2018

Characterizing Health Behavior Information: Developing A Surveillance Text Mining Framework Using Twitter For Diet, Diabetes, Exercise, And Obesity

George Shaw Jr.
University of South Carolina

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

by

George Shaw, Jr.

Bachelor of Science
Charleston Southern University, 2007

Master of Science
North Carolina Agricultural and Technical State University, 2009

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Library and Information Science

College of Information and Communications

University of South Carolina

2018

Accepted by:

Amir Karami, Major Professor

Feili Tu-Keefer, Committee Member

Darin Freeburg, Committee Member

Robert McKeever, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I would like to start by acknowleding two former sharecroppers that invested invaluable time, energy, money, and love in my life thus far. My parents have supported me through this entire process and they have encouraged me more than they know. Thank you, mom and dad! Secondly, I would like to thank my wife for her support and sacrafice. You have been an undeniable shoulder of support for the tough days and a reminder of the responsbilities that come with achieving this academic milestone.

To Dr. Sam Hastigs, thank you for giving me an opportunity to find my love for teaching and cultivating my research abilities. You are more than an admired academic scholar to me. I would like to thank my committee members for sharing their expertise, time, and support during this process – with a special thank you to my dissertation chair Dr. Amir Karami.

I would also like to thank my doctoral colleagues; the awesome work that you all are doing inspire me and drive me to be a better scholar. I would like to thank my SLIS faculty and staff family (past and current) for their insights, kind words, and wedding gifts. Lastly, I would like to thank God for this life changing opportunity and the many blessings that have been associated with this academic achievement.

Abstract

Previous studies have documented the relationship that exists among diabetes, diet, exercise, and obesity. Obesity increases people's risk of developing heart disease and type 2 diabetes. Exercise and proper dieting are modifiable lifestyle behaviors that can help with reducing people's overall weight and risk to 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. 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 results was also conducted. 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 health experiences are prevalent based on Twitter users' sentiments regarding DDEO?, RQ3: How does the performance of the computational tools used for sentiment analysis and topic modeling 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 discovery latent semantic structure in a corpus. In LDA, 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 sixty-eight (or 71%) of the 800 topics were identifiable and related to DDEO. Sentiment prevalence across DDEO includes 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 interrelated health topics from relatively small-scale Twitter data, with the qualitative characterization of those topics. This surveillance text mining framework can assist clinical and allied health professionals with exploring understudied chronic health issues and identify latent risk factors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Obesity Impact

Obesity is a complex issue and a serious health concern in America (Centers for Disease Control and Prevention (CDC) (2016a). By 2030, roughly 78.9% of Americans are expected to be overweight, with 49.9% being obese (Wang, Beydoun, Liang, Caballero, & Kumanyika, 2008). A healthy Body Mass Index (BMI) for men and women is considered to be between 18.5 and 24.9. Individuals overweight have a BMI of 25 to 29.9 and those considered obese have a BMI greater than 35 (Flegal, Carroll, Kit, & Ogden, 2012; Health and Human Services, 2012). Overweight and obesity issues impact 78.8% of Hispanics and 76.7% of blacks in the United States (U.S). The 2009 behavioral risk factor surveillance system (BRFSS) disclosed that obesity among U.S. adults increased 37% from 18.3 to 25.1 between 1998 and 2006 (Finkelstein, Trogdon, Cohen, & Dietz, 2009, p. w823). The National Center for Health Statistics (2017) has also reported a significant increase in obesity among U.S. adults since 1999-2000 through 2013-2014. Although the method of how data are weighted for analysis by the CDC changed since the Finkelstein et al. article (2009), recent reports from the CDC indicate that 20% of adults in all states were identified as obese. Regionally, southern states have the highest prevalence of obesity (Centers for Disease Control and Prevention, 2017).

The estimated medical cost of obesity in the U. S. in 2008 was $147 billion. The medical costs for individuals who are overweight or obese are $1,429 higher than those of normal weight (Finkelstein et al., 2009). A recent systematic review by Kim & Basu (2016) – based on 2014 US dollars – estimated that the medical cost of an obese U.S adult to be $1,901. Medical expenses can be a tremendous burden for individuals that are financially unable to pay this additional cost or individuals who live on a fixed income. To encourage physicians to pay more attention to the condition and address the way health insurance companies pay for various treatments, the American Medical Association recently recognized obesity as a disease (Kim & Basu, 2016)

In addition to individual and national medical costs, obesity is identified as a risk factor for various chronic health issues. Obesity increases the risk of type 2 diabetes, heart disease, reproduction difficulties, and decreases the overall quality of life. Obesity has also been linked to issues of gout and Alzheimer's. Obesity impacts every aspect of our life with regards to these and other chronic health conditions (Health, n.d.; Koh-Banerjee et al., 2004; Menifield, Doty, & Fletcher, 2008).

Proper dieting and exercising are modifiable lifestyle behaviors that can help with reducing obesity and some of the various chronic conditions associated with it, in particular, diabetes. Several behaviors can be discussed, but this research study will focus on the lifestyle behaviors of diet and exercise (Wing et al., 2001). Studies show the strong relationship that exists among diabetes, diet, exercise, and obesity (DDEO) (Flegal et al., 2012; Hartz, Rupley, Kalkhoff, & Rimm, 1983; Wing et al., 2001). Exercise offers benefits for both mental and physical well-being. Improper diet and lack of physical

activity are risk factors that can contribute to type 2 diabetes(Byrne & Byrne, 1993; Turner-McGrievy & Beets, 2015; Wing et al., 2001).

As the severity of the obesity epidemic in the U. S. has increased since the early 90's, traditional and innovative research methodologies have been utilized to gain additional insight and characterize behaviors associated with obesity (Abbar, Mejova, & Weber, 2015). The following sections describe the traditional survey research methods used to collect behavioral risk factor information regarding obesity and new approaches that utilize the Twitter social media platform.

Survey Studies

To obtain an understanding of the population impacted by obesity and the associated chronic diseases, survey research is a commonly used method to collect data for analysis. Surveys provide a quantitative or numeric way to describe trends, identify attitudes or opinions of a population by analyzing a sample of the population data (Creswell, 2014). With the evolution of Information and Communication Technology (ICT) capabilities, researchers can conduct surveys using e-mails, web-based survey platforms, crowdsourcing sites, or social media. Conducting surveys electronically allows researchers access to a larger sample population and provides a cost-effective method for the distribution of survey data. Large-scale surveys like the U.S. National Survey on Drug Use and Health rely on face-to-face interviews that may benefit from clarification of meaning regarding an answer to a question. The use of focus groups for surveying can elicit more in-depth information and free-flowing sources of information with regards to public beliefs and attitudes (Paul & Dredze, 2017)

However, there are limitations in retrieving these data through traditional survey methods. When conducting focus group interviews, groupthink is sometimes considered an issue. Also, in face-to-face interviews, participants may not be as candid to avoid feeling inadequate to the questions asked or being perceived a particular way by the interviewee (Pierannunzi, Hu, & Balluz, 2013). Phone surveys have decreased in their accuracy as the use of landline phones has declined, particularly among low-income people and young adults (Blumberg & Luke, 2007). While recent changes to the Behavioral Risk Factor Surveillance System (BRFSS) – from the CDC - method of sampling includes data from cellphone users, this changes how behavior is reviewed when conducting trend analysis; and future studies will need to include landline and cellphone response in the data analysis process (Alaska Department of Health and Social Services, 2018). As with landline phones, every population is not represented within certain online environments and consideration must be given to time involved with collecting data for a survey conducted online (Creswell, 2014; Schmidt, 1997). Researchers must consider the software used to distribute their survey for accessibility reasons. Lastly, surveys address a limited number of topics that are often designed on the collection of structured text information (Creswell, 2014).

Survey Studies and Health Risk Behavior

A well-known national survey conducted by the CDC is the annual BRFSS. BRFSS is the largest continual national health survey conducted in the world. Using telephone surveys, interviews with U.S. residents are performed to identify their health risk behavior, issues with chronic diseases, and the preventative steps to secure a healthy

future (CDC, 2016b). This survey is informative and provides in-depth information for interdisciplinary research and it is published once a year.

There are also idiosyncrasies of the system and external factors that impact how the data are collected. The state health departments manage the collection of BRFSS data. "The data collected by state health departments are then transmitted to the Behavioral Surveillance Branch of the National Center for Chronic Disease Prevention and Health Promotion at CDC for editing, processing, weighting, and analysis" (Forrest & Lin, 2010 p. 6). For some states, the only source of state data on health risk behavior related to chronic diseases is based on the information provided by the BRFSS. BRFSS allows state and federal agencies the information they need to plan, conduct, and evaluate public health programs and other related activities. That information can be collected by contract with a private company or university, or it can be done internally. The agency that collects information for the state, if not done internally, is required to follow the BRFSS protocols(Center for Disease Control and Prevention, 2016b).

While studies have described the reliability and validity of the BRFSS, there is very little research on reliability and validity for some health topics included in the BRFSS (Fahimi, Link, Mokdad, Schwartz, & Levy, 2008; Pierannunzi et al., 2013). Also, based on information reported from the CDC and current literature, the inclusion of user-generated content through social media has not been utilized by the CDC in its official BRFSS reporting as an additional method to learn about health behaviors. The current landscape – with the increase in computational social science research – provides public health providers and clinical experts with great advantage to utilize the relatively

real-time, huge datasets that are generated through social media (Comito, Pizzuti, & Procopio, 2018).

When examining external factors limiting BRFSS as previously acknowledged in section 1.2, there has been a decrease in landline ownership, particularly among the millennial generation (Fahimi et al., 2008). Participants in the BRFSS survey may also provide embellished information to make their health information favorable to the interviewer and to protect their privacy (Blumberg & Luke, 2014; Tenzer, 2016). Researchers continue to provide empirical evidence of the impact and role that social media can have in the public health sciences (De Choudhury, Counts, & Horvitz, 2013; Paul & Dredze, 2011; Paul & Dredze, 2012).

Public Health and Social Monitoring

Social media monitoring of public health and medical issues is providing a faster and larger coverage of topics. Also, there is debate of whether social media is here to stay or not; can it play a huge role in supporting traditional monitoring methods that are in place (Paul & Dredze, 2017, p. 13). Social media, Twitter specifically, provide researchers with a new opportunity to effectively and economically collect data about health behaviors and health risk factor information. As more Baby Boomers and Generation X (born after the baby boomer generation) increase their use of social media to communicate, conduct business, and share sentiments online, there is an opportunity to dynamically collect information about the sentiments of users' health information and computationally characterize health behavior from user-generated content (Duggan, Ellison, Lampe, Lenhart, & Madden, 2015). When it comes to social media, there is a

massive amount of data pouring in from social media related outlets. Why not look at certain health issues differently and bring other data to bear (Kuehn, 2015)?

Social Monitoring with Twitter

Twitter provides real-time feedback and populates messages based on the most recent timestamp. Of concern to many researchers is the inability to analyze private messages of Twitter users that provides additional in-depth opinions of users that are not normally shared as public posts. However, the argument has been made that Twitter offers large datasets which decreases sampling error (Finfgeld-Connett, 2015). In a recent study, researchers performed a large-scale analysis of 210 thousand Twitter users in the US, analyzing their 502 million tweets. The results from the study showed that the foods tweeted by users were predictive of the national obesity and diabetes statistics. It also showed that friends are likely to show similar interest in food (Abbar et al., 2015). Previous studies have also demonstrated Twitter's ability to monitor adverse side-effects of drugs, tobacco use, and life satisfaction (Aramaki, 2011; Prier, Smith, Giraud-Carrier, & Hanson, 2011). These studies demonstrate Twitter's ability to collect unstructured data that can be used for computational social science research. These studies also demonstrate that the data collected from Twitter can be used to explain and provide solutions to social phenomenon outside of social networking sites (SNS).

Harris et al. (2014) study on reducing childhood obesity stated that limited information regarding evidence-based strategies for communicating evidence-based information is utilized within Twitter. In addition to the ability that Twitter provides for

surveillance, Twitter can be a cost-effective way to gather information from study participants and collect research data at an individual level.

Obesity is a multi-facet issue that requires an array of knowledge to address the pandemic. As noted from the earlier section, obesity is also a link to many chronic conditions (Lukic et al., 2014; Wing et al., 2001) – with diabetes being the focus for this research study. Also, consistent across the literature are the impacts that the behavioral modifications exercise and obesity has with reducing risk to diabetes and lowering one's BMI. While other researchers have studied these health issues individually or in combination with other health issues (Ghosh & Guha, 2013; M. Paul & Dredze, 2011; Turner-McGrievy & Beets, 2015), this is one of the first research studies to examine the relationship of these four health issues using two text mining methods.

Also, this study utilizes small-scale Twitter data to study user generated content regarding obesity and the chronic conditions associated with it. Further, this study computationally characterized the data, performed user evaluation of the computational model and tools used, and described the implications this information has for healthcare providers; particularly with the support it provides with decision-making and informing the design of health interventions. This research study seeks to answer the following questions:

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

RQ2: What health experiences are prevalent based on Twitter users' sentiments and topics regarding DDEO?

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

Chapter 2

Literature Review

This chapter will provide an overview of concepts and a review of the related literature. An understanding of key concepts and terminology will be provided; followed by a review of literature that discuss how Twitter has been used to understand health issues. Selected literature will discuss Twitter's use for surveilling non-health and health-related issues, with a focus on the methodological and computational impacts. Finally, the chapter will conclude by discussing DDEO related health surveillance and the two text mining approaches (sentiment analysis and topic modeling) use in other research.

Data Science, Text Analysis, and Health

Data science and big data are alluring terms with various disciplines – computer science, mathematics, statistics, and information science – contributing to them theoretically and technically. While the early emphasis of these concepts and emergence as a profession has focused on business (Harris & Mehrotra, 2014), it has implications for multiple disciplines. In some instances, the terms are incorrectly thought of as being similar conceptually. However, big data provides researchers and data scientists the ability to gain insights from large sets of data that are stored, processed and analyzed

(Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). Data science refers to the scientific methods, technological tools, information systems, and computational methods used to derive knowledge from large data sets (Carter & Sholler, 2016).

Traditionally, content analysis is performed manually on a corpus of data that has been collected. According to Lacy, Watson, Riffe, and Lovejoy (2015), a manual content analysis involves the process of categorizing data based on human input to answer a more significant research question surrounding the data. In the context of research that utilizes social media, the data under scrutiny are a contextual representation of a larger phenomenon, particularly from the perspective of understanding social behavior through the content generated. With the manual context analysis method, researchers develop a codebook and a list of categories that can be derived from the literature and/or theoretically based. To include a level of objectivity and reduce the biases of subjectivity, inter-coder reliability is reached through training (Krippendorff, 2004). The benefit of this method is that humans are able to identify sarcasm and infer among other researchers the complexity of the data being analyzed (Guo, Vargo, Pan, Ding, & Ishwar, 2016). However, when dealing with big data, this can become very expensive, require a substantial amount of time, and lends itself to increased human error when dealing with a large dataset.

Computationally, there are two approaches to perform content analysis on large datasets or big data: supervised and unsupervised machine learning text analysis. With supervised machine learning, a classifier "learns" how to distinguish classes by analyzing a corpus of pre-labeled data. Classifiers are based on algorithms that automatically assign a label to documents in a corpus. Assignment of labels is based on using an initial

set of data used for testing. After the classifier is trained on the test dataset, the patterns learned are utilized to make inferences about new instances in the future using the data that were not used for testing. Since the "labels are provided as training data, this approach is called supervised machine learning" (Paul & Dredze, 2017, p. 30).

Alternatively, unsupervised machine learning based on clustering can be used. Unlike classifiers, where the categories are known in advance, clustering groups messages into categories and they are grouped together based on their similarities (Paul & Dredze, 2017). There are several studies that have used this approach to discover topics (Paul & Dredze, 2011; Wang, Paul, & Dredze, 2014). The use of these computational approaches can be "a great benefit in our search for knowledge..." (Obole & Welsh, 2012). These machine learning approaches can be applied on data collected from Twitter, where millions of tweets are generated per day (Walker, 2015). Supervised and unsupervised computational text mining methods allow researchers to perform content analysis on millions of tweets (documents) that would normally require significant hours to accomplish manually. These text mining methods also provide an additional method to perform the surveillance of health issues.

CDC refers to surveillance as the systematic, ongoing, collection, management and interpretation of these (health issues) data to public health programs to stimulate public health action (Lupton & Michael, 2017). Through this surveillance, researchers can have access to a variety of data sources. When thinking about social media related outlets, the content generated in these environments is from the general population.

Instagram, Facebook, Twitter, and Reddit all offer unique insight into the emotional and physical well-being of individuals (Lupton, 2016; Robillard, Cabral, Hennessey, Kwon, & Illes, 2015). Twitter has proven to be an abundant source of this type of information (Hill, Merchant, & Ungar, 2013).

Twitter and Health Issues

Researchers have shown that SNS users, particularly Twitter, are more inclined to share their health information within Twitter than they would with their doctor (Paul & Dredze, 2011; Paul & Dredze, 2014). Many of the foundational computational studies using social media, particularly Twitter, began near the end of the first decade during the 21st century (Cooper, 2013; Thelwall, 2014). Several studies have utilized SNS to identify trends and identify real-world events, political opinions, and natural disasters (Paul & Dredze, 2014). SNS such as Facebook, Twitter, and Instagram have grown in their tracking of health issues and show a strong correlation between the trends in social media and the U.S. Centers for Disease Control and Prevention reports. However, only few studies have examined the concerns of Twitter users as it relates to obesity within the United States and the chronic conditions – for example, diabetes – that are associated with this medical condition (Culotta, 2014; Liu, Mei, Hanauer, Zheng, & Lee, 2016; Turner-McGrievy & Beets, 2015). The following sub-sections will explore and discuss:

1. The use of Twitter to surveil non-health related events and behaviors.
2. Twitter's use to surveil health related events and the scholastic contributions to understand and use Twitter as a method for health surveillance.

3. Twitter surveillance systems that have focused on DDEO (which are the primary health issues this study focuses on).
4. Use of sentiment analysis and topics models developed with a focus on addressing the DDEO topics.
5. Finally, a review of selected topic models.

Non-Health Related Surveillance

Emotions can have a significant impact on individual behavior and decision-making (Damasio & Sutherland, 1994). Behavioral economics research focuses on the cognitive, psychological and emotional factors that can impact the economic decisions people make (Bollen, Mao, & Zeng, 2011). Researchers Bollen et al. (2011) were interested in how emotions can influence the stock market by analyzing Twitter data. The researchers utilized two mood tracking tools; OpinionFinder (measures positive versus negative mood) and Google-Profile of Mood States (GPOMS - that measures six different mood dimensions). Using 9.8 million tweets collected from February 28 to December 19 in 2008 that were compared to the Dow Jones Industrial Average, their results showed that mood state can be tracked from large-scale Twitter content (Bollen et al., 2011).

The German election and its mirror of offline sentiments of the general public based on a content analysis explored the potential of Twitter as a surveillance tool. Based on the German federal election, researchers were interested in the political deliberation that took place within the Twitter environment and demonstrating Twitter as an alternative method to predict the popularity of parties and the coalitions that existed in the real world (Tumasjan, Sprenger, Sandner, & Welpe, 2010). Two months of German

politically related Twitter messages sentiments were extracted using the Linguistic Inquiry and Word Count text analysis software (Tausczik & Pennebaker, 2010). Their results showed that the "number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional polls" (Tumasjan et al., 2010, p. 183).

Police enforcement and public safety institutions are benefitting from Twitter surveillance and the predictability that can be gained from Twitter messages. A recent study incorporated sentiment analysis and weather predictors to current crime predictions models (Chen, Cho, & Jang, 2015). Unfavorable weather conditions and aggressive behavior have clear correlations between these factors and criminal activity (Anderson & Anderson, 1984; Chen et al., 2015). By incorporating the use of sentiment analysis to identify the polarity of tweets, weather factors, and the Kernel Density Estimation (Terrell & Scott, 1992) – the ability to make inferences about the population based on a set of data – the researchers provide better accuracy prediction on future crime incidents.

When examining the commonality that exists among the selected research studies and their use of Twitter, they incorporated analysis of users' emotions and feelings through the context of the tweet. As mentioned, these emotions and feelings or sentiments provide insightful prediction when understanding behavior. Moreover, this method is a significant approach when gaining insight, modeling predictability, and characterizing behavior. Additional noticeable implications that can be made from these non-health related uses of Twitter for surveillance and prediction include:

- The need for multiple text mining methods to be coupled together to help with increasing model performance

- User model design subjectively (knowingly or unknowingly) plays a role in the inferences that can be made about the data, and
- Computational methods have benefits to multiple academic disciplines.

Health-Related Surveillance

Public health researchers and information scientists are beginning to explore the potential knowledge that can be derived from various analyses on unstructured Twitter data. It is estimated that roughly 16.6-25.1% of English-language tweets are related to health issues (Hill et al., 2013). Public health surveillance deals with "the continuous, systematic collection, analysis and interpretation of health data" that are gained from using this data source (Paul & Dredze, 2017, p. 11). Although this estimated percentage includes a small portion of all Twitter messages, when the percentage represents several hundred million messages, there is useful information that is collected. From infectious diseases to mental illness, public health surveillance provides public health professionals and relevant stakeholders with an alternative insight than what is provided by traditional surveys. The following subsections will examine how researchers use Twitter data to explore, understand, and surveil seasonal health issues related to communicable disease; a significant portion of the discussion related to communicable diseases will involve the collection of Twitter data for influenza surveillance. The final subsection will examine researchers use of Twitter data to explore, understand and surveil non-communicable diseases.

Health-related surveillance – influenza

Of all the various health related issues that have used public health surveillance, influenza has been the most researched and documented health issue. This includes common cases of influenza and specialized cases like Swine Flu (or H1N1). Using learning methods along with n-gram (word sequencing feature), Twitter data research has demonstrated its ability to predict the Swine Flu epidemic and influenza-like illnesses (Aramaki, 2011; Culotta, 2010; Ritterman, Osborne, & Klein, 2009). Data obtained from search queries were one of the earliest digital methods used to conduct influenza surveillance (Eysenbach, 2006). Polgreen, Chen, Pennock, and Nelson (2008) examined the relationship between searches for influenza and actual influenza occurrences. Using roughly a four-year span of search query logs from Yahoo, they developed a model to predict influenza mortality that could be used for surveillance. After these early initiatives to provide surveillance of flu through search queries, several researchers have developed their own flu models, based largely on the Google Flu Trends Model (Preis & Moat, 2014; Yang, Santillana, & Kou, 2015).

The 2008 Google Flu Trends (GFT) digital disease surveillance system was the result of early query research work and one of the most well-known digital systems. GFT utilizes a simple univariate regression model (Paul & Dredze, 2017, p. 52). However, researchers found that GFT was predicting more than double the proportion of doctor's visits for influenza-like illness than reported from the CDC during 2012-2013. This overestimation by the GFT system can be contributed to the updates of the Google search algorithm and a shift in users searching behavior (Lazer, Kennedy, King, & Vespignani, 2014a). GFT was improved by the constant research of the system and verification of the

model to traditional data surveillance methods and alternative surveillance models that were influenced by GFTs design (Copeland et al., 2013; Stefansen, 2014).

A consistent undertone from the literature is the inclusion of digital surveillance systems such as GFT with additional data support. For example, when combining GFT and CDC data, there was an improvement on the performance of the system to predict influenza activity when compared with GFT alone (Lazer, Kennedy, King, & Vespignani, 2014b). Based on the latest information from Google, "Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns[1]." Researchers can still access the historical estimates that were produced by GFT and Google Dengue Trends. During this period of using Google searches to predict and surveillance the influenza activity, Twitter research became popular as another method to monitor influenza.

There were many approaches taken in the early period of Twitter research to monitor influenza. Chew and Eysenbach (2010) wanted to validate Twitter as a tool to collect sentiments and real-time information concerning the H1N1 outbreak. Using two variations that describe the term ("H1N1" and "swine flu"), they used an open-source surveillance tool to collect 2 million tweets (where 5,395 were used for analysis), created a database query of keywords, and manually coded tweets to perform a content analysis. They identified news and information as the two most commonly tweeted material as it related to H1N1. This study showed that tweets could be used for relatively real-time content and sentiment analysis; and it provided preliminary methods to manually classify

[1] https://www.google.org/flutrends/about/

topics and the potential for automated steps when it comes to analysis (Chew & Eysenbach, 2010).

Around this same time period, another researcher was investigating the messages posted on Twitter and correlating that to influenza data reported by the CDC (very similar to the work that was conducted for GFT services). Culotta's (2010) work used multiple regression models to predict influenza rates based on the frequency of messages that contain influenza-related words. The use of linear and multiple regression modeling assisted with identifying what keywords to monitor (Ginsberg et al., 2009). A bag-of-words classifier was used on 206 messages (positive and negative tweets) to identify erroneous documents and improve the prediction capability of the models. The top-performing model obtains a correlation of .78 with the CDC statistics. His work was pivotal with the use of document classifiers to identify relevant messages and the need to incorporate sophisticated linguistic features, such as n-grams (a sequence of text or speech within a given window for linguistically and semantical analysis), and the need for pre-processing to improve the quality of analysis (Culotta, 2010; Jurafsky & Martin, 2014). As the approaches continued to improve with Twitters surveillance and prediction ability, researchers began to explore the potential of Twitter for additional health related events.

In addition to H1N1 activity, a later research study collected information regarding other health concerns such as disease transmission, disease countermeasures, and consumer concerns regarding the consumption of pork (Signorini, Segre, & Polgreen, 2011). The researchers used a Support Vector Regression (SVR) method to conduct a form of supervised learning classification on the Twitter data used in the study. The

prediction models were trained on the weekly-term frequency statistics based on the influenza-like illness (ILI) reported by the CDC. Results showed that "Twitter traffic can be used not only descriptively, i.e., to track users' interests and concerns related to H1N1 influenza, but also to estimate disease activity in real time, i.e., 1–2 weeks faster" than the current practices used when this study was conducted (Signorini et al., 2011, p. 8).

Signorini et al (2011) work contributed to future studies by providing evidence of the contextual cues that Twitter provides - when conducting sentiment analysis related research - and the need to understand latent variables. These and other foundational digital sources (Corley, Cook, Mikler, & Singh, 2010; Lampos & Cristianini, 2010) can be viewed as the building blocks of computational social science related work and addressing public health issues. This has led to the development of more sophisticated models that include the use of natural language processing, machine learning, data mining, geographic information systems, and additional statistical modeling techniques. These works were also a significant intersection and efforts of researchers from various disciplines to engage in multidisciplinary related work.

Researchers from three different academic departments (Department of Geography, School of Public Health, and Department of Linguistics) sought to improve the monitoring and the spatial and temporal dynamics that are involved with the outbreak of influenza (Allen, Tsou, Aslam, Nagel, & Gawron, 2016). The researchers incorporated spatial filtering methods and machine classification procedures to identify tweets that do not appear to indicate real-world cases of influenza and improve statistical analysis (Nagel et al., 2013). Machine learning measure for quality, that includes recall, precision, and F1 score (the total ability of the test based on the recall and precision

calculations), was used to evaluate the classification model (Powers, 2011). Results demonstrated the new procedures support an advantage to previous studies that have incorporated local, regional, and national ILI reports and GIS twitter data.

To evaluate Twitter's performance for surveillance in multiple English-speaking countries based on traditionally supplied influenza data, researchers used a previously tested influenza detection system (Lamb, Paul, & Dredze, 2013). The algorithm used in this detection system "categorizes individual tweets for relevance to influenza infection and then produces estimates by aggregating the relevant tweets over some time interval (e.g. weekly)" (Paul, Dredze, Broniatowski, & Generous, 2015). Their results showed their Twitter data to be relatively correlated to the country data collected and the importance of using Twitter surveillance in more populated areas.

To further extend the contribution of Twitter with improving Influenza Surveillance, researchers presented a machine learning methodology for their ILI surveillance that incorporated data sources from Google Trends, GFT, Twitter, and hospital visit records from a medical practices management company; they also incorporated data from FluNearYou (participatory surveillance system) (Santillana et al., 2015). Methodically, the researchers chose machine learning algorithms that have unique strengths with combining information; SVM regression, Stacked Linear regression, and Decision Tree Regression (Breiman, 1996; Drucker, Burges, Kaufman, Smola, & Vapnik, 1997; Freund & Schapire, 1997). Their results reveal the prediction ability to combine the various data sources, versus their ability as independent data sources. When considering the impact this study has on social media, the proper combining of social

media and digital data sources can statistically compete with gold standard flu activity data sources, such as the CDC's ILI (Santillana et al., 2015).

Over the past decade, the scholarly contributions to influenza surveillance and prediction of flu-like activities have undoubtedly progressed the field of computational science and computational social science research. The introduction of various machine learning procedures has improved the ability to reduce noise when collecting tweets (Allen et al., 2016), provide the aggregation of multiple data sources for improving Twitter-based influenza surveillance performance (Paul et al., 2014) and the ability to study disease awareness and sentiments (Lamb et al., 2013). The study of additional communicable (infectious diseases) and non-communicable diseases are possible as a result of these efforts to improve the surveillance of influenza and variations of this infectious disease.

Health related surveillance – other communicable diseases.

Although other communicable diseases have not been researched to the extent of influenza, they provide evidence that digital surveillance can be used to track other diseases (Paul & Dredze, 2017). Analogous to the work that was being done with influenza surveillance through search engines in the early portion of the 21st century (Eysenbach, 2006; Ginsberg et al., 2009), trend data were being analyzed to identify outbreaks and opportunities to support traditional data collection methods with other communicable diseases. Researchers explored Google Trends' (GT) ability to be used as a tool to examine Lyme disease (Seifter, Schwarzwalder, Geis, & Aucott, 2010). Their decision to examine Lyme disease is appropriate because similar to influenza, there are

seasonal trends of increased incidents of tick bites; particularly during the spring and summer seasons (Bacon, Kugeler, & Mead, 2008). Using a non-sophisticated technical method of capturing trends through GTs with the search terms of "Lyme disease, tick bite, and cough," GT identified 6 of the cities that were consistent with CDC data on the highest number of cases for Lyme disease. This research contributed to the understanding of search engine bias in this type of research, but most importantly, demonstrated that digital surveillance increases the potential and ability to monitor a broad range of public health concerns.

While there are numerous studies documenting seasonal diseases (Paul & Dredze, 2011) to help public health professionals with Twitter-based surveillance, there was limited scholarly work involving sudden outbreak analysis and control. Researchers Diaz-Aviles and Stewart (2012) sought to study crowd behavior within Twitter during an outbreak. Specifically, they looked at *Enterohemorrhagic Escherichia coli* (EHEC) outbreak – or *E-coli* – outbreak that occurred in Germany by analyzing 456,226 tweets related to the EHEC outbreak; and incorporated a biosurveillance detection method to analyze the tweets (Diaz-Aviles & Stewart, 2012, p. 83). Their surveillance detection method alerted an outbreak earlier than the traditional systems used. Their work was instrumental in demonstrating Twitter's potential for epidemic intelligence when dealing with a sudden outbreak and how users provide the information that can be used for decision-making.

When exploring the possibility of surveillance through social media and search engines, one study (Deiner, Lietman, McLeod, Chodosh, & Porco, 2016) incorporated the work that was done by Santillana et al., (2015) and Ginsberg et al.,

(2009) with improving influenza surveillance through search engine query related data. The researchers compared data from social media (Twitter) and Google searches (GT) to identify the correlation that they have with traditional medical records provided through the hospital system. Data regarding conjunctivitis (or "pink eye") were collected over a two-year period, with a focus on identifying the seasonality of the eye disease. Results from the study show similar patterns among the clinical data, GT searchers, and Twitter posts; it also showed high correlation with understanding the seasonal patterns of the disease and "that data from search engines and social media could serve as a surrogate source of epidemiologic information about infectious eye disease" (Deiner et al., 2016, p. 1028).

While social media (in this case, Twitter) can be viewed as a viable source and alternative to collect information on a variety of events and behaviors, misinformation can happen that may impact Twitter's utility as a surveillance tool. One of the most recent communicable disease outbreaks, Ebola, demonstrated how misinformation can lead to deaths from lack of proper preventative information in a timely matter (Bedrosian et al., 2016). However, citizens' detection of irregularities in the environment is important for identifying outbreaks and being proactive with regards to preventative care and addressing misinformation (Berg, 2013).

An exploration of the recent Zika outbreak[2] demonstrated the advancements made in computational research, that hinges on the work related to influenza and non-health related events (Allen et al., 2016; Brownstein & Freifeld, 2007; Chen et al., 2015). The

[2] http://www.bbc.com/news/health-35370848- this article retrieved from the BBC details the World Health Organization declaration of Zika as a worldwide global health emergency and what everyone should know about the disease.

researchers' surveillance system was used to determine the gender sentiment difference regarding Zika, characterization of tweets (symptoms, transmission, treatment, and prevention), an analysis of the classification performance that was used in the study, and the main topics of each characterization or misconceptions regarding the Zika virus (Miller, Banerjee, Muppalla, Romine, & Sheth, 2017). Results from the sentiment analysis show the negativity associated with the diseases among both genders and the detection of topics using the well-established LDA topic modeling method. This study demonstrates how conversations within Twitter can assist public health professionals with understanding societal concerns regarding a specific disease category and identifying misinformation that also exists within the Twitter environment.

As seen in the studies discussed thus far, seasonality is a major factor with aggregating social media and search engine surveillance data to traditional data sources. The increase in conjunctivitis (Deiner et al., 2016), influenza-like illnesses (Chew & Eysenbach, 2010; Hawelka et al., 2014), and sudden outbreaks such as Ebola and the Zika virus (Miller et al., 2017; Odlum & Yoon, 2015) also demonstrate that non-communicable diseases have physical visual attributes that cue individuals to assess the change in their environment. That can be individually or collectively within a community. However, diseases and health conditions such as cancer, congestive heart failure, bad health addictions, and kidney failure are not as easily identifiable due to the nature of the non-communicable or chronic conditions. A series of bad behavior decisions may not signify a need to visit a primary care physician or hospital, but the continued behavior of this risk factor (i.e. smoking) can drastically increase an individual's susceptibility to lung cancer and other medical conditions related to smoking

(Inoue-Choi et al., 2017; M. J. Paul & Dredze, 2017). The following section will discuss the advancements made to address chronic diseases, provide a detailed scholarly conversation of research to use Twitter as a surveillance method to monitor diet, diabetes, exercise, and obesity, and opportunities to add to this method of social monitoring for public health professionals.

Health related surveillance – non-communicable diseases.

Much of the research conducted that uses social media for surveillance has focused on acute (communicable) related diseases. However, there has been work done to address chronic related diseases and they have contributed to the robustness of the text mining methods use for surveillance. Several studies that look at the search activity regarding cancer originated from the work done by individuals on search engines. Bader & Theofanos (2003) wanted to better understand what lay individuals wanted when searching online for cancer-related information. Using the Ask.com search engine, they used 37 terms to search the Ask.com query log and identified 204,165 instances of cancer-related queries over a three-month (June-August) time frame. They identified 7,500 individual user questions from the related queries that represented 37% of the total three-month pool, with 78.37% represented through 14 different types of cancer. This work was similar to the volume and prevalence of health-related searches on the Web that was conducted by Eysenbach & Kohler (2003). This and similar work like it was instrumental in understanding NLP (Eysenbach, 2006; G. Eysenbach & Kohler, 2003). While NLP related research was important with understanding the retrieval of information in search engines, informing web search information structures, and various

meta-data standards, it has been equally important with understanding what content people want and the language used to obtain that information (Bader & Theofanos, 2003). A study involving multiple qualitative data collection methods found that lay individuals preferred search engines, rather than utilizing traditional or credible medical web sources for health information (Eysenbach & Köhler, 2002). This is consistent with social media sites such as Twitter as sources for consumers to gather, share, and consume health related information (S. Fox, 2009; Westerman, Spence, & Van Der Heide, 2014).

The American Heart Association suggested that one way to reduce the risk for heart disease is the use of *population-level strategies* to reduce and shift the distribution of risk (Lloyd-Jones et al., 2010, p. 589). The current assessment of community-level behavioral and psychological characteristics is difficult. Eichstaedt et al. (2015, p. 160) stated that "social-media based digital epidemiology can support faster response and deeper understanding of public-health threats than can traditional methods". Through a predictive model, the researchers' Twitter-based system outperformed traditional risk factor models that were used. They discovered that a Twitter-based system to track psychological variables is less expensive and generates a more condense pool of information within a shorter time frame. Twitter has also been used to predict atherosclerotic heart disease (ADH) (Eichstaedt et al., 2015).

Research using a discussion social media site, Reddit[3], identified the value in the information that consumers share in these spaces related to cancer. A content analysis of a general cancer forum described the information that was shared by those in the self-characterized illness phase (Eschler, Dehlawi, & Pratt, 2015). Paparrizos, White,

[3] www.reddit.com

and Horvitz (2016) used statistical topic classifiers from Bing to identify health related queries concerning Pancreatic Adenocarcinoma and make interpretations of the searchers' gender. They were able to classify pancreatic cancer-related search queries by analyzing the search log before the date of actual diagnosis of the individual performing the search. Social media, search engines, and online forums related to non-communicable disease are supplemental methods that can be used for surveillance (Allen et al., 2016).

Yin, Fabbri, Rosenbloom, and Malin (2015) further supported this by detecting health issues, based on the health statuses mention on Twitter. Their classifier yielded a .77 precision rate on all 34 health issues that were detected from the users' Twitter status. These studies demonstrate that the advancement made in NLP, machine learning, and various text mining methods are applicable to the surveillance of non-communicable diseases. Surveillance systems extend the ability of public health professionals to monitor changes over an extended period of time, which can be costly with normal surveillance methods and chronic related diseases (Paparrizos et al., 2016). This form of information and surveillance is important as it is provided candidly from users (Yin et al., 2015). However, as discussed in the introduction, reliance solely on the algorithmic detection of topics through surveillance may inadvertently miss information that is relevant to a health professional's decision-making process. This is a notable discussion point in the advancement of alternative surveillance systems (Eschler et al., 2015).

Health related surveillance – diet, diabetes, exercise, and obesity.

Addressing behavioral related medical issues can be difficult when behaviors take place away from the doctor's office and are not always identifiable through traditional survey methods. Self-reported information can be easily fabricated (Fisher & Katz, 2000; Paul & Dredze, 2017). Surveillance of health behaviors or behavioral medicine (study of how people make choices about their health and the impact on their well-being) related to diet, diabetes, exercise and obesity has been an important part of increasing the promptness of care that public health professionals can provide, assist with improving intervention methods, and the implications for advocacy and health related policies. While the BRFSS is internationally recognized as an effective system to collect information on health behaviors such as dieting, exercising, and smoking, social media surveillance can assist with filling gaps not covered through large-scale surveillance systems. The increased use of data points that are provided through social media provides opportunities to improve these gaps (Ayers, Althouse, & Dredze, 2014). The following studies will explain Twitter's ability as a surveillance tool to collect information on diet, diabetes, exercise, and obesity.

As discussed in Chapter 1, obesity is an issue that impacts adults and children. Roughly 24% of teens aged 12-17 use Twitter. Researchers sought to identify the influence of social media on public health and its use to communicate health information (Harris, Moreland-Russell, Tabak, Ruhr, & Maier, 2014). Using metadata hashtags, they selected #childhoodobesity to monitor private person, for-profit, nonprofit, media, government, and education tweets and network connections regarding this topic. By incorporating descriptive statistics, network descriptive statistics and visualization, and network modeling, they were able to identify topics, network ties, and message

engagements based on mentions and retweets of the hashtag. Through social media surveillance, Harris et al. (2014) identified the lack of government, media, and educational sources of childhood obesity being disseminated on Twitter.

Gore, Diallo, and Padilla (2015) investigated the factors that exist in certain geographical locations to obtain insight into how areas within the US experience higher obesity rates than others. Utilizing the previous work done to identify measures that distinguish influenza tweets from other tweets, they adopted this strategy to identify measures related to the variation in obesity rate for a Metropolitan Statistical Area (MSA) and analyzed the geotagged tweets to measure happiness, diet, and physical activity to obesity rates for the MSA (Broniatowski, Paul, & Dredze, 2013). Their work demonstrated the potential of social media to be used as a real-time, population-scale measure of factors related to obesity. Similarly, a study was conducted to identify how social media can be leveraged to provide a greater understanding of the well-being and health behaviors of communities at a granular level (Nguyen et al., 2016). Geo-coordinates of the tweets allowed them to spatially join them to 2010 census tract locations. Their results showed that happy tweets, healthy food references, and physical activity references were less frequent in census tracts with greater economic disadvantage and higher proportions of racial/ethnic minorities and youths. Demonstrating the consistency with results from similar work regarding dietary and exercise behavior (Fried, Surdeanu, Kobourov, Hingle, & Bell, 2014; Gore et al., 2015), the researchers demonstrated that social media could be leveraged to help us better understand the well-being and health behaviors of certain geographical regions.

Turner-McGrievy and Beets (2015) studied the temporal trends in weight loss-related posts. Studies have shown that weight gain typically happens during the holidays and physical activity (PA) is higher in the summer than the winter (Cook, Subar, Troiano, & Schoeller, 2012; Yanovski et al., 2000). This study also centered on identifying the discussion related to weight loss. The researchers showed an increase in posts regarding weight loss during the holidays. Results from this study can add an updated data collection method to earlier work that evaluated search and browsing logs with understanding how people search the web to address their efforts with weight loss (Schraefel, White, André, & Tan, 2009). Understanding seasonal trends and the type of support groups that individuals are searching for can assist public health professionals with identifying and connecting people with sources that are effective in helping people reach their behavioral weight change desires.

The data gathered through Twitter provides a unique understanding of various health behaviors. Gore et al. (2015) used social media to determine the factors of obesity for a MSA and Nguyen et al. (2016) provided an insight into happy emotions and obesity behavior are examples of the insights provided through Twitter data. While sentiments provided are the residual evidence of the assumed behavior, it is an increasing source of information that provides health care professionals with additional information that may not be obtained through traditional surveillance systems. Growing scholarly evidence surrounding the Eichstaedt et al. (2015) statement that "A twitter-based system to track psychological variables is rather less expensive and generates a more condense pool of information within a shorter time frame," supports the ability to track and characterize the DDEO topics.

While these studies illustrate the use of Twitter to provide surveillance and predictive modeling approaches for obesity, they also address the idea of how evidence-based information is dispersed and what this means for studying residual effects of actual behavior (Harris et al., 2014). The echo chambers of information that are created within social media environments (Chandran, 2018) supports the efforts and monetary need by public health professionals to counter misinformation found on social media platforms such as Twitter (Berg, 2013; Miller et al., 2017). The advancement in sentiment analysis through Twitter data is one of the several ways that latent topic discovery provides insight to address issues such as the misinformation regarding vaccinations (Wilson, Atkinson, & Deeks, 2014).

These studies also address the need to question the methodical approach of normalizing the different variations of a word (Ghosh & Guha, 2013). A key element of conducting sentiment analysis across various geographic regions is to maintain the semantic of words by geographic location (Allen et al., 2016). This becomes important in providing location-based services (Ilarri, Illarramendi, Mena, & Sheth, 2016). Although the consideration of geo-location is outside the context of this study, the importance of geographic semantics is important in analyzing and characterizing the topical results. As public health professionals and medical experts continue to identify ways to predict the spread of disease, prevent and/or reduce the number of chronic diseases related cases, and improve the overall healthcare quality in the United States, large-scale Twitter studies still benefit from identifying hidden information; but even more important when the hidden information can be used for practical implementation and use at the local level (Pang & Lee, 2008).

Sentiment Analysis and Topic Modeling DDEO

Two commonly used text mining methods used for social media surveillance research are topic modeling and sentiment analysis, particularly when one is conducting content analysis related research. As noted earlier, unsupervised topic modeling is a common approach used for Twitter data due to the exploratory nature of knowledge discovery that is required when analyzing unstructured textual data. The following section discusses different types of topic models developed for DDEO to discover latent topics; and sentiment analysis development that has contributed to understanding the natural language processing challenges presented with social media related research.

Sentiment analysis

One of the earliest known studies that examined the ability of automated sentiment analysis with comparison to human analysis method was work conducted on the hostile conflicts among American expatriate and Chinese managers in China (Doucet & Jehn, 1997). They used the Dictionary of Affect Language (DAL) database (Sweeney & Whissell, 1984) for the computation method; and an electronic thesaurus consisting of a compiled list of all words related to their core themes of conflict to be used for human analysis. A byproduct of their work provided evidence for the validation and use of automated techniques for sentiment analysis. The automated techniques of DAL showed the computing potential of analyzing an unlimited size of data without the monetary and expenditure of time that would be required by human investment (Leetaru, 2012).

The comparison of human and computational based sentiment analysis was groundbreaking during its time (Doucet & Jehn, 1997), but the researchers provided evidence that sentiment analysis requires in-depth and high intelligence that is hard to determine by human interpretation alone. Understandings of the sentiments expressed are more favorable than finding the overall feelings regarding a corpus of data. Nasukawa and Yi's (2003) work in sentiment analysis examined the issue of how opinions are expressed in texts and determining if they are positive, negative, or both. The goal of their model was to find sentiment expressions that are semantically related to the subject term within the text fragment, and the polarity that exists of the sentiment. While their work did not focus on the DDEO topics, their work was instrumental with developing a model that allows the capture of trends by "comparing sentiments for specific subjects with other subjects" in the data corpora (Nasukawa & Yi, 2003, p. 76).

Around this same time frame in the early 21st century, most models of sentiment analysis assumed the "selection of topics are independent given sentiment labels over words" (Rahman & Wang, 2016, p. 155). No research described how to explicitly model topic coherence and sentiment consistency in an opinionated text document so that one can accurately extract latent aspects and corresponding sentiment polarities. Rahman and Wang (2016) achieved topic coherence by enforcing words in the same sentence to share the same topic assignment and model topic transition between sentences. The researchers develop a unified topic model, or Hidden Sentiment Topic Model (HSTM) that drops the simple mixture assumption found that was used in most sentiment analyses at this time and treat the sentence as the basic structure; therefore, all words included relate to the

topic. However, their work used text generated from online reviews and its performance on Twitter specific data has been limited.

The academic domains of sociology and psychology, which have strong research tradition of using questionnaires and interviews, are also analyzing the data of social media to address psychology related problems. Using a microblogging social media site like the Twitter platform, researchers in China developed a depression detection model based on sentiment analysis (Wang, Zhang, Ji, Sun, & Wu, 2013). The sentiment analysis is based on a comprehensive vocabulary of Chinese words and man-made rules. Their sentiment analysis approach is significant with understanding characteristics of an event, predominantly those dealing with negative emotions.

A recent study with a focus on the DDEO topics demonstrated the importance of characterizing various emotions and the insight that can be gleaned from analyzing them individually and holistically as a dataset (Karami, Dahl, Turner-McGrievy, Kharrazi, & Shaw, 2018). However, their study did not involve domain experts in the development of their sentiment analysis approach. Despite this methodological area that could be improved in their study, their work demonstrated an effective approach for the deep analysis of sentiments that could be applied to other public health issues. The work of Wang et al. (2013) also supported understanding the polarity of ties between people that can help interested stakeholders (in this case psychologist and sociologist) understand sentiments of similar individuals and what topics are normally retweeted by them (Pang & Lee, 2008).

A more closely related study involving sentiment analysis and one of the DDEO topics is an aspect-level approach sentiment analysis for diabetic related tweets (Salas-

Zárate et al., 2017). Their work has contributed to further understanding of precision, recall, and F-measure of a sentiment analysis system. Also, their sentiment analysis aspect-level approach outperformed other similar proposals that focused on the health domain (Biyani et al., 2013; Powers, 2011; Rodrigues, das Dores, Camilo-Junior, & Rosa, 2016). Although this study used a manual approach for labeling the positive, negative and neutral tweets, it also brings into discussion the benefits of this approach has in a sentiment analysis experiment. Human coders infuse their background knowledge when reviewing a document; whereas, computational approaches evaluate the text based on the linguistical input provided (Leetaru, 2012). When performing content analysis, this should be viewed as a benefit and not a methodological inefficiency of infusing these quantitative and qualitative approaches to answer socially related questions from a computational framework.

Topic models

Semantic analysis assists with bringing understanding to a corpus of text (data). However, the unstructured nature of text within social media presents an additional set of challenges (Leetaru, 2012). In addition to the semantic rules that are involved with providing structure to unstructured text, topic extraction further supports the high-level of semantic comparison that is done when dealing with big data sets (Leetaru, 2012). As discussed earlier, topic modeling (for topic extraction) allows researchers to identify how words are related conceptually and provides a higher-level analysis that is complementary to the knowledge discovery that is involved with sentiment content analysis (Paul & Dredze, 2012). There are numerous topic models used for topic

extraction. This section will discuss selected topic models that have focused on health issues and provide detail explanation of the LDA model.

Public health researchers allocate a lot of time and resources to track public health related issues. The use of self-reported real-time information can be huge for health information and public health researchers. Previous studies have combined prediction markets and Twitter to predict certain ailments (Ailment Topic Aspect Model or ATAM) (Paul & Dredze, 2012). ATAM was developed as an extension of the LDA model with a focus on health concepts only. The model creates non-health topics and ailment specific topics. Paul and Dredze's (2011) early work in this area provided model development based on word distribution for topics according to disease names, symptom terms, and treatments.

Building on the discoveries made with this model, researchers Paul and Dredze (2011) proposed a more general model that could be used to predict multiple health ailments. To collect data used for analysis, the researchers used training data from a supervised classifier provided through Mechanical Turk. They also trained SVM with a linear kernel and uni-gram, bi-gram, and tri-gram word sequencing. The objective was to create a model that can discover a range of health topics and not just a single disease; since all topics discussed in Twitter are not clearly defined with regards to ailments. Their model accounted for this by making a category labeled as "Z" topics, where each message contained a distribution over topics. By using a switching variable, the model can determine if it was from an ailment related topic or category "Z." ATAM learns to group symptoms and treatments into latent ailments and regroup those remaining words into health related topics. They were able to show the model's ability to discover ailment

topics, but most importantly, that it has the potential to be used for syndromic surveillance.

Additional work by Paul and Dredze (2011) explored and addressed the linguistic complexity of the Twitter environment. Decoding slang words are important to understand public health outbreaks (Eisenstein, O'Connor, Smith, & Xing, 2010). Referencing their previously developed ATAM and its use of traditional model components, they sought to improve the robustness of the model by including prior knowledge from numerous public health resources to teach the model (ATAM +). Using more general names to identify ailments and ailment clusters, the model improved its ability to understand patients' sentiments. Moreover, this was one of the early works that focused on behavioral risk factors for syndromic surveillance and social media.

Paul and Dredze (2014) also explored different models to exploit the geographic information that is provided in the data corpus. The researchers explored an end-to-end framework (topic modeling framework) for data collection and analysis. The model is used in conjunction with the framework being developed by the authors to test its explanation of geographical and temporal trends with different health topics. The model was compared to the LDA model and ATAM. Using two websites, they collected 20,000 key phrases related to illnesses, symptoms, or treatments that were used for filtering. Results from the study showed that the model was able to detect topics that were often associated with ground truth surveillance and survey data (Paul & Dredze, 2014).

Another study by Hong and Davison (2010) suggested that with many research studies conducted, researchers only focus on the messages or the author, but never simultaneously. Their work focused on training a standard topic model within a short

text environment. The Author-Topic Model (AT model) is an extension of the LDA where each word {w} in a document is associated with two latent variables: an author (x) and a topic (z). The results from their research showed how topic models could be conducted for short text environments (Hong & Davison, 2010). They demonstrated that when trained for shortened messages, the topic model can be used as a standalone model or complementary to other topic models. While it was not a direct implication of the study, their work provided an exploration of assigning certain behaviors to health topics to identify the probability of the behavior with the chronic disease (or health topic).

Comito, Pizzuti, and Procopio (2017) focused on developing a model that helped with detecting health topics and providing an understanding of how people talk about health on Twitter. Previous topic models such as LDA, Doc-p, and SFPM (Soft Frequent Pattern Mining), do not perform well with identifying topics within tweets with low word frequency, and when they are related to a health issue. Their HealthS-tweet model accounts for these drawbacks. The researchers applied their HealthS-Tweet algorithm to secondary Twitter data collected from September 2015 to April 2016. Their study demonstrated the additional insight that can be gained from low-frequency topical data. Comito et al. (2017) showed that their model outperformed LDA and Doc-p with identifying low-frequency topics; mainly for flu and influenza-related topics.

With the significant increase in data from Twitter as a source of health information since its inception in 2006 (Finfgeld-Connett, 2015) and modeling various behavior from this data source, managing large datasets has become necessary. Unsupervised text mining approaches, such as LDA, are exploratory in nature and discover patterns in unlabeled or unstructured datasets. A new approach, Fuzzy Latent

Semantic Analysis (FLSA), is a clustering method that accounts for the ambiguity that exists in language and language association (Karami, Gangopadhyay, Zhou, & Karrazi, 2015a; Karami, Gangopadhyay, Zhou, & Kharrazi, 2015b; Karami, Gangopadhyay, Zhou, & Kharrazi, 2017). This study is one of the first to explore this fuzzy clustering approach for topic modeling and addressing the issue of redundancy that is found in the LDA model. While the quantitative experiment in the research show that FLSA produced better performance for document classification, document clustering, document modeling, and execution time, FLSA has mostly been utilized within the health domain for images and the Twitter dataset used in this study is based on unstructured data collected from Twitter.

Latent Dirichlet Allocation Model

For this research, the Latent Dirichlet Allocation model (LDA) model will be used to conduct topic modeling from Twitter data. The LDA allows for the probabilistic model of a corpus. Therefore, each topic can be characterized by a probabilistic distribution over a set of documents. The model can be trained to discover topics for a group of sentences, documents, or unstructured words (Blei, Ng, & Jordan, 2003). Documents are represented within random mixtures in NLP. Depending on the purpose of the study, the attributes of what defines a document can be arbitrary (see Buckland's (1997) article for a detailed discussion of what is a document). Documents for this study are defined as individual tweets. Essentially, a corpus is made up of a collection of related tweets or documents – based on the query terms used.

As explained, LDA distributes documents over topics and topics over the words that allow for semantically, coherent word sets. Since all topics found on Twitter do not reference ailments, there is potential to argue the impact the ATAM model has with ailment topic discovery (Blei et al., 2003; Paul & Dredze, 2011; Paul & Dredze, 2014). Statistically, the ATAM model did not have significantly different results than the commonly used LDA topic model (Paul & Dredze, 2014, p. 9).

While these studies did not specifically look at the concerns of users regarding obesity, they provide useful insight into keywords used by Twitter users in relation to weight management concerns and the chronic health issues that are likely associated with obesity. Results from these studies also show that LDA maintains its relevancy as a topic model for discovering health related topics.

In addition to using the LDA model to topically learn and infer behavior, it is important to evaluate the models when describing the usefulness of the information discovered. In an often-referenced study of using human evaluation (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009), the authors presented new quantitative and qualitative methods to measure semantic meaning from the topic model results. Human evaluators were tasked with selecting topics inferred by the topic model and word intrusion detection to understand the semantic coherence of a topic. Using the services of Amazon Mechanical Turk, users were assigned to a word intrusion task and topic intrusion task (Chang et al., 2009).

For the word intrusion task, participants selected the word that did not belong to the formation of the topic listed. For the topic intrusion, they were given a set of four topics, with a short blurb from the article or text, and they were instructed to remove the

topic that did not inform or represented the document. Results showed that although the models were good at achieving prediction perplexity, they underperformed and were less interpretable in latent spaces. The human evaluators utilized a different logic to assigned topics when the documents contain multiple disparate topics (Chang et al., 2009). In addition to the knowledge that is derived from computational methods for data collection and content analysis, there is a need to qualitatively evaluate the information and its utility for decision-making.

This research is significant because it provides a framework to dynamically identify the obesity-related medical concerns associated with chronic diseases and health behaviors based on Twitter users' information. Also, the research incorporates a qualitative evaluation approach during the topic analysis process. Qualitative evaluation or characterization of the topics is not always used because it is labor intensive and often subjective.

Chapter 3

Methodology

Framework Overview

The surveillance framework for this study uses a multi-component structure to retrieve relatively real-time information that is used for the assessment of the four health issues and the discovery of chronic health conditions. The framework is predicated on its ability to assist clinical and allied health professionals with decision-making that is faster and more economical than traditional health surveillance systems. A recent multi-component framework, that uses sentiment analysis and the LDA model for topic discovery, demonstrated the ability to characterize common health issues associated with DDEO (Karami et al., 2018). This multi-component framework used topic modeling, natural language processing, and linguistics analysis to identify positive/negative emotions and topics. The sentiment analysis tool used in their framework approach allows you to identify topics to a set of corpus based on sentiments (positive or negative) and linguistically linked commonly used words to health behaviors (Tausczik & Pennebaker, 2010).

The systematic steps in constructing this surveillance framework include data collection, data cleaning, sentiment analysis, topic discovery, topic content analysis, and evaluation (Figure 3.1). These steps were applied to the corpora of English tweets collected to identify and assess the concerns regarding DDEO.

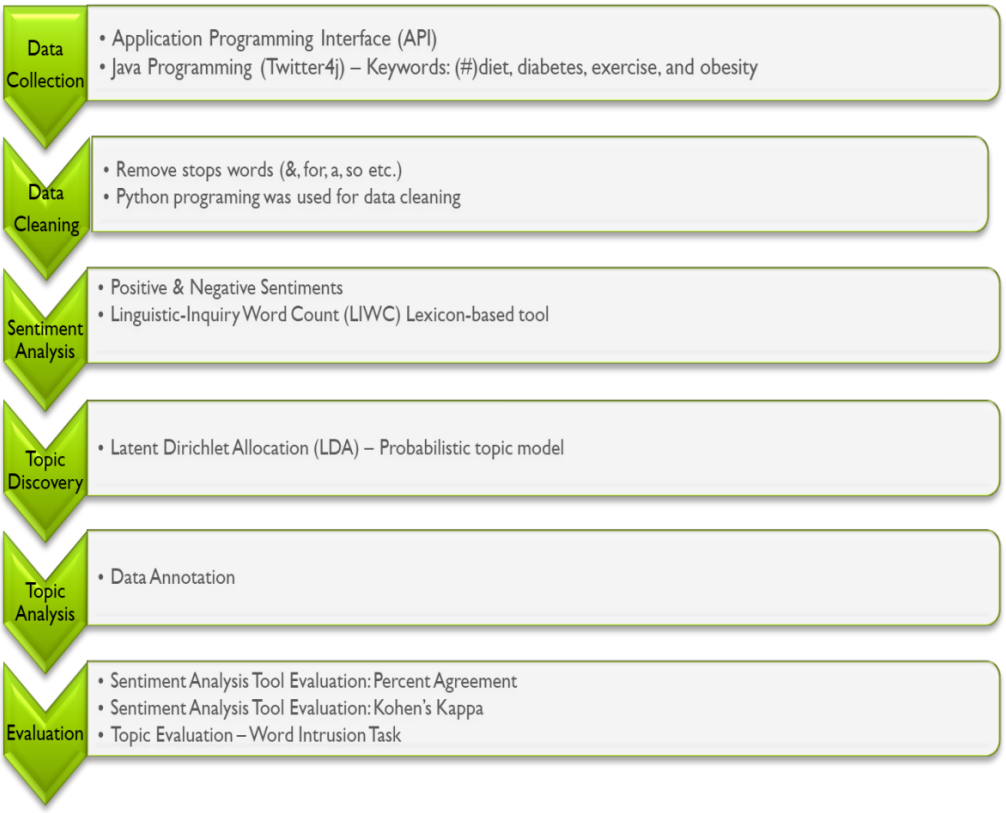


Figure 3.1: Analytical Framework

Data Collection

The Twitter API – with its real-time data collection ability – was used to collect data through the Java programming (Twitter4j) software tool.[4] The Twitter API allows you to stream roughly 10% of the publicly available tweets. Therefore, data were streamed using the keywords in Table 3.1 from June 2016 to August 2016. The traditional spelling and hashtag version of DDEO was used for data collection. Paul and Dredze (2011) work involving the obesity ailment demonstrated the connection among

[4] http://twitter4j.org/en/

DDEO and the use of DDEO for effective query searching. A total of 15 million tweets were collected during this three-month period.

Passive monitoring does not require active participation from the users, while allowing researchers to identify what individuals are thinking, feeling, and doing. Passive monitoring is also a low-cost and easy approach for data collection (Komito, 2011; Paul & Dredze, 2017, p. 24). It is important to clarify that collecting this type of data is not based on observed behavior. However, passive monitoring allows us to gain insight into behaviors. Making claims about observing behavior from the data collected goes beyond the scope of this study and the computational approaches applied in this framework.

Twitter data in its rawest form is unstructured and a complex structure of various metadata. Computational text mining methods, like sentiment analysis tools, were developed with the purpose of extracting information from unstructured, user-generated content (Pang & Lee, 2008). However, there is uncertainty about what the data from Twitter can tell us and applying long-standing methodologies and frameworks to this form of social media data (Paul & Dredze, 2017). In-spite of the uncertainties that do exist, Twitter data is generated for the purpose of being shared and created by people:

1. Twitter data can tell us that Friends on Twitter are likely to share similar demographic variables and food interests (Abbar et al., 2015).
2. Twitter data can provide us insight into the exercising prevalence (or lack thereof) for various communities (Nguyen et al., 2016).
3. Twitter data can be used by doctors to improve patient-to-doctor communication and improve the delivery of healthcare (Hawn, 2009).

4. Twitter data can be used for general-purpose (Abbasi et al., 2014) and specific health related (Nguyen et al., 2016) analysis.

The Twitter data in this study focuses on DDEO health issues for analysis. The Twitter data regarding DDEO in this study will provide an alternative approach for understanding the experiences of users (Shaw & Karami, 2017). To collect the necessary data, it is important to identify the correct semantics to capture the data regarding DDEO.

To verify the relevancy of data collection in this study based on the query search terms used (see Table 3.1), a combination of by-hand and keyword searching using the Twitter web search interface[5] was used to verify that the data collection strategy support the research study. Ten messages per search query were identified through both approaches to verify the correct information was being collected, and that enough data could be collected to mine and analyze. Figure 3.2 shows an example – paraphrase for privacy – of the collected tweets.

Table 3.1: Queries used to search Twitter API for Tweets

DDEO TOPICS	TWITTER API QUERIES
DIABETES	#diabetes OR diabetes
DIET	#diet OR diet
EXERCISE	#exercise OR exercise
OBESITY	#obesity OR obesity

[5] https://twitter.com/search-home

Figure 3.2: A Sample of Collected Tweets

Data Cleaning

Obtaining and cleaning data is an often-overlooked process in the data analysis process. Preparation with regards to cleaning and filtering the data impacts the remaining portions of the analytical process. Data cleaning also reduces the level of noise or text that has no meaning (Leetaru, 2012)).

During the data cleaning process using a Python-based program, additional Retweets (RT) and the tweets having a URL for avoiding spams were removed – this assist with improving the quality of the data and focuses on personal experiences of the users. Data cleaning was also conducted to remove stop words. For example, words like

of, *the*, and *in* were removed using the program and special characters were also removed as part of this process.

Sentiment Analysis

Sentiment analysis is a text mining method to find favorable and unfavorable opinions towards specific subjects. This text mining method can determine if a sentiment is positive, negative, or both (Finfgeld-Connett, 2015). This portion of the research will answer, what are the positive and negative experiences of Twitter users regarding DDEO? Positive sentiments regarding healthy behaviors (i.e. food choices) show that those individuals within that geographical location are more likely to be healthier overall (Nguyen et al., 2016). Performing sentiment analysis also assists with addressing the question what additional health conditions are prevalent based on the Twitter users' sentiments regarding DDEO?

There are two main approaches to sentiment analysis: a learning-based and lexicon-based approach. Machine learning techniques build classifiers from data to infer positive and negative polarity of the words (Signorini, Segre, & Polgreen, 2011). The lexicon-based approach uses a pre-defined dictionary of the positive and negative words to find the frequency of positive and negative words (Medhat, Hassan, & Korashy, 2014). By using the Linguistic Inquiry and Word Count (LIWC) software, the research utilizes the lexicon-based machine learning approach.

First, the cleaned corpora were uploaded into the LIWC software based on DDEO to detect sentiments. To focus only on the positive and negative sentiments, all additional behavioral variables within the software were disabled (Figure 3.3).

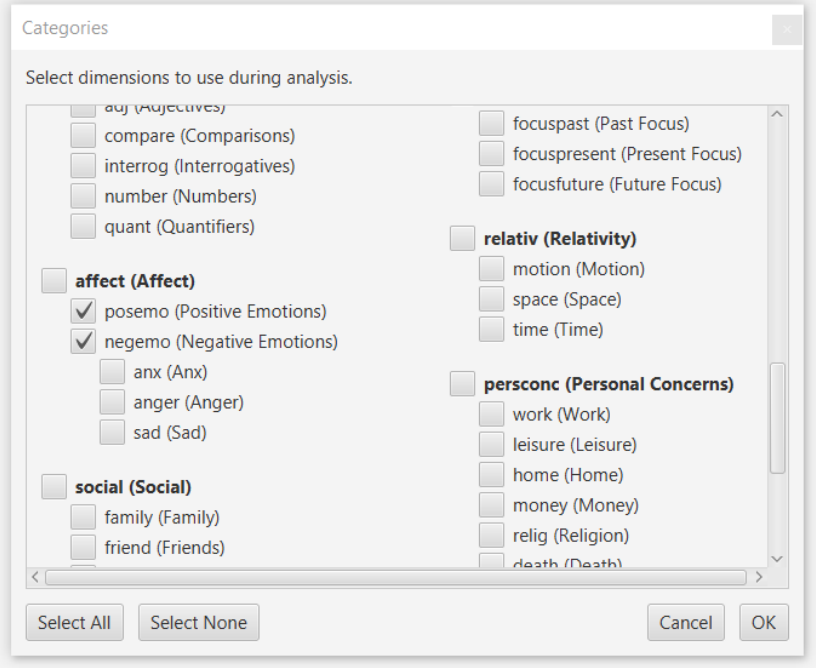


Figure 3.3: Dimension Selection Interface for LIWC Sentiment Tool

To identify an overall direction the different corpora of data were indicating, each corpus of data representing the DDEO topics was identified as negative or positive. To determine whether a tweet was positive or negative, the difference between the positive value and negative value was calculated. If the overall value was positive, the tweet was labeled positive. Otherwise, if the overall value was negative, it was labeled negative. Tweets that fell into the category of neutral were removed as the research questions guiding this portion of the research study focus on positive and negative sentiments of people regarding DDEO.

Topic Discovery

To discover the hidden semantic structure in the corpora of negative and positive tweets, topic modeling was used. This step of the framework addresses the research question, what are the positive and negative health experiences with respect to DDEO? While there are numerous models used for topic modeling of unstructured text data (Chang et al., 2009; Fisher, Amico, Fisher, & Harman, 2008; Paul & Dredze, 2012), LDA is the most well-known model topic model (Blei et al., 2003; Comito et al., 2018; M. Paul & Dredze, 2011). Topic modeling has been used to discover relevant clinical concepts and structure in patients' health records, find patterns that exist within genetics data, and has impacted the development of other topic model approaches (Arnold, El-Saden, Bui, & Taira, 2010; Comito et al., 2018).

LDA semantically cluster related words in a fuzzy way. LDA makes the probabilistic assumption that semantically related words reflect latent or undiscovered topics (Wallace, Paul, Sarkar, Trikalinos, & Dredze, 2014). Therefore, each document (or in this case each tweet) can be assigned to each of the topics in a corpus with a varying degree of weights (Figure 3.4).

To identify the optimum number of topics, a similar approach used by Chang et al. (2009) was adapted for this study. The number of topics was set at 25, 50, and 100 to explore the topic composition and determine which one provided an adequate representation of the topics. Previous studies have incorporated prior knowledge in the development of various topic models (Paul & Dredze, 2011). Deriving topics without prior knowledge provides a more inductive approach to discovering topics and allows an exploratory analysis of the topics identified.

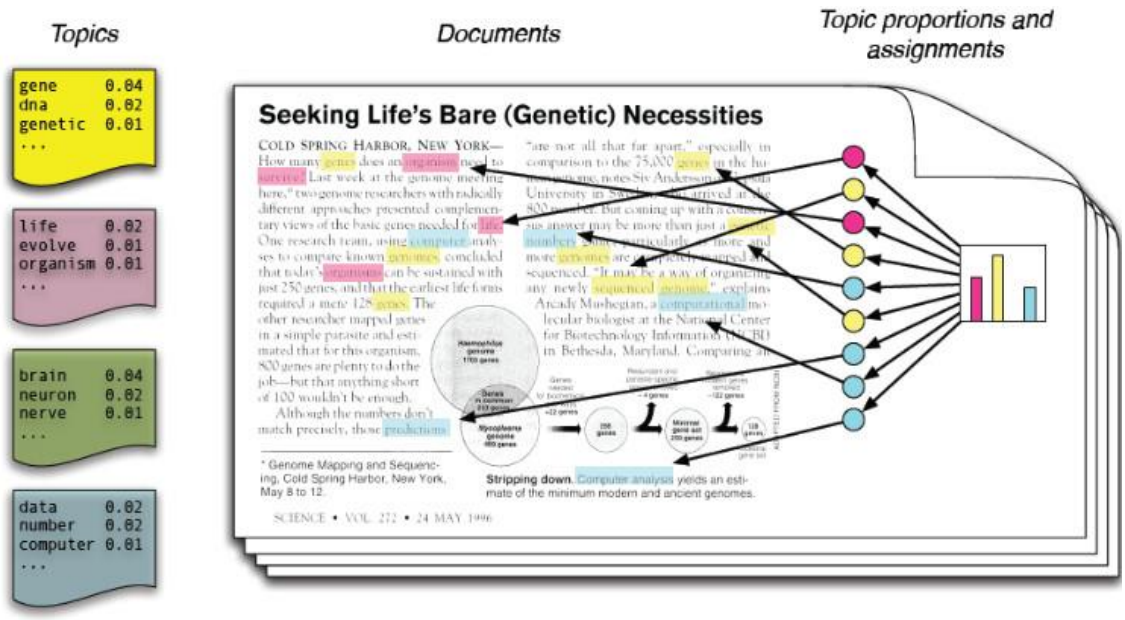


Figure 3.4: Example of Latent Dirichlet Allocation Model (Blei, 2012)

This approach also lends itself to the strength of the LDA model as an unsupervised machine learning method that allows for the identification of insightful patterns (Blei et al., 2003; Chang et al., 2009).

Topic Analysis

A multi-step approach was taken to perform data annotation on the topics. To determine if topics were health-related, they were first labeled as related or unrelated. Once the health-related topics were detected, the next step was characterizing (labeling) the topic content. This study adopted a similar approach conducted by Miller, Banerjee, Muppalla, Romine, and Sheth (2017) where they categorized communicable disease (Zika virus) – this method is applicable with identifying granular topics that are represented within the topics. While pre-defined labels for the categorization of topics

were used in the Miller et al., (2017) study, this research utilized a more descriptive, qualitative process to designate and characterize them.

The data annotation approach allows the qualitative labeling of topics while characterizing the topics (PyData, 2017). Performing this characterization of topics is labor intensive, but humans perform a complex level of understanding when interpreting the topics. Probabilistic models arrange words in a topic based on their probable association with the topic.

Evaluation – Sentiment Analysis

The final step in the analytical framework is the evaluation of the sentiment analysis tool (LIWC) and the topic model used in the study (LDA). Previous studies have incorporated computational and human evaluation methods (Chang et al., 2009) – this study used evaluation methods that are not often conducted in topic modeling studies when it comes to evaluating topic models (PyData, 2017). Since Twitter was used for data collection, it was important to use participants that would represent the general population. Therefore, using workers provided by Amazon Mechanical Turk was appropriate for this study; this includes population representation (discussed later) and ability to complete the task required for sentiment analysis evaluation.

To evaluate the sentiment analysis tool, a preliminary test was conducted with two raters to determine agreement and receive feedback on areas to improve the test. Convenience sampling was used to identify the participants for the preliminary test. Two Library and Information Science graduate students consented to participating in the study. The preliminary test – study ID: Pro00078907 – was approved by the University

of South Carolina (USC) Institution Review Board (IRB) (the preliminary test would be updated later for additional IRB review). First, a sample of twenty-five tweets was conveniently selected from each of the topics (100 total). The tweets were selected based on the following criteria:

1. A variation of short and long semantic structure with a minimum of three words
2. Tweets categorized by the sentiment analysis tool as positive or negative. Since neutral tweets were outside the scope of the originally accepted proposal, they were not analyzed during the data annotation process[6] (Ahmed, 2016).
3. Absent of multiple shorthand words – words like bcuz (because), wyd (what are you doing), ttyl (talk to you later) – that would reduce the comprehension of the tweet by the coders.

Any modification of the tweets was kept to a minimum to retain the authenticity of the tweets. The sentiment analysis tool only recognizes text during the analysis process; therefore, hashtags, ampersands, exclamation marks and other special characters were not recognized during the analysis process. Also, the grammatically correct spelling of the word was provided within parentheses for tweets that included shorthand words.

Using Google Forms, the tweets were randomly arranged in their respective category. To increase participation, participants were offered a $5 Amazon gift. The raters were informed that their responses would be stored on a password protected, limited access, departmental desktop located at the institution site (USC). For each

[6] Appendix A contains the complete study invitation provided to the participants and paraphrase examples of the tweets participants evaluated. Paraphrasing the tweets addresses the ethical concerns of using Twitter data – particularly with the difficulty of obtaining informed consent when collecting data for large datasets.

category, the raters were instructed to label the tweet as positive or negative in relation to the category represented by the tweet.

After analyzing the preliminary results and consultation with the participants, adjustments were made to the questions and four tweets were replaced that were more semantically coherent. This included removing tweets that coders identified as difficult to read and replacing shorthand words that were not replaced during the selection process. Participants in the preliminary study also discussed their frustration of not be able to identify tweets that they were unsure about or deemed neutral. As a result, an additional option of neutral was added for the raters to select.

After corrections were made based on the preliminary sentiment evaluation test, raters were recruited using Amazon's Mechanical Turk (MTurk)[7]. MTurk is a marketplace where businesses and researchers can recruit workers to perform tasks that require human intelligence (Human Intelligence Task). MTurk has increasingly been used by researchers who require human intelligence and has been developed as a gold-standard for data studies involving labeling (Berinsky, Huber, & Lenz, 2010; Deng et al., 2009). MTurk workers for this study were limited to residents of the United States (U.S.). A 60% or greater rating was an additional parameter used to select MTurk workers. The approval rating is based on the worker's accuracy and completion of tasks. Premium qualifications options (age, income, employment status etc.) offered by the crowdsourcing marketplace were not included. Demographic information that includes age, sex, race and socio-economic status etc., were not criteria that determined participation in the study. This allows for anyone to participate in the study, as the

[7] https://www.mturk.com/

Twitter environment is only restricted to individuals reporting to be 13 years of age or older; MTurk users must be 18 years of age or older. A total of eleven raters were recruited through the crowdsourcing marketplace.

After the responses were collected from the MTurk workers, they were aggregated into an excel spreadsheet with their respective results. To perform the rater agreement calculation, a binary classification approach was utilized. A value of one is allocated if the coder and sentiment analysis tool identified the tweet as positive, and a value of 0 was provided if the tweet was labeled as negative. To determine the level of agreement between the rater and the tool, initial proportion of agreement was calculated (Figure 3.5) and Cohen's kappa (McHugh, 2012b).

Var#	Raters		Difference
	Mark	Susan	
1	1	1	0
2	1	0	1
3	1	1	0
4	0	1	–1
5	1	1	0
6	0	0	0
7	1	1	0
8	1	1	0
9	0	0	0
10	1	1	0
Number of Zeros			8
Number of Items			10
Percent Agreement			80

Figure 3.5 Example of the method used to calculate percent agreement (McHugh, 2012)

While both methods have their strengths and limitations, they both offer insight into the data. Percent agreement provides the direct interpretation of the data, but kappa accounts for the possibility of chance among the raters (McHugh, 2012b).Cohen's kappa measure is determined by the amount of agreement over and above the agreement expected by chance (Di Eugenio & Glass, 2004). Cohen's kappa is visually described in the formula below, where Pr(a) represents the actual observed agreement and Pr(e) represents the amount of chance agreement (Mchugh, 2012a; McHugh, 2012b):

$$\text{kappa (k)} = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

When conducting Cohen's kappa, the crosstab table (or cross tabulation) must be the same. Therefore, your matrix table must be a 2x2, 3x3, 4x4, etc. For example, Rater 1 and Rater 2 (Figure 3.6) marked 222 observations as normal or abnormal.

		Rater 1		Row Marginals	
		normal	abnormal		
Rater 2	normal	147	3	150	rm^1
	abnormal	10	62	72	rm^2
Column Marginals		157	65	222	n
		cm^1	cm^2		

Raw % Agreement

$$\frac{147 + 62}{222} = .94$$

Figure 3.6: Example of data used for Cohen's kappa calculation (McHugh, 2012b)

They both have the option of selecting normal and abnormal. If Rater 2 was given the option to select normal, somewhat normal, and abnormal, but not Rater 1, Cohen's kappa could not be used to calculate interrater reliability (McHugh, 2012b). While previous research using this method has identified the two coders (raters) as people (McHugh, 2012a), this study identifies the computational tool as a rater. Cohen's kappa was computed using IBM's SPSS software package.

Evaluation – Topic Evaluation

The topic evaluation for this study used a familiar approach to measure how well the inferred topics match human understanding of the topics. The word intrusion task presents the person with the probabilistic ordered words representing the topics. Their task was to select the word that is most likely to be unassociated with the word cluster (Chang et al., 2009). Using the crowdsourcing marketplace of Amazon Mechanical Turk workers, MTurk workers were presented with 40 topics. MTurk workers from the U.S. with an approval rating of 80% or better were selected. Again, certain demographic factors were not necessary to be a participant. DDEO was represented by ten topics (Diet = 10, Diabetes = 10, Exercise = 10, and Obesity = 10 for a total of 40 topics). The ten topics consisted of give positive and five negative topics (Diet – Positive = 5, Diet – Negative = 5, Diabetes – Positive = 5, etc.) Each one of the 40 topics contains five words. Each of the five words presented to the MTurk worker included one word that was the unassociated (intruding) word. The unassociated word has a high probability in another topic, but a low probability in the topic the MTurk workers were evaluating.

To analyze the results, a similar analysis approach conducted by Paul & Dredze (2014, p. 8) was used. While their approach focused on the labeling of clusters for the ailments (or topics), the approach used in this study focused on word intrusion. For this study, the purpose of word intrusion is not to understand agreement among the participants but gain an understanding of humans' ability to detect words that are semantically incoherent to a topic. This study was approved by the University of South Carolina IRB (ID: Pro00078907).

Additional signs of chronic disease can be identified through a user's tweets. Considering ethical and privacy concerns, this offers researchers or healthcare providers the ability to track patients and behavior. This framework allows health professionals and/or health care providers to be proactive in addressing emerging health concerns. As stated, Twitter has a desirable property of being a real-time data source, in contrast to surveys and surveillance networks that can take weeks or even years to deliver information (Paul & Dredze, 2014).

When using social media as the primary tool for data collection, there is the concern of excluding large segments of populations. Older Americans and low-income families that do not own a mobile device are typically the populations identified in this argument. There is also the possibility that topics from user sentiments will be included in both positive and negative sentiments. As noted in Chapter 2, there is no direct observation of behavior that happens within SNSs, but the use of probabilistic models allows us to infer from the content analyzed. These concerns will be addressed in Chapter 5.

Chapter 4

Results

This chapter provides a detailed explanation of the research study results. Moreover, each of the research questions guiding this study will be analyzed individually to provide deeper insight into the results. Each query topic and the respective sentiments will be addressed individually with brief discussion included in the reporting of the results. Chapter 5 will provide a discussion of my positionality during the topic analysis step of the framework.

Within the practice of analytics, a crucial component of the analysts or data scientist's job is deriving knowledge from the results. According to Gary M. King - Director for the Institute for Quantitative Social Science at Harvard University – "The data itself isn't likely to be particularly useful….The question is whether you can make it useful."[8] While analyzing the data output from computational models and making sense of the patterns, it is important to recognize that the data annotation method used in this study is relatively subjective. Unlike the trained model that is used in supervised machine learning, the unsupervised machine learning model text mining method (program with the correct parameters) is predicated on its ability to capture the semantics representing a topic (Chang et al., 2009). During the data annotation process, the

[8] https://gking.harvard.edu/

characterization ("labeling" is the often-used term in the machine learning community) of the topics is based on the analysis of me as the researcher as informed by the literature.

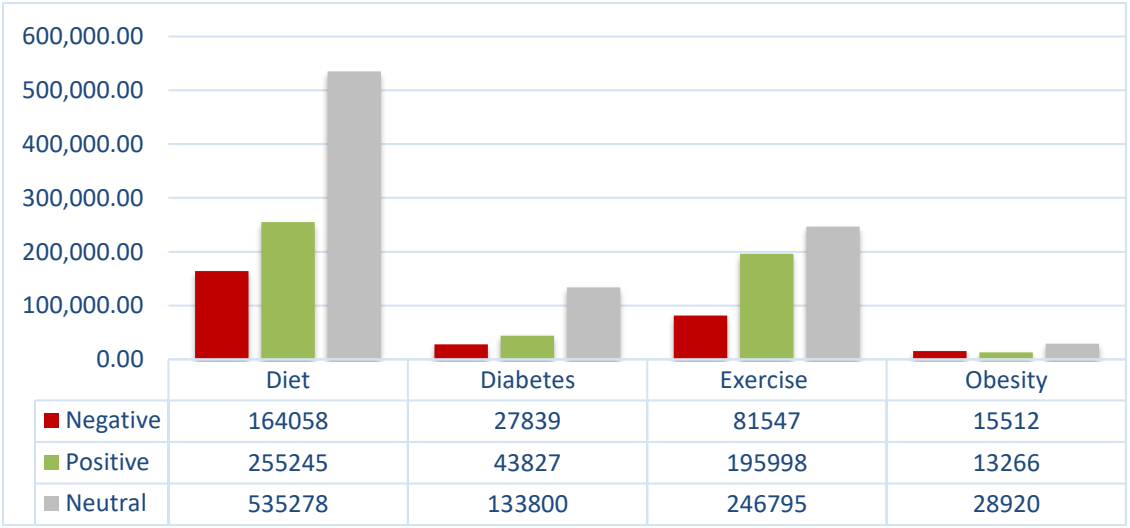
While this research study uses a quantitative methodology, the nuance of data science research can raise concerns for some qualitive researchers and the idea of complete objectivity in the research process. Algorithmic bias in machine learning and natural language processing affects policies aimed at different genders, interview decisions, and communities that consists of minorities and low-income people (Knight, 2017). The tools that were used as part of this study are developed based on well-defined mathematical, statistical, social science, and computer theory (Blei et al., 2003; Paul & Dredze, 2011; Tausczik & Pennebaker, 2010); but the algorithms and dictionary development are ultimately affected by human design decisions (Tausczik & Pennebaker, 2010). Therefore, it is important to acknowledge the limits (address in sections of chapter 4 and extensively in chapter 5) of the statistical and computational tools used in the steps of the analytical framework.

After conducting the data cleaning in step two of the framework on the 15 million tweets collected (this included removing retweets), 1.7 million tweets were used for sentiment analysis. Of the 1.7 million tweets, 288,956 of the tweets were negative, 508,336 were positive, and the remaining 944,793 were neutral (Figure 4.1). This smaller number of tweets supports Finfgeld-Connett's (2015) position on the need to explore Twitter for small-scale studies. The three-month data collection period for this study provides the opportunity for more condensed and in-depth analysis as it covers one temporal season, a typical increase in indoor and outdoor exercise activity and change in eating habits (Turner-McGrievy & Beets, 2015). Figures 4.2a and 4.2b provide an

overview of the individual positive and negative topics that were discovered during the topic analysis step.

Positive and Negative Sentiments

Finfgeld-Connett (2015) highlighted three health-related messages that are dominant on Twitter; (1) commentaries and opinions, (2) highly personal, moment-to-moment sentiments and emotions, and (3) informational, such as vaccine information. The first and the second points of Finfgeld-Connett's (2015) findings work support the importance of Twitter with understanding the positive and negative sentiments of the public, and primary representation of the tweets collected for this research. However, this does not withhold the notion that some of the latent topics discovered may support Finfgeld-Connett's (2015) third dominant message represented on Twitter.



	Diet	Diabetes	Exercise	Obesity
Negative	164058	27839	81547	15512
Positive	255245	43827	195998	13266
Neutral	535278	133800	246795	28920

Figure 4.1: Sentiment Polarity for DDEO

Health related habits of people - food consumption and exercising – further show Twitters' ability to be comparable to the BRFSS identification of risk behaviors. Five hundred and sixty-eight (or 71%) of the 800 topics were identifiable and related to DDEO. Sixty-seven percent (n=266) of the positive topics and 75% (n=302) of negative topics, from the total 400 topics for each sentiment, were identifiable. These numbers also include the tweets that were unrelated to DDEO but representing a topic. Negative sentiments – exercise

Overall, of the topics generated by applying the topic model approach previously outlined to the negative exercise corpus, 90% were identified as actual topics during the data annotation process (Table 4.2). Of the total dataset, 54% were identified as health topics directly related to exercise and the remaining 46% of the topics were unrelated to the exercise topic – with 10% representing inconclusive results from the total dataset. Inconclusive results (or topics) are defined as those words representing a topic with no coherent – after being analyzed – meaning. As seen in Figure 4.1a, results from the data annotation process show that the model identified 24 individual topics.

The three dominant health topics - as determined in step two of the data annotation process – were physical activity, chronic conditions, and diet. The cluster of words representing physical activity includes *bike*, *walk*, *fit*, *dance*, and *cardio*. Table 4.2 provides an example of the top five topics identified through the data annotation process. While lifestyle was one of the least represented topics (2), it has the highest topic probability among the corpora of data with 6.5%.

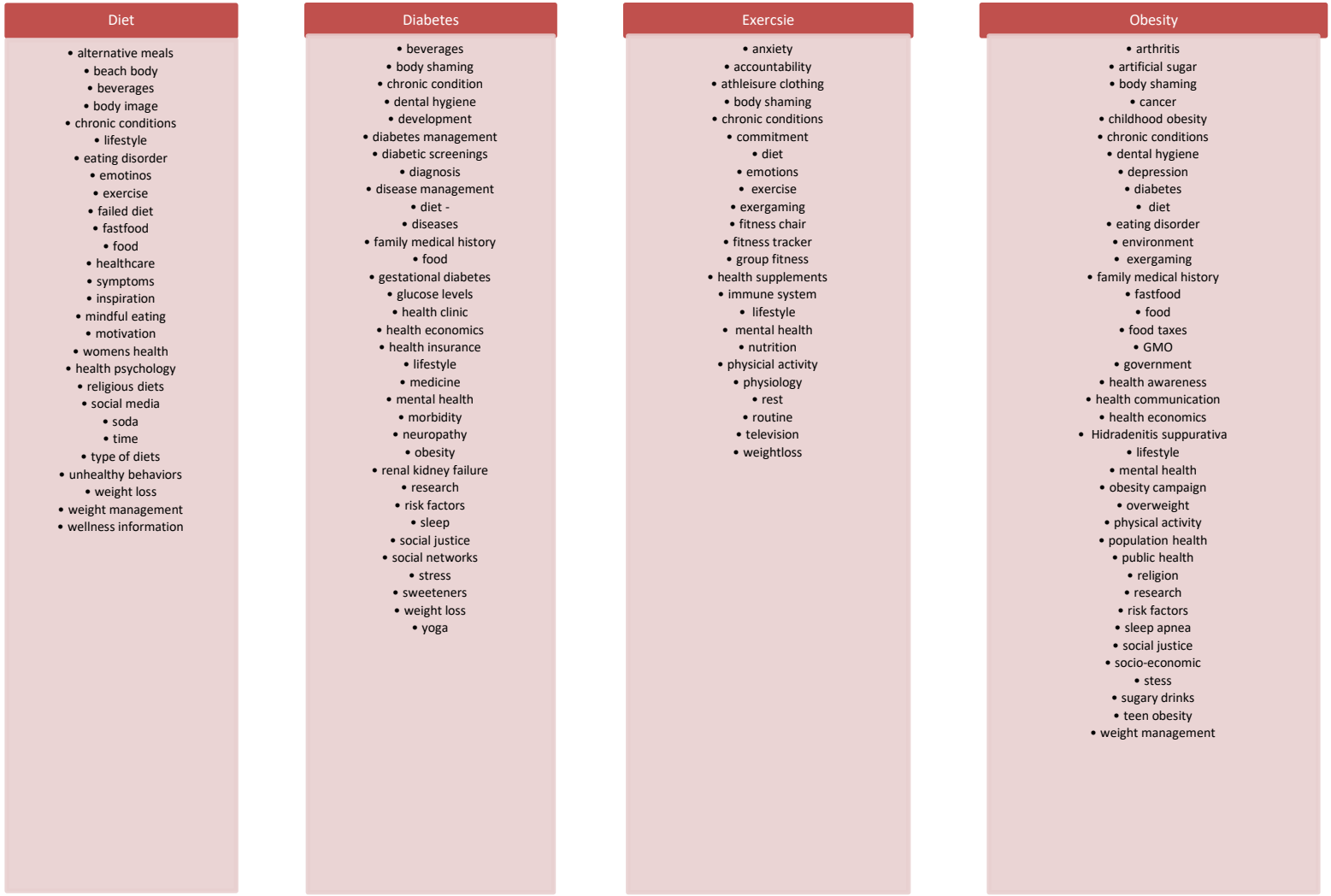


Figure 4.2A: Complete List of Negative Health Topics

Figure 4.2B: Complete List of Positive Health Topics

This is a possible opportunity for public health professionals and health psychologist to work with experts in the gaming community to build better gaming environments that consider the environment and psychological impacts that may be unbeknown to that community of experts.

Table 4.1: Top Five Negative Health Topics for Exercise

	Physical Activity 8 %	**Chronic Condition 5%**	**Diet 5%**	**Weight loss 3%**	**Lifestyle 2%**
T[1]	exercise room dance dancing song listen songs	exercise pain back knee hurts leg chronic	exercise calories low fat calorie carb carbs	exercise count cost problems regrets cutting loser	diet exercise regime results routine wishes lifestyle
T[2]	low impact high intensity exercises aerobic swimming	blood exercise pressure stress high diabetes cholesterol	weight exercise lose diet losing body muscle	Lost lbs lose pounds weeks months gained	exercise time hard life struggle hours motivation
T[3]	Fitness exercise workout health gym cardio training	exercise back bad hell painful keeping problems	exercise eat water cut food sleep carbs	fat body lose belly cardio stomach diet	

Exergaming appeared as a negative topic for exercising among Twitter users. If progression towards more augmented reality gaming continues as expected, game

designers must determine how they will incorporate safety parameters that will minimize the risk of the user (Heaven, 2016).

Positive sentiments – exercise.

After identifying the top positive exercising activities, overall, 57% of the positive data corpus represented exercise related topics. Exercises such as walking, weight training, and the attending the gym were the most represented words when describing the physical activity topic. As noted in Table 4.1, exergaming was one of the top five positive health topics for exercise. The rise of mobile gaming applications – like Pokemongo – have positive sentiments with addressing issues of obesity and increasing physical activities like walking. These types of applications are shifting the way that people and researchers think about how we engage in physical activity (LeBlanc & Chaput, 2017).

Two intriguing topics that were only represented once but pertain specifically to women were women's health and menstrual cramps. The cluster of words characterizing the topic menstrual cramps include *helped*, *good*, and *love*. This latent topic may support the arguments for exercising that helps to alleviate cramp related pain. Based on the words constructing the topic for women's health, pregnancy seems to be the health issue represented. Studies have focused on women's health as it pertains to breast cancer awareness and social media (Thackeray, Burton, Giraud-Carrier, Rollins, & Draper, 2013; Vraga et al., 2018); but the identification of these two latent topics may speak to the additional research that is needed on women's health issues.

As expected, physical activity was the most represented topic. Therefore, focusing on the words characterizing the topic will provide deeper perception and understanding of this topic. In particular, it is important to focus on the type of physical activities that are being performed and how they relate to the other topics represented. Time and lifestyle are two topics that appear here, but also in the additional topics to follow. Therefore, it is necessary to address the subtle differences between the two topics. As you see in Table 4.1 above, time is a word probability in the topic Time and Lifestyle. When performing an initial analysis of the word cluster, the decision may be to also annotate the lifestyle topic as time.

When you examine the word cluster for lifestyle further, time is given a lower probability and it also includes possible barriers (sore, back, tired) to engaging in physical activity. The higher probability of positive words such as good and great indicate there is a benefit of exercising. While there are certain barriers that may prevent someone from adopting a physically active lifestyle, there are benefits to engaging in exercise.

Negative sentiments – obesity

Sixty-six percent of the topics were identified as relating to the obesity topic. Nine percent of the topics were inconclusive based on the word probability for the topic and a total of 34% were unrelated to obesity. Some of the unrelated topics associated with the negative sentiments of obesity include gun violence, social justice, news, government, healthcare system, racism, and feminism.

Table 4.2: Top Five Positive Health Topics for Exercise[9].

	Physical Activity 16%	Lifestyle 3%	Mental Health 3%	Exergaming 3%	Time 3%
T^1	exercise mind yoga meditation breathing activity peace	exercise good time morning weeks yesterday day	exercise health physical mental benefits improve positive	exercise pokemon playing play pokemongo people walk	exercise day week minutes days times hours
T^2	exercise muscles good core breathing abs hold	style practice happiness life success control mind	social mental health increases energy wellbeing levels	Poka mon playing benefits athlete good positive	years today good time weeks gym months
T^3	strength workout training weight squats weights legs	exercise health life food daily lifestyle regular	helps brain energy function stress levels system	people fun game playing walking app egg	exercise time day life good find work

People's negative sentiments on feminism were found through the following tweet: "*I don't get why feminists encourage obesity*." For a topic such as gun violence, a parallel is often made between gun violence, the obesity epidemic, and the tools

[9] The words for the topics consist of low and high probability words to provide a representative sample.

contributing to the epidemics[10]. It was difficult to understand where some of the other unrelated topics – smoking, pool party, tanning, and sex-trafficking – fit into the explanation of the negative topics for obesity

Childhood obesity was identified as the topic that most users have negative sentiments about. Based on the probability of the word cluster for T^1 in Figure 4.1, exergaming is a sub-level topic that can be identified. In this example, the words *pokemon, pokemongo,* and *game* are also high probability words that are related to the exergaming topic (Table 4.3). The previous section, on positive topics for exercising, listed exergaming as one of the top five topics. There are benefits of exergaming as an exercise alternative, but the negative topic of exergaming may speak to the concern that people have as a solution to the obesity epidemic. Many public health researchers and medical experts have identified the obesity epidemic as the leading concern entering the 21st century. Childhood obesity also has connection to the other topics overweight and population health concerns. Although childhood obesity it is a concern among users, a group often left from the obesity conversation is teenagers. Researchers have identified a steady increase in obesity-related cases among teenagers (M. Fox, 2016; Ogden et al., 2006). Teen obesity was identified once among the 66 obesity-related topics.

Health economics was another major topic that was not labeled as a positive obesity topic. This negative sentiment of people supports the concerns of Finkelstein et al. (2009) that obesity is a significant contributor to medical cost. Treatments aimed at addressing obesity – medical and surgical options – are not the main drivers of cost but treating the diseases that obesity promotes (Finkelstein et al., 2009). Based on the word

[10] http://www.allgeneralizationsarefalse.com/why-guns-are-not-like-spoons/

probability for health economics (Table 4.3), much of the negative sentiments centers on socio-economic disparities and the role of the government in addressing this issue. The health economic topic coincides with the socio-economic, food taxes, and social justice topics that were discovered. When examining the connection that the negative topics for obesity have with the other query topics, negative topics such as diabetes, chronic conditions, fast food, genetically modified food, Hidradenitis Suppurativa (skin condition) and physical activity demonstrate the topical connections that can be made.

Positive sentiments – obesity

Immediately, when viewing the top positive topics for obesity, childhood obesity is a topic that was also identified. However, the words that have a high probability of childhood obesity in this section are aimed at addressing the epidemic. Words such as *reducing, combat, michelle, campaign, government,* and *program* notates the efforts to address the issue. This supports the slight reduction in childhood obesity that has been identified by researchers (M. Fox, 2016). The topics of weight management and obesity prevention are two areas that Twitter users think can address the epidemic. Food was not an unexpected topic and the words that characterize this topic consist of food high in calories and sugar. People have positive sentiments or may enjoy eating these types of foods, but they also contribute to improper dieting issues associated with obesity. This presents a challenge for health communication campaigns aimed at addressing the issue.

Table 4.3: Top Five Negative Health Topics for Obesity

	Childhood obesity 5%	Lifestyle 4%	Over-weight 4%	Health Economics 3%	Population health 3%
T^1	sad epidemic america children national scary reason	Morning friday week battle gym mcdonald tv	obese bmi lower person overwght level number	Health problems public poverty cuts economic government	countries global world rise malnutrition middle east
T^2	obesity kids parents child blame fat poor	obesity times diabetes makes plans questionable day	people struggle obesity morbid issue lazy understd	Problem obesity issue nation biggest society poor	countries poverty world states highest rate obesity
T^3	child death leading issue fact hate issue	life threatening obesity exercise style sedentary lifestyle	talk abt obesity common human morbid media	nhs costs money crisis spending dollars economy	world hunger starvation people crazy poverty dying

This topic conjures questions such as; how can we frame health communication messages that address the health concern – consumption of unhealthy foods – but are mindful of an individual right to food choice? What are the effective communication strategies to deliver the core messages of our health campaign?

Interpreting the chronic conditions topics for obesity can be difficult. The idea of people not managing their hypertension or glucose levels to increase their risk of diabetes is not a positive sentiment (Ali et al., 2013). However, I think the positive sentiment

represented here for obesity is people's understanding of various chronic conditions and the role that obesity plays. This perspective is supported by other topics identified that include obesity awareness, health education, interventions, obesity prevention, weight management, and exercising.

Table 4.4: Top Five Positive Health Topics for Obesity

	Chronic Conditions 7%	**Childhood obesity 6%**	**Food 5%**	**Weight Management 4%**	**Obesity Prevention 3%**
T[1]	obesity gut diabetes cancer research preventing cells	Obesity michelle childhood combat reducing pokemongo child	love obesity free pizza lunch fries chocolate	Weight prevent reduce help obesity maintain control	treat act reduce preventing parks seniors cosponsoring
T[2]	science workout inflammatory autism asthma measures aware	strategy childhood health plan government tackle reduce	ice cream calories cookie donuts salad chips	Fitfam easy positive reduction cardio stopobesity nutrition	eating calories disorder metabolism program healthy combat
T[3]	reduces treatment arthritis likelihood rheumatoid clinic early	obesity childhood tackle campaign program government parents	pizza snack rolls pepperoni entire bag sizes	bmi weight healthy normal determined body factors	advise treated obesitysummit responses coverage existent obesity

Seventy-three percent of the topics were identified related to the positive sentiments of obesity. Only 9% of the total topics were inconclusive topics. Of the twenty-seven

unrelated obesity topics (this includes the nine inconclusive topics), 62% were annotated with a topic.

One of the unrelated obesity topics was gun violence. However, gun control was identified as a positive topic. Two consistent words that were given a high probability for gun violence and gun control were *gun* and *spoon*. While the gun debate is unrelated to the obesity epidemic, the semantical nuance of this topic may give an indication of how we try to conceptually relate epidemics. Other unrelated topics include global warming, meeting, celebrities, animals, and fashion.

Negative sentiments – diabetes

Of the 72 total topics identified for diabetes, 34 were negative. I identified 19% of the negative topics as inconclusive. Gun violence, politics, and sports were some of the topics identified that were unrelated to diabetes. Unlike the other query topics, diabetes included a significant number of Spanish language text in the topic. For example, one of the inconclusive topics contains the following words: *diabetes la para tension mi medicinas hay con por medicina los las medicamentos*. While these words do refer to medical/medication concerns relating to diabetes, the data collection was aimed at English tweets. Possible explanations for this are provided in the discussion section.

Chronic condition and symptoms were the two topics represented the most. Similar to the previous query topics (obesity), chronic conditions are a concern (Figure 4.5). In addition to diabetes, the topic shows that there are negative sentiments towards conditions like cystic fibrosis, hypertension, and arthritis (Figure 4.5). Previous

researchers have indicated the link that exists between diabetes and hypertension Hypertension in obese type 2 diabetes patients is associated with increases in insulin resistance and IL-6 cytokine levels. Potential targets for an effective preventive intervention and national health agencies such as the CDC have documented the health consequences that can be attributed to obesity that includes diabetes, arthritis, and high cholesterol (Kim & Basu, 2016, p. 603). With some chronic conditions, there are no immediate warning signs or symptoms that prompt people to seek medical advice. When someone gets "pink eye," some physical signs and symptoms indicate there is a deviation from the normal physical condition of their eye (Deiner et al., 2016). Deiner et al., (2016) also demonstrated the seasonality of pink eye issues. However, many chronic conditions such as hypertension, diabetes, kidney disease, and cancer do not have immediate symptoms and can go unnoticed until diagnosed conditions are severe. It is important for people to know their family medical history. There are numerous factors that increase people's risk to chronic conditions. This includes food choices, environment, and family medical history. [11] From the negative sentiment findings, Family Medical History represented 3% of the topics related to diabetes.

A prior study that focused on the negative sentiments of users and the DDEO health concerns also identified genetics as a word probability (Shaw & Karami, 2017). Therefore, I think the high prevalence of these topics supports the links of these conditions (diabetes, hypertension, and diabetes) and the negative concerns people have with identifying them early. Medication was another negative concern of users. Metformin is a word that was identified twice among the medication topic. Metformin is

[11] https://ghr.nlm.nih.gov/primer/inheritance/familyhistory

a medication that is often used to treat type 2 diabetes, but Metformin has also been prescribed to patients that are considered prediabetic. Researchers have raised concern regarding the Metformin drug and its link to neurodegenerative effects (Moore et al., 2013). The potential side effects of this may indicate users have negative sentiments for medication and this drug in particular.

Table 4.5: Top Five Negative Health Topics for Diabetes

	Chronic Condition 6%	**Symptoms 6%**	**Food 5%**	**Medication 4%**	**Weight Loss 3%**
T^1	diabetes fibrosis cystic celiac mellitus disease problems	warning eye signs general general symptoms exam	meat fruit antibiotic s veggies vegan livestock dairy	Insulin medical price drug medicine price metformin	losing lose weight feet pounds lost lbs
T^2	risk type diabetes higher obesity increase hypertension	lost dad year ago diagnosed type diabetic	chicken diabetes fried wing mac dinner cheese	Metformin medication pharmacy effects devil hypoglycemi a obesity	loss diabetes weight diet exercise surgery plan
T^3	cancer chronic pain depression arthritis disease thyroid	long term diabetes opinion fatal mistake memory	chocolate candy cake bar sugar nutella cookies	Anti meds drugs companies parma addicted doctor	lbs great blood pressure battle legs diabetes

Diabetes management was identified as a negative sentiment that further supports the issue of dealing with chronic conditions such as diabetes and losing weight. Other notable negative topics include mental health, diet, diabete screenings, and neuropathy.

Positive sentiments – diabetes

For the negative sentiments, diabetes management represented 3% of the topics. Diabetes management was a significant topic for positive sentiments of diabetes, representing 12% of the topics. Therefore, while people may have concerns with managing their diabetes, overall, it seems that people have positive sentiments with their ability to manage their diabetes. Positive topics of diabetes awareness and research indicate that people are concerned with creating a mindfulness of the disease. From an emotional and behavioral aspect, people are using emotions to deal with their management of diabetes. When studying various population groups and chronic conditions, the senior citizen population (often identified as => 55 or =>62) is at greatest risk of being diagnosed and managing chronic conditions. Humor may be an effective strategy to address the cognitive-behavior of people suffering from chronic conditions to increase their life satisfaction and reduce pain (Tse et al., 2010).

Food is a topic for the negative and positive feelings people have regarding diabetes. Insulin intolerance is improved by maintaining a proper diet and preventing sudden increases in a person's glucose levels (Lukic et al., 2014). As the topics from Tables 4.5 and 4.6 exhibit for food, I think people battling diabetes are indulging in *fried* foods, having a *pizza,* or consume *cookies* while trying to maintain a proper nutritional diet.

The positive sentiments of foods and drinks high in sugar – or high fructose corn syrup – indicate that people are consuming foods which negatively impact their blood glucose levels. I identified topics that include sugary drinks, sugar, and sweeteners. The American Heart Association has discussed the "demonstrated association with obesity" and sugar-sweetened beverages when it comes to monitoring cardiovascular health (Lloyd-Jones et al., 2010, p. 595). In the previous negative topics for diabetes, I identified beverages as a topic that contains word clusters of *sweet*, *soda*, and *coke*. Researchers have documented the association that exists with artificial sugar, obesity, and diabetes but it appears to remain a difficult issue for people to manage.

Sixty-seven percent of the topics were related to diabetes with 9% focusing on the awareness and research of diabetes. Nineteen percent of the topics were inconclusive. Consistent with the findings from the negative inconclusive topics for diabetes, there were numerous Spanish words – *la obesidad diabetes amor mi familia* - that are seen among the topic. Exercise, that is also one of the query topics used, was a positive topic for diabetes. Other notable topics include diabetic kid camps, mountain climbing, population health, and anemia.

Negative sentiments – diet

As seen with three of the six sentiments – representing three of the query topics – lifestyle is a top five topic. It represented 18% of the 100 topics, followed by 8% of the topics representing unhealthy behavior.

Table 4.6: Top Five Positive Health Topics for Diabetes

	Diabetes Management 12%	Diabetes Awareness 5%	Research 4%	Emotions 4%	Food 3%
T^1	control insulin diet sugar weight healthy helps	pm free tomorrow session program health team	article manage understand writer film patient kickstarter	humor chronicillnes s pokemongo sense hahaha colitis fun	fried fair state pizza oreos cinnamon food
T^2	diabetes glucose monitor meter downs call dexcom	type raise money awareness charity diabeteswee k funds	american diabetes foundation association research endocrine scientific	love happy cool diabetes real joke life	ice cream chocolate sugar coffee milk caramel
T^3	blood test gestational doctor time glucose normal	coming hospital tonight community neuropahty discuss diabetic	researcher s medical display medicres models join meeting	fun jokes joke lmao people man big	white rice brown bread intestines baked egg

Based on the negative sentiments of diet surrounding lifestyle, people are having a difficult time with developing a regimen or cutting foods low in nutritional value. Topic T^3 shows that those who may be on a diet are secretly consuming foods that are high in fat with no nutritional value (Lloyd-Jones et al., 2010). It appears that those struggling to develop a positive dieting lifestyle are not able to engage in mindful eating. Mindful

eating is an approach that normally helps with developing a healthy and nutritional diet. “When exposed to negative emotions, some people tend to use avoidant or impulsive styles of coping and often overeat in response to stress, consuming excess calories in an automatic and dissociative fashion” (Dalen et al., 2010, p. 263). The Dalen et al. (2010) statement is supported by the emotion, fastfood, and mental health topics identified. There were also negative sentiments regarding the type of diets.

Looking at Table 4.7, the type of diets represented included the Atkins diet, Ketogenic diet, and vegan/vegetarian diet. The type of diets topic is consistent with negative topic findings from Shaw and Karami's (2017) characterization of negative health sentiments. Focusing on the ketogenic diet, this newly hyped diet plan has been used by doctors for several years. Someone on this diet plan consumes more food high in fat and the cells that normally use blood sugar that comes from carbohydrates to fuel the body (provide energy), now use the fat in molecules that are called ketones. While this diet has been documented to address weight loss issues, there is no information on the long-term effects of this diet. Also, there is a concern about the amount of red meat and other foods that are high in cholesterol (Campos, 2017). The religious diets topic I identified could have been included as types of diets, however, I think it is important to understand how peoples’ faith also impacts their dieting behaviors independently of mainstream diet plans.

Twenty-seven percent of the topics were identified as unrelated to dieting, with 6% being inconclusive. Seventy-three percent were related to the diet query topic. Notable additional topics include soda, women’s pregnancy, eating disorder, body image, and beach body.

Table 4.7: Top Five Negative Health Topics for Diet

	Lifestyle 18%	Unhealthy Behaviors 8%	Mindful Eating 5%	Fastfood 4%	Type of Diets 3%
T[1]	bad good diet decisions pretzel justify breaking	days weeks whiskey couple fourteen totie hotdogs	night morning chocolate bed late pm ate	bad diet craving habit whataburger atm combo	vegan based vegetarian animals veganism harm cruelty
T[2]	plan workout strict follow lazy frustrated regimen	diet alcohol unhealthy smoking quit drunk weed	crap eating piece loads bunch healthy piss	coke mcdonald regular mcdonalds large drive ordered	low fat high carb ketogenic saturated intake
T[3]	diet needed badly secret krispy kreme window	stress poor sleep exercise choices schedule hormones	food eat junk unhealthy soft poisoning chinese	diet big cheese burger mac fries mcdonalds	diet evil borke atkins fighing hungry cheats

A negative topic of diet that was also a negative topic for exercise and diabetes was weight loss. This is consistent with the literature and the importance that proper diet and exercise have with weight loss initiatives to reduce one's risk of type 2 diabetes (Fahimi et al., 2008; Lloyd-Jones et al., 2010).

Positive sentiments – diet.

The total percentage of topics relating to diet was 69% - of the 31% topics that were not related to diet, 14% were identified as inconclusive. Once more, lifestyle was identified as a topic and represented 8% of all topics relating to the query topic of diet (Table 4.8). The importance of developing a lifestyle of proper dieting and exercising (see Table 4.1) is supported by numerous clinical and non-clinical health related research (Flegal et al., 2012; Hall, Johnson-Turbes, & Williams, 2010; Lloyd-Jones et al., 2010; Ogden et al., 2006; Wing et al., 2001). Lifestyle was also a significant portion of the negative topics for diet, during the analysis, lexical semantics and uncovering the story of the topic is not often considered in data science research.

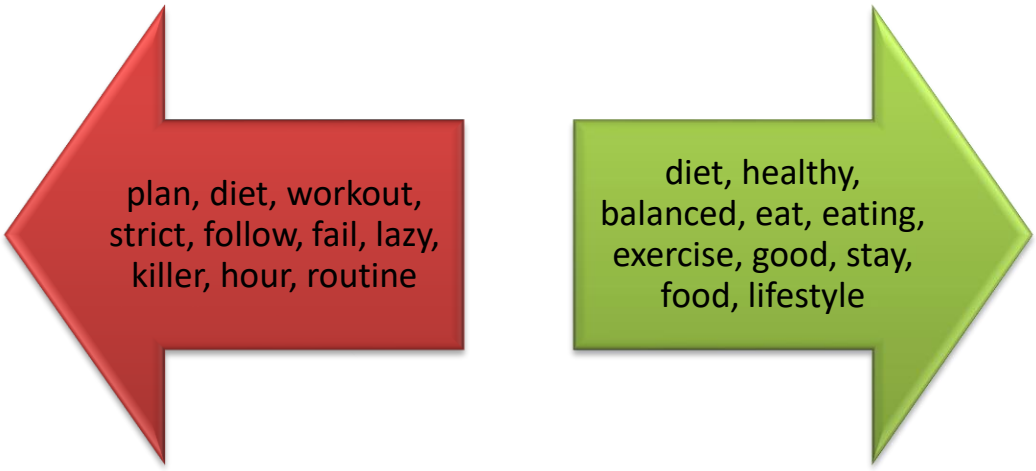


Figure 4.3: Negative and Positive Lifestyle Topic for Diet

In Figure 4.3, the negative (left pointing arrow) and positive (right pointing arrow) words for the lifestyle topic are represented. This figure contains ten of the twenty words that represent the topic; they are listed based on their probability of being associated with the topic. The words represented have closely related probability of being related to the lifestyle topic. Based on the pre-determined dictionary of the LIWC tool,

collectively, the words in Figure 4.3 represent a more positive or negative sentiment. The words *lazy, fail,* and *strict* (left pointing arrow) are normally associated with a negative connotation. While those words can be used for inspiration or an internal desire to "be better," that may not be the initial emotional intentions when those words are written or spoken by someone. Moreover, the LIWC tool – from a psychological dimension – identifies these words as negative. On the contrary, I think the words *healthy, balanced,* and *good* conjures positive feelings towards developing a positive dieting lifestyle. Again, this is computationally determined by the LIWC tool classifying these words as positive. The critical analysis of computational output to this point speaks to the understanding that "holistic data science requires that we understand the context of data…clearly communicate what a dataset can and cannot tell us about the world" (Blei & Smyth, 2017, p. 8691).

Unlike the other negative diet topics associated with lifestyle, the additional positive topics appear to focus on food and beverages. In addition to the negative sentiments surrounding the diets in the previous section, paleo and gluten-free diets are identified from the topic. People feel that meal planning is essential with developing healthy meals that impact "lifestyle." Beverages included a variety of different drinks, ranging from alcohol to water. This may indicate that people enjoy a variety of beverage types and present a unique challenge for health professionals with communicating information about the consumption of empty calories. Soda was also identified as a topic but was not included with beverages.

Table 4.8: Top Five Positive Health Topics for Diet

	Lifestyle 8%	Food 6%	Type of Diets 6%	Healthy Meal 4%	Beverages 4%
T^1	healthy eating habits lifestyle clean exercise changing	hot cheetos dog smoothi e spinach grapes fruit	plant diet based vegetarian food planet benefits	meat healthy fruits veggies rich fresh vegetables	water coke sweet coffee juice beer vodka
T^2	balanced healthy exercise lifestyle life maintain balance	chicken fries cheese pizza salad bacon chips	high fat protein carbs keto calories energy	rice consists chicken potatoes soup peanut cheese	Tea sweet soda sugar drinks Snapple coffee
T^3	august dietingforlife fit lifeextension worth life consciousness	cream ice chocolat e eating cookies summer cake	years ago pale change diet started glad	health protein vitamins fiber iron nutrients foods	diet red wine glass chocolate almond coffee

The words representing the topic included specific brands *mountain dew, diet coke*, and *pepsi*. Other notable topics that were discovered include exercise, children diets, dental hygiene, religious diets, and social networks.

Representation of additional Chronic Diseases

When deciding on a topic model to use, an essential component of the evaluation process is determining if the model has captured the internal structure of the corpus. This portion of the research study seeks to identify additional health experiences prevalent based on Twitter users' sentiments and topics regarding DDEO (RQ2). While the other topics may have chronic conditions present – like hypoglycemia found in the negative medication topics for diabetes – the primary focus on the health issues in this study are chronic related. Also, an in-depth insight is more feasible with a small sample to analyze.

Within the chronic conditions, cancer is one of the known chronic diseases that is associated with obesity. Other expected conditions that were identified include arthritis, diabetes, and cardiovascular conditions. The words pregnancy, infertility, and cramps signals women's health related conditions that are likely associated with obesity and diabetes. An intriguing word identified was constipation. Coupled with other conditions and factors associated with diabetes, an individual can experience constipation. The issue of chronic constipation may also indicate that the individual has irritable bowel syndrome (IBS) or hypothyroidism (The National Academy of Hypothyroidism, 2017). Even though these conditions are not considered the observed chronic condition from the health topics under surveillance, they can provide other possible conditions to consider.

The word kidney was found which may provide indication towards the chronic condition of kidney disease. Words such as hypothyroidism, hypercalcemia, and Alzheimer were also associated with the chronic condition topic omitted by the tool. Hypercalcemia is often a condition that is associated with cancer but inclusion of the term within the chronic condition topic may suggest otherwise. Autism was another word that

was part of the topic for chronic conditions. Research has suggested that the prevalence of obesity in this population is just as prevalent as children overall – which warrants the need for research on the additional factors that impact this population regarding this health issue (Curtin, Anderson, Must, & Bandini, 2010). The inclusion of these words within the chronic condition topics provides insight into additional chronic health experiences of users; it also supports related studies and the complexity of the obesity medical condition.

Evaluation - sentiment analysis.

To evaluate the performance of the sentiment analysis tool and answer RQ3, the results of each rater (Mturk workers) were compared to the computational tool in determining their agreement with the computationally identified positive and negative sentiments. When the rater selected neutral for their response, that case was removed. The crosstabulation must be the same (i.e. 2x2, 3x3 and so forth) to conduct Cohen's kappa on the dataset. The average number of valid cases among the raters was 80.5, and the average number for missing cases was 19.6. When examining the different cases where raters selected neutral tweets, Rater 7 (R7) had the least number of cases where neutral was selected – R10 had the highest number of neutral tweets selected. Given the limited sample, a correlational analysis between the number of missing cases and valid cases would not be appropriate (Laerd Statistics, 2015).

For the raters to identify the tweets as positive or negative, raters were provided instructions on the purpose of the research. Each rater was presented the same number of tweets, in the same order (see Appendix A). A mixer of positive and negative tweets

represented DDEO. DDEO was presented to the raters in separate sections within the Google Forms. Each rater determined whether the tweet was positive, negative, or neutral. As noted in Chapter 3, the additional option for raters to select neutral was provided after feedback from the preliminary testing of the sentiments. If a rater determined that the tweet was neutral, it was classified as a missing case. Based on my observation, missing cases were due to three factors:

1. The rater was not spending enough time to determine if the tweet was positive or negative (It took raters roughly six minutes to complete the human intelligence task (HIT)),

2. As the semantic structure for tweets increase – containing components of negative or positive sentiments – it was difficult for some raters to identify a dominant sentiment.

3. The semantic structure that the raters are using may be different from the computational tool and does not lend itself to a binary classification approach. While there may be other factors that contributed to the missing cases, these were observational factors that I noticed while aggregating the data and conducting the analysis. After removing any participant that did not fully complete the survey, the final number of participants for the sentiment analysis evaluation task was ten (n=10).

The initial agreement between raters and the sentiment analysis tools was conducted using percent agreement. This direct observation allows you to determine the amount of data that are erroneous – only one rater is considered correct when there is disagreement (McHugh, 2012a) The average agreement for raters across each case was

83.6%. Therefore, on average, 16.4% of the tweets misrepresent the positive and negative sentiments conveyed. R10 had the highest percentage agreement, but they also had the highest number of missing cases (Figure 4.4). When looking at each case in the complete list for percent agreement (Figure 4.6), the assumption can be made that the sentiment recognition that is determined by the participants (humans) is closely accurate to the computation tool (i.e. the raters and sentiment analysis tool agreed on most of the negative and positive sentiments represented by the tweets). The percent agreement provides an easy method to calculate and gain insight regarding interrater reliability, but it does not account for the chance that the rater and computational tool agree. With percentage agreement, typically, there is no gold standard criterion for the correctness of the decisions made (McHugh, 2012b, p. 278).

To determine if there was agreement by chance between the sentiment analysis tool and the raters' judgment on positive and negative sentiments, Cohens' kappa interrater reliability test was conducted. When examining the agreement between individual raters and the sentiment analysis tool, R10 and the sentiment analysis tool agreed on 23 of the negative tweets correctly. They agreed on 33 of the positive sentiments. However, the sentiment analysis tool identified four tweets as negative, but R10 identified them as positive; and the sentiment analysis tool identified one as positive where the R10 suggested that it was negative (Table 10).

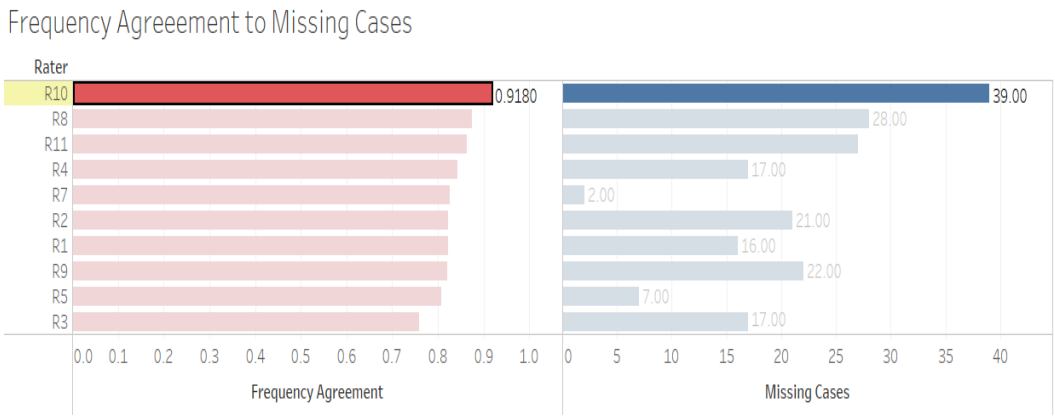


Figure 4.4: Number of Missing Cases to Frequency Agreement for R10 (Based on Percentage Agreement)

R10 had very good agreement with the sentiment analysis tool, k = .832, 95% CI [.697 to .973], p < .001). Although percent agreement and Cohen's kappa are two different interrater reliability testing methods, there is consistency with a high level of agreement between R10 and the research tool. R3 had the lowest rater agreement among the 10 cases, k = .508, 95% CI [.330 to .686], p < .001.

Table 4.9: Complete List of Percent Agreement for Each Rater and Sentiment Analysis Tool

Rater	Valid Cases	Missing Cases	Frequency Agreement
R1	84	16	82%
R2	79	21	82%
R3	83	17	76%
R4	83	17	84%
R5	93	7	81%
R7	98	2	83%
R8	72	28	88%
R9	78	22	82%
R10	61	39	92%
R11	73	27	86%

Table 4.10: Crosstabulation of R10 to the Sentiment Analysis Tool

R10 * ST

			ST		
			0	1	Total
R10	0	Count	23	1	24
		Expected Count	10.6	13.4	24.0
	1	Count	4	33	37
		Expected Count	16.4	20.6	37.0
Total		Count	27	34	61
		Expected Count	27.0	34.0	61.0

Based on the strength of agreement guidelines from Altman (1990), there was a moderate level of agreement in the case of R3 and the sentiment analysis tool. They agreed on 40 of the negative tweets and 23 of the positive tweets. The sentiment analysis tool identified 4 tweets negative where R3 identified them as positive and the sentiment analysis tool identified 16 tweets as positive where R3 identified them as negative (Table 11). "However, the value of kappa (κ) is heavily dependent on the marginal distributions, which are used to calculate the level (proportion) of chance agreement. As such, the value of kappa (κ) will differ depending on the marginal distributions" (Laerd Statistics, 2015). Therefore, caution is warranted with comparing one of Cohen's kappa to another – unless the marginal distributions are the same. Although each interrater reliability test in this study is based on individual cases, by looking at each of the remaining cases, each one has a k value that would classify it as good (Figure 4.10). There is disagreement on what should be considered acceptable and unacceptable for a k value (Di Eugenio & Glass, 2004; McHugh, 2012b). Additional discussion regarding the kappa value findings and the inference that can be suggested from these findings will be revisited in Chapter 5.

Table 4.11: Crosstabulation of R3 to the Sentiment Analysis Tool

R3 * ST

			ST		
			0	1	Total
R3	0	Count	40	16	56
		Expected Count	29.7	26.3	56.0
	1	Count	4	23	27
		Expected Count	14.3	12.7	27.0
Total		Count	44	39	83
		Expected Count	44.0	39.0	83.0

Evaluation - topic model.

A total of 31 Mturk workers (raters) completed the HIT of identifying the intrusion word represented in the topic.

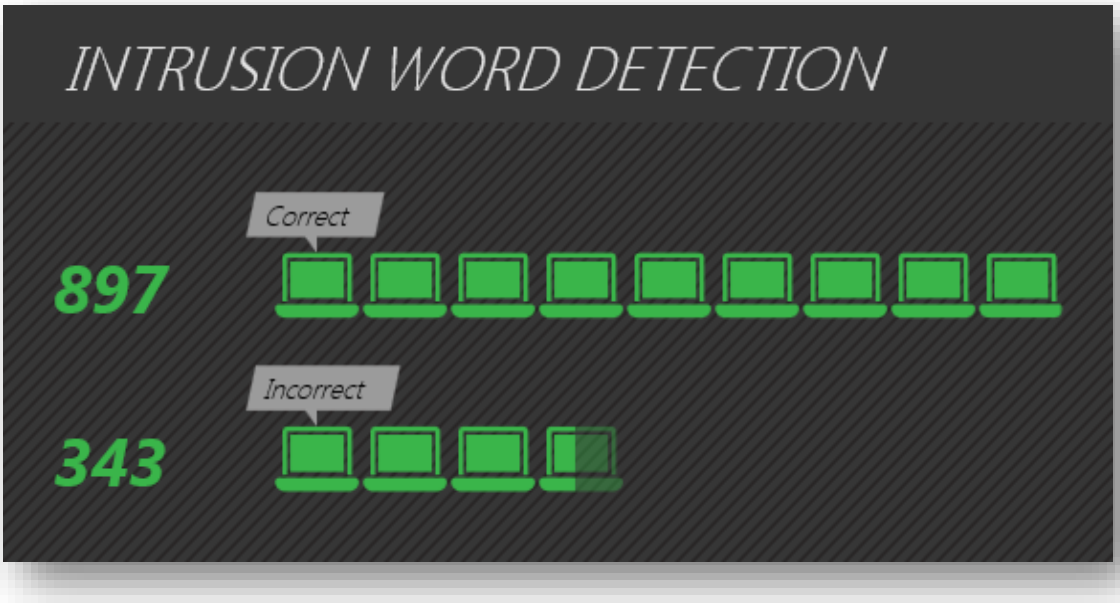


Figure 4.5: Number of Total Intrusion Tasks Correctly Completed

From the total 1,240 word intrusion tasks (31 x 40), 73% of the intruding words were correctly identified among the 31 raters (Figure 4.7). When examining the average number of tasks selected correctly among the raters, the average number of intruding words detected was 29%. Only one-quarter of the raters identified less than 75% of the intrusion word when completing their HIT.

Raters 25, 4, and 21 are the only raters that identified less than 10 tasks correctly. This is likely attributed to the participants randomly clicking through items to complete the task. The median number of tasks completed correctly was 31. Majority of the raters were able to identify the correct intrusion words for diet while the raters struggled with identifying the intrusion word for diabetes. This may be due to the difficulty of understanding the topic from a series of complex words that represented diabetes. The decrease in word detection may also be the result of how the raters add to the complexity of words. In addition to the logical understanding of the text, users may interject past experiences and former knowledge structures when analyzing the text.

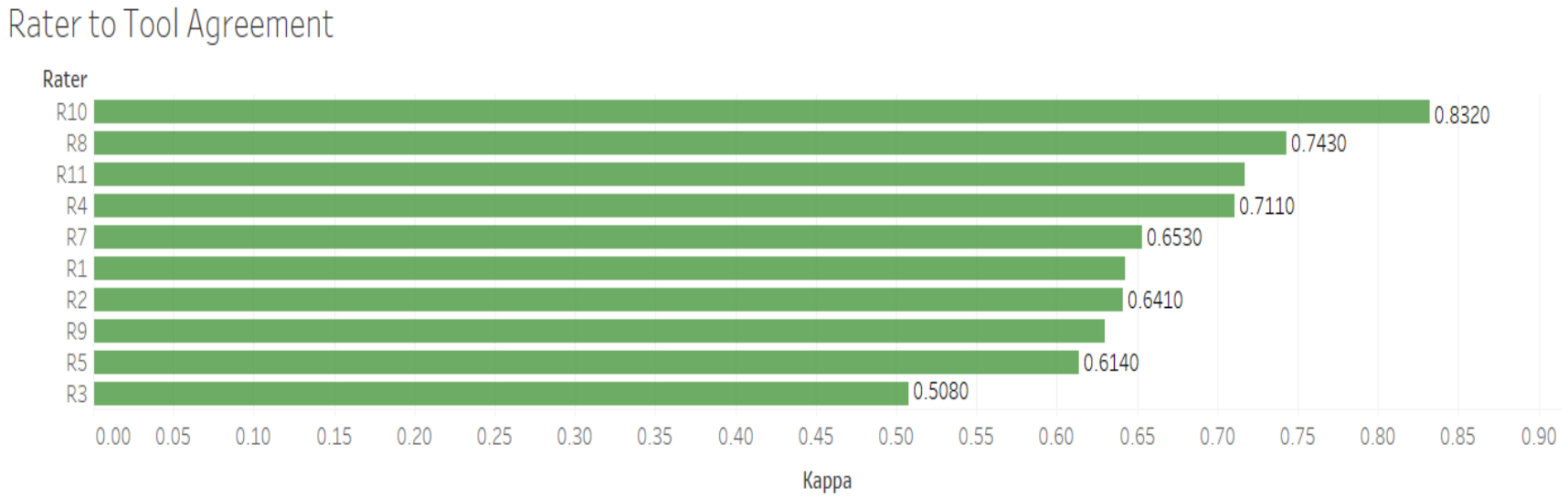


Figure 4.6: Bar Chart Representing the Cohen's kappa Value for each rater in Descending Order

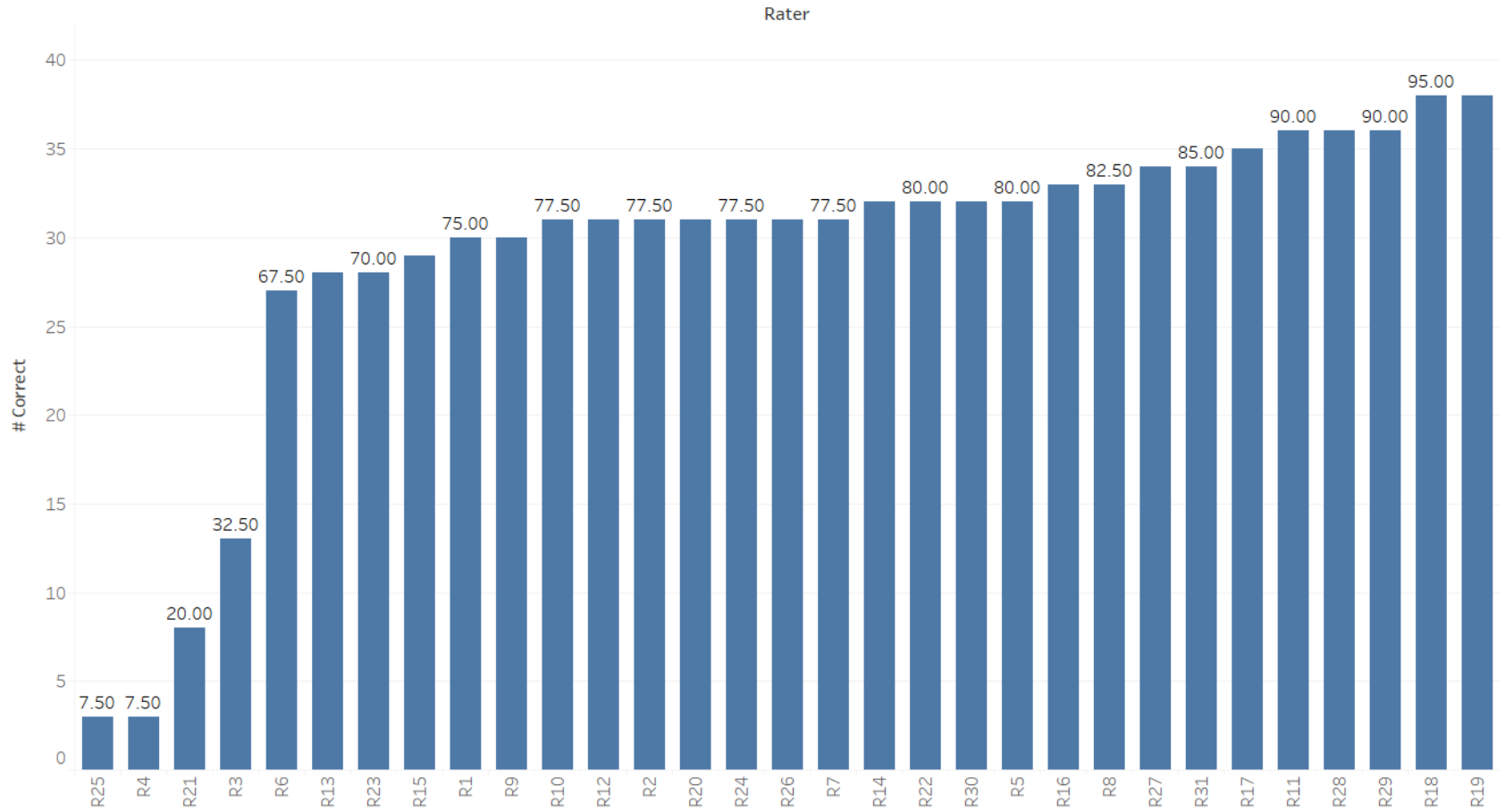


Figure 4.7: Word Intrusion Detection Percentage by each Rater

Chapter 5

Conclusion

Discussion

This study was conducted to demonstrate the development and use of my analytical framework; and how the information gleaned from the framework can assist public health professionals. This includes understanding four related health issues from Twitter users' experiences, the characterization of obscure and non-obscure topics, and performing qualitative methodological strategies to characterize a computational output. This chapter will provide a discussion regarding important components of the text mining framework, limitations of the framework and Twitter data, and future directions for this research.

Discussion - topic analysis.

The framework discussed in Chapter 3 allowed for the characterization of users' sentiments by condensing millions of individual tweets in statistically distributed topics. Although each component of the framework is important, it is within the topic analysis component that we discover how the data assist with understanding the four health related topics and the knowledge discovered. More importantly, has it assisted the researcher with understanding the semantic structure of the data corpus? An important

aspect of the analytical process that the topic analysis cultivated was creating a story, based on the topic.

The celebrity topic identified in the positive topics for diet provides an example of why human cognition and culture is important in the analytical process. The cluster of words – *diet, men, gucci, women, mane, prison, joe, sentence* – is an example of the eight words in the celebrity topic. If you are unaware of the celebrity mentioned within the topic, this topic would not mean anything. However, to those who follow hip-hop, Gucci Mane is a well-known rapper that was overweight and boasted about his improper dieting and exercise behaviors. After he was released from prison, fans and entertainment outlets noticed that he lost 50 pounds. It was a result of him changing to a low-carb diet and adopting a stringent workout regimen (Bristout, 2016). A salient point in his interview with media outlet Revolt was the importance of removing drug substances from his body and the psychological impact this had with his decision-making (Bristout, 2016). To a social worker or public health professional working with teens that identify with a celebrity figure like Gucci Mane, this information may have value to them. Health communication research has shown that through narratives using persuasive communication strategies, audiences can be impacted when they look up to or desire to be more like a character (or in this case, a celebrity) (Hoeken, Kolthoff, & Sanders, 2016; Moyer-Gusé, 2008). While many of these strategies are aimed at health education and awareness campaigns, using these strategies with actual celebrities may be a different approach for certain persuasive health communication. The computational methods used

in this study identified this positive topic, but more importantly, the positive story of why dieting is essential and the impacts this information has for addressing social-determinant factors for a population group adds additional insight into the topic.

Humans understand culture and shared experiences, which is a trait that is difficult to capture within an algorithm. However, computational models use scientific approaches and established methodologies to help us derive knowledge from large datasets that humans are not able to rapidly and efficiently process. This study shows the value of human cognition and computational social science methods in the knowledge discovery process. Whether you are pro-data or anti-data, the use of computational tools helps us with sifting through these massive datasets – structured or unstructured – to identify what exists and address data privacy issues that can lay dormant if not probed (Zhou, Zhang, Yang, & Wang, 2018).

Discussion - sentiment analysis.

A possible concern that individuals may have is the appearance of a topic for positive and negative sentiments. These conerns were consistently seen in Table 4.7 and Table 4.8 for the topic lifestyle. Lifestyle was the most represented topic for both sentiments. I think this addresses the difficulty and notion that sentiments are not finite. The sentiments expressed also speak to the temporal and spatial related concerns. Someone can express a negative sentiment for exercising in one month, and the next month, have positive sentiment to engaging in exercise. The idea of what I call dual-sentiments is challenging to account for in this type of research. This research focuses on

the big picture that exists from the dataset to make interpretations and use the information to form questions that are not being discussed.

The sentiment evaluation portion of the research was revealing but also raises additional questions. When examining each individual case to decide on how well the sentiment analysis tool is performing, a premature conclusion could be reached. Based on how Cohen suggested that the kappa results be interpreted, most cases would be grouped into moderate level of agreement categorically (Laerd Statistics, 2015; McHugh, 2012b). When you look at the percent agreement for R10 and R3, they are either at or above 75% percent agreement. The k-value places the level of agreement much lower. This is consistent with the k-values for the other raters (Figure 4.11). There is a difference in the equivalence levels. As the expected chance agreement between increases, the lower the kappa. The results indicate that majority of the cases are moderate with the tool and rater in identifying the polarity of the sentiments; however, this means that roughly 30 to 40 of the data could be considered erroneous. For health policies and health communication campaigns, this level of inaccurate data can have significant consequences if policies or allocation of resources are based on those decisions.

There is contention regarding interpreting Cohen's kappa coefficient and stating that the value or agreement is "good." The sample impacts the expected standard error. Therefore, a higher sample size is likely to produce a small Confidence Interval that would likely estimate the agreement to be precise (Di Eugenio & Glass, 2004; Ludbrook, 2002; McHugh, 2012a). Ludbrook (2002) expressed his concerns against the kappa statistics and its ability to detect bias among the two raters. A counter argument to this notion is that the computational tool is not biased and remains objective. This resurfaces

the conversation of who determines the training data that teaches the algorithm. For this study, I think the inclusion of a direct observation percentage agreement method and the kappa statistic is acceptable with understanding the agreement between the raters and the sentiment analysis tool. The raters were not trained as part of the research process and there may have been a significant level of just guessing.

Using the unsupervised topic model and sentiment analysis, additional chronic conditions and short-term conditions are identified. The framework also assisted with understanding the complexity of obesity and how lifestyle behaviors impact obesity and diabetes. Mental Health was a topic that appears across the spectrum of topics. It is uncertain what this may imply for the development of interventions and health campaigns. Mental health only emerged as a significant topic once for exercise (positive sentiment) but at least once for obesity (positive and negative), diabetes (negative), and exercise (negative)

Discussion – positionality.

A "quantitative" research design was the guiding focus on this research, but it is important to address my positionality within the topic analysis component of the framework. While the topics were computationally derived, the deterministic factor of characterizing the topics was conducted by me. The topics selected were based on scholarly related literature. However, my cultural, social, and research positioning could have impacted the topics assigned. Where I have assigned chronic conditions to the positive topic for diabetes in figure 4.1b, another scholar may assign the topic medical conditions. It is difficult to truly measure how and why specific topics are assigned, but

it is a subjective measure that we should try to evaluate (Lyra, 2017). As a result, this can also be considered a limitation of this study.

In studies that use a similar approach, annotators characterizing the topics have a predefined list of topics to characterize the topics (Mowery, Bryan, & Conway, 2015), or they have users label a sample of words that represent the topic use in charactering the topics (Paul & Dredze, 2014). Using a predefined set of labels would restrict the inductive process that took place during this study. Misspelled words can be difficult for the computation tool to recognize. An example of this is found in the hypoglycemic topic identified that was omitted by using the LIWC tool. Searching the misspelled term (hypoglecic) in the consumer health services of MedlinePlus.gov, hypoglycemic was the suggested term. I think instances like the problem mentioned describe why the topic analysis step is a crucial component of this text mining framework.

Before addressing the implications of this study, I think it is necessary to discuss the limitations. The use of interrater agreement on the characterizing topics could improve topic recognition. While the unsupervised topic model was trained on the corpora of data, the use of a supervised classification model could be trained on a small corpus of data to determine the correct topics to use.

Since social media data can be viewed as a collection of self-reported data, there is valid concern when interpreting the results from user generated content. Active surveillance that requires interaction with participants – that include well-established objective measures – is likely to yield different results. An example of this was seen with the development of the GFT. Mowery (2016) reported that roughly 40% of tweets describing the flu are misrepresented. When policies and clinical decisions can be based

on this information, this large misrepresentation of the flu disease can be consequential to various stakeholders. The twitter data used in this study is based on self-perception (Jurges, 2007) and that the conclusions derived from this data would be different than traditional medical collected data.

RQ2 focuses on the agreement between tool and each individual rater; performing Cohen's kappa between raters would provide the ability to see the agreement between raters. However, caution is warranted with trying to make a comparison between the two interrater reliability groups (rater to toll vs. rater to rater). In addition to Cohen's kappa statistical limitations, it is likely that healthcare professionals would have evaluated the sentiments differently than the general population (or raters). Even though healthcare professionals are not experts in sentiments analysis, they have professional understanding with respect to DDEO, risk factors, and health behaviors.

Using the Fleiss kappa statistical test would improve the interpretation that could be derived from RQ3. DDEO were separate sections during the experiment for RQ3. With Fleiss kappa, the raters could be broken into multiple groups and each group would rate a different query topic (DDEO). This is a unique aspect of Fleiss kappa, as the level of agreement would be determined among non-unique raters (Fleiss, Levin, & Paik, 2003). This research study does not address spatial and time related issues. The geographic location of the tweets may considerably impact the interpretation of the topics. The inclusion of time series analysis based on geo-tagged may be a better way to explore chronic conditions on a granular level. As noted in Chapter 2, GFT's is a perfect case study of when the surveillance system does not perform or predict as it should (Lazer et al., 2014a). However, LDA and many probabilistic techniques are used to

explore the phenomena under scrutiny; not to provide evidence-based findings. I think probabilistic models are good methods to help researchers explore questions they have not considered and identify latent variables. Another limitation is the concern of underrepresented populations. Topics are likely impacted by demographics represented in this social media platform (Ghosh & Guha, 2013). As an example, demographic concerns may be address through future studies using this framework and focusing on geographic locations that are predominately represented by minorities.

Conclusion

From a practical perspective, the framework in this research has the potential to assist public health professionals with exploring understudied chronic conditions. For example, various forms of hypercalcemia can result in chronic kidney disease or too much vitamin D in the system. For healthcare professionals and researchers interested in studying the condition, the use of this framework could assist them with exploring what is discussed regarding the condition. "Disease Surveillance will benefit from social data more in locations and populations without robust surveillance, and for diseases without so many resources" (Paul & Dredze, 2017, p. 95). As a next step in this research, this framework can be coupled with traditional data sources like the BRFSS or the National Health and Nutrition Examination Survey. Social media surveillance systems may not replace traditional surveillance systems, but they can provide another data source to understand health behaviors and risk factors. Initial testing of this framework with conventional sources may include analyzing BRFSS data from a longitudinal perspective to understanding what the framework can and cannot answer. Incorporating local data

sources will provide an additional layer to this process as well. More work is warranted with understanding what public health professionals need or even want from this form of data source.

An intriguing result during the sentiment analysis and word intrusion task, which was not part of the original data analysis, was the frequency with which raters selected neutral tweets. In many cases, as the complexity of the lexical semantics (or understanding the relationship of words) increased in difficulty, participants were unsure of the sentiments the tweet represented. Also, participants seem more likely to identify the sentiment of the tweets related to diet and exercise than obesity and diabetes. Another intriguing point to examine is how knowledge is constructed – within the context of health – for people in comparison to another knowledge domain. Within the framework of topic detection, what are humans searching for? How does the experience of humans impact their interaction with text in different domains? Is it possible to computationally replicate the investigation processes? A comparison of the general public and health experts would also provide insight into how these populations interpret sentiments concerning DDEO. I think these are intriguing questions to explore, particularly with the design of characterizing (or labeling) topics for use in supervised machine learning methods.

The methodological applications of this study support some of the research using mixed-methods or quantitative lead design to conduct surveillance research (Nakhasi et al., 2012). The topics identified through this framework can be used to train supervised machine learning models. These results can help with building better classifiers and be used to compare those already in place that focus on the detection of chronic conditions.

To my knowledge, this research is one of the first attempts to understanding interrater agreement between the LIWC sentiment analysis tool using Cohen's kappa. The robustness of this study could benefit from a one-sample t-test to compare the mean score for all coders based on a hypothesized rating of correctness.

Lastly, through this surveillance text mining framework, the subjectively that is part of the topic modeling process deserves more conversation within the data science community. Addresing this subjectivey can include the development of a measure to account for this in the analytic process. As the United States once again becomes home to the world's fastest supercomputer[12], "it is the human perspective that reveals how aspect of the data analysis process, such as metadata, data provenance, data analysis work-flow, scientific reproducibility, are critical to modern scientific research" (Blei & Smyth, 2017, p. 3). This research makes a small attempt to understand how computational tools and human cognition work in tandem to better understand the world around us.

[12] https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

References

Abbar, S., Mejova, Y., & Weber, I. (2015). You Tweet what you eat: Studying food consumption through Twitter. *CHI '15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3197–3206. https://doi.org/10.1145/2702123.2702153

Abbasi, A., Adjeroh, D., Dredze, M., Paul, M. J., Zahedi, F. M., Zhao, H., … Ross, A. (2014). Social media analytics for smart health. *IEEE Intelligent Systems*, *29*(2), 60–80. https://doi.org/10.1109/MIS.2014.29

Ahmed, W. (2016). Ethical challenges of using Twitter as a data source. *Multimedia Information and Technology*, *42*(1), 31–32.

Alaska Department of Health and Social Services. (2018). BRFSS overview. Retrieved from http://dhss.alaska.gov/dph/Chronic/Pages/brfss/default.aspx

Ali, M. K., Bullard, K. M., Saaddine, J. B., Cowie, C. C., Imperatore, G., & Gregg, E. W. (2013). Achievement of goals in U.S. diabetes care, 1999–2010. *New England Journal of Medicine*, *368*(17), 1613–1624. https://doi.org/10.1056/NEJMsa1213829

Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., & Gawron, J.-M. (2016). Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. *PLOS ONE*, *11*(7), e0157734. https://doi.org/10.1371/journal.pone.0157734

Anderson, C. A., & Anderson, D. C. (1984). Ambient temperature and violent crime:

Tests of the linear and curvilinear hypotheses. *Journal of Personality and Social Psychology*, *46*(1), 91–97. https://doi.org/10.1037/0022-3514.46.1.91

Aramaki, E. (2011). Twitter catches the flu : Detecting influenza epidemics using Twitter *Computational Linguistics*, *2011*, 1568–1576. Retrieved from http://www.aclweb.org/anthology/D11-1145

Arnold, C. W., El-Saden, S., Bui, A. A. T., & Taira, R. (2010). Clinical cased-based retrieval using latent topic analysis. In *AMIA Fall Symp* (pp. 26–30). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041464/

Ayers, J. W., Althouse, B. M., & Dredze, M. (2014). Could behavioral medicine lead the web data revolution? *JAMA*, *311*(14), 1399. https://doi.org/10.1001/jama.2014.1505

Bacon, R. M., Kugeler, K. J., & Mead, P. S. (2008). *Surveillance for Lyme disease --- United States, 1992--2006*. *MMWR Surveillance Summaries* (Vol. 57). Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/ss5710a1.htm

Bader, J. L., & Theofanos, M. F. (2003). Searching for cancer information on the internet: Analyzing naturallanguage search queries. *Journal of Medical Internet Research*, *5*(4), 80–108. https://doi.org/10.2196/jmir.5.4.e31

Bedrosian, S., Young, C. E., Smith, L. A., Cox, J. D., Manning, C., Pechta, L., … Daniel, K. L. (2016). Lessons of risk communication and health promotion — West Africa and United States. *MMWR Suppl.*, *3*, 68–74. https://doi.org/http://dx.doi.org/10.15585/mmwr.su6503a10

Berg, N. (2013). How citizens have become sensors | GreenBiz. Retrieved from https://www.greenbiz.com/news/2013/03/20/how-citizens-have-become-sensors

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2010). Using Mechanical Turk as a subject

recruitment tool for experimental eesearch. Retrieved from http://qipsr.as.uky.edu/sites/default/files/Berinsky.Using Mechanical Turk as a Subject Recruitment Tool for Experimental Research.pdf

Biyani, P., Caragea, C., Mitra, P., Zhou, C., Yen, J., Greer, G. E., & Portier, K. (2013). Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13* (pp. 413–417). New York, New York, USA: ACM Press. https://doi.org/10.1145/2492517.2492606

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf

Blei, D. M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, *114*(33), 8689–8692. https://doi.org/10.1073/pnas.1702076114

Blumberg, S. J., & Luke, J. V. (2007). Coverage bias in traditional telephone surveys of low-income and young adults. *Public Opinion Quarterly*, *71*(5), 734–749. https://doi.org/10.1093/poq/nfm047

Blumberg, S. J., & Luke, J. V. (2014). *Wireless substitution: Early release of estimates from the National Health Interview Survey, July–December 2013*. Retrieved from http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201407.pdf

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*, 49–64. Retrieved from https://link.springer.com/content/pdf/10.1007/BF00117832.pdf

Bristout, R. (2016). Gucci Mane reveals the turning point that led to weight loss. Retrieved from https://revolt.tv/stories/2016/07/21/gucci-mane-reveals-turning-point-led-weight-loss-7c59b74310

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE*, *8*(12), e83672. https://doi.org/10.1371/journal.pone.0083672

Brownstein, J. S., & Freifeld, C. C. (2007). HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *12*(11), E071129.5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18053570

Buckland, M. (1997). What is a document? *Journal of the American Society for Information Science*, *48*(9), 804–809. Retrieved from http://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/what-is-a-document.pdf

Byrne, A., & Byrne, D. G. (1993). The effect of exercise on depression, anxiety and other mood states: A review. *Journal of Psychosomatic Research*, *37*(6), 565–574. https://doi.org/10.1016/0022-3999(93)90050-P

Campos, M. (2017). Ketogenic diet: Is the ultimate low-carb diet good for you? Retrieved

from https://www.health.harvard.edu/blog/ketogenic-diet-is-the-ultimate-low-carb-diet-good-for-you-2017072712089

Carter, D., & Sholler, D. (2016). Data science on the ground: Hype, criticism, and everyday work. *Journal of the Association for Information Science and Technology*, *67*(10), 2309–2319. https://doi.org/10.1002/asi.23563

Center for Disease Control and Prevention. (2016a). Adult obesity causes & consequences. Retrieved from http://www.cdc.gov/obesity/adult/causes.html

Center for Disease Control and Prevention. (2016b). Behavioral risk factor surveillance system. Retrieved from http://www.cdc.gov/brfss/

Centers for Disease Control and Prevention. (2017). Adult obesity prevalence maps | Overweight & Obesity | CDC. Retrieved from https://www.cdc.gov/obesity/data/prevalence-maps.html

Chandran, N. (2018). Obama to David Letterman: Media is dividing Americans. Retrieved from https://www.cnbc.com/2018/01/12/former-president-barack-obama-warns-on-polarizing-media-us-electoral-system.html

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22*, 288--296. https://doi.org/10.1.1.100.1089

Chen, X., Cho, Y., & Jang, S. Y. (2015). Crime prediction using Twitter sentiment and weather. In *2015 Systems and Information Engineering Design Symposium* (Vol. 00, pp. 63–68). IEEE. https://doi.org/10.1109/SIEDS.2015.7117012

Chew, C., & Eysenbach, G. (2010). *Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. PLOS ONE*, (5).

https://doi.org/https://doi.org/10.1371/journal.pone.001

Comito, C., Pizzuti, C., & Procopio, N. (2018). How people talk about health ? Detecting health topics from Twitter streams. In *BDIOT* (pp. 1–6). https://doi.org/http://dx.doi.org/10.1145/12345.67890

Cook, C. M., Subar, A. F., Troiano, R. P., & Schoeller, D. A. (2012). Relation between holiday weight gain and total energy expenditure among 40- to 69-y-old men and women (OPEN study). *The American Journal of Clinical Nutrition*, *95*(3), 726–731. https://doi.org/10.3945/ajcn.111.023036

Cooper, B. B. (2013). The surprising history of Twitter's hashtag origin. [web log comment]. Retrieved from https://blog.bufferapp.com/a-concise-history-of-twitter-hashtags-and-how-you-should-use-them-properly

Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). Google disease trends: An update. *Nature*, *457*(Cdc), 1012--1014. Retrieved from http://research.google.com/pubs/archive/41763.pdf

Corley, C., Cook, D., Mikler, A., & Singh, K. (2010). Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*, *7*(2), 596–615. https://doi.org/10.3390/ijerph7020596

Creswell, J. W. (2014). Quantitative Methods. In *Research Design:Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed., pp. 155–182). Thousand Oaks: Sage Publications.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages (pp. 1–8). Washington, DC: ACM.

Culotta, A. (2014). Estimating county health statistics with twitter. In *Proceedings of the*

32nd annual ACM conference on Human factors in computing systems - CHI '14 (pp. 1335–1344). Toronto, Ontario, Canada: ACM Press. https://doi.org/10.1145/2556288.2557139

Curtin, C., Anderson, S. E., Must, A., & Bandini, L. (2010). The prevalence of obesity in children with autism: a secondary data analysis using nationally representative data from the National Survey of Children's Health. *BMC Pediatrics*, *10*(1), 11. https://doi.org/10.1186/1471-2431-10-11

Dalen, J., Smith, B. W., Shelley, B. M., Sloan, A. L., Leahigh, L., & Begay, D. (2010). Pilot study: Mindful Eating and Living (MEAL): Weight, eating behavior, and psychological outcomes associated with mindfulness-based intervention for people with obesity. *Complementary Therapies in Medicine*, *18*, 260–264. https://doi.org/10.1016/j.ctim.2010.09.008

Damasio, A. R., & Sutherlnad, S. (1994). Descartes' error: Emotion, reason, and the human brain. *Nature*, *372*(6503), 287.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social Media as a measurement tool of depression in populations. *WebSci '13 Proceedings of the 5th Annual ACM Web Science Conference*. https://doi.org/10.1145/2464464.2464480

Deiner, M. S., Lietman, T. M., McLeod, S. D., Chodosh, J., & Porco, T. C. (2016). Surveillance tools emerging from search engines and social media data for determining eye disease patterns. *JAMA Ophthalmology*, *134*(9), 1024–1030. https://doi.org/10.1001/jamaophthalmol.2016.2267

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer*

Vision and Pattern Recognition (248–255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Di Eugenio, B., & Glass, M. (2004). Squibs and discussions the Kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95–101. Retrieved from https://www.mitpressjournals.org/doi/pdfplus/10.1162/089120104773633402

Diaz-Aviles, E., & Stewart, A. (2012). Tracking Twitter for epidemic intelligence: Case study. In *Web Science Conference* (82–85).

Doucet, L., & Jehn, K. A. (1997). Analyzing harsh words in a sensitive setting: American expatriates in communist China. *Journal of Organizational Behavior*, *18*(S1), 559–582. https://doi.org/10.2307/3100265

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In *Advances in Neural Information Processing Dystems* (Vol. 1, pp. 155–161). https://doi.org/10.1.1.10.4845

Duggan, M. Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). *Social media update 2014*.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169. https://doi.org/10.1177/0956797614557867

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, (October), 1277–1287. https://doi.org/10.1038/nrm2900

Eschler, J., Dehlawi, Z., & Pratt, W. (2015). Self-Characterized Illness Phase and Information Needs of Participants in an Online Cancer Forum. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 101–109). Retrieved from http://ai2-s2-pdfs.s3.amazonaws.com/bec4/c2023bbd2ec296e22d56103162d7bdbd0c66.pdf

Eysenbach. (2006). Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 244–248. https://doi.org/PMC1839505

Eysenbach, G., & Kohler, C. (2003). What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, 225–229. https://doi.org/D030003690 [pii]

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ (Clinical Research Ed.), 324*(7337), 573–577. https://doi.org/10.1136/BMJ.324.7337.573

Fahimi, M., Link, M., Mokdad, A., Schwartz, D. A., & Levy, P. (2008). Tracking chronic disease and risk behavior prevalence as survey participation declines: statistics from the behavioral risk factor surveillance system and other national surveys. *Preventing Chronic Disease*, *5*(3), A80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18558030

Finfgeld-Connett, D. (2015). Twitter and health science research. *Western Journal of Nursing Research, 37*(10), 1269–1283. https://doi.org/10.1177/0193945914565056

Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., & Dietz, W. (2009). Annual medical spending attributable to obesity: payer-and service-specific estimates. *Health Affairs*, *28*(5), w822–w831. https://doi.org/10.1377/hlthaff.28.5.w822

Fisher, J., Amico, K. R., Fisher, W., & Harman, J. J. (2008). *The Information-Motivation-Behavioral Skills Model of Antiretroviral Adherence and Its Applications*. Storrs.

Fisher, R. J., & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. *Psychology and Marketing*, *17*(2), 105–120. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<105::AID-MAR3>3.0.CO;2-9

Flegal, K. M., Carroll, M. D., Kit, B. K., & Ogden, C. L. (2012). Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. *JAMA : The Journal of the American Medical Association*, *307*(5), 491–497. https://doi.org/10.1001/jama.2012.39

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). Statistical methods for rates and proportions. John Wiley & Sons.

Forrest, K. Y., & Lin, Y. (2010). *Analyses of Behavioral Risk Factor Surveillance System Data for Rural Health Outcomes*.

Fox, M. (2016). America's obesity epidemic hits a new high. NBC News. Retrieved from https://www.nbcnews.com/health/health-news/america-s-obesity-epidemic-hits-new-high-n587251

Fox, S. (2009). The social life of health information. *Pew Research Center*. Retrieved from http://www.pewresearch.org/fact-tank/2014/01/15/the-social-life-of-health-information/

Freund, Y., & Schapire, R. E. (1997). A desicion-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1007/3-540-59119-2_166

Fried, D., Surdeanu, M., Kobourov, S., Hingle, M., & Bell, D. (2014). Analyzing the language of food on social media. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 778–783). IEEE. https://doi.org/10.1109/BigData.2014.7004305

Ghosh, D., & Guha, R. (2013). What are we "tweeting" about obesity? Mapping tweets with Topic Modeling and Geographic Information System. *Cartogr Geogr Information Science*, *40*(2), 90–102. https://doi.org/10.1080/15230406.2013.776210

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Gore, R. J., Diallo, S., & Padilla, J. (2015). You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. *PLOS ONE*, *10*(9), e0133505. https://doi.org/10.1371/journal.pone.0133505

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big Social Data Analytics in Journalism and Mass Communication. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. https://doi.org/10.1177/1077699016639231

Hall, I., Johnson-Turbes, A., & Williams, K. (2010). *The Potential of Black Radio to Disseminate Health Messages and Reduce Disparities* (Vol. 7). Retrieved from http://www.cdc.gov./pcd/issues/2010/

Harris, J. G., & Mehrotra, V. (2014). Getting value from your data scientists. *MIT Sloan*

Management Review, *56*(1), 15–18. Retrieved from http://sloanreview.mit.edu/article/getting-value-from-your-data-scientists/

Harris, J. K., Moreland-Russell, S., Tabak, R. G., Ruhr, Li. R., & Maier, R. C. (2014). Communication About Childhood Obesity on Twitter. *American Journal of Public Health*, *104*(7), e62–e69. https://doi.org/10.2105/AJPH.2013.301860

Hartz, A. J., Rupley, D. C., Kalkhoff, R. D., & Rimm, A. A. (1983). Relationship of obesity to diabetes: Influence of obesity level and body fat distribution. *Preventive Medicine*, *12*(2), 351–357. https://doi.org/10.1016/0091-7435(83)90244-X

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271. https://doi.org/10.1080/15230406.2014.890072

Hawn, C. (2009). Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. *Health Affairs*, *28*(2), 361–368. https://doi.org/10.1377/hlthaff.28.2.361

Health and Human Services, D. N. I. of H. (2012). Overweight and obesity statistics. Retrieved from http://www.niddk.nih.gov/health-information/health-statistics/Pages/overweight-obesity-statistics.aspx

Health, H. S. of P. (n.d.). Obesity Consequences: Health Risk. Retrieved January 1, 2016, from https://www.hsph.harvard.edu/obesity-prevention-source/obesity-consequences/health-effects/

Heaven, D. (2016). Beyond Pokémon Go: augmented reality is set to transform gaming | New Scientist. Retrieved May 22, 2018, from

https://www.newscientist.com/article/2100412-beyond-pokemon-go-augmented-reality-is-set-to-transform-gaming/

Hill, S., Merchant, R., & Ungar, L. (2013). Lessons Learned About Public Health from Online Crowd Surveillance. *Big Data*, *1*(3), 160–167. https://doi.org/10.1089/big.2013.0020

Hoeken, H., Kolthoff, M., & Sanders, J. (2016). Story Perspective and Character Similarity as Drivers of Identification and Narrative Persuasion. *Human Communication Research*, *42*(2), 292–311. https://doi.org/10.1111/hcre.12076

Hong, L., & Davison, B. D. (2010). Empirical Study of Topic Modeling in Twitter. In *In 1st Workshop on Social Media Analytics* (pp. 1–9). Washington, DC: ACM.

Ilarri, S., Illarramendi, A., Mena, E., & Sheth, A. (2016). Semantics in Location-Based Services The Importance of Location-Based Services The Importance of Managing Semantics in LBS. *IEEE Internet Computing*, *15*(6), 1–9.

Inoue-Choi, M., Liao, L. M., Reyes-Guzman, C., Hartge, P., Caporaso, N., & Freedman, N. D. (2017). Association of long-term, low-intensity smoking with all-cause and cause-specific mortality in the national institutes of health-AARP diet and health study. *JAMA Internal Medicine*, *177*(1), 87–95. https://doi.org/10.1001/jamainternmed.2016.7511

Jurafsky, D., & Martin, J. H. (2014). N-Grams. In *Speech and Language Processing* (pp. 2–7). Retrieved from https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. Health economics, 16(2), 163-178.

Karami, A., Dahl, A. A., Turner-McGrievy, G., Kharrazi, H., & Shaw, G. (2018). Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management*, *38*(1), 1–6. https://doi.org/10.1016/j.ijinfomgt.2017.08.002

Karami, A., Gangopadhyay, A., Zhou, B., & Karrazi, H. (2015). FLATM: A fuzzy logic approach topic model for medical documents. In *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC)* (pp. 1–6). IEEE. https://doi.org/10.1109/NAFIPS-WConSC.2015.7284190

Karami, A., Gangopadhyay, A., Zhou, B., & Kharrazi, H. (2015). A Fuzzy Approach Model for Uncovering Hidden Latent Semantic Structure in Medical Text Collections. In *iConference* (pp. 1–5). Retrieved from http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/6222/version/1

Kim, D. D., & Basu, A. (2016). Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis. *Value in Health*, *19*(5), 602–613. https://doi.org/10.1016/J.JVAL.2016.02.008

Knight, W. (2017). Biased algorithms are everywhere, and no one seems to care - MIT Technology Review. Retrieved July 26, 2018, from https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/

Koh-Banerjee, P., Wang, Y., Hu, F. B., Spiegelman, D., Willett, W. C., & Rimm, E. B. (2004). Changes in body weight and body fat distribution as risk factors for clinical diabetes in US men. *American Journal of Epidemiology*, *159*(12), 1150–1159.

https://doi.org/10.1093/aje/kwh167

Komito, L. (2011). Social media and migration: Virtual community 2.0. *Journal of the American Society for Information Science and Technology*, *62*(6), 1075–1086. https://doi.org/10.1002/asi.21517

Krippendorff, K. (2004). *Content Analysis: An Introduction to its Methodology. Sage Publications* (2nd Editio). Thousand Oaks: Sage Publications. https://doi.org/10.2307/2288384

Kuehn, B. M. (2015). Twitter Streams Fuel Big Data Approaches to Health Forecasting. *JAMA : The Journal of the American Medical Association*, *314*(19), 2010–2012.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and Best Practices in Content Analysis. *Journalism & Mass Communication Quarterly*, *92*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of NAACL-HLT 2013* (pp. 789–795). https://doi.org/10.2196/jmir.2534

Lampos, V., & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. In *2010 2nd International Workshop on Cognitive Information Processing, CIP2010* (pp. 411–416). IEEE. https://doi.org/10.1109/CIP.2010.5604088

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *52*(2), 21–32. https://doi.org/10.0000/PMID57750728

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014a). The parable of Google Flu: Traps in Big Data Analysis. *Www.Sciencemag.Org SCIENCE*, *343*(March), 1203–

1206. Retrieved from http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014b). The parable of Google Flu: Traps in Big Data Analysis. *Www.Sciencemag.Org SCIENCE*, *343*(March), 1203–1206.

LeBlanc, A. G., & Chaput, J.-P. (2017). Pokémon Go: A game changer for the physical inactivity crisis? *Preventive Medicine*, *101*, 235–237. https://doi.org/10.1016/j.ypmed.2016.11.012

Leetaru, K. H. (2012). *Data mining methods for the Content Analyst: An introduction to the computational analysis of content*. New York, New York, USA; Routledge Taylor and Francis Group.

Liu, Y., Mei, Q., Hanauer, D. A., Zheng, K., & Lee, J. M. (2016). Use of Social Media in the Diabetes Community: An Exploratory Analysis of Diabetes-Related Tweets. *JMIR Diabetes*, *1*(2), 1–8. https://doi.org/10.2196/diabetes.6256

Lloyd-Jones, D. M., Hong, Y., Labarthe, D., Mozaffarian, D., Appel, L. J., Van Horn, L., … Rosamond, W. D. (2010). Defining and Setting National Goals for Cardiovascular Health Promotion and Disease Reduction: The American Heart Association's Strategic Impact Goal Through 2020 and Beyond. *Circulation*, *121*(4), 586–613. https://doi.org/10.1161/CIRCULATIONAHA.109.192703

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clinical and Experimental Pharmacology and Physiology*, *29*(7), 527–536. https://doi.org/10.1046/j.1440-1681.2002.03686.x

Lukic, L., Lalic, N. M., Rajkovic, N., Jotic, A., Lalic, K., Milicic, T., … Gajovic, J. S.

(2014). Hypertension in obese type 2 diabetes patients is associated with increases in insulin resistance and IL-6 cytokine levels: potential targets for an efficient preventive intervention. *International Journal of Environmental Research and Public Health*, *11*(4), 3586–3598. https://doi.org/10.3390/ijerph110403586

Lupton, D. (2016). Digitized health promotion: Risk and personal responsibility for health in the Web 2.0 era. In J. Davis & A. M. Gonzales (Eds.), *To fix or to heal: Patient care, public health, and the limits of biomedicine* (pp. 152–176). New York, New York, USA: University Press.

Lupton, D., & Michael, M. (2017). "For Me, the Biggest Benefit Is Being Ahead of the Game":The Use of Social Media in Health Work. *Social Media + Society*, *3*(2), 205630511770254. https://doi.org/10.1177/2056305117702541

Mchugh, M. L. (2012). Interrater reliability: the kappa statistic Importance of measuring interrater reliability Theoretical issues in measurement of rater reliability. *Biochem Med (Zagreb). Oct;*, *22*(3), 276–282. Retrieved from http://pl8cg5fc8w.search.serialssolutions.com/?ctx_ver=Z39.88-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF-8&rfr_id=info%3Asid%2Fsummon.serialssolution...

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Menifield, C., Doty, N., & Fletcher, A. (2008). Obesity in America. *The ABNF Journal*, 83–88.

Miller, M., Banerjee, T., Muppalla, R., Romine, W., & Sheth, A. (2017). What Are People Tweeting About Zika? An Exploratory Study Concerning Its Symptoms, Treatment, Transmission, and Prevention. *JMIR Public Health and Surveillance*, *3*(2), e38. https://doi.org/10.2196/publichealth.7157

Moore, E. M., Mander, A. G., Ames, D., Kotowicz, M. A., Carne, R. P., Brodaty, H., … AIBL Investigators, the A. (2013). Increased risk of cognitive impairment in patients with diabetes is associated with metformin. *Diabetes Care*, *36*(10), 2981–2987. https://doi.org/10.2337/dc13-0229

Mowery, D. L., Bryan, C., & Conway, M. (2015). Towards Developing an Annotation Scheme for Depressive Disorder Symptoms: A Preliminary Study using Twitter Data. In *2nd Workshop on Computational Linguistics and Clinical Psychology* (pp. 89–98). Denver.

Mowery, J. (2016). Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns. *Online Journal of Public Health Informatics*, *8*(3), e198. https://doi.org/10.5210/ojphi.v8i3.7011

Moyer-Gusé, E. (2008). Toward a Theory of Entertainment Persuasion: Explaining the Persuasive Effects of Entertainment-Education Messages. *Communication Theory*, *18*(3), 407–425. https://doi.org/10.1111/j.1468-2885.2008.00328.x

Nagel, A. C., Tsou, M.-H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., … Sawyer, M. H. (2013). The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *Journal of Medical Internet Research*, *15*(10), e237. https://doi.org/10.2196/jmir.2705

Nakhasi, A., Passarella, R., Bell, S. G., Paul, M. J., Dredze, M., & Pronovost, P. J.

(2012). Malpractice and Malcontent : Analyzing Medical Complaints in Twitter. *In International Conference on Weblogs and Social Media (ICWSM).*, 1–2. Retrieved from https://www.aaai.org/ocs/index.php/FSS/FSS12/paper/viewFile/5572/5857

Nasukawa, T., & Yi, J. (2003). Sentiment analysis. In *Proceedings of the international conference on Knowledge capture - K-CAP '03* (pp. 70–77). New York, New York, USA: ACM Press. https://doi.org/10.1145/945645.945658

Nguyen, Q. C., Kath, S., Meng, H. W., Li, D., Smith, K. R., VanDerslice, J. A., … Li, F. (2016). Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, *73*, 77–88. https://doi.org/10.1016/j.apgeog.2016.06.003

Obole, A., & Welsh, K. (2012). The danger of big data: Social media as computational social science. *First Monday*, *17*(7). https://doi.org/10.5210/fm.v17i7.3993

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control*, *43*(6), 563–571. https://doi.org/10.1016/j.ajic.2015.02.023

Ogden, C. L., Carroll, M. D., Curtin, L. R., McDowell, M. A., Tabak, C. J., & Flegal, K. M. (2006). Prevalence of Overweight and Obesity in the United States, 1999-2004. *JAMA*, *295*(13), 1549. https://doi.org/10.1001/jama.295.13.1549

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. (J. Callan & F. Sebastiani, Eds.), *Foundations and Trends® in Information Retrieval* (Vol. 2). Hanover: now Publishers Inc. https://doi.org/10.1561/1500000011

Paparrizos, J., White, R. W., & Horvitz, E. (2016). Screening for Pancreatic Adenocarcinoma Using Signals From Web Search Logs: Feasibility Study and

Results. *Journal of Oncology Practice*, *12*(8), 737–744. https://doi.org/10.1200/JOP.2015.010504

Paul, M., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media* (pp. 265–272). Spain: Association for the Advancement of Artificial Intelligence.

Paul, M. J., & Dredze, M. (2012). *A Model for Mining Public Health Topics from Twitter*.

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0103408

Paul, M. J., & Dredze, M. (2017). *Social Monitoring for Public Health*. (G. Marchionini, Ed.), *Synthesis Lectures on Information Concepts, Retrieval, and Services* (Vol. 9). Morgan & Claypool Publishers. https://doi.org/10.2200/S00791ED1V01Y201707ICR060

Paul, M. J., Dredze, M., Broniatowski, D. A., & Generous, N. (2015). Worldwide Influenza Surveillance through Twitter. In *The World Wide Web and Public Health Intelligence: Papers from the 2015 AAAI Workshop* (pp. 6–11).

Pierannunzi, C., Hu, S. S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Medical Research Methodology*, *13*(1), 49. https://doi.org/10.1186/1471-2288-13-49

Polgreen, P. M., Chen, Y., Pennock, D. M., & Nelson, F. D. (2008). Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, *47*(11), 1443–1448. https://doi.org/10.1086/593098

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://doi.org/10.1.1.214.9232

Preis, T., & Moat, H. S. (2014). Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, *1*(2), 1–5. https://doi.org/10.1098/rsos.140095

Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011). Identifying health-related topics on twitter an exploration of tobacco-related tweets as a test topic. In *In International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (Vol. 6589 LNCS, pp. 18–25). Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19656-0_4

Rahman, M. M., & Wang, H. (2016). Hidden Topic Sentiment Model. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16* (pp. 155–165). Quebec, Canada. https://doi.org/10.1145/2872427.2883072

Ritterman, J., Osborne, M., & Klein, E. (2009). Using prediction markets and Twitter to predict a swine flu pandemic. In *1st International Workshop of Mining Social Media* (pp. 9–17). Edinburgh.

Robillard, J. M., Cabral, E., Hennessey, C., Kwon, B. K., & Illes, J. (2015). Fueling Hope: Stem Cells in Social Media. *Stem Cell Reviews and Reports*, *11*(4), 540–546. https://doi.org/10.1007/s12015-015-9591-y

Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., & Rosa, T. C. (2016). SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, *85*(1), 80–95.

https://doi.org/10.1016/j.ijmedinf.2015.09.007

Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017). Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach. *Computational and Mathematical Methods in Medicine*, *2017*, 5140631. https://doi.org/10.1155/2017/5140631

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., & Brownstein, J. S. (2015). Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology*, *11*(10), 1–15. https://doi.org/10.1371/journal.pcbi.1004513

Schmidt, W. (1997). World Wide Web survey research: benefits, potential problems, and solutions. *Behav Res Methods Instrum Comput*, *29*(2), 274–279.

schraefel, m. c., White, R. W., André, P., & Tan, D. (2009). Investigating web search strategies and forum use to support diet and weight loss. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09* (p. 3829). New York, New York, USA: ACM Press. https://doi.org/10.1145/1520340.1520579

Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J. (2010). The utility of "Google Trends" for epidemiological research: Lyme disease as an example. *Geospatial Health*, *4*(2), 135–137. https://doi.org/10.4081/gh.2010.195

Shaw, G., & Karami, A. (2017). Computational content analysis of negative tweets for obesity, diet, diabetes, and exercise. *Proceedings of the Association for Information Science and Technology*, *54*(1), 357–365.

https://doi.org/10.1002/pra2.2017.14505401039

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, *6*(5), e19467. https://doi.org/10.1371/journal.pone.0019467

Stefansen, C. (2014). Google AI Blog: Google Flu Trends gets a brand new engine. Retrieved from https://ai.googleblog.com/2014/10/google-flu-trends-gets-brand-new-engine.html

Sweeney, K., & Whissell, C. (1984). A Dictionary of Affect in Language: I. Establishment and Preliminary Validation. *Perceptual and Motor Skills*, *59*(3), 695–698. https://doi.org/10.2466/pms.1984.59.3.695

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tenzer, J. (2016, March 14). Technically, Millennials Are Way Ahead! *Huffington Post*. Retrieved from http://www.huffingtonpost.com/jamee-tenzer/technically-millennials-a_b_9453480.html

Terrell, G. R., & Scott, D. W. (1992). Variable Kernel Density Estimation. *Source: The Annals of Statistics The Annals of Statistics*, *20*(3), 1236–1265. Retrieved from http://www.jstor.org/stable/2242011

Thackeray, R., Burton, S. H., Giraud-Carrier, C., Rollins, S., & Draper, C. R. (2013). Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. *BMC Cancer*, *13*(1), 508. https://doi.org/10.1186/1471-2407-13-508

The National Academy of Hypothyroidism. (2017). Hypothyroidism, Irritable Bowel Syndrome (IBS) and Your Health. Retrieved July 8, 2018, from https://www.nahypothyroidism.org/hypothyroidism-irritable-bowel-syndrome-and-your-health/

Thelwall, M. (2014, June). A brief history of altmetrics. https://doi.org/10.1371/journal.pone.0064841

Tse, M. M. Y., Lo, A. P. K., Cheng, T. L. Y., Chan, E. K. K., Chan, A. H. Y., & Chung, H. S. W. (2010). Humor therapy: relieving chronic pain and enhancing happiness for older adults. *Journal of Aging Research*, *2010*, 343574. https://doi.org/10.4061/2010/343574

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Fourth International AAAI Conference on Weblogs and Social Media* (pp. 178–185). https://doi.org/10.1074/jbc.M501708200

Turner-McGrievy, G. M., & Beets, M. W. (2015). Tweet for health: Using an online social network to examine temporal trends in weight loss-related posts. *Translational Behavioral Medicine*, *5*(2), 160–166. https://doi.org/10.1007/s13142-015-0308-1

Vraga, E. K., Stefanidis, A., Lamprianidis, G., Croitoru, A., Crooks, A. T., Delamater, P. L., … Jacobsen, K. H. (2018). Cancer and Social Media: A Comparison of Traffic about Breast Cancer, Prostate Cancer, and Other Reproductive Cancers on Twitter and Instagram. *Journal of Health Communication*, *23*(2), 181–189. https://doi.org/10.1080/10810730.2017.1421730

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, *21*(6), 1098–1103. https://doi.org/10.1136/amiajnl-2014-002711

Wang, S., Paul, M. J., & Dredze, M. (2014). Exploring Health Topics in Chinese Social Media : An Analysis of Sina Weibo. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 20–23).

Wang, X., Zhang, C., Ji, Y., Sun, L., & Wu, L. (2013). A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 201–213).

Wang, Y., Beydoun, M. A., Liang, L., Caballero, B., & Kumanyika, S. K. (2008). Will All Americans Become Overweight or Obese? Estimating the Progression and Cost of the US Obesity Epidemic. *Obesity*, *16*(10), 2323–2330. https://doi.org/10.1038/oby.2008.351

Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social Media as Information Source: Recency of Updates and Credibility of Information. *Journal of Computer-Mediated Communication*, *19*(2), 171–183. https://doi.org/10.1111/jcc4.12041

Wilson, K., Atkinson, K., & Deeks, S. (2014). Opportunities for utilizing new technologies to increase vaccine confidence. *Expert Review of Vaccines*, *13*(8), 969–977. https://doi.org/10.1586/14760584.2014.928208

Wing, R. R., Goldstein, M. G., Acton, K. J., Birch, L. L., Jakicic, J. M., Sallis, J. F., … Surwit, R. S. (2001). Behavioral Science Research in Diabetes: Lifestyle changes related to obesity, eating behavior, and physical activity. *Diabetes Care*, *24*(1), 117–

123. https://doi.org/10.2337/diacare.24.1.117

Yang, S., Santillana, M., & Kou, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(47), 14473–14478. https://doi.org/10.1073/pnas.1515373112

Yanovski, J. A., Yanovski, S. Z., Sovik, K. N., Nguyen, T. T., O'Neil, P. M., & Sebring, N. G. (2000). A Prospective Study of Holiday Weight Gain. *New England Journal of Medicine*, *342*(12), 861–867. https://doi.org/10.1056/NEJM200003233421206

Yin, Z., Fabbri, D., Rosenbloom, S. T., & Malin, B. (2015). A scalable framework to detect personal health mentions on Twitter. *Journal of Medical Internet Research*, *17*(6), e138. https://doi.org/10.2196/jmir.4305

Zhou, L., Zhang, D., Yang, C. C., & Wang, Y. (2018). Harnessing social media for health information management. *Electronic Commerce Research and Applications*, *27*, 139–151. https://doi.org/10.1016/j.elerap.2017.12.003

Appendix A

Sentiment Analysis Evaluation – Call for Participants

Dear Participant,

You are invited to participate in this study to evaluate a collection of tweets representing four health related categories. Sentiment analysis or the positive and negative opinions regarding an event, provides insight into peoples' feelings toward a political event, newly release movie, or healthcare issue. This research seeks to characterize health behaviors related to obesity from social data using computational methods. Also, this research uses qualitative methods to evaluate the computation model outputs and reliability.

The expected time to complete the task of labeling the tweets is 15-20 minutes. If you feel that the tweets are obtrusive, offensive, and/or are causing emotional harm to you personally, you are not obligated to continue the task; however, you will not be compensated if you end the task for this reason. Should you have any questions, concerns, complaints, or think the research has hurt you, contact the primary investigator George Shaw, Jr. at gshaw@email.sc.edu

This research project is part of the primary investigator's dissertation and all task contributors' answers will be used towards the understanding and advancement of research with regards to complementary data sources for health risk behavior regarding obesity. The study was reviewed and approved by the Institutional Review Board (IRB ID: Pro00078907) at the University of South Carolina.

Evaluation of Tweets – Diet

- I have a bad diet
- Include fresh goods in your diet
- I want more veggies and fish in my diet. I also want to go back to tea rather than coffee. I did very well in the past.
- I am unsure if I will have a cheat meal later or stay on the diet

Evaluation of Tweets – Diabetes

- Diabetes is annoying
- I can't quit drinking sweet tea, my brother told me I am going to get diabetes
- Health fair teaches healthy eating, diabetes prevention
- Free from diabetes

Evaluation of Tweets – Exercise

- My bike means so much to me than words can express: freedom, health, me, time, exercise, living
- Diet, exercise, and outlook are key to self-improvement.
- I really need some exercise to lose my weight
- 20 minutes of exercise in the morning and afternoon is perfect

Evaluation of Tweets – Obesity

- We need to keep keys active. Obesity is a keyword. We need to keep funding
- I fear obesity
- Burgers are $.99 and salads are $4.99: here is why we have an obesity problem in America
- Morbid obesity is a serious health problem