@ColditzJB
#SBM2016
Use of Twitter to Assess Sentiment toward Waterpipe Tobacco Smoking

Jason B. Colditz, MEd

Maharsi Naidu, Class of 2018
Noah A. Smith, PhD
Joel Welling, PhD
Brian A. Primack, MD, PhD
Goals

• Summarize known harms related to waterpipe tobacco smoking (WTS)

• List ways in which Twitter trends are currently being used in public health and medicine

• Define “machine learning” and describe how it can be used to automate large-scale data classification

• Compare Western and Eastern hemispheres with regard to overall sentiment toward WTS
Background: WTS

- Waterpipe Tobacco Smoking (WTS)
  - *Hookah, Shisha, Narghile* [nar·ghee·leh]

**Head / Bowl:**
- Flavored tobacco mixture
- Charcoal to maintain heat

**Hose / Mouthpiece:**
- Shared by smokers
- Typically not filtered

**Base:**
- Filled with water or flavored liquid
- Smoke is cooled as it bubbles through
Background: WTS & Health

• Typical toxicants from tobacco combustion
  – Additional toxicants from charcoal
  – Carbon monoxide and second-hand smoke
  – High volume of smoke

• Addictive potential
  – From social to habitual use
  – Transitioning to other tobacco products
Background: WTS Epidemiology

• Traditional and widely prevalent in Eastern global cultures
  – Widespread public health concerns of addiction and preventable disease

• Novel and gaining popularity in Western global cultures
  – Fun social activity / cultural immersion
  – Seen as relatively harmless vs. “smoking”
Background: Twitter & Health

• **Twitter for “Big Data”**
  – Used by nearly a third of young adults
  – Access to large scale data via Twitter’s Application Programming Interface (API)

• **Twitter for Public Health infodemiology:**
  – Natural disaster relief
  – Foodborne illness / Communicable diseases
  – E-cigarette sentiment & marketing
Background: Twitter Data

• Characteristics
  – 140 characters includes text, links, and...
    • Hashtags: #SBM2016 #DataScience
    • Emoji: 💖😊😞

– Basic location metadata:

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Prevalence</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geo-location</td>
<td>~ 1%</td>
<td>Calculated &amp; exact</td>
</tr>
<tr>
<td>Time Zone</td>
<td>Common</td>
<td>Self-reported &amp; broad</td>
</tr>
<tr>
<td>Location from user profile</td>
<td>Very Common</td>
<td>Self-reported &amp; aberrant</td>
</tr>
</tbody>
</table>
Background: Machine Learning

*Machine Learning* is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.

- Computers are adept at discovering patterns in large sets of data.

- Researchers can *train* computers to look for particularly *useful* patterns.
Methods: Data Collection

• Twitter stream for 48 weekend hours:
  – From Friday, 11/14/2014, 17:00 GMT through Sunday, 11/16/2014, 16:59 GMT

• Filters:
  – English language
  – Search terms: hookah, hooka, shisha, sheesha, narghile

Tweets: \( N = 43,155 \)
Methods: Human Coding

- Random subset of 2,000 tweets
  - Independently double-coded

- Coding:
  - Relevant?
    - No: False positive
      - Marijuana
      - Marketing
      - Pop-culture
    - Yes: WTS Sentiment:
      - Positive?
      - Negative?
Methods: Machine Learning

- Supervised learning
  - Natural Language Toolkit (NLTK) for *Python*
  - Human coding as gold standard
  - Trained *Naïve Bayes* classifiers for WTS sentiment
    - Testing model’s *Accuracy*, *Precision*, and *Recall*
    - 3:1 training to testing ratio:

- *Unigram* parameters
  - Individual words
  - Emoji

Coded as WTS-relevant

\[ n = 1,345 \]

Sentiment classification:

- Training Data
  \[ n = 1,008 \ (75\%) \]

- Testing Data
  \[ n = 337 \ (25\%) \]
Results: Human Coding

• 655 (33%) Tweets excluded
  • Not WTS related
  • Marketing or pop-culture references

• 1,345 Tweets considered relevant:
  • 54% Positive sentiment
    – Cohen’s $K = 0.74$
    – Agreement = 87%
  • 21% Negative sentiment
    – Cohen’s $K = 0.71$
    – Agreement = 92%
  • Disagreements manually adjudicated by coders to provide overall consensus
## Results: Machine Learning

- **Positive sentiment:**
  - Precision: 71%* & 76%† Recall: 84%* & 60%†
  - Overall accuracy: 73%

- Exemplar predictive features:

<table>
<thead>
<tr>
<th><em>Is positive:</em></th>
<th>†Is not positive:</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.9</td>
<td>13.7 “starter”</td>
</tr>
<tr>
<td>7.6</td>
<td>12.9 “cigarettes”</td>
</tr>
<tr>
<td>5.9 “chill”</td>
<td>5.5 “hit”</td>
</tr>
<tr>
<td>4.8</td>
<td>4.9 “lounges”</td>
</tr>
<tr>
<td>3.4</td>
<td>3.5 🙄🙄</td>
</tr>
</tbody>
</table>
Results: Machine Learning

- **Negative sentiment:**
  - Precision: 41%* & 75%† Recall: 93%* & 60%†
  - Overall accuracy: 70%

- Exemplar predictive features:

  *Is negative:*
  - 23.1 “cigarettes”
  - 20.1 “shit”
  - 18.6 “tar”
  - 8.7 “ban”
  - 6.9 😞

  †Is not negative:*
  - 6.7 “lads”
  - 6.4 “tonight”
Results: Hemispheres

- Coded WTS tweets had time zone data
  - 66% (n = 890)

- Western $n = 727$
  - 56% positive*
  - 24% negative

- Eastern $n = 163$
  - 31% positive*
  - 23% negative

* $\chi^2 = 32.0$, $p < .001$
Limitations / Considerations

• Twitter data biases
  – English language
  – Timeframes

• Keyword search parameters
  – Broad terms like “smoke” increase recall (sensitivity), but decrease precision (specificity)

• Classifier sophistication
  – Unigrams vs. $n$-grams (bigrams, trigrams, etc.)

• Human coding is time and labor intensive
  – Crowdsourcing (e.g., Mechanical Turk)
Discussion

- Waterpipe tobacco smoking (WTS) has serious health risks and is gaining popularity in the US.
- Twitter provides opportunities for researchers and public health advocates to tap into online discourse and assess sentiment toward health behaviors.
- Machine learning methods allow for infodemiology: large-scale data categorization using geographic metadata, words, and symbols (e.g., emoji).
- Initial appraisal of our Twitter data indicated proportionately higher positive sentiment toward WTS in the western hemisphere—This warrants further investigation.
Thank You!

Jason B. Colditz, M.Ed.
jbc28@pitt.edu

@ColditzJB

~

Center for Research on Media, Technology, and Health

@CRMTH_Pitt