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.

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

**Weight**
- fat
- weight
- gym
- eat
- body
- lose
- healthy
- eating
- eat
- work out

**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:

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Substance</th>
<th>Yak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use, neutral</td>
<td>Alcohol</td>
<td>Who else is already several beers deep?</td>
</tr>
<tr>
<td>Use, positive</td>
<td>Marijuana</td>
<td>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.</td>
</tr>
<tr>
<td>Use, negative</td>
<td>LSD</td>
<td>Did acid and think it might’ve messed with me. Would not recommend it for the feeble minded</td>
</tr>
<tr>
<td>Solicitation</td>
<td>Alcohol</td>
<td>I would like someone to buy me booze</td>
</tr>
<tr>
<td>Social group</td>
<td>Marijuana</td>
<td>Anyone want to smoke? I got the weed and now I want some company</td>
</tr>
<tr>
<td>Addiction</td>
<td>Tobacco</td>
<td>It’s a funny realization when you realize that you’ll just never quit smoking cigarettes</td>
</tr>
<tr>
<td>Info-seeking</td>
<td>Marijuana</td>
<td>What is it like to be high? I’ve never smoked pot.</td>
</tr>
</tbody>
</table>
## Substance Use

<table>
<thead>
<tr>
<th></th>
<th>Alcohol</th>
<th>Tobacco</th>
<th>Marijuana</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Code Distribution</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td>50.6%</td>
<td>30.0%</td>
<td>44.4%</td>
<td>55.0%</td>
</tr>
<tr>
<td>Solicitation</td>
<td>17.3%</td>
<td>3.3%</td>
<td>23.1%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Social groups</td>
<td>17.3%</td>
<td>23.3%</td>
<td>14.5%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Addiction</td>
<td>6.2%</td>
<td>26.7%</td>
<td>2.6%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Info-seeking</td>
<td>8.6%</td>
<td>16.7%</td>
<td>15.4%</td>
<td>10.0%</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>81</td>
<td>30</td>
<td>117</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Alcohol</th>
<th>Tobacco</th>
<th>Marijuana</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Use: Positive or Negative Experiences</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>4.9%</td>
<td>0.0%</td>
<td>7.7%</td>
<td>9.1%</td>
</tr>
<tr>
<td>Negative</td>
<td>14.6%</td>
<td>0.0%</td>
<td>5.8%</td>
<td>18.2%</td>
</tr>
<tr>
<td>Neutral</td>
<td>80.5%</td>
<td>100.0%</td>
<td>86.5%</td>
<td>72.7%</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>41</td>
<td>9</td>
<td>52</td>
<td>11</td>
</tr>
</tbody>
</table>
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