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 Giraud-Carrier[2], and Carl L. Hanson[1]

[1] Department of Health Science
[2] Department of Computer Science
Brigham Young University, Provo UT 84602, USA
kyle.prier@byu.edu, smitty@byu.edu, cgc@cs.byu.edu, carl_hanson@byu.edu

Abstract. Public health-related topics are difficult to identify in large conversational datasets like Twitter. This study examines how to model and discover public health topics and themes in tweets. Tobacco use is chosen as a test case to demonstrate the effectiveness of topic modeling via LDA across a large, representational dataset from the United States, as well as across a smaller subset that was seeded by tobacco-related queries. Topic modeling across the large dataset uncovers several public health-related topics, although tobacco is not detected by this method. However, topic modeling across the tobacco subset provides valuable insight about tobacco use in the United States. The methods used in this paper provide a possible toolset for public health researchers and practitioners to better understand public health problems through large datasets of conversational data.

Keywords: Data Mining, Public Health, Social Media, Social Networks, LDA, Tobacco Use, Topic Modeling

1 Introduction

Over recent years, social network sites (SNS) like Facebook, Myspace, and Twitter have transformed the way individuals interact and communicate with each other across the world. These platforms are in turn creating new avenues for data acquisition and research. Such web-based applications share several common features [3], and while there are slight variations in actual implementations, each service enables users to (1) create a public profile, (2) define a list of other users with whom they share a connection, and (3) view and discover connections between other users within the system. Since SNS allow users to visualize and make public their social networks, this promotes the formation of new connections among users because of the social network platform [3, 10]. Not only do SNS enable users to communicate with other users with whom they share explicit social connections, but with a wider audience of users with whom they would not have otherwise shared a social connection. Twitter, in particular, provides a medium whereby users can create and exchange user-generated content with a potentially larger audience than either Facebook or Myspace.

Twitter is a social network site that allows users to communicate with each other in real-time through short, concise messages (no longer than 140 characters), known as "tweets." A user's tweets are available to all of his/her "followers," i.e., all others who choose to subscribe to that user's profile. Twitter continues to grow in popularity. As of August 2010, a rough estimate of 54.5 million people used Twitter in the United States, with 63% of those users under the age of 35 years (45% were between the ages of 18 and 34) [17]. In addition, the US accounts for 51% of all Twitter users worldwide.

Because Twitter status updates, or tweets, are publicly available and easily accessed through Twitter's Application Programming Interface (API), Twitter offers a rich environment to observe social interaction and communication among Twitter users [15]. Recent studies have begun to use Twitter data to understand "real," or offline, health-related behaviors and trends. For example, Chew and Eysenbach analyzed tweets in an effort to determine the types and quality of information exchanged during the H1N1 outbreak [6]. The majority of Twitter posts in their study were news-related (46%), with only 7 of 400 posts containing misinformation. Scanfield et al. explored Twitter status updates related to antibiotics in an effort to discover evidence of misunderstanding and misuse of antibiotics [14]. Their research indicated that Twitter offers a platform for the sharing of health information and advice. In an attempt to evaluate health status, Culotta compared the frequency of influenza-related Twitter messages with influenza statistics from the Centers for Disease Control and Prevention (CDC) [7]. Findings revealed a .78 correlation with the CDC statistics suggesting that monitoring tweets might provide cost effective and quicker health status surveillance. These studies are encouraging. It is our contention that user-generated content on Twitter does indeed provide both public and relevant health-related data, whose identification and analysis adds value to researchers, and may allow them to tailor health interventions more effectively to their audiences.

In this paper, we address the problem of how to effectively identify and browse health-related topics within Twitter's large collection of conversational data. Although recent studies have implemented topic modeling to process Twitter data, these studies have focused on identifying high frequency topics to describe trends among Twitter users. Current topic modeling methods prove difficult to detect lower frequency topics that may be important to investigators. Such methods depend heavily on the frequency distribution of words to generate topic models. Public health-related topics and discussions use less frequent words and are therefore more difficult to identify using traditional topic modeling. Our focus, here, is on the following 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?

We choose to test our topic modeling effectiveness by focusing on tobacco use in the United States. Although we are interested in devising a method to

identify and better understand public health-related tweets in general, a test topic provides a useful indicator through which we can guide and gauge the effectiveness of our methodology. Tobacco use is a relevant public health topic to use as a test case, as it remains one of the major health concerns in the US. Tobacco use, primarily cigarette use, is one of the leading health indicators of 2010 as determined by the Federal Government [11]. Additionally, tobacco use is considered the most preventable cause of disease and has been attributed to over 14 million deaths in the United States since 1964 [16, 1]. Also, there remain approximately 400,000 smokers and former smokers who die each year from smoking-related diseases, while 38,000 non-smokers die each year due to second-hand smoke [1, 12].

In the following sections, we discuss our data sampling and analysis of tweets. We used Latent Dirichlet Allocation (LDA) to analyze terms and topics from the entire dataset as well as from a subset of tweets created by querying general, tobacco use-related terms. We highlight interesting topics and connections as well as limitations to both approaches. Finally, we discuss our conclusions regarding our research questions as well as possible areas of future study.

2 Methods and Results

In this section, we introduce our methods of sampling and collecting tweets. Additionally, we discuss our methods to analyze tweets through topic modeling. We used two distinct stages to demonstrate topic modeling effectiveness. First, we model a large dataset of raw tweets in order to uncover health-related issues. Secondly, we create a subset of tweets by querying a raw dataset with tobacco-related terms. We run topic modeling on this subset as well, and we report our findings.

2.1 Data Sampling and Collection

In order to obtain a representative sample of tweets within the United States, we chose a state from each of the nine Federal Census divisions through a random selection process. For our sample, we gathered tweets from the following states: Georgia, Idaho, Indiana, Kansas, Louisiana, Massachusetts, Mississippi, Oregon, and Pennsylvania.

Using the Twitter Search API, recent tweets for each state were gathered in two minute intervals over a 14-day period from October 6, 2010 through October 20, 2010. We used the Twitter Search API's "geocode" parameter to retrieve the most recent tweets from areas within each of the states, regardless of the content of the tweets. Since the "geocode" parameter requires a circle radius length, latitude and longitude, we gathered tweets from multiple specified circular areas that spanned each state's boundaries. To prepare the dataset for topic modeling, we remove all non-latin characters from the messages, replace all links within the dataset with the term "link," and only include users that publish at least two tweets. This process results in a dataset of 2,231,712 messages from 155,508 users. We refer to this dataset as the "comprehensive" dataset.

2.2 Comprehensive Dataset Analysis

We analyze the comprehensive dataset by performing LDA on it to produce a topic model. LDA is an unsupervised machine learning generative probabilistic model, which identifies latent topic information in large collections of data including text corpora. LDA represents each document within the corpora as a probability distribution over topics, while each topic is represented as a probability distribution over a number of words [2, 9]. LDA enables us to browse words that are frequently found together, or that share a common connection, or topic. LDA helps identify additional terms within topics that may not be directly intuitive, but that are relevant.

We configure LDA to generate 250 topic distributions for single words (unigrams) as well as more extensive structural units (n-grams), while removing stopwords. We began at 100 topics and increased the topic number by increments of 50. At 250 topics, the model provided topics that contained unigrams and n-grams that exhibited sufficient cohesion. Additionally, we set LDA to run for 1,000 iterations. We suspected that this topic model would provide less-relevant topics relating to public health and specifically tobacco use, since such topics are generally less frequently discussed. The LDA model of the comprehensive dataset generally demonstrates topics related to various sundry conversational topics, news, and some spam as supported by Pear Analytics [5]. However, the model provides several topics, which contain health-related terms as shown in Table 1. The table includes a percent of the tweets that used any of the n-grams or unigrams within each topic. Although counting tweets that have any of the topic terms may be considered generous, we decided that within the study's context this measurement provides a relatively simple metric to evaluate the frequency of term usage.

Several themes are identifiable from the LDA output: physical activity, obesity, substance abuse, and healthcare. Through these topics we observe that those relating to obesity and weight loss primarily deal with advertising. Additionally, we observe that the terms relating to healthcare refer to current events and some political discourse. Such an analysis across a wide range of conversational data demonstrates the relative high frequency of these health-related topics. However, as we suspected, this analysis fails to detect lower frequency topics, specifically our test topic, tobacco use. In order to "bring out" tobacco use topics, we create a smaller, more focused dataset that we will refer to as the "tobacco subset."

2.3 Tobacco Subset Analysis

To build our "tobacco subset," we query the comprehensive dataset with a small set of tobacco-related terms: "smoking," "tobacco," "cigarette," "cigar," "hookah," and "hooka." We chose these terms because they are relatively unequivocal, and they specifically indicate tobacco use. We chose the term "hookah" because of an established trend, particularly among adolescents and college students, to engage in hookah, or waterpipe, usage [8]. The recent emergence of

Topic	Most Likely Topic Components (n-grams)	%
44	gps app, calories burned, felt alright, race report, weights workout, christus schumpert, 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	0.01
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	0.01
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	0.04
Topic	**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	13.63

Table 1. Comprehensive Dataset Topics Relevant to Public Health. The last column is the percent of tweets that used any of the n-grams or unigrams within each topic.

hookah bars provides additional evidence of the popularity of hookah smoking [13].

Querying the comprehensive dataset with these terms, results in a subset of 1,963 tweets. The subset is approximately 0.1% of the comprehensive dataset. After running LDA on this subset, we found that there were insufficient data to determine relevant tobacco-related topics. We thus extended the tobacco subset to include tweets that were collected and preprocessed in the same manner as the comprehensive dataset. The only difference is that we used tweets collected from a 4-week period between October 4, 2010 and November 3, 2010. This resulted in a larger subset of 5,929,462 tweets, of which 4,962 were tobacco-related tweets (approximately 0.3%). Because of the limited size of the subset, we reduce the number of topics returned by LDA to 5, while we retain the output of both unigrams and n-grams.

The topics displayed in Table 2, contain several interesting themes relating to how Twitter users discuss tobacco-related topics. Topic 1 contains topics related not only to tobacco use, but also terms that relate to substance abuse including marijuana and crack cocaine. Topic 2 contains terms that relate to addiction recovery: quit smoking, stop smoking, quitting smoking, electronic cigarette, smoking addiction, quit smoking cigarettes, link quit smoking, and link holistic remedies. Topic 3 is less cohesive, and contains terms relating to addiction recovery: quit smoking, stop smoking, secondhand smoke, effective steps. Additionally, it contains words relating to tobacco promotion by clubs or bars: drink specials, free food, ladies night. Topic 4 contains terms related to

Topic	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 ready, twenty years people	0.16
2	quit smoking, stop smoking, 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 smoke, link quit smoking naturally, smoking pot, youtube video, link quit smoking, link holistic remedies, chronical protect	0.27
3	cigarette smoke, dont smoke, quit smoking, stop 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	0.22
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	0.25
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	0.06

Table 2. Tobacco Subset Topics. The most likely topic components for each of the five topics generated by LDA. The last column is the percent of tweets that used any of the n-grams within each topic.

both promotion by bars or clubs (ladies free, piedmont cir, hookahs great food, smoking room, halloween party, million people) and marijuana use (smoking weed, pot smoking). Topic 5 contains several terms that relate to anti-smoking and addiction recovery themes: stopped smoking, smoking kills, chain smoking, ban smoking, people die, damn cigarette, hate cigarettes. In this case, topic modeling has helped to understand more fully how users are using tobacco-related tweets.

3 Discussion and Conclusion

In this study, we directly addressed the problem of how to effectively identify and browse health-related topics on Twitter. We focus on our test topic, tobacco use, throughout the study to explore the realistic application and effectiveness of LDA to learn more about health topics and behavior on Twitter. As expected, we determined that implementing LDA over a large dataset of tweets provides

very few health topics. The health topics that LDA does produce during this first stage suggests the popularity of these topics in our dataset. The topic relating to weight loss solutions indicate a high frequency of advertisements in this area. The topic relating to healthcare, Obama, and other current health issues indicate a current trend in political discourse related to health. Additionally, the high frequency of marijuana-related terms indicates a potentially significant behavior risk that can be detected through Twitter. While this method did not detect lower frequency topics, it may still provide public health researchers insight into popular health-related trends on Twitter. This suggests that there is potential research needed to test LDA as an effective method to identify health-related trends on Twitter. The first method can provide answers to research questions regarding what topics in general are most discussed among Twitter users. The second method we used to identify tobacco-related topics appeares to be most promising to identify and understand public health topics. Although this method is less automated and requires us to choose terms related to tobacco use, the results indicated this method may be a valuable tool for public health researchers.

Based on the results of our topic model, Twitter has been identified as a potentially useful tool to better understand health-related topics, such as tobacco. Specifically, this method of Twitter analysis enables public health researchers to better monitor and survey health status in order to solve community health problems [4]. Our results suggest that chronic health behaviors, like tobacco use, can be identified and measured over shorter periods of time. However, our results do not verify the extent to which short term health events like disease outbreaks can or should be surveyed with this methodology, since tobacco use was used as a test topic. Also, we suspect that the demographics of Twitter users may affect the extent to which topics are discussed on Twitter. We suspect that different public health-related topics may be more or less frequently used on Twitter, and we recommend additional study in this area. Because LDA generates relevant topics relating to tobacco use, we are able to determine themes and also the manner in which tobacco is discussed. In this way, irrelevant conversations can be removed, while tweets related to health status can be isolated. Additionally, by identifying relevant tweets to monitor health status, public health professionals are able to create and implement health interventions more effectively. Researchers can collect almost limitless Twitter data in their areas that will provide practitioners with useful, up-to-date information necessary for understanding relevant public health issues and creating targeted interventions.

Finally, the results from the second method suggest that researchers can better understand how Twitter, a popular SNS, is used to promote both positive and negative health behaviors. For example, Topic 4 contains terms that indicate that establishments like bars, clubs, and restaurants use Twitter as a means to promote business as well as tobacco use. In contrast, Topic 2 contains words that relate to addiction recovery by promoting programs that could help individuals quit smoking.

The use of LDA in our study demonstrates its potential to extract valuable topics from extremely large datasets of conversational data. While the method

proves a valuable outlet to automate the process of removing irrelevant information and to hone in on desired data, it still requires careful human intervention to select query terms for the construction of a relevant subset, and subsequent analysis to determine themes. Research is required to further automate this process. In particular, new methods that can identify infrequent, but highly relevant topics (such as health) among huge datasets will provide value to public health researchers and practitioners, so they can better identify, understand, and help solve health challenges.

References

1. Armour, B.S., Woolery, T., Malarcher, A., Pechacek, T.F., and Husten, C. (2005). Annual Smoking-Attributable Mortality, Years of Potential Life Lost, and Productivity Losses. *Morbidity and Mortality Weekly Report*, **54**: 625-628.
2. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**:993-1022.
3. Boyd, D.M., and Ellison, N.B. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, **13**:210-230.
4. Centers for Disease Control and Prevention. http://www.cdc.gov/od/ocphp/nphpsp/essentialphservices.htm
5. Pear Analytics. http://www.pearanalytics.com/blog/2009/twitter-study-reveals-interesting-results-40-percent-pointless-babble/
6. Chew, C.M., Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of "tweets" During the 2009 H1N1 Outbreak. *Public Library of Science*, **5**(11):e14118. (Paper presented 09/17/09 at *Medicine 2.0*, Naastricht, NL).
7. Culotta, A. (2010). Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In *Proceedings of the KDD Workshop on Social Media Analytics*.
8. Eissenberg, T., Ward, K.D., Smith-Simone, S., and Maziak, W. (2008). Waterpipe Tobacco Smoking on a U.S. College Campus: Prevalence and Correlates. *Journal of Adolescent Health*, **42**:526-529
9. Griffiths, T.L., and Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences*, **101**:5228-5235
10. Haythornthwaite, C. (2005). Social Networks and Internet Connectivity Effects. *Information, Communication, & Society*, **8**:125-47.
11. Healthy People 2010. http://www.healthypeople.gov/lhi/
12. Mokdad, A.H., Marks, J.S., Stroup, D.F., and Gerberding, J.L. (2004). Actual Causes of Death in the United States. *Journal of the American Medical Association*, **291**:1238-1245.
13. Primack, B.A., Aronson, J.D., and Agarwal, A.A. (2006). An Old Custom, a New Threat to Tobacco Control. *American Journal of Public Health*, bf 96:1339.
14. Scanfield, D., Scanfield, V., and Larson, E. (2010). Dissemination of Health Information through Social Networks: Twitter and Antibiotics. *American Journal of Infection Control*, **38**:182-188.
15. Twitter API documentation. http://dev.twitter.com/doc
16. U.S. Department of Health and Human Services (2004). The Health Consequences of Smoking: A Report for the Surgeon General. Report, USDHHS, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.
17. Quantcast. http://www.quantcast.com/twitter.com