Journal of Biomedical Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter

Ramakanth Kavuluru [a,b,*], A.K.M. Sabbir [b]

[a] Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, 230E MDS Building, 725 Rose Street, Lexington, KY 40536, USA
[b] Department of Computer Science, University of Kentucky, Davis Marksbury Building, 329 Rose Street, Lexington, KY 40506, USA

ARTICLE INFO

Article history:
Received 9 September 2015
Revised 4 February 2016
Accepted 4 March 2016
Available online xxxx

Keywords:
Electronic cigarettes
Text mining
Text classification

ABSTRACT

Background: Electronic cigarettes (e-cigarettes or e-cigs) are a popular emerging tobacco product. Because e-cigs do not generate toxic tobacco combustion products that result from smoking regular cigarettes, they are sometimes perceived and promoted as a less harmful alternative to smoking and also as means to quit smoking. However, the safety of e-cigs and their efficacy in supporting smoking cessation is yet to be determined. Importantly, the federal drug administration (FDA) currently does not regulate e-cigs and as such their manufacturing, marketing, and sale is not subject to the rules that apply to traditional cigarettes. A number of manufacturers, advocates, and e-cig users are actively promoting e-cigs on Twitter.

Objective: We develop a high accuracy supervised predictive model to automatically identify e-cig "proponents" on Twitter and analyze the quantitative variation of their tweeting behavior along popular themes when compared with other Twitter users (or tweeters).

Methods: Using a dataset of 1000 independently annotated Twitter profiles by two different annotators, we employed a variety of textual features from latest tweet content and tweeter profile biography to build predictive models to automatically identify proponent tweeters. We used a set of manually curated key phrases to analyze e-cig proponent tweets from a corpus of over one million e-cig tweets along well known e-cig themes and compared the results with those generated by regular tweeters.

Results: Our model identifies e-cig proponents with 97% precision, 86% recall, 91% *F*-score, and 96% overall accuracy, with tight 95% confidence intervals. We find that as opposed to regular tweeters that form over 90% of the dataset, e-cig proponents are a much smaller subset but tweet two to five times more than regular tweeters. Proponents also disproportionately (one to two orders of magnitude more) highlight e-cig flavors, their smoke-free and potential harm reduction aspects, and their claimed use in smoking cessation.

Conclusions: Given FDA is currently in the process of proposing meaningful regulation, we believe our work demonstrates the strong potential of informatics approaches, specifically machine learning, for automated e-cig surveillance on Twitter.

1. Introduction

Electronic cigarettes (e-cigarettes or simply e-cigs) were introduced in the United States (US) in 2007 [1] and are currently a popular emerging tobacco product across the world. An e-cig essentially consists of a battery that heats up liquid nicotine available in a cartridge into a vapor that is inhaled by the user [2]. E-cig users are termed vapers and the process of using an e-cig is called vaping. E-cigs are similar to conventional tobacco cigarettes with regards to visual, sensory, and behavioral aspects and hence were observed to reduce craving [3]. Owing to their recent introduction, there are very few studies on e-cig safety, risk of abuse, and their efficacy as a smoking cessation aid especially about long term use effects. In fact, currently the search phrase `electronic nicotine delivery systems OR e-cigarette OR electronic cigarette` with its plural/hyphenated variants yields 1794 articles in the PubMed search system out of which 1459 (≈81%) had dates of publication in 2014 or 2015. Because e-cigs do not generate toxic combustion products that are produced with tobacco cigarettes, they are perceived and also sometimes marketed as

* Corresponding author at: Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, 230E MDS Building, 725 Rose Street, Lexington, KY 40536, USA.
E-mail addresses: ramakanth.kavuluru@uky.edu (R. Kavuluru), akm.sabbir@uky.edu (A.K.M. Sabbir).

http://dx.doi.org/10.1016/j.jbi.2016.03.006
1532-0464/

suitable alternatives for smoking cessation [4]. However, scientific research to verify these claims is limited and is often inconclusive. On one hand there are studies that indicate comparable or superior effectiveness of e-cigs in smoking cessation [5,6]. However, there are also results [7,8] that show no such associations exist between e-cig use and quitting or reduced conventional cigarette consumption. Another recent effort [9] also indicates that passive exposure to e-cigs increases the desire to smoke both regular cigarettes and e-cigs. Nevertheless, current research seems to indicate that they are less harmful than traditional cigarettes [10].

The ongoing healthy scientific debate around e-cigs is welcomed by the society, especially by regular smokers who are interested in quitting or adopting less harmful alternatives. However, lack of FDA regulation (except for therapeutic use) has heavily increased marketing of e-cigs on the Web [11] and through television ads [12] even if individual states have recently started to enact their own regulations to limit sales, marketing, and use [13]. According to a 2013 Centers for Disease Control and Prevention (CDC) report [14], e-cig consumption doubled in middle and high school students from 2011 to 2012. Furthermore, 9.3% of middle and high school ever e-cig users in 2012 have never smoked conventional cigarettes. Alarmingly, this percentage goes up to 20.3% when considering only middle school students. A more recent CDC report [15] shows that e-cig use tripled from 2013 to 2014 among middle and high school students. Since long term safety of e-cigs has not been thoroughly studied yet, the prospect of adolescents developing nicotine dependence could be detrimental to public health in future generations. When considering adult smokers, however, the significant increase in e-cig awareness has reduced their perception of e-cigs as being less harmful compared with regular cigarettes [16]. Since Web based advertising and discussion still plays a major role in e-cig marketing and use and given one in four online US teenagers uses Twitter [17], we believe it is critical to study the landscape of e-cig messages and their authors on Twitter.

Although e-cig message themes and author classification might be highly granular, in this pilot project we take a simpler approach to tweet author classification – each tweeter is either a "proponent" or not for our purposes. Proponents are tweeters who represent e-cig sales or marketing agencies, individuals who advocate e-cigs, or tweeters who specifically identify themselves as vapers in their profile bio. Essentially these tweeters are generally more inclined to support e-cigs regardless of their specific motivation (e.g., business, lobbying, smoking cessation). In this paper, based on a hand-labeled dataset of 1000 tweeter profiles, we build machine learned models to automatically identify proponents. We subsequently use this model to analyze the content of tweets generated by proponents in comparison with other tweeters along several well known e-cig themes (e-cig flavors, harm reduction, smoke-fee aspect, and smoking cessation) using straightforward text processing. We demonstrate that proponents are many times as likely to highlight the attractive (and sometimes scientifically not yet verified) aspects of e-cigs compared with regular tweeters. To our knowledge this is the first attempt in identifying proponents and a first step in building a framework for automatic surveillance of e-cig related chatter.

2. Background and related work

Since its introduction in 2006, Twitter has grown into one of the top 15 visited websites [18] in the world with 100 million daily active users who generate over 500 million tweets per day [19]. The asymmetric network structure of Twitter inherently supports information diffusion and given that a recent study [20] reveals that over 95% of Twitter profiles are public, mining tweets is a practical tool to measure user engagement with various events and products. Since users are not required to publicly declare personal information, several recent studies have focused on identifying user demographic attributes such as age groups and life stages [21], gender [22], and race and ethnicity [23].

In the context of public health, Twitter based automatic syndromic surveillance has been shown to have high correlation with traditional surveillance methods [24,25] with the added advantage of near real time access to trends, especially in the early epidemic stages [26,27]. Recent efforts also noted Twitter's suitability for promoting health literacy [28], encouraging fitness activity [29], and monitoring drug safety [30]. In the context of tobacco control advocacy, researchers found significant reach through Twitter in obtaining signatures for an online petition to drop tobacco sponsorship for an international music concert in Indonesia [31]. Another recent Twitter based study focused on emerging tobacco products [32] found high prevalence of positive sentiment for hookah and e-cig. It also successfully demonstrated the application of machine learning methods in automatically identifying tobacco related tweets, constituent themes, and sentiments.

The most relevant effort in the context of our paper is from Huang et al. [33] who automatically identify "commercial" tweets from a corpus of nearly 73,000 tweets collected in the months of May and June in 2012. For their purposes tweets that contain links to sales websites and promotional messages are all commercial regardless of who posts them (regular tweeters vs e-cig marketers/advocates). They use DiscoverText, a cloud based commercial text analytics software program, to semi-automatically (Naive Bayes with additional heuristics) classify tweets as commercial or not. They report that 90% of their tweets are commercial and nearly 10% of such tweets mention smoking cessation. In our effort, instead of identifying commercial tweets, we identify e-cig proponent Twitter profiles using supervised machine learning. We believe this is a more direct approach that aids in electronic surveillance efforts needed in the immediate future to monitor e-cig marketing/publicity practices and as such complements Huang et al.'s effort. Furthermore, compared with 2012, there is an order of magnitude increase in the number of e-cig tweets (based on an official quote from Twitter Inc.) and our experiments are conducted on a corpus of over one million e-cig related tweets. After identifying proponent profiles, we conducted analyses along well known e-cig themes to see differences in tweeting behaviors of proponents and regular tweeters.

3. Datasets and annotation

We used two different datasets of e-cig related tweets: the first set of 224,000 tweets was obtained using rate limited Twitter streaming API during the months of September to December 2013 and a second dataset of nearly one million tweets (purchased from the exhaustive Twitter firehose[1]) for the month of March 2015 that match the query terms: `electronic-cigarette`, `e-cig`, `e-cigarette`, `e-juice`, `e-liquid`, `vape-juice`, and `vape-liquid`. Variants of these terms with spaces instead of hyphens or just without the hyphens (for matching hashtags) were also used in the query. These terms were chosen in consultation with a faculty member in the College of Nursing at the University of Kentucky (UKY) who works on tobacco policy research. They are specific enough and empirically shown to result in a 99% match to actual e-cig related tweets [33]. The juice/liquid terms represent the liquid nicotine cartridges that need to be refilled for the vaping devices.

[1] Twitter's terms of use for purchased datasets allow for reporting of aggregate analyses but not presentation of full tweets. Hence all examples shown in this paper are from datasets collected through their rate limited API calls.

Please cite this article in press as: R. Kavuluru, A.K.M. Sabbir, Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on Twitter, J Biomed Inform (2016), http://dx.doi.org/10.1016/j.jbi.2016.03.006