Informatics in Medicine Unlocked 8 (2017) 8–12

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

A tale of two tools: Reliability and feasibility of social media measurement tools examining e-cigarette twitter mentions

Amelia Burke-Garcia[a,*], Cassandra A. Stanton[a,b]

[a] Westat, 1600 Research Blvd, Rockville, MD 20850, United States of America

[b] Department of Oncology, Georgetown University Medical Center/Cancer Prevention and Control Program, Lombardi Comprehensive Cancer Center, United States of America

ARTICLE INFO

Keywords:
E-cigarettes
Vaping
Twitter
Tweets
Social media

ABSTRACT

Given 70% of Americans are seeking health information online, social media are becoming main sources of health-related information and discussions. Specifically, compounding rising trends in use of e-cigarettes in the US, there has been a rapid rise in e-cigarette marketing – much of which is happening on social media. Public health professionals seeking to understand consumer knowledge, attitudes and beliefs about e-cigarettes should consider analyzing social media data and to do so, there are numerous free and paid tools available. However, each uses different sources and processes, which makes data validation challenging. This exploratory study sought to understand the reliability and feasibility of two social media data tools analyzing e-cigarette tweets. Twitter mentions were pulled from two different industry standard tools (GNIP and Radian6) and data were evaluated on six measures, e.g. Cost, Feasibility, Ease of Use, Poster Type (individual/organization), Context (tweet content analysis), and Valence (positive/negative). Findings included similarities amongst the data sets in terms of the content themes but differences in cost and ease of use of the tools themselves. These findings align with prior research, notably that e-cigarette marketing tweets are most common and public health-related content is noticeably absent. Findings from this exploratory study can inform future social media studies as well as communication campaigns seeking to address the emerging issue of e-cigarette use.

1. Introduction

Electronic cigarettes, or 'e-cigarettes' as they are commonly referred to, are battery-powered products that typically deliver nicotine in the form of an aerosol [24]. Data from the most recent 2013–2014 Population Assessment of Tobacco and Health (PATH) Study, which is a large, nationally representative, longitudinal study of tobacco use and health in the United States funded by the National Institute on Drug Abuse and the Food and Drug Administration, found that more than a quarter (27.6%) of adults were current users of at least one type of tobacco product and a total of 8.9% of youths had used a tobacco product in the previous 30 days with 1.6% of youths being daily users [13]. Findings from this study included that approximately 40% of all tobacco users use multiple tobacco products simultaneously [13]. The most common combination was cigarettes and e-cigarettes [13]. Higher rates of tobacco use were reported amongst young adults ages 18 to 24, male adults and youths, members of racial minorities, and members of sexual minorities.

Compounding this issue is the increasing use of Twitter and other social media by the tobacco industry to market e-cigarettes and related products. This is primarily driven by the substantial growth of social media use over the last ten years, with nearly two-thirds of American adults (65%) now using social networking sites [22] and almost half of all social media users interacting on social media for a direct response to a problem or issue [21]. Moreover, much of this social media use is also focused on health-related information dissemination and engagement [22], with data suggesting that social media contribute to "facilitating, sharing, and obtaining health messages" ([17], p. 7). Pew Research Center data validate this finding that social media are becoming main sources of health information and related discussions [7] and that they are increasingly considered more important than health care providers as an information source [14]. Moreover, numerous studies have examined health promotion through social media [30,8], the findings from which indicate potential for using blogs, Twitter and other online communication channels, not only for increasing awareness but also to influence decision making.

Related to e-cigarette marketing, research as recent as 2016 suggests that Twitter communication about e-cigarettes has increased fivefold since 2012 with users overwhelmingly exposed to messages that favor e-cigarettes and most of these have been related to the

* Corresponding author.
E-mail address: ameliaburke-garcia@westat.com (A. Burke-Garcia).

http://dx.doi.org/10.1016/j.imu.2017.04.001
Received 6 January 2017; Received in revised form 20 April 2017; Accepted 22 April 2017
Available online 25 April 2017

marketing of e-cigarettes with public health messages notably absent from the conversation [10,27]. Moreover, Huang et al.'s [10] study examining Twitter mentions of e-cigarettes found that 90% of tweets analyzed were related to the commercialization of e-cigarettes. Van der Tempel et al. [27] found that in a sample of 300 tweets from high-authority users, commercial tweets were positively valenced (supporting e-cigarette use) and the most prevalent message themes were marketing, news, and first-person experiences with e-cigarettes. Finally, Yamin, Bitton, and Bates' [29] work supports this claim, finding that "the presence of e-cigarettes on the Internet, including in Web searches, virtual user communities, and online stores where people sell e-cigarettes on commission, is increasing rapidly" (p. 607).

Such engagement levels in social media – and specifically related to e-cigarette promotion – has resulted in a plethora of social media data. Specifically, Google reports that there are 2 million search requests every minute and Facebook reports that there are almost 700,000 pieces of content posted every minute [6]. For businesses, these data can provide insight in the overall online reputation of their brands, competitors, products and services [25]. Marketing, on the other hand, can use the insight to understand feedback on campaigns and respond appropriately [25]. However, this wealth of activity can also serve other purposes, e.g. it can help public health professionals understand emerging health-related trends like those related to e-cigarette marketing practices.

To understand e-cigarette marketing trends, social data can be collected and analyzed; but the options for mining and analyzing these data are ubiquitous. There are currently more than 200 social media monitoring tools available, which makes it difficult for researchers and practitioners to make an educated choice about which tool to use [25]. Key issues with identifying the right tool for mining social data include quality and validity of the data, cost and usability of the tool [25]. For example, each tool delivers results based on individual algorithms that employ a mix of keywords, users, and geotags, but there is a lack of documentation on these algorithms, which makes data validation challenging [18,26]. These tools vary in price and ease-of-use but evaluating these differences can be hard to decipher for the inexperienced researcher.

Finally, Twitter's enterprise application program interface (API), GNIP, provides full access to tweets which are not available directly from Twitter [1,23,28,2], GNIP appears to be the most frequently referenced Twitter data mining tool in the literature. It is even the choice for archiving Twitter data for the Library of Congress [11,32]. This abates investigation into other data mining options. Moreover, the nuances of GNIP data are often not discussed. For instance, GNIP currently offers three levels of data – a 1% of the data; the "Gardenhose" (otherwise known as the "Decahose") or 10% of the data; and the "Firehose", or access to the full form of GNIP [5], all of which have different price points and provide differing coverage levels of conversations [18,26]. This has resulted in limited discussion of appropriate methods of social media data collection and analysis.

Given the value of these social media data as well as the issues that exist with mining them, understanding the reliability and feasibility of available tools is an important area of research. However, there is little prior research in this area. Hofer-Shall [9] assessed social media mining tools based on three criteria: current offering (services and features offered), strategy (how they address enterprise-level needs) and market presence. Yet, Stavrakantonakis et al. [25] found these criteria insufficient. Therefore, the researchers built on this prior work by conducting their own study where they evaluated 10 social media data mining tools on three measures: the main concepts related to social media monitoring (analysis, insights, engagement, workflow management and influence); the technology used by the tools; and the user interface [25].

The aim of this exploratory study is to compare two tools for data mining the content of social media using e-cigarette tweets in order to better understand the implications of e-cigarette content in social media. This formative study is an initial step to expand research in this area to inform a larger study of measurement, as well as the use of social media for public health communication. In pursuit of these goals, the following research questions were posited:

RQ1. How do two leading social media data collection tools, e.g. Radian6 vs. GNIP, differ in terms of cost, ease-of-use and feasibility?

RQ2. How does Twitter coverage of e-cigarette-related conversations differ by data source (e.g. Radian6 vs. GNIP)?

2. Methods

This pilot study was designed to evaluate process rather than outcome measures and to better understand how a set of tweets from Twitter's GNIP "Firehose" service compares to a set of tweets pulled from Salesforce's Radian6 tool based on the same search parameters. A qualitative approach was utilized to collect and analyze data.

2.1. Sample

Using the key words, "e-cigarettes OR vaping" OR "e-cigarettes health" OR "vaping health", 500 mentions were collected from each tool over a 0:30 period of time (12:57 pm EST on August 7, 2015) for a total of 1000 mentions. This sample was identified based on data made available to the researchers during this time period. It therefore was a convenience sample used to analyze feasibility and validity of the two social media mining tools. Table 1 includes example tweets from the two data sets.

2.2. Procedures

Six measures were used in this analysis. Tools were compared on Cost, Feasibility, and Ease of Use; and mentions were compared on Poster (individual/organization), Context (tweet content analysis), and Valence (positive/negative).

2.2.1. The original coder identified 12 content themes

To assess the tools on the measures of Cost, Feasibility, and Ease of Use, and the content on the measures of Poster (individual/organization), Context (tweet content analysis) and Valence (positive/negative), an original content code frame was developed. The code frame was adapted from a code frame used in unpublished work under the National Institutes of Health's (NIH) National Children's Study [4]. The original coder identified 12 themes that the content was coded on. These were: (1) General Health; (2) Health Consequence; (3) Cessation; (4) Product Characteristics – Brand; (5) Product Characteristics – Flavor; (6) Product Characteristics – Other; (7) Marketing/Sales; (8) Consumer Purchases; (9) Utilization Patterns; (10) Policy; (11) Endorsement; and (12) Other. Given the low number of codes in "General Health", these were combined with the "Health Consequence" codes creating a combined code called, "Health/Consequence". As well, the three individual "Product Characteristics" themes did not have enough codes each to warrant separate themes; therefore, they were combined into one code called "Product Characteristics". This resulted in 9 themes for final analysis. These were: (1) Health/Consequence; (2) Cessation; (3) Product Characteristics; (4) Marketing/Sales; (5) Consumer Purchases; (6) Utilization Patterns; (7) Policy; (8) Endorsement; and (9) Other.

2.2.2. Analysis and coding verification

Tool and content analyses were conducted to evaluate these two data sets. The content was then analyzed through March 26, 2016. Two independent coders used this frame to code the two sets of tweets and then the codes were compared. Coding verification was conducted on March 27, 2016. For this, every other post from the sampled content was selected for testing. Results were compared and intercoder reliability was 90%.