Published in final edited form as:

Tob Control. 2015 March ; 24(2): 136–138. doi:10.1136/tobaccocontrol-2013-051243.

Applying linguistic methods to understanding smoking-related conversations on Twitter

Ashley Sanders-Jackson^{1,2}, Cati G Brown^{1,2}, and Judith J Prochaska²

¹Center for Tobacco Control Research and Education, University of California San Francisco, San Francisco, California, USA

²Department of Medicine, Stanford Prevention Research Center, Stanford University, Stanford, California, USA

Abstract

Introduction—Social media, such as Twitter, have become major channels of communication and commentary on popular culture, including conversations on our nation's leading addiction: tobacco. The current study examined Twitter conversations following two tobacco-related events in the media: (1) President Obama's doctor announcing that he had quit smoking and (2) the release of a photograph of Miley Cyrus (a former Disney child star) smoking a cigarette. With a focus on high-profile individuals whose actions can draw public attention, we aimed to characterise tobacco-related conversations as an example of tobacco-related public discourse and to present a novel methodology for studying social media.

Methods—Tweets were collected 11–13 November 2011 (President Obama) and 1–3 August 2011 (Miley Cyrus) and analysed for relative frequency of terms, a novel application of a linguistic methodology.

Results—The President Obama data set (N=2749 tweets) had conversations about him quitting tobacco as well as a preponderance of information on political activity, links to websites, racialised terms and mention of marijuana. Websites and terms about Obama's smoke-free status were most central to the conversation. In the Miley Cyrus data (N=4746 tweets), terms that occurred with the greatest relative frequency were positive, emotional and supportive of quitting (eg, love, and please), with words such as 'love' most central to the conversation.

Conclusions—People are talking about tobacco-related issues on Twitter, and semantic network analysis can be used to characterise on-line conversations. Future interventions may be able to harness social media and major current events to raise awareness of smoking-related issues.

Copyright Article author (or their employer) 2013.

Correspondence to Dr Ashley Sanders-Jackson, Center for Tobacco Control Research and Education, University of California San Francisco, 530 Parnassus Avenue, Ste. 366, San Francisco, CA 94143, USA; asnsande@stanford.edu.

Contributors AS-J: conceived the idea of this article, led the analysis and drafted the first version of this article. CGB: provided guidance on cleaning the linguistic data, participated in the design of the analyses and substantially revised this article. JJP: provided guidance on all steps of the process, including idea inception and discussion of analytic strategies, and she completed substantial revisions of this article.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

INTRODUCTION

The decline in adult smoking prevalence in the USA has stalled, and innovative approaches are needed to engage smokers and support them with quitting smoking for good.¹ There is increasing interest in social media as a promising new channel to promote cessation and health-related behaviour change.²

Twitter is a social media microblogging site (each message, or 'tweet', is 140 characters or less) accessible across a wide variety of platforms (eg, web and mobile). The tweets are published to the sender's network almost instantaneously, enabling rapid information exchange. Twitter users can identify their comments as belonging to a particular topic by using a hashtag (#) as part of the tweet, for example, #welovemiley. Twitter use continues to grow in the USA. In 2012, 15% of on-line adults used Twitter³ compared with 11% in 2011⁴ and 4% in 2010.⁵

Twitter is particularly promising as a tool for smoking cessation due to its accessibility, increasing popularity and the social and informational nature of tweeting. Twitter seems well suited for disseminating evidence-based treatment recommendations widely within a supportive social network consisting of people going through a similar life changes (ie, quitting smoking) in shared time. However, we are aware of only one study to date that has reported on the use of Twitter for quitting smoking.⁶ It reported growth in quit smoking-related user accounts from 2007 to 2010 with much of the content commercial in nature and lacking strong evidence base. What has not been explored is how smoking-related issues are communicated more broadly on Twitter.

With a focus on high-profile individuals whose actions can draw public attention, the current study aimed to characterise tobacco-related conversations on Twitter to describe on-line rhetoric. Both conversations were preceded by a precipitating event. First, former Disney Star (Miley Cyrus) was photographed smoking a cigarette. Second, President Barack Obama's doctors' announced that he had quit smoking. To describe the Twitter data, we completed a relative frequency analysis of the most common terms. Applying the linguistic methodology of comparing relative frequency of terms, we are able to describe what is most often discussed in these on-line conversations. The current study is novel in its examination of naturally occurring (ie, organic) conversations related to both youth smoking (prevention) and adult cessation (treatment) with a focus on individuals of national recognition: the US president and a television and movie star popular among youth.

METHODS

Data collection

Functioning as important opinion leaders and national figures, President Obama, the 44th and current US president, had over 32 million followers on Twitter and Miley Cyrus, an actress and singer, had numerous Billboard top 10 musical hits and over 12 million Twitter followers. A smoker for roughly 30 years, President Obama had been attempting to quit since 2006.⁷⁸ Miley Cyrus, who had previously maintained her Disney-mandated image of not engaging in high-risk behaviours, was caught in a photograph smoking cigarettes.

Although Cyrus was later photographed engaging in numerous high-risk behaviours, this was a major turning point in how Cyrus's fans perceived her behaviour.

Tweets concerning President Obama quitting smoking were collected 11–13 November 2011 with search terms such as 'Obama and smoke-free', 'President and tobacco' and 'Barack and smoker'. Data were collected using *Archivist*, a shareware tool for tweet capture. The tweets concerning Miley Cyrus being caught smoking were collected between 1 and 3 August 2011 with search terms such as 'Miley Cyrus and smoking' and 'Hannah Montana and smoking'. Terms for the Cyrus data were collected using the Twitter application programming interface (API) directly. Additional terms (eg, hashtags occurring >5 times in a search) were added as the data collection progressed. The searches ended when there were <5 new tweets per hour.

Data cleaning

Non-English tweets were excluded, capitalisation was not retained and high-frequency stopwords (eg, *the* and *an*) were removed. A cluster analysis of word relationships was completed in UCInet⁹ and used in the creation of synonym sets for the Obama data, and those synonyms were used for both data sets. For example, 'Prez' and 'President' were treated as equivalent terms. For the Miley Cyrus data, due to the sheer number of messages and the number of spelling errors, the cluster analysis was difficult to interpret. Instead of an automated cluster analysis based on the Miley Cyrus data, synonym sets from the Obama data were applied to the Miley Cyrus data. Additionally, obvious synonyms and variants of terms were paired: for example, 'cig,' 'cigs,' 'cigarette' and 'cigarettes' were all treated as synonyms. In the Obama data set, websites were assigned the value 'http', whereas in the Cyrus data set they were assigned the value 'web'. Usernames and the term 'rt' (short for 'retweet', which is the verbatim forwarding of another's tweet) were removed from the messages except where they were used as content within the tweet, for example, '@Miley Not Smoke'. Redundant tweets and redundant retweets (we retained the first retweet only) were dropped from the analysis leaving a final sample of N=2749 unique President Obama tweets and retweets and N=4746 unique Miley Cyrus tweets and retweets. Thus, the relative frequency analysis reflected the unique contribution of terms as opposed to aggregate frequency of terms across messages that were repeated. This approach is particularly appropriate for studies of message features (such as specific words used or the presence of hashtags) as opposed to studies of content saturation in a mediated environment.

Following data cleaning, relative frequencies of the most common terms were produced using Automap.⁹

RESULTS

Table 1 shows the terms with the greatest relative frequency for both data sets. Overall, the President Obama data set had more consistent use of terms than the Miley Cyrus data set, likely due to more subconversations and greater use of slang by Cyrus' followers, which decreased conversational consistency. The most frequently used words (excluding the search terms) in the President Obama data set had a relative frequency of 0.60 compared with 0.25 in the Miley Cyrus data set. In the President Obama data, terms were mixed across topics,

‘cain’ (a political rival), ‘http’ (the coding for the presence of a website) and ‘weed’ (a reference to marijuana) occurred with some significant frequency along with words related to smoking (five terms in the top 17) and health-related information (three terms in the top 17). Hashtags were not used with enough consistency to appear in the terms with the greatest relative frequency in the President Obama data set. With the Miley Cyrus data set, frequent terms were largely supportive in nature (eg, ‘love’ and ‘heart13’) and included hashtags (eg, ‘#welovemileynomatterwhat’).

A semantic network analysis was completed using Automap.¹⁰ A semantic network analysis is similar to a social network analysis, but compares the strength of connections between words as opposed to people.¹¹ Semantic network analysis has been applied in other tobacco control¹² and social media contexts.^{13,14} However, it has not been applied to a study of social-mediated tobacco-related information. The results indicated that the majority of conversations were tobacco related (see online supplementary appendix 1).

DISCUSSION

The current study applied a novel methodology to characterise on-line social media tobacco-related conversations concerning two high-profile individuals (Miley Cyrus and President Obama) with purportedly different audiences that may have different demographic and regional characteristics. The conversations, which were precipitated by events covered by the mainstream media, occurred on Twitter and in relatively high volume and speed (2700–4700 tweets in 3 days) with most of the conversation centred on tobacco use.

In the President Obama data, the most central terms were about smoking cessation. However, there were also many frequent terms political in nature. In the Miley Cyrus data set, the types of terms that were most central were also the most frequent (positive and emotional), and most terms were about smoking cessation. The conversations were different across data sets, perhaps reflecting not only the difference in news (cessation vs initiation of tobacco use) but also likely reflective of the demographic of the on-line followers. Notably, the volume of tweets was 73% greater for Cyrus versus Obama despite a 2.7-fold greater number of followers to the president's Twitter account. Perhaps utilisation of Twitter between the two groups differed such that Miley followers used Twitter as a platform for social networking and two-way communication, whereas Obama followers used Twitter as a news source and one-way form of communication.

The study findings indicate that large groups of Twitter followers can be engaged in conversation about smoking cessation and prevention. Although it may not be feasible (or advisable) for public health organisations to mimic the language used in naturally occurring on-line conversations, the findings highlight the value of focusing on current events and information that is topically relevant to particular populations. As a public health strategy, it may be helpful to employ tweeters experienced with producing media for a population of interest, for example, youth. Further, our results describe the emotional and cognitive perspectives of the public towards tobacco and may suggest the potential for crowdsourcing as a method to engage large groups in national health-related campaigns (eg, banning

tobacco advertising directed towards youth) by tapping opinion leaders online in support of campaigns.

These data are cross-sectional in nature, and no attempt was made to observe changes in on-line conversations longitudinally. Further, the data used in this analysis are not representative of Twitter or social media as a whole, and although it appeared that Cyrus' fans were younger, Twitter use tends to skew to a middle-aged or older population. Additionally, we chose to drop redundant tweets and retweets in order to capture the breadth of information discussed. This allowed us to give equivalent strength to each tweet in the analysis and prevented the analysis from being highly skewed towards a small subset of messages; however, it did affect the interpretation of the relative frequency analysis. Finally, both of these case studies occurred in the USA with English language tweets, further limiting the generalisability of the findings.

Relative frequency analysis provided an automated tool for understanding the contents of these Twitter conversations and an ability to understand how particular on-line communities use language. Understanding how on-line communities use language may provide researchers with a better way to intervene, both by inducing conversations on health-related topics (perhaps with a focus on celebrities) and by providing health practitioners and scholars with language that may be more likely to be of interest to targeted groups. This study provides two methodological approaches for studying tobacco-related discourse online that can be applied across websites and social media (eg, Facebook, Twitter and Tumblr) contexts both in tobacco control and other arenas.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding This research was supported in part by grants from the National Institutes of Health, including #T32 CA113710, #R01 MH083684, #P50 DA09253 and #CA-113710, and by the State of California Tobacco-Related Disease Research Program #17RT-0077.

REFERENCES

- Centers for Disease Control and Prevention. Adult Cigarette Smoking in the United States: current Estimate.. Secondary Adult Cigarette Smoking in the United States: Current Estimate. 2013. Published June 2013. http://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/ (accessed Jun 2013)
- Neiger BL, Thackeray R, Burton SH, et al. Evaluating social media's capacity to develop engaged audiences in health promotion settings use of twitter metrics as a case study. *Health Promot Pract*. 2013; 14:157–62. [PubMed: 23271716]
- Pew Internet and American Life. Pew Internet: full Report. Secondary Pew Internet: Full Report 2012. Published 2012. <http://www.pewinternet.org/Commentary/2012/March/Pew-Internet-Social-Networking-full-detail.aspx> (accessed Dec 2012)
- Pew Internet and American Life. Twitter Update 2011.. Secondary Twitter Update 2011. Published 2011. <http://www.pewinternet.org/Reports/2011/Twitter-Update-2011.aspx> (accessed Dec 2012)
- Edison Research. Twitter Usage in America 2010.. Secondary Twitter Usage in America 2010. Published 2010. http://www.edisonresearch.com/twitter_usage_2010.php (accessed Dec 2012)

6. Prochaska JJ, Pechmann C, Kim R, et al. Twitter=quitter? An analysis of Twitter quit smoking social networks. *Tob Control*. 2012; 21:447–9. [PubMed: 21730101]
7. Roberts, H. Secondary [internet] 2011. Available November, 2011. <http://www.dailymail.co.uk/news/article-2055912/He-tobacco-free-Obamas-doctor-confirms-president-quit-smoking.html> (accessed May 2013)
8. The Telegraph [internet]. Barak Obama quits smoking after 30 years.. Secondary Barak Obama quits smoking after 30 years. Available November, 2011. <http://www.telegraph.co.uk/news/worldnews/barackobama/8314049/Barack-Obama-quits-smoking-after-30-years.html> (accessed May 2013)
9. Borgatti, SP.; Everett, MG.; Freeman, LC. Ucinet for windows: software for social network analysis. Analytic Technologies; Harvard, MA: 2002.
10. Carley KM, Diesner J. AutoMap: software for Network Text Analysis. 2005
11. Ferrer ICR, Sole R. The small world of human language. *Proc R Soc London Biol*. 2001; 268:2261–5.
12. Linkovich-Kyle TL, Schreiner AM, Dunn ME. Modeling the activation of tobacco smoking expectancies in memory in relation to use patterns. *Addict Behav*. 2012; 37:528–32. [PubMed: 22178600]
13. Web science 2.0: Identifying trends through semantic social network analysis. IEEE; 2009.
14. Gonçalves B, Perra N, Vespignani A. Modeling users' activity on Twitter networks: validation of Dunbar's number. *PLoS One*. 2011; 6:e22656. [PubMed: 21826200]

What this article adds

- ▶ This article presents a data analytic method that can be applied across large data sets of text to answer tobacco-related and health-related questions.
- ▶ This article describes two real-world on-line conversations about tobacco-related issues that can inform how researchers and public health practitioners produce and disperse smoking-related information.
- ▶ Tobacco-related issues are currently being discussed on Twitter, public health professionals may be able to obtain information about what is topically relevant to populations of interest using Twitter.

Table 1

Terms with the greatest relative frequency in both data sets

President Obama		Miley Cyrus	
Term	Relative frequency	Term	Relative frequency
Co	0.60	Love	0.21
Http	0.60	Never	0.12
President	0.37	Heart13	0.10
Tobacco	0.27	Know	0.09
Free	0.27	Welovemileynomatterwhat	0.08
Quit	0.15	Please	0.08
Cain	0.14	People	0.08
Fit	0.12	Want	0.08
Doctor	0.12	3	0.07
Barack	0.11	She	0.07
Physical	0.09	Life	0.07
Weed	0.09	Care	0.06
Cigarettes	0.09	Matter	0.05
Health	0.09	Now	0.05
Ad	0.08	Watching	0.05
Smoker	0.07	Mileypleasestopsmoking	0.05

The name of the person being searched or a version of that name (eg, Miley or Obama) was the most frequent term and occurred in all tweets. It was therefore removed. 'Smoke' occurred with the second greatest frequency, an artefact of the searching procedure, and it too was removed. 'heart13' is an emoticon translated into ascii text; 'welovemileynomatterwhat' and 'mileypleasestopsmoking' are both hashtags. The presence of 'http' denotes a website that was substituted for the presence of a URL or website; 'co' was often used instead of the word company in a variety of different contexts and the term 'cain' is in reference to Herman Cain, a possible Republican presidential candidate at the time.