1

Identifying Health Related Topics on Twitter

An Exploration of Tobacco Related Tweets as a Test Topic

Kyle W. Prier[1], Matthew S. Smith[2], Christophe G. Giraud-Carrier[2], and Carl L. Hanson[3]

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

Identifying Health Related Topics on Twitter

2

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

Identifying Health Related Topics on Twitter

3

Overview

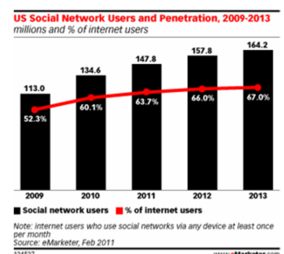
- Background
  - Online Social Networks
    - Prevalence
    - Limitations
  - Twitter
    - Twitter as a Data Source
  - Tobacco as a Test Topic
- Problem
- Methods
  - Data Collection
  - Latent Dirichlet Allocation
- Data Analysis
- Results & Discussion

Identifying Health Related Topics on Twitter

4

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



Identifying Health Related Topics on Twitter

5

Twitter

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

Identifying Health Related Topics on Twitter

6

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 Health Related Topics on Twitter

7

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

Identifying Health Related Topics on Twitter

8

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

Identifying Health Related Topics on Twitter

9

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?

Identifying Health Related Topics on Twitter

10

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
  - McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

Identifying Health Related Topics on Twitter

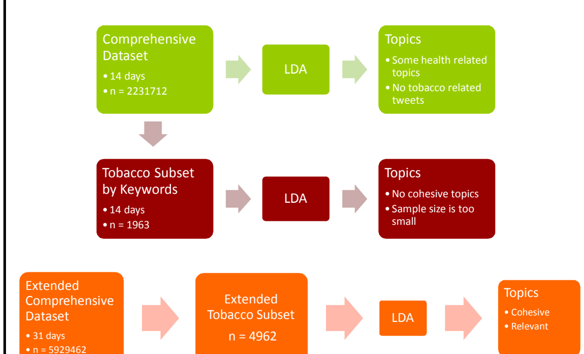
11

Data Sampling

- 9 Randomly selected states from the Federal Census divisions
  - Georgia, Idaho, Indiana, Kansas, Louisiana, Massachusetts, Mississippi, Oregon, and Pennsylvania
- 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 5,929,462 tweets
- 3 subsets generated for analysis
  - 2 subsets were created using tweets that included the terms: smoking, tobacco, cigarette, cigar, hookah, and hooka

Identifying Health Related Topics on Twitter

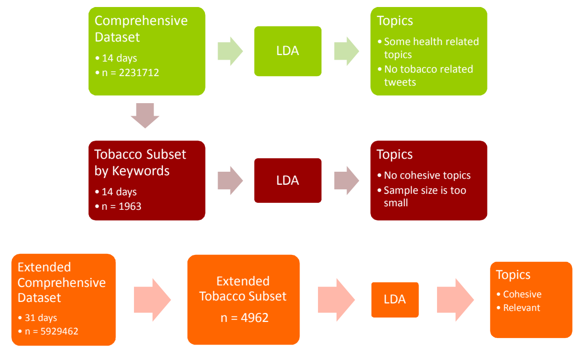
12



Identifying 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, chrissie wellington, started cycling, schwinn airdyne, core fitness, vff twitter acct, mc gold team, fordironman ran, fetcheveryone ive, logyourrun iphone, elite athletes, lester greene, big improvement, myrtle avenue
45	alzheimers disease, breast augmentation, compression garments, sej nuke panel, weekly newsletter, lab result, medical news, prescription medications, diagnostic imaging, accountable care, elder care, vaser lipo, lasting legacy, restless legs syndrome, joblessness remains, true recession, bariatric surgery, older applicants, internships attract, affordable dental
131	weight loss, diet pills, acai berry, healthy living, fat loss, weight loss diets, belly fat, alternative health, fat burning, pack abs, organic gardening, essential oils, container gardening, hcg diet, walnut creek, fatty acids, anti aging, muscle gain, perez hilton encourages
#	**Most Likely Topic Components (unigrams)**
18	high, smoke, shit realwizkhalifa, weed, spitta, currensy, black, bro, roll, yellow, man, hit, wiz, sir, kush, alot, fuck, swag, blunt



Extended Tobacco Subset: Topics

#	Most Likely Topic Components (n-grams)
1	smoking weed, smoking gun, smoking crack, stop smoking, cigarette burns, external cell phones, hooka bar, youre smoking, smoke cigars, smoking kush, hand smoke, im taking, smoking barrels, hookah house, hes smoking, ryder cup, dont understand, talking bout, im reading, twenty years people
2	quit smoking, stop moking, cigar guy, smoking cigarettes, hookah bar, usa protect, quitting smoking, started smoking, electronic cigarette, cigars link, smoking addiction, cigar shop, quit smoking cigarettes, chronical green smoking, link quit smoking naturally, smoking pot, youtube video, link quit smoking, link holistic remedies, chronical protect
3	cigarette smoke, dont smoke, quit smoking, smoking pot, im gonna, hookah tonight smoking ban, drink specials, free food, ladies night, electronic cigarettes, good times, smoking session, cigarette break, secondhand smoke, everythings real, effective steps, smoking cigs, smoking tonight
4	smoking weed, cont link, ladies free, piedmont cir, start smoking, hate smoking, hookahs great food, cigarette butts, thingswomenshouldstopdoing smoking, lol rt, sunday spot, cigarettes today, fletcher knebel smoking, pot smoking, film stars, external cell, fetishize holding, smoking room, halloween party, million people
5	smoke cigarettes, smoking hot, im smoking, smoking section, stopped smoking, chewing tobacco, smoking kills, chain smoking, smoking area, ban smoking, people die, ring ring hookah ring ring, love lafayette, link rt, damn cigarette, healthiest smoking products, theyre smoking, hate cigarettes, world series, hideout 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
  - Prier, K., Smith, M., Giraud-Carrier, C., Hanson, C. (2011). Identifying Health-Related Topics on Twitter: An Exploration of Tobacco-Related Tweets as a Test Topic. International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP11), March.