Published in final edited form as:
Cartogr Geogr Inf Sci. 2013 ; 40(2): 90–102. doi:10.1080/15230406.2013.776210.

What are we ‘tweeting’ about obesity? Mapping tweets with Topic Modeling and Geographic Information System

Debarchana (Debs) Ghosh[*] and
Department of Geography, University of Connecticut, Storrs, CT 06040

Rajarshi Guha
(guhar@mail.nih.gov), NIH Center for Advancing Translational Science, Rockville, MD, 20850

Abstract

Public health related tweets are difficult to identify in large conversational datasets like Twitter.com. Even more challenging is the visualization and analyses of the spatial patterns encoded in tweets. This study has the following objectives: How can topic modeling be used to identify relevant public health topics such as obesity on Twitter.com? What are the common obesity related themes? What is the spatial pattern of the themes? What are the research challenges of using large conversational datasets from social networking sites? Obesity is chosen as a test theme to demonstrate the effectiveness of topic modeling using Latent Dirichlet Allocation (LDA) and spatial analysis using Geographic Information System (GIS). The dataset is constructed from tweets (originating from the United States) extracted from Twitter.com on obesity-related queries. Examples of such queries are ‘food deserts’, ‘fast food’, and ‘childhood obesity’. The tweets are also georeferenced and time stamped. Three cohesive and meaningful themes such as ‘childhood obesity and schools’, ‘obesity prevention’, and ‘obesity and food habits’ are extracted from the LDA model. The GIS analysis of the extracted themes show distinct spatial pattern between rural and urban areas, northern and southern states, and between coasts and inland states. Further, relating the themes with ancillary datasets such as US census and locations of fast food restaurants based upon the location of the tweets in a GIS environment opened new avenues for spatial analyses and mapping. Therefore the techniques used in this study provide a possible toolset for computational social scientists in general and health researchers in specific to better understand health problems from large conversational datasets.

Keywords

mapping; social media; topic models; GIS; text mining; obesity

Introduction

Over recent years, social networking sites (SNS) such as Twitter, Facebook, MySpace, FriendFeed, and GooglePlus have significantly transformed the way individuals interact and communicate with each other across the world. Moreover, these platforms are creating new

[*] Corresponding Author: debarchana.ghosh@uconn.edu.

avenues for real-time data and research. They have been a gold mine for scholars in fields such as linguistics, sociology, economics, health, and psychology that are looking for real-time *conversational* or *text* data for behavioral analyses, sentiment analyses, trend analyses, information dissemination, and health surveillance.

These web-based micro-blogging applications share several common features (Boyd and Ellison 2008). While there are slight variations in actual implementations, each service enables users to 1) create a public profile, 2) post and read short messages, 3) define a list of other users with whom they share a connection, 4) view and discover connections between other users within the system, 5) communicate with users with whom they share explicit social connections, and 6) interact with a wider audience of users with whom they would not have otherwise shared a social connection (Haythornthwaite 2005; Boyd and Ellison 2008).

We focus on one such SNS, Twitter.com, which in particular provides a medium whereby users can create and exchange content with a potentially larger audience than either Facebook or MySpace. Twitter.com allows users to communicate with each other in real-time through short, concise messages (no longer than 140 characters), known as ‘tweets’. A user’s tweets are available to all of his/her ‘followers’ and ‘friends’, i.e. all others who choose to subscribe to that user’s profile and are publicly searchable by default (Butts and Acton 2011; Twitter 2011). It is needless to mention that Twitter.com among all other SNS, continues to grow in popularity. As of June 2011, a rough estimate of 100 million individuals used Twitter.com in the world of which 70 million people used in the United States (US). Approximately 63% of the Twitter users in the US are under the age of 35 years (45% are between the ages 18 and 37 years) (Quantcast 2011). The conversations on Twitter.com through tweets are publicly available and easily accessed through Twitter’s Application Programing Interface (API) and therefore offer a rich opportunity to observe social interaction and communication among users (Twitter 2011). Two characteristics of such conversational data from Twitter.com are specifically appealing to researchers, i.e. its *immediacy* and *immensity*. Instead of relying on traditional questionnaires and other laborious and time-consuming methods of data collection, social scientists can now simply take advantage of Twitter’s stream of (virtually limitless) textual data on a variety of topics.

Recent studies have begun to investigate health-related behaviors, awareness, and surveillance trends by using Twitter’s real-time or archived data. For example, Salathe and Khandelwal (Salathe and Khandelwal 2011) used publicly available twitter data over a period of six months to measure the spatiotemporal sentiment towards a new vaccine for pandemic influenza (HINI) using social network analysis and Geographic Information Systems (GIS). The approach was validated by identifying a strong correlation between sentiments expressed online and Center for Disease Control and Prevention’s (CDC) estimated vaccination rates by states. In an attempt to evaluate health status, Culotta compared the frequency of influenza-related twitter messages with influenza statistics from CDC (Culotta 2010). Findings revealed a 0.78 correlation with the CDC statistics suggesting that tracking tweets might provide a cost effective and quicker way to monitor health status. Chew and Eysenbach analyzed tweets in an effort to determine the types and quality of information exchanged during the H1N1 outbreak (Chew 2010). The significant proportion of Twitter posts in this study was news-related (46%) with very few posts contained misinformation. Scanfield *et al*. explored twitter's status updates related to antibiotics in an effort to discover evidence of misunderstanding and misuse of antibiotics (Scanfield et. al. 2006). Their research indicated that Twitter.com offers a platform for sharing of health information and advice. More recently Paul and Drezde considered a broader range of public health applications for Twitter. They applied the recently introduced Ailment Topic Aspect Model to over one and a half million health related tweets and discovered mentions of over a dozen ailments, including allergies, obesity and insomnia. The models showed quantitative correlations with public health data and qualitative evaluations of model output suggested that Twitter.com has broad applicability for public health research (Paul and Dredze 2011).

While these studies are encouraging, the use of Twitter data in health research could be extended much further by using additional attribute data from the tweets such as the *geographic location* of the users or the location from which the tweets are posted. Although sharing of geographic location is optional, ubiquitous mobile geolocation services have enabled users of SNS to disclose their geographic location very easily. The Twitter API exposes geolocation information for those users that have specifically allowed such sharing (Twitter 2011). The geographic location of the twitter users allows us to address new questions in public health, viz., whether '*place*' or '*where we* live' affects our choices, healthy behavior, or access to health care. The location data also allows linking twitter data to other datasets such as the US Census, economic surveys, health surveys, and disease surveillance. It is our contention that user-generated conversation on Twitter.com along with their geolocation information provide relevant health-related data, whose identification and analyses adds value to health researchers. It may allow tailoring health interventions more effectively to their audiences and where they live.

In this paper, we address a research problem of how to effectively identify, mine, and spatially analyze public health related topics within Twitter's large conversational data. Unlike other topics such as politics, sports, and celebrity, public health related topics and discussions typically use less frequent words and are therefore more difficult to identify using traditional topic modeling techniques. Even more challenging is the visualization and analysis of the spatial patterns encoded in such tweets. Our focus, here, is on the following questions.

1. How can topic modeling be used to most effectively identify relevant public health topics such as obesity on Twitter.com?
2. What are the common obesity related themes?
3. What is the spatial pattern of the identified themes?
4. What are the research challenges of using large conversational datasets from social networking sites?

We test the proposed methodology by focusing on the 'obesity' epidemic in the US. Although we are interested in a technique to identify and better understand the spatial distribution of public health related tweets in general, a test topic provides a useful indicator through which we can guide and gauge the effectiveness of our methodology.

Obesity is now the most critical health challenge in the US in terms of both increasing rates and rising health expenditures. During the past 20 years, there has been a dramatic increase in obesity in the US and rates still remain high. In 2009-2010, almost 41 million women and more than 37 million men aged 20 or over were obese (35.7%). Among children and adolescents aged 2-19 years, more than 5 million girls and approximately 7 million boys were obese (17%) (CDC 2012). A recent study conducted by ‘Partnership to Fight Chronic Disease’, indicates that if current trends continue, 43% of US adults will be obese and spending will quadruple to $344 billion by 2018 (Thorpe 2009). On a positive note, if obesity rates are instead held at current levels and cost effective treatments are used, the US would save nearly $200 billion in health care costs (Thorpe 2009). Most critical is the rising rates of obesity among children. Since 1980, obesity prevalence among children and adolescents has almost tripled and now 1 of 7 low-income, preschool-aged children is obese (CDC 2012).

In the following sections, we discuss our data sampling and analysis of tweets. We developed topic models based on Latent Dirichlet Allocation (LDA) to analyze topics from a dataset of tweets collected by searching for obesity-related search terms. We also highlight important topics and themes and their spatial pattern using GIS. Finally, we discuss our results, challenges of using large conversational datasets from SNS as well as possible areas for future research.

Methodology and Results

The workflow of this paper is divided into six major steps (Figure 1). Each of the six steps along with their results is described in the following sections.

Data Collection

Using the Twitter Search API and based on obesity-related search terms, a geographically representative sample of tweets from the US was gathered at 1-hour intervals over a period of five months from October 2011 to March 2012. A 1-hour interval was chosen because obesity related tweets were not truly real-time events and therefore it was unlikely that there would be significant changes in the search results within an hour. The ‘representativeness’ of the tweets was measured by evaluating the proportion of geocoded tweets to the entire set of tweets obtained during a given time period. With no filtering or search terms applied, we observed that 0.8 percent of the tweets were geocoded. On inspecting the obesity related tweets, we observed a similar proportion of geocoded tweets.

The obesity-related search terms were: 1) ‘childhood AND obesity’, 2) ‘eat AND right’, 3) ‘farm AND policy’, 4) ‘food AND desert’, 5) ‘high AND calorie’, 6) ‘fructose’, 7) ‘obesity AND corn’, 8) ‘soft AND drink’, 9) ‘weight AND gain’, 10) ‘McDonalds AND obesity’, 11) McDonalds and overweight’, 12) ‘obesity’, 13) ‘overweight’, 14) ‘obesity AND diabetes’, and 15) ‘overweight AND diabetes’. These search terms were chosen based upon the domain knowledge of obesity and its risk factors in the US. Also, we manually inspected each of the terms to confirm whether obesity-relevant tweets were retrieved. After combining all the tweets with the above search terms and deleting duplicate tweets, the dataset contained 2,581,283 unique tweets. The relevant columns in the dataset were 1) tweet_ID (this is the unique identifier of the tweets, 2) User_from_ID (ID of the user who posted the tweet, 3) User_to_ID (ID of the user to whom the tweet was posted), 4) tweet (The content of the tweet with a maximum of 140 characters), 5) Geog_X (The X coordinate of the geographic location from which the tweet was posted), 6) Geog_Y (The Y coordinate of the geographic location from which the tweet was posted), 7) Pname (Place name such as city and state from which the tweet was posted, and 8) Time (The day and time at which the tweet was posted). Lets us call this dataset the '*master dataset*'. In addition, a related dataset was also created for all the unique User_from_ID with columns 1) ID (ID of the users in the User_from_ID column in the master dataset), 2) Screen_Name (profile name of the twitter user), 3) Name (Name of the user), 4) Location (Location of the user), 5) UTC_offset (This is the difference in hours and minutes from Coordinated Universal Time (UTC) for a particular time), and 6) Time_zone (US time zones). Let us call this as the '*user dataset*'. Table 1 describes the structure and the relationship of the master and the user datasets. In the next step we geocoded the tweets.

Geocoding

The tweets were geocoded in six stages. In stage one, the tweets from the master dataset with both Goeg_X and Geog_Y values were added as a XY event layer and then exported to a point shape file using the functions in ESRI's ArcGIS 10.0 (ESRI 2012). In stage two, the tweets from the same master dataset with city and state names in the 'Pname' column were geocoded using the Address Locator in ArcGIS 10.0 (ESRI 2012). In stage three, the tweets with only city name and missing state name in the 'Pname' column were geocoded case by case by using ancillary information from the 'UTC_offset' and 'Time_Zone' columns. At this stage, the tweets, which were still ambiguous, were not geocoded. In stage four, the locations of the twitter users in the user dataset were geocoded using the city and the state names from the 'Location' column. As with stage two, the Address Locator in ArcGIS 10.0 was used for geocoding. In stage five, the Twitter users with only city name and missing state names in the 'Location' column were geocoded by using supporting information from the 'UTC_offset' and 'Time_Zone columns. The locations of users, which were still undetermined at this stage such as with only state names, atypical abbreviations, unknown places and non-English words were not geocoded. In stage six, the locations of the twitter users which were successfully geocoded were populated in the master dataset based on the common columns, 'User_from_ID (master dataset) and 'ID' (user dataset). After completing these six stages of geocoding, 455,981 tweets were successfully geocoded for the US. Table 2 reports the breakdown of geocoded and non-geocoded tweets by the obesity related search terms mentioned above. In the following step, the geocoded tweets were mapped to visualize spatial patterns.

Visualization

The point density maps in Figure 2 show the spatial distribution of obesity related tweets for 1) all search terms (Figure 2A), 2) 'obesity' (Figure 2B), 'childhood AND obesity' (Figure 2C), and 4) 'McDonalds AND obesity' (Figure 2D). The density maps were generated using the 'Point Density Tool' from ArcGIS 10.0 (ESRI 2012). The tool calculates the density of point features (location of tweets) around each output raster cell. A neighborhood is defined around each raster cell center, and the number of points that fall within the neighborhood is totaled and divided by the area of the neighborhood. In overall the density of obesity-related geocoded tweets are higher in number in the eastern part of the country. A closer inspection of the maps shows higher density of tweets around bigger cities such as New York, Washington DC, Chicago, San Francisco, San Diego, Chicago, Seattle, and Atlanta, and lower density of tweets in the central region of the country, states near the Gulf coast (Mississippi, Alabama, and Georgia), Appalachian, and northernmost states (North Dakota, Montana, Wyoming, and Idaho). This pattern of spatial distribution of tweets is consistent over all the obesity related search terms although the density of tweets varies significantly by search terms. One obvious plausible reason for such pattern is that people from cities have higher rates of using SNS, however, this pattern also indicates that people in large numbers are conversing with their friends and followers on Twitter about the rising rates of obesity, childhood obesity, high calorie foods, and health behaviors that increases the risk of obesity. In that sense, Twitter might be an effective way to spread the awareness of a new obesity prevention program or promoting health behavior to reduce the rates of obesity among large sections of population in a very short period (Prier et al. 2011; Salathe and Khandelwal 2011). We further evaluated a statistical correlation between the prevalence of obesity rates among US adults (with BMI >= 30) and the percentage of all obesity-related tweets at the state level. The 2010 obesity prevalence data was obtained from CDC's Behavioral Risk Factor Surveillance System (CDC 2012). The obesity-related tweets were aggregated at the state level and then normalized by the number of Internet users at the state level in 2010. The estimated number of Internet users was based on figures from US Census surveys (IWS 2012). The correlation value was −0.356 at a 0.001 level of significance. Even though the correlation was weak the negative association was interesting. Twitter users are typically educated and relatively aware of the growing concern of the rising rates of obesity. Since several studies have already shown that the prevalence of obesity is negatively correlated to education, it is likely that people talking about obesity and related issues will be located in regions with higher education levels and consequently, lower obesity rates. Also, the obesity-related conversations or tweets revolve around ways to prevent obesity, risk factors of obesity, and other related health concerns. However additional analyses are required for explaining this negative association. After this initial visualization of the location of the tweets, in the next step we mine the text of the tweets, which is not more than 140 characters.

Text Mining

The textual content of tweets was analyzed in R 2.15.1, using the 'tm' package (Feinerer, Hornik, and Meyer 2008; Feinerer 2012). Prior to topic modeling the text of the tweets were first imported in the package as a 'corpus', a data structure to manage and analyze a collection of documents. Once the corpus was created, the texts were then passed through a pre-processing pipeline consisting of the following steps. First, the raw data were cleaned by removing URLs (words starting with 'http://') and HTML entities (e.g., '&quot;', '&amp;', etc.). Second, we removed punctuation characters and converted all text to lower case. Finally, stop words (i.e., common words and words that we specifically wished to ignore such as "I, and, the, then, to etc.") were removed from the text of individual tweets. During these procedures we ensured that user identifiers (of the form @username) were retained, allowing us to explore interactions between users who are not necessarily friends or followers.

Following cleaning, we then applied the Porter stemming algorithm (Porter 1980) to convert words to a base form. Thus the words, “eat”, “eating”, “eater” would all be converted to the word “eat”. This ensures that different morphological variants of a word, all of which have similar or identical semantics, are still considered as equivalent for the purpose of topic modeling.

Next, we performed tokenization, in which a piece of text is broken into smaller components, called n-grams. For example, when n = 1, a piece of text is decomposed into individual words. In this study we focused on bigrams (n=2), so that a sentence such as “I love McDonalds burger and fries”, would be decomposed into the following series of bigrams: “love McDonalds”, “McDonalds burger”, “burger fries”. Note the stop words in this sentence such as “I”, and “and” were already removed and therefore will not be included in the list of bigrams. Bigrams were more informative than individual words as they provide some degree of context. While the use of larger n than bigrams would provide more contexts, the resultant dataset would become very sparse (i.e., most 3-grams, 4-grams and higher would occur in fewer tweets). The final step was to construct a document term matrix (DTM), in which the rows correspond to documents (tweets) and the columns correspond to the bigrams (considered over all tweets). If the j’th bigram is present in the i’th tweet, then element (i.j) of the DTM is set to 1, otherwise it is set to 0. This approach to constructing the DTM can lead to a large number of columns and thus a very a sparse DTM. To reduce the sparsity, we only considered terms that occurred in three or more documents. This obviously reduces the number of terms (i.e., columns). But as a result, some tweets end up not containing any of the identified bigrams. We removed these tweets or rows from further consideration, resulting in a final DTM with 366,230 rows and 1,009 columns.

Topic Modeling – Latent Dirichlet Allocation (LDA) models

Using the DTM described above, we then generated a topic model, using the LDA approach (Blei 2003). Fundamentally, a topic model allows one to algorithmically identify “topics” within a collection of documents based on the words contained in each document. For this study, documents are tweets and thus we employ a topic model to identify multiple topics, which can be used to group the individual tweets, based on the words in each tweet. Fundamentally, the LDA model repeatedly samples (based on a multinomial distribution) the words from a collection of tweets to identify which words tend to associate with each other. The result of this sampling process is that words are assigned to multiple topics (the number of which must be specified a priori) with differing probabilities. The topics identified by the model are not explicit; rather they are identified in a probabilistic manner, such that certain terms (and therefore their associated documents) are more likely to appear in a certain topic than another. Therefore, topic modeling can be considered a form of clustering (so that topics are conceptually similar to clusters). Having generated a set of assignments for the words, the LDA model then determines the assignment of documents to individual topics. Importantly, an LDA model will usually assign a document to multiple topics, with differing probabilities. The result of this is that one can consider a given document (which in our case is a tweet) to be primarily assigned to one topic, but in addition be related to a lesser degree to other topics. In this sense the LDA model is a generative model that attempts to capture the (unseen) random process that generates the observed documents (tweets). The reader is referred to Blei's article on probabilistic topic model (Blei 2012) for a high level overview of topic modeling and to Blei *et. al's* book chapter in 'Text mining, classification, clustering and applications' (Blei and Lafferty 2009) for a more technical review.

We developed topic models using the implementation in the *topicmodels* package (Bettina and Hornik 2011), which is a freely available package for the open source R statistical programming environment. We employed the default settings in the package, only specifying k, the number of topics desired. We developed a number of models with k ranging from 5 to 100 in increments of 5, and selected a final model using the perplexity [Blei,, et. al. 2003], defined as

$$2^{-\Sigma_{i=1}^{n}\left(\frac{1}{n}\right)log_2\, q(x_i)}$$

where q is the model, x_i is the i'th test document and n is the number of documents in the corpus. A good model will assign higher probabilities, $q(x_i)$, to the test documents, leading to a lower perplexity value. In our experiments, the model with k = 50, exhibited the thus lowest perplexity (158.61) and the rest of this work used this model to explore the thematic structure of this collection of tweets. The output from the selected LDA model is summarized in Table 2. The column 1 shows the topic number; column 2 lists the most likely topic components or *terms* with higher frequency for each of the topics in column 1, column 3 defines the obesity-related themes based on the terms, and column 4 is the percentage of tweets that used any of the terms within each topic.

Mapping of obesity-related themes

The topics displayed in Table 3 contain several interesting themes relating to how Twitter users discuss obesity-related topics. The themes are as follows.

"Theme 1 – Childhood obesity and schools"—Topics 2 and 20 contain a higher number of terms related not only to childhood obesity but also terms that relate to a recent incident where the US Congress passed a revised agriculture appropriations bill, essentially making it easier to count pizza sauce as a serving of vegetables in schools. These terms are: pizza vegetable, pizza tomato, school lunches, tomato sauce, tomato paste, vending machines, etc., comprising 4.79% (n = 21,887) of the total number of tweets. We also examined the temporal distribution of the tweets associated with this theme to find that all these tweets were posted within a month from the time the bill was passed (November 14th 2011). Hence we defined topics 2 and 20 as '*childhood obesity and schools*' theme.

A cluster analysis of the tweets associated with this theme was conducted using the hot-spot analysis tool in ArcGIS 10.0 to identify geographic regions with higher and lower density of tweets (ESRI) (Ghosh et al. 2011). The hot-spot tool identifies spatial clusters of variables with statistically significant high or low values. Given a weighted variable, in this case, density of tweets related to childhood obesity and schools by US counties, this tool delineates clusters of counties with higher than expected density of tweets. These clusters are called hot spots. The tool also delineates spatial clusters of lower than expected density of tweets. These clusters, called cold spots, are counties with significantly low density of tweets. The hot spot analysis tool also provided Z-scores and associated p-values, which were plotted to geographically locate the clusters of counties with hot and cold spots of tweets (Figure 3). Very high (> +1.96) and very low (< −1.96) values of Z-score are indicators of statistically significant hot and cold spots respectively. In this sense, Figure 3 shows five regions of hot spots where the density of tweets associated with Topics 2 and 20 (Childhood obesity and school theme), are significantly higher than the surrounding regions. These hot spots are in and around the cities of San Francisco, southern California including San Diego, Los Angeles, Austin, Chicago, and I-95 corridor including Washington DC, Philadelphia, New York City, cities in Connecticut and Massachusetts including Boston, and Providence. There were clusters of counties with lower density of tweets, however their Z-scores (lowest values was −1.460) indicated that the clusters were not statistically significant.

"Theme 2 – Obesity Prevention"—The topics 7, 13, and 17 contain topic components or bigram terms that relate to different ways of obesity prevention: calorie diet, calorie burn, obesity program, eating habit, eat rule, obesity awareness, weight loss, body mass, program prevent, eat salad, food stamp, achieve healthy, fight obesity, health news, avoid holiday, burn calorie, curb weight, food industry, avoid weight, food desert, and health center. There are approximately 32,000 tweets (6.96%) associated with this theme. We further summarized these tweets by states and compared that with the number of state policies related to obesity, nutrition, and physical activity, which are on going, introduced, or enacted in the year 2011. The information on these policies was obtained from the Center for Disease Prevention and Control's database of chronic disease state policy tracking system (CDC 2012). The bar graphs in Figure 4 shows this comparison. The state of California had highest number of tweets with 100 obesity related policies where as Arkansas had only 158 tweets with 13 state policies. The states such as Indiana, Arizona, West Virginia, North Dakota, South Dakota, and Wyoming have only *one* state policy for obesity, nutrition, and physical activity. Interestingly these are also the states with lower number of Twitter users messaging conversations related to prevention of obesity. The result was validated by identifying a positive correlation between the percentage of tweets (normalized by the number of Internet users at the state level in 2010) and the number of state policies of 2010-2011 with R^2 value of 0.69 at a 0.001 significance level (CDC 2012; IWS 2012).

"Theme 3 – Obesity and food habits"—The topics 16, 21, 25, and 29 contain wide range of terms related to high calorie food such as corn syrup, fructose corn, corn sugar, coca cola, McDonald, fries, junk food, ice cream, fat junk, eat chocolate, soft drink eat candy, and Chinese food. The other terms with higher frequency in these topics associate obesity with specific holidays or social events such as thanksgiving dinner, video game, holiday weight, food night, and super bowl. To understand the spatial implication of this theme we conducted a proximity analysis of tweets with location of McDonalds using tools from ArcGIS 10.0 (ESRI 2012). The term "McDonalds", in the context of obesity, was the topic component with highest frequency in Topics 16, 21, 25, and 29. According to InfoUSA database there are 14,007 McDonalds fast food restaurants in the US (InfoUSA 2012). The spatial distribution of the location of tweets (n = 44,230) in this theme and the location of McDonalds restaurants (n = 14,007) are shown in Figure 5. Based on these point shape files, we used the 'Near' tool from the proximity toolset of ArcGIS 10.0 to measure the Euclidean distance from each of the location of a tweet (Figure 5A) to the nearest location of a McDonalds (Figure 5B) within a specified search radius (ESRI). The search radius ranged from 0.25 miles to 5 miles with 0.5-mile increment. The output from the 'Near' tool is summarized in Table 4. The first column lists the search radius; column 2 is the percentage of tweets with at least one McDonald within the search radius, column 3 shows the average distance to the nearest McDonald in miles, column 4 is the average number of McDonalds in the search radius, and column 5 shows the average distance to all McDonalds in miles within the search radius. Within only quarter of a mile from the location of a tweet there are on average two McDonalds and the average distance to the nearest McDonalds is *only* 0.15 miles. Within the search radius of 1 mile from a location a tweet containing terms related to high calorie food there are on average three McDonalds and the distance to the nearest one is less than half a mile. To the opposite end of the spectrum, according to USDA's research on access to healthy food, an estimated total of 13.6 million people have low access to a supermarket or a large grocery store-that is, they live more than 1 or 10 miles from a supermarket or large grocery store. Of these 13.6 million people, 82.2% are in urban areas (USDA 2012). Within a search radius of 5 miles from the locations of 77 percent of the tweets related to this theme there are on average 17 McDonalds restaurants. The average distance to the nearest McDonalds is still less than a mile. Thus the results from the proximity analysis show a strong relation between the locations from where a Twitter user is tweeting about high calorie food and obesity and location of a McDonald.

Discussion and Conclusion

In this study, we focus on our test topic, obesity in the US, to address the following research questions: How can topic modeling be used to most effectively identify relevant public health topics on twitter? What are the common obesity related themes? What is the spatial pattern of the common obesity related themes? What are the important research challenges of using large conversational datasets from social networking sites?

The combined use of topic modeling and GIS in our study demonstrated its potential to extract themes from large datasets of conversational or textual data. After completing time consuming steps of geocoding in ArcGIS 10.0 we were able to successfully georeference 455,981 tweets. A rigorous text mining protocol followed by topic modeling identified 11 coherent topics from the best model. Based upon the most likely topic components or in other words bi-gram terms with higher frequency, these 11 topics were further grouped into four themes of: 1) childhood obesity and schools, 2) obesity prevention, 3) obesity and food habits, and 4) obesity and health. Needless to mention, these topics seems to be very common and important obesity-related issues in the US as well as in the world. In overall the density maps of geocoded tweets show higher density of tweets around bigger cities on the eastern and western coasts, and lower density of tweets in the central region of the country, states near the Gulf coast, Appalachian, and northernmost regions. This pattern of spatial distribution of tweets is consistent over all the obesity related search terms although the density of tweets varies significantly by specific search terms. This pattern indicates that people in large numbers are conversing with their friends and followers on Twitter about the different issues with obesity and it might be a quick and easy way to spread the awareness of a new prevention program or promoting health behavior to reduce the rates of obesity. Thus this methodology, in overall, proves a very effective way to automate the process of removing irrelevant information and to hone in on desired outcomes such as obesity-related themes mentioned above.

The topics (Number 2 and 20) with the theme of childhood obesity and school, effectively captured the sentiments around the revised agriculture appropriations bill passed by the United States' Congress essentially making it easier to count pizza sauce as a serving of vegetables in November 14th 2011. This is a controversial bill and raised serious concerns by health practitioners, nutritionists, teachers, fresh food advocates, and social workers. The topic model also suggested that researchers could understand how Twitter, a popular SNS, is used to promote both positive and negative health behaviors. For example, Topics 16, 21, 25, and 29 (theme: Obesity and food habits) contain high frequency terms that indicate that fast food businesses use Twitter as a means to promote access to eating high calorie, sweet, and junk food. In contrast Topics 7, 13, and 17 (theme: Obesity prevention) contain top words that could help individuals prevent obesity by eating healthy foods, dieting, physical activity, and joining weight loss programs.

Further, using the geographic location of the tweets opens up new avenues of research for health scientists to better monitor, understand, and survey health status in order to solve localized health problems. The mapping of tweets not only allows us to know '*what we are tweeting about obesity'* but also from '*from where we are tweeting about obesity'*. As expected the spatial pattern shows that individuals are referring more to obesity (both positive and negative sentiments) from the coasts and big cities and much less from the heartland of the US. This pattern is similar for most of the search terms and very prominent for search terms such as 'obesity', 'childhood AND obesity', 'McDonalds AND obesity', 'fructose', and 'obesity AND diabetes'. The geographic location of the tweets also allows unique spatial analysis by combining tweets with other datasets through GIS. We demonstrated this by conducting proximity analysis between the location of tweets in the 'Obesity and food habit' theme and the location of McDonalds restaurants. The results from the analysis show a strong correlation between the locations from where a Twitter user is tweeting about high calorie food and obesity (either positive or negative sentiment) and locations of McDonalds. When number of tweets in the theme, 'Obesity prevention', was summarized by state and compared to the number of on-going obesity related state policies the correlation was 0.66. This finding further points to the fact that Twitter users who are talking about ways to prevent the obesity epidemic in the US might be influenced by the different obesity related policies in progress in their respective states.

Finally we also discuss some challenges when we use large conversational datasets from SNS such as Twitter.com. First, the textual data from any SNS are noisy. This requires a careful and extensive cleaning process. While keyword based processing is able to eliminate much of the noise, more sophisticated natural language processing (NLP) tools could be used to enhance this step. For example, we did not consider parts of speech tagging, which could be used to remove irrelevant verbs and adverbs. Second, it is optional for Twitter users to enable the geographic location information (latitude and longitude) when they sign up for the account online or when using the Twitter application from their smartphones. Given that sharing geolocation data is not the default option, this results in a relatively small proportion (approximately 0.78%) of tweets being geocoded, compared to the total number of tweets collected. Even then, many tweets contained nonsensical place names "I stay on the moon" or ambiguous locations, such as only city names or state names. Such tweets cannot be considered for geocoding and therefore further lowers the number of tweets with geographic locations. Another potential problem is that the tweet locations may not be necessarily corresponding to user locations due to traveling. In our study we have found that the percentage of tweets with location information (both Twitter provided and inferred via manual geocoding) to all tweets is 2.5 to 3 percent and this value is consistent over all the obesity-related search terms. We would like to caution the readers that although the quality of the geographic locations of tweets provided by Twitter is accurate, the process of manual geocoding using city, state, time zone, and UTC-offset information, is coarse. For example, when a tweet is manually geocoded to the centroid of San Diego city, the original location is uncertain because we cannot precisely say where exactly in the city the tweet is located. Third, the development of topic models is obviously dependent on the textual content of the tweets. While our search terms focused on obesity related topics, it is possible that tweets containing these terms, but used in a different sense were also included. Given the unsupervised nature of the current study, there was no way to exclude such tweets. One approach to alleviating this problem is to manually build a collection of labeled (obesity related and non obesity related) tweets, which could be used to train a topic model. Such a manual approach has been described by various workers (Paul and Dredze 2011), and obviously limits the size of the training set that can be composed. Similar approaches could be used to restrict tweets to only those that are written in English, though language prediction models may not give reliable predictions based on 140 or fewer characters of text. Fourth, it is evident that while one can specify an arbitrary number of topics, they do not all make clear sense. Indeed, some topics may appear to be a random collection of terms, whereas other topics can contain a more cohesive collection of terms. Ultimately, the inspection and selection of topics from a given model is a manual, and thus subjective task. Fifth, while topic modeling and GIS allows us to know '*what we are tweeting about obesity'* and '*from where we are tweeting about obesity',* we do not know for sure "*who is tweeting about obesity*". In other words, there is no socioeconomic information of the Twitter users. We suspect that the demographics of Twitter users may affect the extent to which topics are discussed on Twitter. Future research is required to address these challenges for health researchers to better identify, understand and help solve health issues.

Acknowledgments

The authors gratefully acknowledge the assistance of Patrick Gallagher, a Master's student in the Department of Geography, University of Connecticut. He was responsible for geocoding the tweets. The authors also thank the editors and the anonymous reviewers for helpful comments and suggestions. Responsibility for the opinions expressed herein is solely that of the authors.

References

Grun B, Hornik K. topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software. 2011; 40(13):1–30.

Blei, D.; Lafferty, J. Topic Models. In: Srivastava, A.; Sahami, M., editors. Text mining, classification, clustering and applications. Chapman & Hall; 2009.

Blei DM. Probabilistic Topic Models. Communications of the ACM. 2012; 55(4):77–84.

Blei DM, Andrew Y.Ng. Jordan Michael I. Latent Dirichlet Allocation. Journal of Machine learning Research. 2003; 3:993–1022.

Boyd DM, Ellison NB. Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication. 2008; 13:210–230.

Butts, CT.; Acton, RM. Spatial modeling of social networks. In: Nyergers, TL.; Couclelis, H.; McMaster, R., editors. The Sage Handbook of of GIS and Society. SAGE; Thousand Oaks, CA: 2011. p. 222-250.

CDC. Center for Disease Control and Prevention. Retrieved May 31th 2012, from http://www.cdc.gov/obesity/

Chew CM, Eysenbach G. Pandemics in the Age of Twitter: Content Analysis of "tweets" During the 2009 H1N1 Outbreak. Public Library of Science. 2010; 5(11):e 14118.

Culotta, A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages; Paper read at KDD Workshop on Social Media Analytics; 2010;

ESRI. Environmental Sciences Research Institute. Retrieved June 11th 2012, from http://www.esri.com/

Feinerer, I. R package version 0.5-7.1. 2012. tm: Text Mining Package.

Feinerer I, Hornik K, Meyer D. Text Mining Infrastructure in R. Journal of Statistical Software. 2008; 25(5)

Ghosh D, Sterns A, Drew B, Hamera E. Geospatial Analysis of Psychiatric Mental Health Advanced Practice Nurses in the United States. Psychiatric Services. 2011; 62:1506–1509. [PubMed: 22193800]

Haythornthwaite C. Social Networks and Internet Connectivity Effects. Information, Communication, & Society. 2005; 101:5228–5235.

InfoUSA. Retrieved June 11th 2012, from http://home.infousa.com/

IWS. Internet World Stats. Retrieved December 15th 2012, from http://www.internetworldstats.com/

Paul MJ, Dredze M. You are what you tweet: Analyzing Twitter for Public Health. Association for the Advancement of Artificial Intelligence. 2011

Porter MF. An algorithm for suffix stripping. Program. 1980; 14(3):130–137.

Prier, KW.; Smith, MS.; Giraud-Carrier, C.; Hanson, CL. Identifying Health-Related Topics on Twitter: An Exploration of Tobacco-Related Tweets as a Test Topic; Paper read at International Conference on Social Computing, Behavioral Modeling, and Prediction; Maryland, USA. 2011;

Quantcast. Retrieved June 11th 2012, from http://www.quantcast.com/twitter.com

Salathe M, Khandelwal S. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. PLoS Computational Biology. 2011; 7(10)

Scanfield D, Scanfield V, Larson E. Dissemination of Health Information through Social Networks: Twitter and Antibiotics. American Journal of Infection Control. 2006; 38:182–188.

Thorpe, K. The Future Cost of Obesity: National and State Estimates of the Impact of Obesity on Direct Health Care Expenses. United Health Foundation; 2009.

Twitter. Twitter API documentation. Retrieved October 10th 2011, from https://dev.twitter.com/docs

Twitter. What is Twitter?. Retrieved October 10th 2011, from https://business.twitter.com/en/basics/what-is-twitter/

USDA. United States Department of Agriculture. Retrieved June 11th 2012, from http://www.ers.usda.gov/data-products/food-desert-locator/go-to-the-locator.aspx]

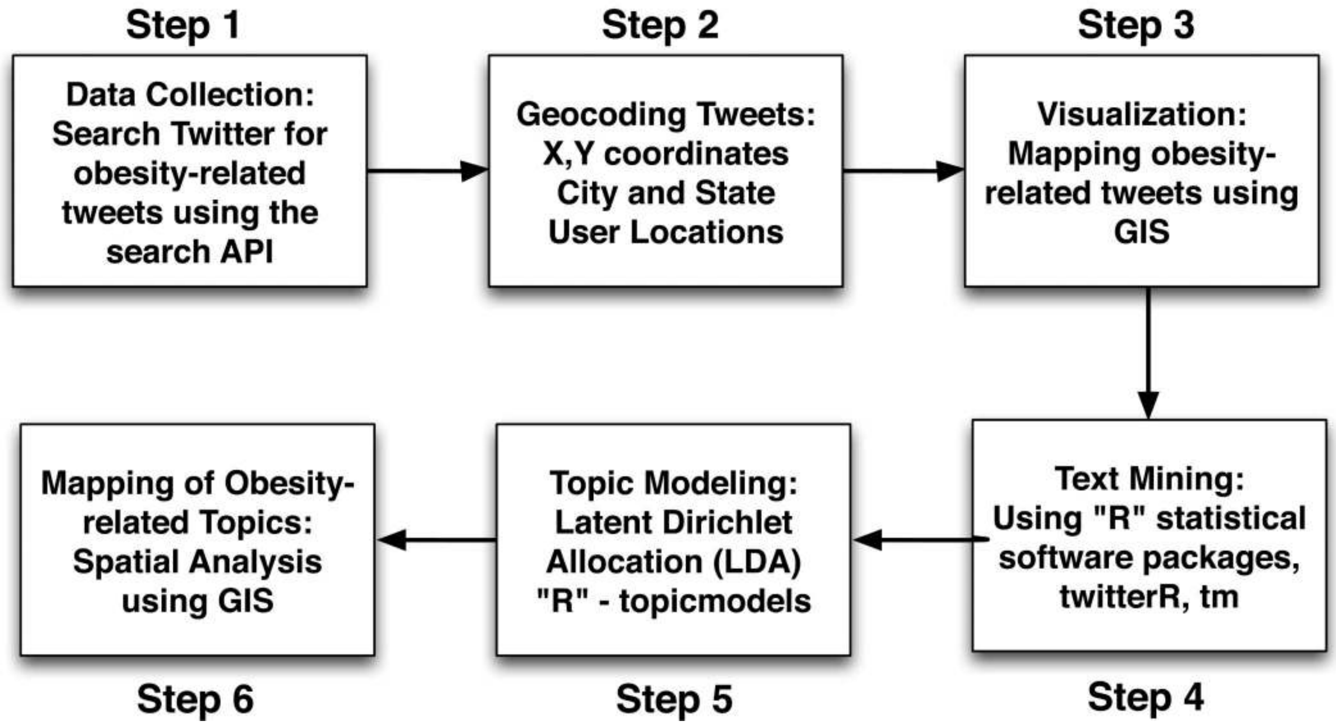
Step 1
Data Collection: Search Twitter for obesity-related tweets using the search API
Step 2
Geocoding Tweets: X,Y coordinates City and State User Locations
Step 3
Visualization: Mapping obesity-related tweets using GIS
Text Mining: Using "R" statistical software packages, twitterR, tm
Step 4
Topic Modeling: Latent Dirichlet Allocation (LDA) "R" - topicmodels
Step 5
Mapping of Obesity-related Topics: Spatial Analysis using GIS
Step 6

Figure 1. Major steps in the workflow

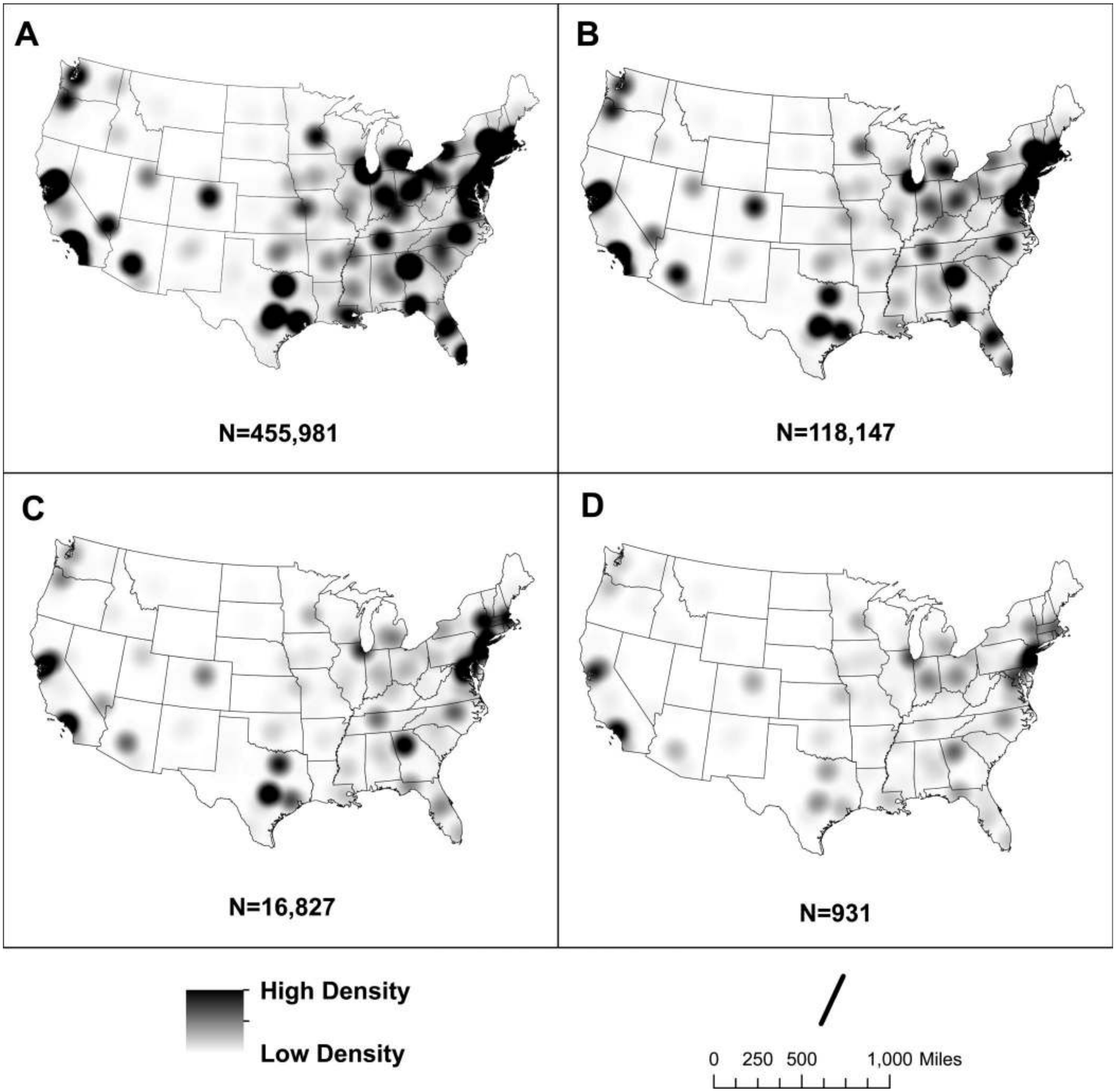

Figure 2. Visualization of geocoded tweets based on obesity-related search terms A: All search terms, B: 'Obesity', C: 'Childhood AND Obesity', D: 'McDonalds AND Obesity'

Note: A: All search terms with 455,981 tweets, B: 'Obesity' with 118,147 tweets, C: 'Childhood AND Obesity' with 16,826 tweets, and D: 'McDonalds AND Obesity' with 931 tweets.

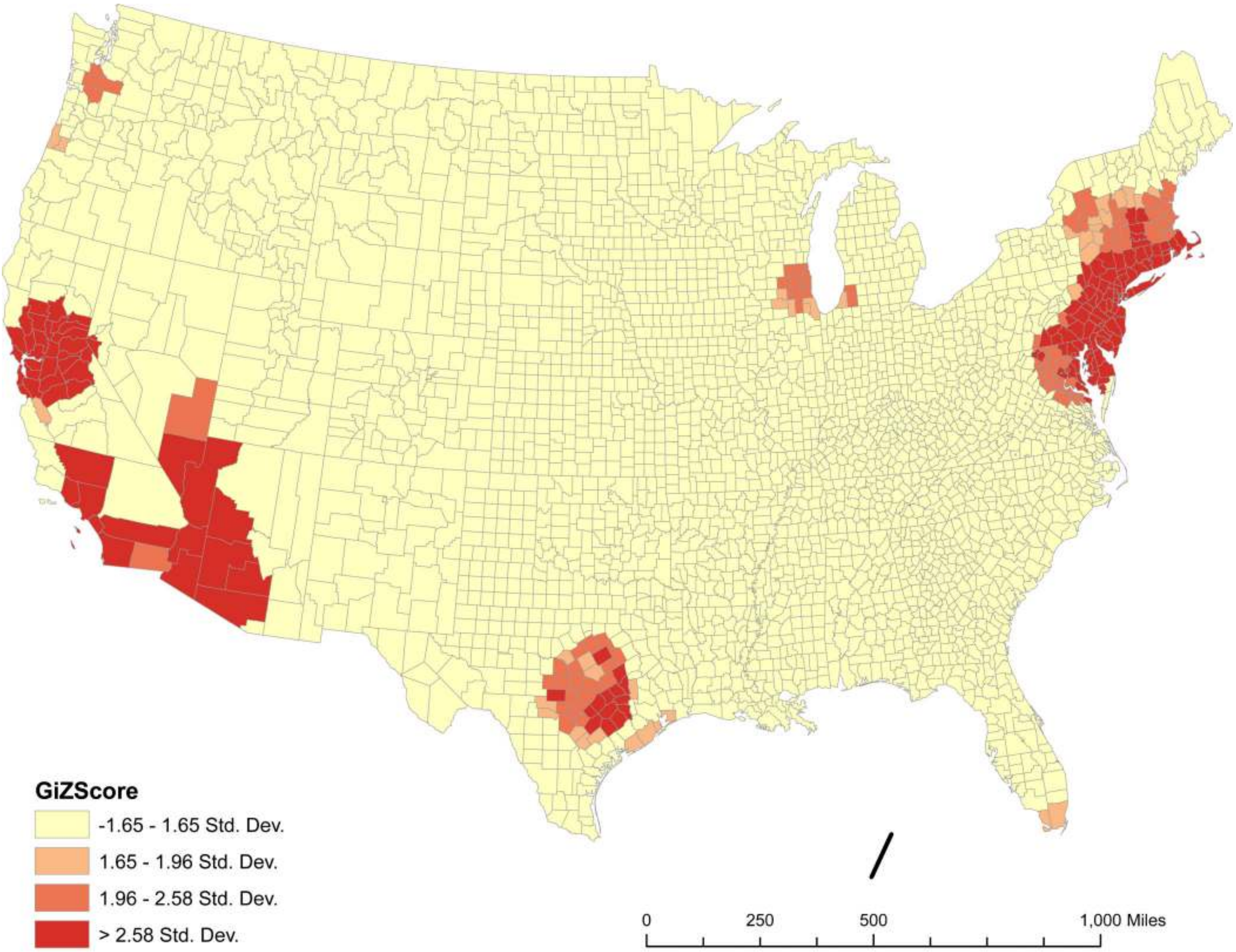

Figure 3. Hot-Spot analysis of tweets associated with topics 2 and 20 and theme 'Childhood obesity and Schools'

Note: There are approximately 21,842 tweets (4.79 percent) related to this theme.

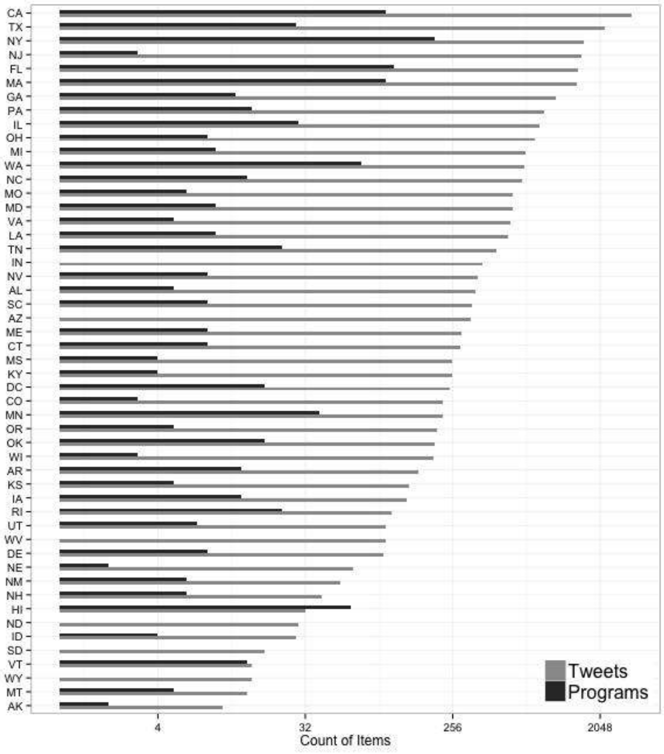

Figure 4. Bar graph showing the comparison of tweets related to obesity prevention theme and number of policies related to obesity, nutrition, and physical activity by state

Note: Here count of items are number of tweets and obesity-related programs by states. The scale of x-axis is logarithmic.

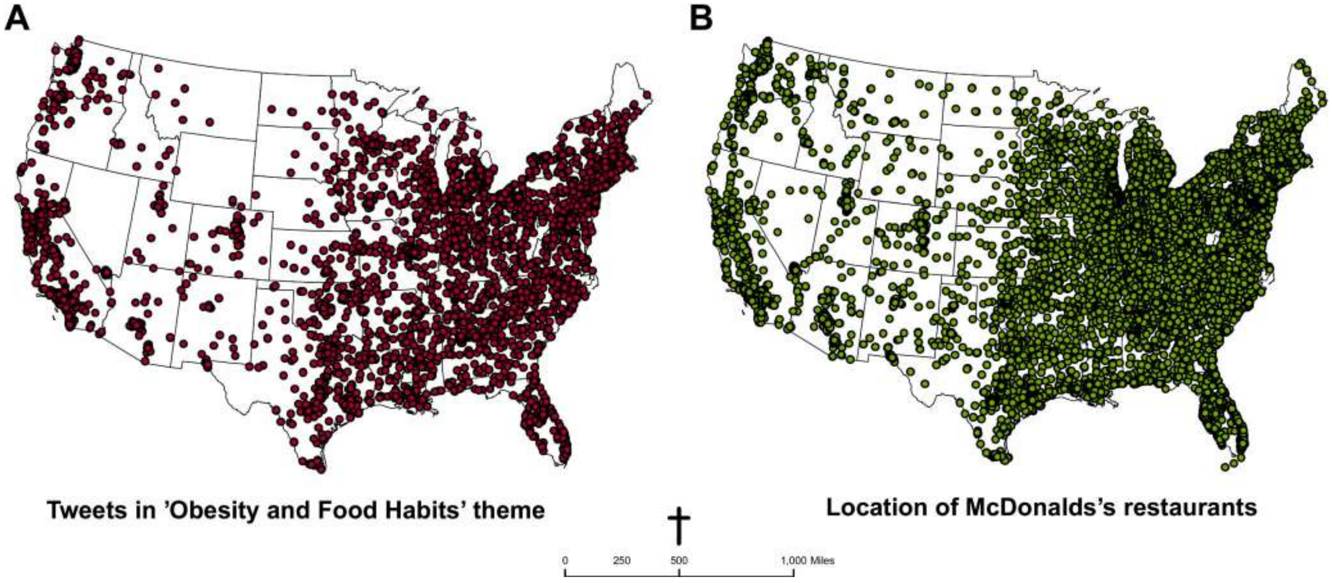
A
B
Tweets in 'Obesity and Food Habits' theme
Location of McDonalds's restaurants
0
250
500
1,000 Miles

Figure 5. Location of Tweets in 'Obesity and food habits' theme (A) and Location of McDonalds restaurants (B)

Note: The number of tweets in this theme is approximately 44,230. The number of McDonald restaurants is 14,007 (data source: InfoUSA).

Table 1

Structure of Master Dataset and User Dataset

Master Dataset							
tweet_ID	User_from_ID*	User_to_ID	Geog_X	Geog_Y	Pname	Time	Text
1188442900205400 00	38460 8898	176729	42.9656	−85.6738	Grand Rapids, MI	Wed, 28 Oct 2011 00:27:41 +0000	“@mcdonalds your not helping obesity in america by putting monopoly pieces on large fries”
1192254758549010 00	40203 8982		33.6875	−84.0009		Thu,29 Nov 2011 01:42:23 +0000	“stop childhood obesity”
1196329944441200 00	16618 951					Fri, 30 Oct 2011 04:41:43 +0000	“my friend s 7 year old niece was just rushed to the hospital amp diagnosed with type 2 diabetes wake up call childhood obesity is not cute”

User Dataset					
ID*	Screen_Name	Name	Location	UTC_offset	Time_Zone
359642511			Atlanta, GA	−18000	Eastern Time (US & Canada)
165983173	stacinthesun	Staci Reiter	Monterey, CA	−28800	Eastern Time (US & Canada)
25361231	nicolesaysroar	Nicole Norris	Baltimore, MD	−18000	Eastern Time (US & Canada)

*Notes: The two datasets were joined based on the User_from_ID (Master dataset) and the ID (User dataset) columns.

Table 2
Number of geocoded and non-geocoded tweets by obesity related search terms

Search Terms	Geocoded Count	Non-Geocoded Count
childhood AND obesity	16,827	57,508
eat AND right	79,211	733,590
farm AND policy	9	15
food AND deserts	1,949	5,531
high AND calorie	2,520	12,589
fructose	9,473	28,230
obesity AND corn	246	681
soft AND drink	7,294	14,234
weight AND gain	92,322	476,745
McDonalds AND obesity	931	2,117
McDonalds AND overweight	985	2,765
obesity	118,147	498,396
overweight	63,869	304,400
obesity AND diabetes	4,465	19,812
overweight AND diabetes	924	3210
TOTAL	**455,981**	**2,159,823**

Table 3
Obesity-related themes from the LDA model

Topic	Most Likely Topic Components (bi-grams)	Theme	Percentage Tweets
2, 20	Childhood obesity, school lunches, pizza vegetable, pizza tomato, diet soda, vending machine, parent education, tomato paste, obese kid, fight childhood, help childhood, physical activity	**Childhood obesity and schools**	**4.79**
7, 13, 17	Calorie diet, calorie burn, obesity program, eating habit, eat rule, obesity awareness, weight loss, body mass, program prevent, eat salad, food stamp, achieve healthy, fight obesity, health news, avoid holiday, burn calorie, curb weight, food industry, avoid weight, food desert, health center	**Obesity prevention**	**6.96**
16, 21, 25, 29	Corn syrup, fructose corn, corn sugar, coca cola, food addict, get fat, lay bed, thanksgiving dinner, video game, holiday weight, McDonald, fries, junk food, ice cream, fat junk, food effect, eat chocolate, food epidemic, soft drink, grocery store, food night, super bowl, eat candy, Chinese food	**Obesity and food habits**	**9.68**

Notes: Column 1 shows the topic numbers, column 2 shows the most likely topic components or bi-gram terms from each topic, column 3 describes the topic themes, and the last column is the percent of tweets that used any of the bi-gram words within each theme.

Table 4

Proximity Analysis between Location of Tweets related to 'Obesity and food habit' theme and location of McDonalds restaurants

Search Radius in miles	Percentage of Tweets with at least one McDonald	Average Distance to the Nearest McDonald in miles	Average Number of McDonalds	Average Distance to all McDonalds in miles
0.25	17.84 (7,890)	0.15	2	0.17
0.5	36.00 (15,923)	0.26	3	0.32
1.0	55.93 (24,736)	0.44	3	0.51
1.5	68.92 (30,483)	0.59	4	0.82
2.0	73.53 (32,521)	0.66	6	1.12
2.5	75.43 (33,362)	0.70	7	1.41
3.0	76.05 (33,636)	0.72	9	1.70
3.5	76.31 (33,754)	0.73	11	2.00
4.0	76.64 (33,899)	0.74	13	2.25
4.5	76.83 (33,982)	0.75	15	2.54
5.0	77.00 (34061)	0.76	17	2.81

Note: The first column lists the search radius; column 2 is the percentage of tweets with at least one McDonald within the search radius in column 1, column 3 shows the average distance to the nearest McDonald in miles, column 4 is the number of McDonalds in the search radius, and column 5 shows the average distance to all McDonalds in miles within the search radius.