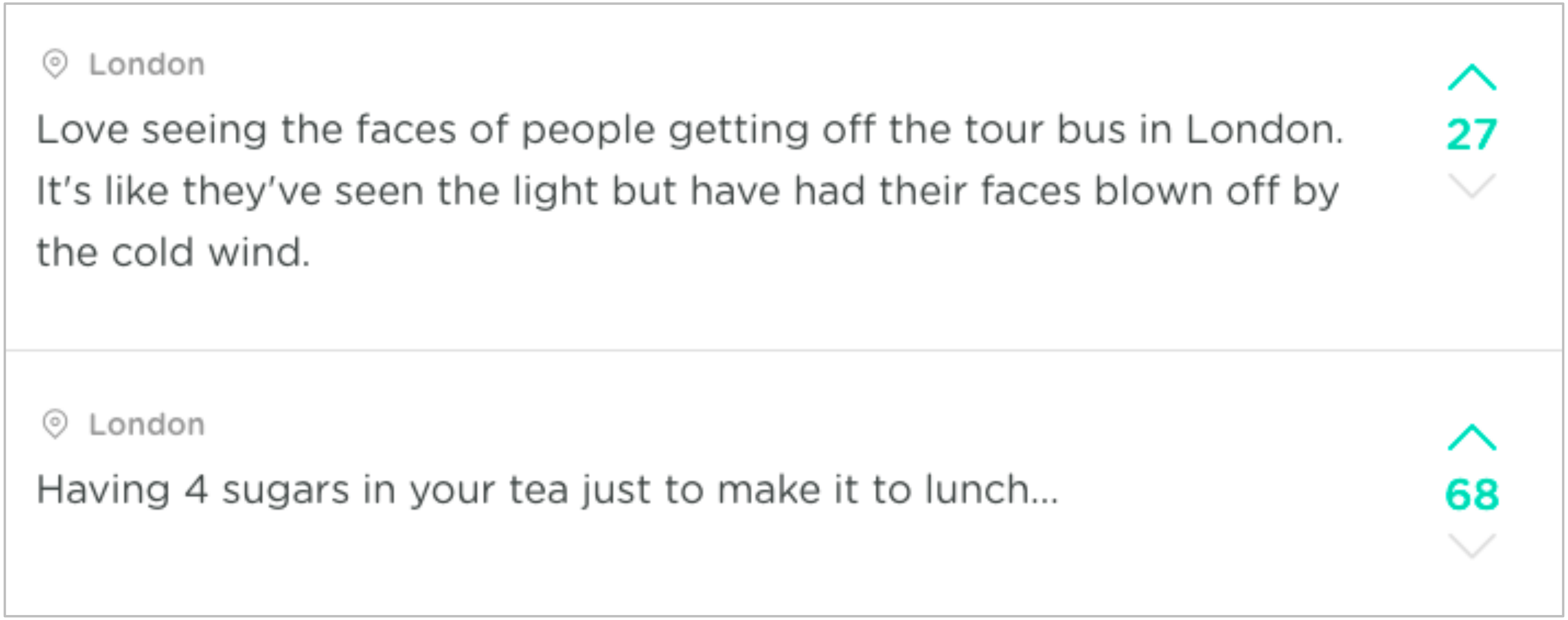
- User identifiers
  - Real names (Facebook)
  - Pseudonyms (Twitter)
- Target audience
  - Social network (friends, peers)
  - General public? (for public figures)

Yik Yak

- Social media platform launched in 2013
- Over 3 million active monthly users
- Popular with younger users

Yik Yak

- Short messages called “yaks”
- Messages are anonymous



London

Love seeing the faces of people getting off the tour bus in London. It’s like they’ve seen the light but have had their faces blown off by the cold wind.

27

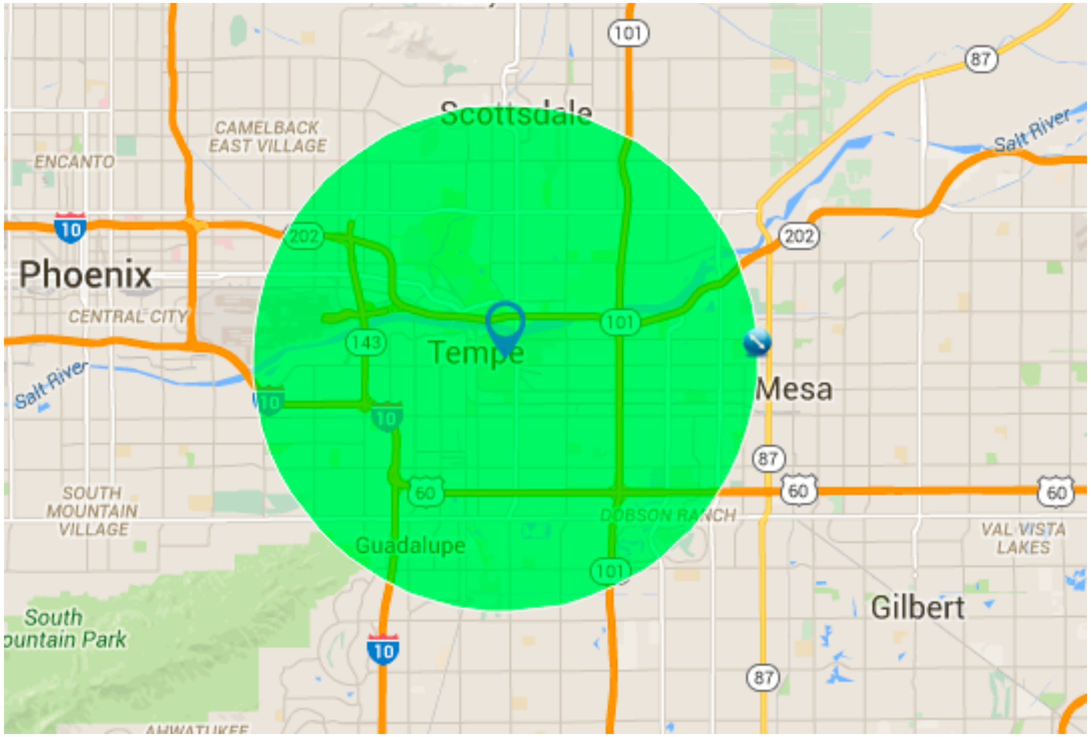
London

Having 4 sugars in your tea just to make it to lunch...

68

Yik Yak

- Messages are only viewable within geographic proximity to author
  - 5-mile radius



Health in Social Media

Yik Yak:

- No user identifiers
  - Fully anonymous (same property as 4chan)
- Target audience
  - Geographic network
  - Students

Yik Yak

Research Question 1:
What health topics are discussed on an **anonymous** platform?

Hypothesis:
Users will be more willing to discuss **stigmatizing** health issues

Yik Yak

Research Question 2:
What health topics are discussed near **college campuses?**

We can filter for messages near specific locations

Data Collection

- Crawler spoofs the geo-coordinates of the agent
  - Can collect data within radiuses that we specify
- Crawled data from **120** college campuses
  - Google Maps API used to define center point of campus
- Data crawled continuously from June 12, 2015 - July 14, 2015
- Dataset size: **122,179** yaks
  - plus replies

Health Topics

What topics are discussed in the dataset?

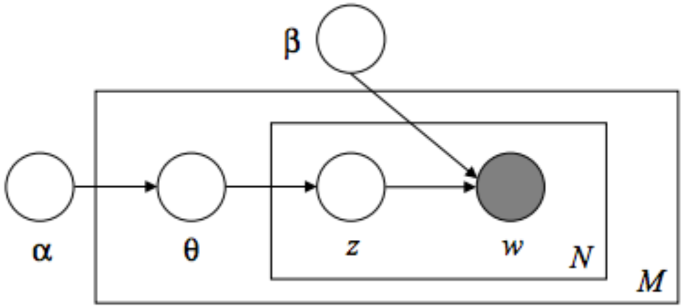
- We trained a topic model on the yaks
  - Latent Dirichlet Allocation (LDA)
  - 50 topics

Note: the Ailment Topic Aspect Model (ATAM) did not work well on this dataset because of low representation of health topics in the data

Health Topics

Latent Dirichlet Allocation (LDA)

- Probabilistic model
- Learns to associate documents with topics
- Learns to associate topics with words
  - Each topic is interpreted as a cluster of related words
- Used to understand common themes in text data



Health Topics

Drugs/Smoking

weed
smoke
drugs
smoking
drug
doctor
high
take
anxiety
got

Eating

eat
food
pizza
good
eating
cheese
chicken
chipotle
like
want

Drinking

drink
drunk
coffee
beer
drinking
water
alcohol
wine
milk
starbucks

Health Topics

Weight

fat
weight
gym
eat
body
lose
healthy
eating
im
workout

Sex

sex
like
girl
get
girls
guys
guy
time
want
feel

Hygiene

smell
use
like
shower
water
teeth
wash
skin
hair
face

Health Topics

For comparison, example health topics in Twitter:

Influenza-like Illness	Insomnia & Sleep Issues	Diet & Exercise	Cancer & Serious Illness	Injuries & Pain	Dental Health
better	night	body	cancer	hurts	dentist
hope	bed	pounds	help	knee	appointment
ill	body	gym	pray	ankle	doctors
soon	ill	weight	awareness	hurt	tooth
feel	tired	lost	diagnosed	neck	teeth
feeling	work	workout	prayers	ouch	appt
day	day	lose	died	leg	wisdom
flu	hours	days	family	arm	eye
thanks	asleep	legs	friend	fell	going
xx	morning	week	shes	left	went

Health Topics

- 9 out of 50 topics identified as relevant to health
- No topics about illness (despite common in Twitter)
- Topics about sensitive issues
  - sex, drugs, bathroom habits

Substance Use

Opportunity to study substance use on campuses

  - Not commonly discussed in public social media

- Could give insights into interest, awareness, attitudes toward drugs
  - Especially important for novel drugs

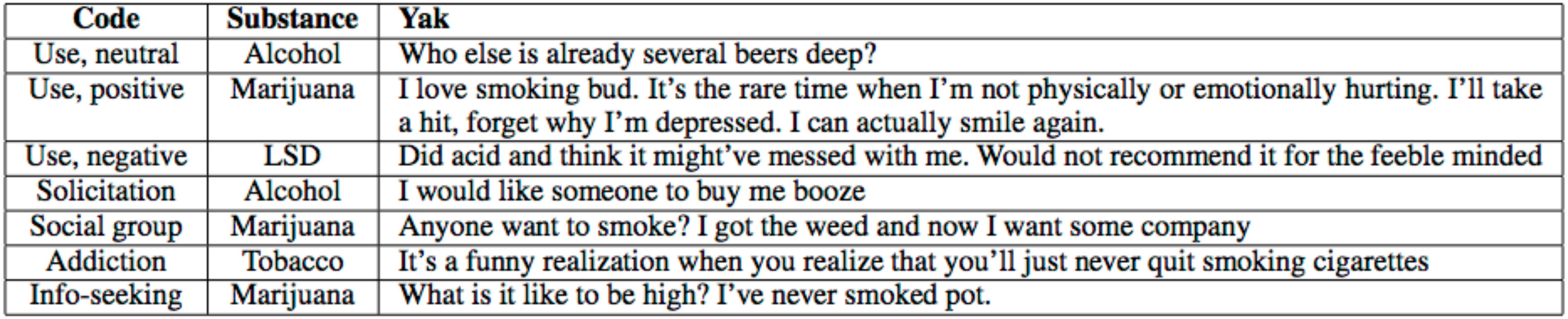
Substance Use

- Filtered yaks for drug-related keywords
- Annotated those yaks for relevance
- Drug-relevant dataset: **2,047** yaks

- We coded 500 yaks for fine-grained information
    - Will code more in future work

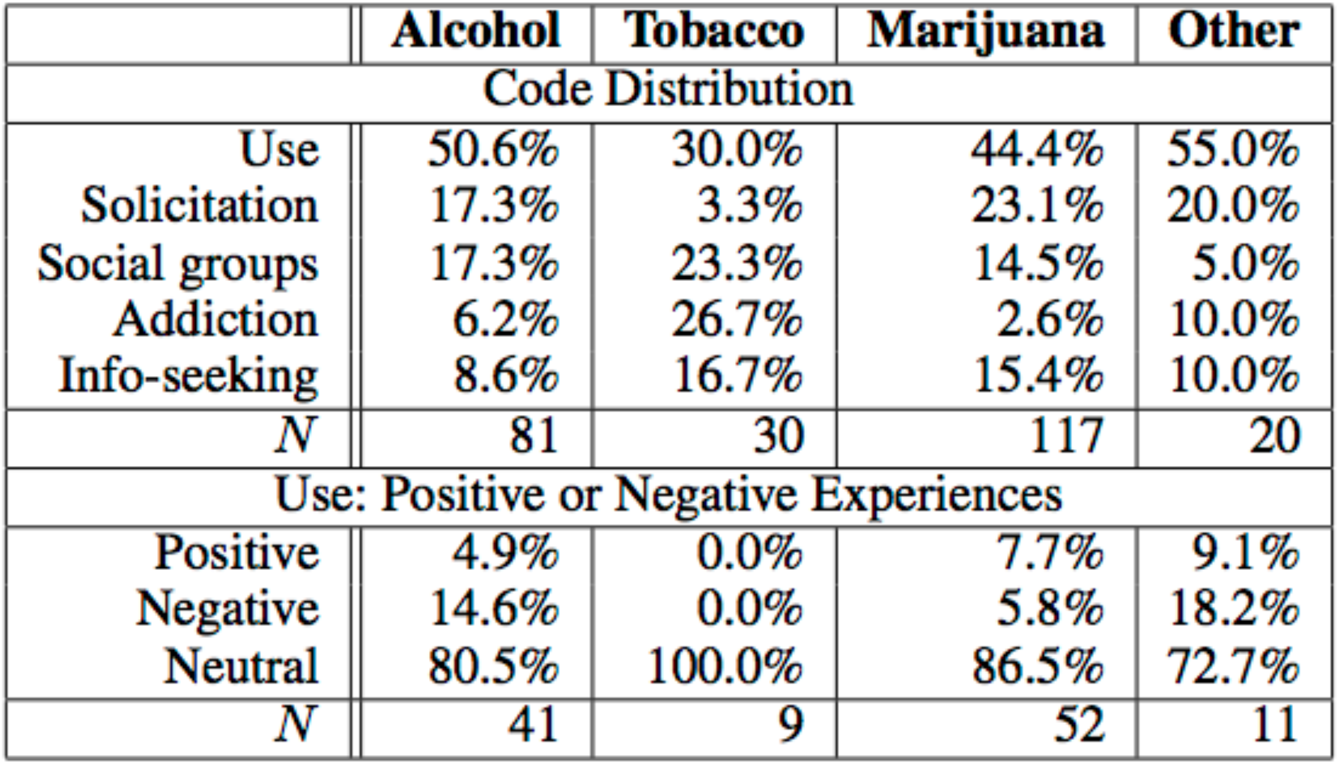
Substance Use

Codes (with examples)



Code	Substance	Yak
Use, neutral	Alcohol	Who else is already several beers deep?
Use, positive	Marijuana	I love smoking bud. It's the rare time when I'm not physically or emotionally hurting. I'll take a hit, forget why I'm depressed. I can actually smile again.
Use, negative	LSD	Did acid and think it might've messed with me. Would not recommend it for the feeble minded
Solicitation	Alcohol	I would like someone to buy me booze
Social group	Marijuana	Anyone want to smoke? I got the weed and now I want some company
Addiction	Tobacco	It's a funny realization when you realize that you'll just never quit smoking cigarettes
Info-seeking	Marijuana	What is it like to be high? I've never smoked pot.

Substance Use



	Alcohol	**Tobacco**	**Marijuana**	**Other**
Code Distribution				
Use	50.6%	30.0%	44.4%	55.0%
Solicitation	17.3%	3.3%	23.1%	20.0%
Social groups	17.3%	23.3%	14.5%	5.0%
Addiction	6.2%	26.7%	2.6%	10.0%
Info-seeking	8.6%	16.7%	15.4%	10.0%
N	81	30	117	20
Use: Positive or Negative Experiences				
Positive	4.9%	0.0%	7.7%	9.1%
Negative	14.6%	0.0%	5.8%	18.2%
Neutral	80.5%	100.0%	86.5%	72.7%
N	41	9	52	11

Substance Use

- People mostly use Yik Yak simply to describe use
- Requesting to buy substances is common
  - Offering to sell is uncommon
- Addiction discussion is highest for tobacco
- Sentiment is generally negative

Conclusion

- Anonymous social media has potential as a data source for understanding high-stigma health issues

- Substance use is commonly disclosed in Yik Yak
    - in contrast to Twitter

- Limitation: anonymity makes it hard to infer demographic attributes