OJPHI

ISDS 2014 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Combining Text Mining and Data Visualization Techniques to Understand Consumer Experiences of Electronic Cigarettes and Hookah in Online Forums

Annie T. Chen[2], Shu-Hong Zhu[1] and Mike Conway*[1]

[1]Department of Family and Preventative Medicine, University of California San Diego, La Jolla, CA, USA; [2]School of Information and Liberty Science, UNC, Chapel Hill, NC, USA

Objective

Our aim in this work is to apply text mining and novel visualization techniques to textual data derived from online health discussion forums in order to better understand consumers' experiences and perceptions of electronic cigarettes and hookah.

Introduction

Since their introduction to the US market in 2007, electronic cigarettes (e-cigarettes) have posed considerable challenges to both public health authorities and government regulators, especially given the debate – in both the scientific world and the community at large – regarding the potential *advantages* (e.g. helping individuals quit smoking) and *disadvantages* (e.g. renormalizing smoking) associated with the product[1]. Similarly, hookah – a kind of waterpipe used to smoke flavored tobacco – has increased in popularity in recent years, is known to be particularly popular among younger people, and has prompted a range of regulatory responses[2]. One important – and currently largely unexplored – area of research involves exploring consumer perceptions and experiences of these emerging tobacco products. In this work, we use online health discussion forums in conjunction with text mining and novel data visualization techniques to investigate consumer perceptions and experiences of e-cigarettes and hookah, focusing on the automatic identification of *symptoms* associated with each product, and consumer *motivations* for product use.

Previous related research has focused on using text-mining to analyze e-cigarette or hookah related Twitter posts[3,4] and on the qualitative identification of e-cigarette related symptoms from online discussion forums[5]. The research reported in this abstract is – to the best of our knowledge – the first time that text mining techniques have been used with online health forums to understand e-cigarette or hookah use.

Methods

Data were automatically crawled from three different sources: VaporTalk (www.vaportalk.com – a forum devoted to e-cigarettes), Hookah Forum (www.hookahforum.com – a hookah discussion forum), and Reddit (www.reddit.com – a popular general forum with *stopsmoking*, *e-cig*, and *hookah* "subreddits"). We used two broad approaches to text mining the data. First, we iteratively developed bespoke lexicons representing dimensions of health behavior – e.g. *symptoms*, *cost*, *quitting* – and calculated the proportion of posts in which words from a particular category occurred, allowing us to compare across forums. Second, we used *topic modeling*[6] – a set of techniques drawn from the Natural Language Processing and Machine Learning communities that allow for the automatic identification of topics, as represented by popular key words – to analyze the text. As part of this work, we have developed a novel, interactive visualization system (implemented in Python and D3) for the analysis and summarization of forum data.

Results

Several unique findings indicate the usefulness of text mining online forum data in conjunction with the use of sophisticated visualization techniques. For example, our analysis indicates that e-cigarette users have a tendency to focus on e-cigarette equipment (vaporizers, liquid types, etc.) hookah users often discuss the sensory experience of smoking (e.g. optimizing "buzz"). Further, our analysis of VaporTalk Health & Safety forum, and the Reddit stopsmoking subreddit indicates key differences in symptom reporting between the two forums, with Vapor|Talk concentrating overwhelmingly on physical symptoms associated with e-cigarette use (e.g. headache, coughing), and the Reddit stopsmoking forum more focused on psychological symptoms (e.g. craving, anxiety).

Keywords

Natural Language Processing; Text Mining; Informatics

Acknowledgments

This work was supported by a grant from the National Cancer Institute U01 CA154280.

References

1. WHO Framework Convention on Tobacco Control. Electronic Nicotine Delivery Systems. 2014
2. Grekin E, Ayna D. Waterpipe smoking among college students in the United States: a review of the literature. J Am Coll Health 2012;60:244-9
3. Myslín M, Zhu SH, Chapman W, Conway M.Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. J Med Internet Res 2013;15(8):e174
4. Huang J, Kornfield R, Szczypka G, Emery S. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tob Control 2014;23:iii26-iii30
5. Hua M, Alfi M, Talbot P. Health-Related Effects Reported by Electronic Cigarette Users in Online Forums. J Med Internet Res 2013;15(4):e59
6. Blei D, Ng A, Jordan M. Latent dirichlet allocation. Journal of Machine Learning Research. 2003;3:993–1022.

*Mike Conway
E-mail: mconway@ucsd.edu