ISDS 2012 Conference Abstracts

Content Analysis of Tobacco-related Twitter Posts

Mark Myslín*, Shu-Hong Zhu and Michael Conway

UC San Diego, San Diego, CA, USA

Objective

We present results of a content analysis of tobacco-related Twitter posts (tweets), focusing on tweets referencing e-cigarettes and hookah.

Introduction

Vast amounts of free, real-time, localizable Twitter data offer new possibilities for public health workers to identify trends and attitudes that more traditional surveillance methods may not capture, particularly in emerging areas of public health concern where reliable statistical evidence is not readily accessible. Existing applications include tracking public informedness during disease outbreaks [1].

Twitter-based surveillance is particularly suited to new challenges in tobacco control. Hookah and e-cigarettes have surged in popularity, yet regulation and public information remain sparse, despite controversial health effects [2,3]. Ubiquitous online marketing of these products and their popularity among new and younger users make Twitter a key resource for tobacco surveillance.

Methods

We collected 7,300 tobacco-related Twitter posts at 15-day intervals from December 2011 to July 2012, using ten general keywords such as cig* and hookah. Each tweet was manually classified using a tri-axial scheme, capturing genre (firsthand experience, joke, news, …), theme (underage usage, health, social image, …), and sentiment (positive, negative, neutral). Machine-learning classifiers were trained to detect tobacco-related vs. irrelevant tweets as well as each of the above categories, using Naïve Bayes, k-Nearest Neighbors, and Support Vector Machine algorithms. Finally, phi correlation coefficients were computed between each of the categories to discover emergent patterns.

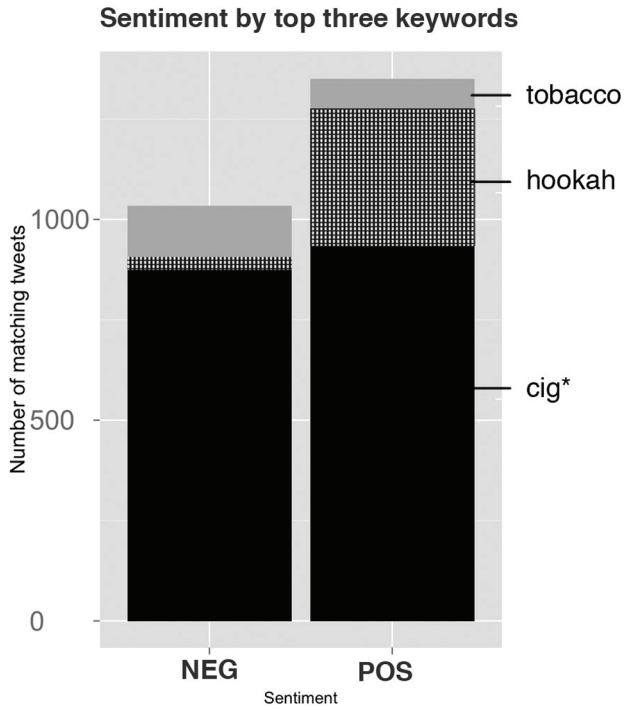
Results

The most prevalent genre of tweets was personal experience, followed by categories such as opinion, marketing, and news. The most common themes were hookah, cessation, and social image, and sentiment toward tobacco was more positive (26%) than negative (20%). The most highly correlated categories were social image–underage, marketing–e-cigs, and personal experience–positive sentiment. E-cigarettes were also correlated with positive sentiment and new users (even excluding marketing posts), while hookah was highly correlated with positive sentiment, pleasure, and social relationships. Further, tweets matching the term "hookah" reflected the most positive sentiment, and "tobacco" the most negative (Figure 1). Finally, negative sentiment correlated most highly with social image, disgust, and non-experiential categories such as opinion and information.

The best machine classification performance for tobacco vs. non-tobacco tweets was achieved by an SVM classifier with 82% accuracy (baseline 57%). Individual categories showed similar improvements over baseline.

Conclusions

Several novel findings speak to the unique insights of Twitter surveillance. Sentiment toward tobacco among Twitter users is more positive than negative, affirming Twitter's value in understanding positive sentiment. Negative sentiment is equally useful: for example, observed high correlations between negative sentiment and social image, but not health, may usefully inform outreach strategies. Twitter surveillance further reveals opportunities for education: positive sentiment toward the term "hookah" but negative sentiment toward "tobacco" suggests a disconnect in users' perceptions of hookah's health effects. Finally, machine classification of tobacco-related posts shows a promising edge over strictly keyword-based approaches, allowing for automated tobacco surveillance applications.



Sentiment in "hookah" tweets is disproportionately more positive than in "cig" and especially "tobacco" tweets.

Keywords

social media; surveillance; Twitter; tobacco; NLP

References

[1] Chew, C. & Eysenbach, G. Pandemics in the age of twitter: Content analysis of tweets during the 2009 H1N1 outbreak. PLoS One. 2010; 5(11):e14118.

[2] Ayers, J. W., Ribisl, K. M., & Brownstein, J. S. Tracking the rise in popularity of electronic nicotine delivery systems (Electronic cigarettes) using search query surveillance. Am J Prev Med. 2011; 40(4):448–453.

[3] Grekin, E. R. & Ayna, D. Waterpipe smoking among college students in the United States: A review of the literature. J Am Coll Health. 2012; 60(3):244–249.

*Mark Myslín
E-mail: mmyslin@gmail.com