International Journal of Information Management 38 (2018) 1–6

Contents lists available at ScienceDirect

International Journal of Information Management

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

Characterizing diabetes, diet, exercise, and obesity comments on Twitter

Amir Karami[a,*], Alicia A. Dahl[b], Gabrielle Turner-McGrievy[b], Hadi Kharrazi[c], George Shaw Jr.[a]

[a] *University of South Carolina, School of Library and Information Science, United States*
[b] *University of South Carolina, Arnold School of Public Health, United States*
[c] *Johns Hopkins University, Bloomberg School of Public Health, United States*

ARTICLE INFO

Keywords:
Health
Diabetes
Diet
Obesity
Exercise
Topic model
Text mining
Twitter

ABSTRACT

Social media provide a platform for users to express their opinions and share information. Understanding public health opinions on social media, such as Twitter, offers a unique approach to characterizing common health issues such as diabetes, diet, exercise, and obesity (DDEO); however, collecting and analyzing a large scale conversational public health data set is a challenging research task. The goal of this research is to analyze the characteristics of the general public's opinions in regard to diabetes, diet, exercise and obesity (DDEO) as expressed on Twitter. A multi-component semantic and linguistic framework was developed to collect Twitter data, discover topics of interest about DDEO, and analyze the topics. From the extracted 4.5 million tweets, 8% of tweets discussed diabetes, 23.7% diet, 16.6% exercise, and 51.7% obesity. The strongest correlation among the topics was determined between exercise and obesity ($p < .0002$). Other notable correlations were: diabetes and obesity ($p < .0005$), and diet and obesity ($p < .001$). DDEO terms were also identified as subtopics of each of the DDEO topics. The frequent subtopics discussed along with "Diabetes", excluding the DDEO terms themselves, were blood pressure, heart attack, yoga, and Alzheimer. The non-DDEO subtopics for "Diet" included vegetarian, pregnancy, celebrities, weight loss, religious, and mental health, while subtopics for "Exercise" included computer games, brain, fitness, and daily plan. Non-DDEO subtopics for "Obesity" included Alzheimer, cancer, and children. With 2.67 billion social media users in 2016, publicly available data such as Twitter posts can be utilized to support clinical providers, public health experts, and social scientists in better understanding common public opinions in regard to diabetes, diet, exercise, and obesity.

1. Introduction

The global prevalence of obesity has doubled between 1980 and 2014, with more than 1.9 billion adults considered as overweight and over 600 million adults considered as obese in 2014 (World Health Organization Fact Sheet, 2016). Since the 1970s, obesity has risen 37% affecting 25% of the U.S. adults (Flegal, Carroll, Kit, & Ogden, 2012). Similar upward trends of obesity have been found in youth populations, with a 60% increase in preschool aged children between 1990 and 2010 (Harvard HSPH, 2017). Overweight and obesity are the fifth leading risk for global deaths according to the European Association for the Study of Obesity (World Health Organization Fact Sheet, 2016). Excess energy intake and inadequate energy expenditure both contribute to weight gain and diabetes (Hill, Wyatt, & Peters, 2012; Wing et al., 2001).

Obesity can be reduced through modifiable lifestyle behaviors such as diet and exercise (Wing et al., 2001). There are several comorbidities associated with being overweight or obese, such as diabetes (Kopelman, 2000). The prevalence of diabetes in adults has risen globally from 4.7% in 1980 to 8.5% in 2014. Current projections estimate that by 2050, 29 million Americans will be diagnosed with type 2 diabetes, which is a 165% increase from the 11 million diagnosed in 2002 (Boyle et al., 2001). Studies show that there are strong relations among diabetes, diet, exercise, and obesity (DDEO) (Association et al., 2004; Barnard et al., 2009; Hartz, Rupley, Kalkhoff, & Rimm, 1983; Wing et al., 2001); however, the general public's perception of DDEO remains limited to survey-based studies (Tompson et al., 2012).

The growth of social media has provided a research opportunity to track public behaviors, information, and opinions about common health issues. It is estimated that the number of social media users will increase from 2.34 billion in 2016 to 2.95 billion in 2020 (Statista, 2017). Twitter has 316 million users worldwide (Olanoff, 2015) providing a unique opportunity to understand users' opinions with respect to the most common health issues (Mejova, Weber, & Macy, 2015). Publicly available Twitter posts have facilitated data collection and leveraged the research at the intersection of public health and data science; thus,

[*] Corresponding author.
E-mail addresses: karami@sc.edu (A. Karami), adahl@email.sc.edu (A.A. Dahl), mcgrievy@mailbox.sc.edu (G. Turner-McGrievy), kharrazi@jhu.edu (H. Kharrazi), gshaw@email.sc.edu (G. Shaw).

http://dx.doi.org/10.1016/j.ijinfomgt.2017.08.002
Received 28 July 2017; Accepted 8 August 2017

informing the research community of major opinions and topics of interest among the general population (Nasukawa & Yi, 2003; Wiebe et al., 2003; Zabin & Jefferies, 2008) that cannot otherwise be collected through traditional means of research (e.g., surveys, interviews, focus groups) (Eichstaedt et al., 2015; Wartell, 2015). Furthermore, analyzing Twitter data can help health organizations such as state health departments and large healthcare systems to provide health advice and track health opinions of their populations and provide effective health advice when needed (Mejova et al., 2015).

Among computational methods to analyze tweets, computational linguistics is a well-known developed approach to gain insight into a population, track health issues, and discover new knowledge (Moreland-Russell, Tabak, Ruhr, & Maier, 2014; Paul & Dredze, 2011, 2012; Zhao et al., 2011). Twitter data has been used for a wide range of health and non-health related applications, such as stock market (Bollen, Mao, & Zeng, 2011) and election analysis (Tumasjan, Sprenger, Sandner, & Welpe, 2010). Some examples of Twitter data analysis for health-related topics include: flu (Culotta, 2010; Lampos & Cristianini, 2010, 2012; Lampos, De Bie, & Cristianini, 2010; Ritterman, Osborne, & Klein, 2009; Szomszor, Kostkova, & De Quincey, 2010), mental health (Coppersmith, Dredze, Harman, & Hollingshead, 2015), Ebola (Lazard, Scheinfeld, Bernhardt, Wilcox, & Suran, 2015; Odlum & Yoon, 2015), Zika (Fu et al., 2016), medication use (Buntain & Golbeck, 2015; Hanson, Cannon, Burton, & Giraud-Carrier, 2013; Scanfeld, Scanfeld, & Larson, 2010), diabetes (Harris, Mueller, Snider, & Haire-Joshu, 2013), and weight loss and obesity (Dahl, Hales, & Turner-McGrievy, 2016; Ghosh & Guha, 2013; Harris et al., 2014; Turner-McGrievy & Beets, 2015; Vickey, Ginis, & Dabrowski, 2013).

The previous Twitter studies have dealt with extracting common topics of one health issue discussed by the users to better understand common themes; however, this study utilizes an innovative approach to computationally analyze unstructured health related text data exchanged via Twitter to characterize health opinions regarding four common health issues, including diabetes, diet, exercise and obesity (DDEO) on a population level. This study identifies the characteristics of the most common health opinions with respect to DDEO and discloses public perception of the relationship among diabetes, diet, exercise and obesity. These common public opinions/topics and perceptions can be used by providers and public health agencies to better understand the common opinions of their population denominators in regard to DDEO, and reflect upon those opinions accordingly.

2. Methods

Our approach uses semantic and linguistics analyses for disclosing health characteristics of opinions in tweets containing DDEO words. The present study included three phases: data collection, topic discovery, and topic-content analysis.

2.1. Data collection

This phase collected tweets using Twitter's Application Programming Interfaces (API) (Twitter, 2017). Within the Twitter API, diabetes, diet, exercise, and obesity were selected as the related words (Wing et al., 2001) and the related health areas (Paul & Dredze, 2011). Twitter's APIs provides both historic and real-time data collections. The latter method randomly collects 1% of publicly available tweets. This paper used the real-time method to randomly collect 10% of publicly available English tweets using several pre-defined DDEO-related queries (Table 1) within a specific time frame. We used the queries to collect approximately 4.5 million related tweets between 06/01/2016 and 06/30/2016. The data will be available in the first author's website. Fig. 1 shows a sample of collected tweets in this research.

Table 1
DDEO queries.

Health issue	Queries	Number of tweets	Percentage
Diabetes	diabetes OR #diabetes	353,655	8.0%
Diet	diet OR #diet OR dieting	1,045,374	23.7%
Exercise	exercise OR #exercise OR exercising	734,118	16.6%
Obesity	obesity OR #obesity OR fat	2,283,517	51.7%

2.2. Topic discovery

To discover topics from the collected tweets, we used a topic modeling approach that fuzzy clusters the semantically related words such as assigning "diabetes", "cancer", and "influenza" into a topic that has an overall "disease" theme (Karami, 2015; Karami, Gangopadhyay, Zhou, & Kharrazi, 2017). Topic modeling has a wide range of applications in health and medical domains such as predicting protein-protein relationships based on the literature knowledge (Asou & Eguchi, 2008), discovering relevant clinical concepts and structures in patients' health records (Arnold, El-Saden, Bui, & Taira, 2010), and identifying patterns of clinical events in a cohort of brain cancer patients (Arnold & Speier, 2012).

Among topic models, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is the most popular effective model (Lu, Mei, & Zhai, 2011; Paul & Dredze, 2011) as studies have shown that LDA is an effective computational linguistics model for discovering topics in a corpus (Hong & Davison, 2010; Mcauliffe & Blei, 2008). LDA assumes that a corpus contains topics such that each word in each document can be assigned to the topics with different degrees of membership (Karami & Gangopadhyay, 2014; Karami, Gangopadhyay, Zhou, & Kharrazi, 2015a, 2015b).

Twitter users can post their opinions or share information about a subject to the public. Identifying the main topics of users' tweets provides an interesting point of reference, but conceptualizing larger subtopics of millions of tweets can reveal valuable insight to users' opinions. The topic discovery component of the study approach uses LDA to find main topics, themes, and opinions in the collected tweets.

We used the Mallet implementation of LDA (Blei et al., 2003; McCallum, 2002) with its default settings to explore opinions in the tweets. Before identifying the opinions, two pre-processing steps were implemented: (1) using a standard list for removing stop words, that do not have semantic value for analysis (such as "the"); and, (2) finding the optimum number of topics. To determine a proper number of topics, log-likelihood estimation with 80% of tweets for training and 20% of tweets for testing was used to find the highest log-likelihood, as it is the optimum number of topics (Wallach, Murray, Salakhutdinov, & Mimno, 2009). The highest log-likelihood was determined 425 topics.

2.3. Topic content analysis

The topic content analysis component used an objective interpretation approach with a lexicon-based approach to analyze the content of topics. The lexicon-based approach uses dictionaries to disclose the semantic orientation of words in a topic. Linguistic Inquiry and Word Count (LIWC) is a linguistics analysis tool that reveals thoughts, feelings, personality, and motivations in a corpus (Karami & Zhou, 2014a, 2014b, 2015). LIWC has accepted rate of sensitivity, specificity, and English proficiency measures (Golder & Macy, 2011). LIWC has a health related dictionary that can help to find whether a topic contains words associated with health. In this analysis, we used LIWC to find health related topics.