Characterizing health behavior information: Developing a surveillance text mining framework using Twitter for diet, diabetes, exercise, and obesity

George Shaw, Jr.
School of Library and Information Science
University of South Carolina
Columbia, SC 29208
Voice: (803) 565-1432
gshaw@email.sc.edu

The medical costs for individuals who are overweight or obese are $1,429 higher than those of healthy weight. Obesity increases one's risk of developing type 2 diabetes, heart disease, reproduction difficulties, and decreases the overall quality of life. Few studies have utilized a small-scale Twitter study to retrieve user-generated content regarding diet, diabetes, exercise, and obesity to characterize the topics associated with them; a human evaluation of the sentiment analysis and topic cluster results was also conducted. Previous studies show the relationship that exists among diabetes, diet, exercise, and obesity. Exercise and proper dieting are modifiable lifestyle behaviors that can help with reducing obesity and various chronic conditions like diabetes.

A national survey conducted by the Centers for Diseases Control and Prevention (CDC) is the annual Behavioral Risk Factor Surveillance Survey (BRFSS). Twitter provides researchers with a new opportunity and alternative data source to collect information regarding health behaviors using real-time data. Previous studies have demonstrated Twitter's ability to monitor adverse side-effects of drugs, tobacco use, and life satisfaction. Twitter can be a cost-effective way to gather information from study participants and collect population-level research data. The research questions guiding this study are:

RQ1: What are the positive and negative sentiments of Twitter users regarding diet, diabetes, exercise, and obesity (DDEO)?

RQ2: What additional health conditions are prevalent based on Twitter users' sentiments regarding DDEO?

RQ3: How does the performance of the computational tools used for sentiment analysis and topic recognition compare to the use of human performance?

The systematic steps in constructing this surveillance framework include data collection, data cleaning, sentiment analysis, topic discovery, topic analysis, and evaluation. Nearly 15 million tweets were collected through the Twitter API from June 2016 – August 2016. Sentiment analysis and the Latent Dirichlet Allocation (LDA) topic modeling text mining methods were used to answer **RQ1** and **RQ2**. The LDA model allows for the probabilistic model of a corpus. Therefore, each topic can be characterized by a probabilistic distribution over a set of documents – paired with linguistic analysis to capture individual's positive and negative sentiments. Eight-hundred topics were analyzed (100 for each query term and each sentiment)

through the topic analysis step of the framework. Percentage agreement and Cohen's kappa statistics were used to address **RQ3.**

Five hundred and twenty-three (or 65%) of the 800 topics were identifiable and related to DDEO. Sentiment prevalence across DDEO include topics of lifestyle, childhood obesity, food, and type of diets. Hypothyroidism, dementia, and diabetic retinopathy are additional chronic conditions identified through this framework. An essential aspect of the analytical process that this framework supports is a different approach to understanding topics from unsupervised machine learning, with the qualitative characterization of those topics. This surveillance text mining framework can assist public health professionals and medical social workers with promoting healthy behaviors and identify latent risk factors.