Identifying Public Health Related Topics on Twitter

Kyle W. Prier¹, Matthew S. Smith², Christophe G. Giraud-Carrier², and Carl L. Hanson³

¹Dept. of Health, Behavior & Society, Johns Hopkins Bloomberg School of Public Health
²Dept. of Computer Science, Brigham Young University
³Dept. of Health Science, Brigham Young University

Presenter Disclosures

Kyle Prier

(1) The following personal financial relationships with commercial interests relevant to this presentation existed during the last 12 months:

No relationships to disclose

Overview

Online Social Networks

- Growth
- Prevalence of Internet Access (via ITU)
  - U.S.: 77.3%
  - Developed: 68.6%
  - World: 29.7%
  - Developing: 21.1%
- Changing how we interact and communicate in both online and offline settings

Twitter

- “Microblogging” & “Tweets” – say it all in 140 characters or less
- 54.5 million people have used Twitter within the United States
- Demographics (Quanvacast.com)
  - 45% between 18 and 35
  - 63% under 35 years

Twitter: As a Public Health Data Source

- Twitter API
  - Open and relatively easy to access
- Twitter used to understand offline health behaviors and trends
  - Accuracy of tweets regarding H1N1 (Chew & Eysenbach, 2009)
    - 46% News related, 7% containing misinformation
  - Tweets regarding antibiotic use (Scanfield et al., 2010)
  - Influenza related tweets by symptom keywords (Culotta, 2010)
    - .78 correlation with CDC
Identifying Public Health Related Topics on Twitter

1. **Tobacco as a Test Topic**
   - Approximately 19% of adults smoke in the US.
   - Tobacco is attributed to over 14 million deaths in the US since 1964.
   - 400,000 smokers and former smokers die each year from tobacco.
   - 38,000 non-smokers die from secondhand smoke each year.
   - Source: CDC.

2. **Problem**
   - Increasing data available from social media that contain potentially relevant conversations for public health issues.
   - It is difficult to effectively identify and browse relevant health related topics among such large datasets.
   - It is difficult to isolate conversations relevant to topics that occur less frequently.

3. **Research Questions**
   - How can topic modeling be used to most effectively identify relevant public health topics on Twitter?
   - Which public health related topics, specifically tobacco use, are discussed among Twitter users?
   - What are common tobacco related themes?
   - How do the number of tweets influence LDA output?

4. **Latent Dirichlet Allocation**
   - Unsupervised machine learning generative probabilistic model.
   - Identifies latent topic information from text corpora.
   - Each topic represented as a probability distribution over a number of words.
   - First proposed by Blei, Ng, & Jordan in 2002.
   - We used the LDA implementation in MALLET.

5. **Data Sampling**
   - 9 Randomly selected states from the Federal Census divisions.
   - Twitter Search API used with the "geocode" parameter.
   - Recent tweets were gathered.
   - Frequency: every 2 minutes.
   - Duration: 31 day period from 4 Oct 2010 – 3 Nov 2010.
   - Resulted in 5,029,462 tweets.
   - 3 subsets generated for analysis.

6. **Extended Data Sampling**
   - Comprehensive Dataset: 2 subsets.
     - Smoking, tobacco, cigarette, cigar, hookah, and hooka.
   - Tobacco Subset by Keywords: 4 subsets.
     - No cohesive topics, sample size too small.
   - Extended Comprehensive Dataset: 3 subsets.
     - Cohesive, relevant topics.
   - Extended Tobacco Subset: 1 subset.
     - Cohesive, relevant topics.

---

**Additional Information**

- Source: CDC.
- Problem:
- Research Questions:
- Latent Dirichlet Allocation:
- Data Sampling:
- Additional Data Sampling:
- Additional Dataset Sampling:
Identifying Public Health Related Topics on Twitter

Comprehensive Dataset: Relevant Topics

# Most Likely Topic Components (n-grams)

44
gps app, calories burned, felt alright, race report, weights workout, christus shumpert, workout named, christus wellington, started cycling, schrader andy, core fitness, vff better acts, mc-gold barn, frontenacian roar, meicolahyline lee, loginsun phone, edie aflaire, leader greene, lag.

45
demographics disease, breast augmentation, compression garments, sq ruba pere, weekly newsletter, lab result, medical/news, prescription medications, diagnostic imaging, accountable care, elder care, osteoporosis, breast cancer, lego arthritis, true recessions, bariatric surgery, osteoarthritis, affordable dental

131
weight loss, diet pills, acai berry, healthy living, fat loss, weight loss diets, belly fat, alternative health, fat burning, pack aloe, organic gardening, essential oils, container gardening, hcg diet, natural creek, hca acids, anti-aging, muscle gain, press habit encouraged

# Most Likely Topic Components (unigrams)

18
high, smoke, still needashakira, weed, spitter, career, block, trn, jilly, yellow, more tat, we, air, eul, tick, burn, burst

Extended Tobacco Subset: Topics

# Most Likely Topic Components (n-grams)

1
smoking weed, smoking gun, smoking crack, stop smoking, cigarette burns, external cell phones, hubble bar, phone smoking, smoke, cough, hand smoke, im taking, smoking/bad, hookah, house, free, smoking, tyler cup, dont understand, talking book, im reading, twenty years people

2
quit smoking, stop smoking, quit, quit smoking, smoking, cigarette, cigar, hookah, tobacco, smoking, addiction, cigar shop, quit smoking cigarettes, chronic lung disease, smoking, smoking, stop smoking, smoking,撤出, hongkong, new york, smoking

3
cigarette smoke, dont smoke, smoking, smoking pot, im gone, huck, tonight smoking, smoking, five specials, free food, beed, right, electronic cigarettes, good times, smoking cessation, cigarette brand, secondhand smoke, smoking, smoking, smoking, smoking, smoking

4
smoking weed, cont link, bad, free, pavement, sir, start smoking, hate smoking, hooah, great hooah, hooah, hooah, các, smoking, cough, hooah, smoking, fletcher, smoking, still, smoking, smoking, external cell, smoking, smoking ban, smoking, smoking, smoking, smoking, smoking, smoking, smoking, smoking, smoking, smoking, smoking, smoking

5
smoke cigarettes, smoking hot, im smoking, smoking section, stopped smoking, chewing tobacco, smoking kills, clean smoking, smoking area, ban smoking, people for, ring ring hooah, ring ring beer, smoke containing, link t, damad, cigarette, healthiest smoking products, theyre smoking, hate cigarettes, dont smoke, smoking, apartment

Discussion

Analyzing a large conversational dataset provides few health related topics
High frequency of marijuana-related terms
Querying a sample of random tweets is less automated and requires us to manually identify relevant terms
These results were relevant
Chronic behaviors vs. short term events
LDA and similar algorithms may enable researchers to more efficiently monitor and survey health statuses among communities

More info

Contact information:

Kyle Prier
kprier@jhsph.edu

Original paper


Extended Tobacco Subset: n = 4962

Comprehensive Dataset: Relevant Topics

Tobacco Subset by Keywords

Topics

Extended Tobacco Subset: n = 4962

Topics

Extended Tobacco Subset: n = 4962

Topics

Extended Tobacco Subset: n = 4962

Topics

Extended Tobacco Subset: n = 4962

Topics