Original Paper

Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity

Quynh C Nguyen[1], PhD; Dapeng Li[2], MS, PhD; Hsien-Wen Meng[1], MS; Suraj Kath[3], MS; Elaine Nsoesie[4], MS, PhD; Feifei Li[3], PhD; Ming Wen[5], MA, MS, PhD

[1]Department of Health, Kinesiology, and Recreation, University of Utah College of Health, Salt Lake City, UT, United States

[2]Department of Geography, University of Utah, Salt Lake City, UT, United States

[3]School of Computing, University of Utah, Salt Lake City, UT, United States

[4]Department of Global Health, University of Washington, Seattle, WA, United States

[5]Department of Sociology, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Quynh C Nguyen, PhD
Department of Health, Kinesiology, and Recreation
University of Utah College of Health
1901 E. South Campus Drive
Annex Room 2124
Salt Lake City, UT,
United States
Phone: 1 801 585 5134
Fax: 1 801 585 3646
Email: quynh.ctn@gmail.com

Abstract

Background: Studies suggest that where people live, play, and work can influence health and well-being. However, the dearth of neighborhood data, especially data that is timely and consistent across geographies, hinders understanding of the effects of neighborhoods on health. Social media data represents a possible new data resource for neighborhood research.

Objective: The aim of this study was to build, from geotagged Twitter data, a national neighborhood database with area-level indicators of well-being and health behaviors.

Methods: We utilized Twitter's streaming application programming interface to continuously collect a random 1% subset of publicly available geolocated tweets for 1 year (April 2015 to March 2016). We collected 80 million geotagged tweets from 603,363 unique Twitter users across the contiguous United States. We validated our machine learning algorithms for constructing indicators of happiness, food, and physical activity by comparing predicted values to those generated by human labelers. Geotagged tweets were spatially mapped to the 2010 census tract and zip code areas they fall within, which enabled further assessment of the associations between Twitter-derived neighborhood variables and neighborhood demographic, economic, business, and health characteristics.

Results: Machine labeled and manually labeled tweets had a high level of accuracy: 78% for happiness, 83% for food, and 85% for physical activity for dichotomized labels with the *F* scores 0.54, 0.86, and 0.90, respectively. About 20% of tweets were classified as happy. Relatively few terms (less than 25) were necessary to characterize the majority of tweets on food and physical activity. Data from over 70,000 census tracts from the United States suggest that census tract factors like percentage African American and economic disadvantage were associated with lower census tract happiness. Urbanicity was related to higher frequency of fast food tweets. Greater numbers of fast food restaurants predicted higher frequency of fast food mentions. Surprisingly, fitness centers and nature parks were only modestly associated with higher frequency of physical activity tweets. Greater state-level happiness, positivity toward physical activity, and positivity toward healthy foods, assessed via tweets, were associated with lower all-cause mortality and prevalence of chronic conditions such as obesity and diabetes and lower physical inactivity and smoking, controlling for state median income, median age, and percentage white non-Hispanic.

Conclusions: Machine learning algorithms can be built with relatively high accuracy to characterize sentiment, food, and physical activity mentions on social media. Such data can be utilized to construct neighborhood indicators consistently and cost effectively. Access to neighborhood data, in turn, can be leveraged to better understand neighborhood effects and address social

XSL•FO
RenderX

determinants of health. We found that neighborhoods with social and economic disadvantage, high urbanicity, and more fast food restaurants may exhibit lower happiness and fewer healthy behaviors.

(JMIR Public Health Surveill 2016;2(2):e158) doi: 10.2196/publichealth.5869

KEYWORDS

social media; Twitter messaging; health behavior; happiness; food; physical activity

Introduction

There is an increasing recognition that health is determined by a myriad of factors, including where you live, play, and work [1-5]. Poor access to healthy food [6-10], abundance of fast food chains [11], lack of recreational facilities [12,13], and higher crime rates [7,14] have been shown to predict higher obesity rates. Environmental exposure to toxins, noise, and violence can be detrimental to health [15,16]. Conversely, neighborhood resources such as playgrounds for children, grocery stores, and gyms can be beneficial to health [17]. Adverse neighborhood conditions converge in poor, minority neighborhoods [18-21], thereby increasing health disparities.

Social environments can offer social and emotional support that buffers stressful life events [22]. Johns and colleagues found that neighborhoods with higher social cohesion had lower posttraumatic stress disorder [23]. Higher community happiness levels are linked with lower obesity, hypertension, and suicide rates as well as increased life expectancy [24-29]. Evidence also suggests that emotional states such as happiness, optimism, depression, or suicidality can spread through social networks [30-33]. The social environment can offer opportunities for social control in regulating unhealthy behaviors and facilitating the social learning of healthy behaviors but can also promote risky behaviors. Health behaviors, such as food consumption, health screening, smoking, alcohol consumption, drug use, and sleep have also been observed to spread through social networks [34-37].

The extreme scarcity of neighborhood data greatly limits research on neighborhood effects. Some places [38,39] have extensive neighborhood data collected on them, but they are the anomaly rather than the rule, and it is difficult to make comparisons across geographies because available measures vary greatly across them. Neighborhood data collection is expensive and time consuming and only available for certain time periods [40]. Widespread usage of the Internet and open recording of many transactions (eg, Yelp reviews, Foursquare check-ins, and reporting of personal opinions and behaviors through social media) has led to the availability of massive amounts of data that enable understanding of previously hidden local area interactions. Researchers are increasingly utilizing social media and user-generated data to track health behaviors and perform health surveillance (eg, for outbreak detection) [41-45]. Others have used social media to track sleep issues [46], personal health status disclosed by Twitter users [47,48], and patient-perceived quality of care [49].

In this study, we explored the utility of building a national neighborhood database from geotagged Twitter data to characterize well-being and health behaviors. We validated our machine learning algorithm for constructing indicators of happiness, food, and physical activity by comparing machine-generated values to values generated by human labelers. In addition, we explored associations between Twitter-derived neighborhood variables and neighborhood demographic and economic characteristics. This project makes significant, relevant contributions to the field because neighborhood environments are increasingly linked to an array of important health outcomes and this project addresses the limits to research resulting from the lack of neighborhood data by providing new, cost-efficient data resources and methods for characterizing neighborhoods. To our knowledge, our study was the first to attempt to create a national neighborhood database from Twitter data, with indicators constructed for public health researchers. The only other type of neighborhood data that is consistently available for local areas is census data on the compositional characteristics of neighborhoods. Twitter is uniquely suited to characterize the social environment, including prevalent sentiment and health behaviors.

Methods

Social Media Data Collection

From February 2015 to March 2016, we utilized Twitter's streaming application programming interface (API) to continuously collect a random 1% sample of publicly available tweets with latitude and longitude coordinates. Given that neighborhood researchers differ in their use and interest in data at the census tract and zip code level, we constructed neighborhood indicators at both levels thereby increasing the flexibility of our dataset to address the potential data needs of other researchers. In total, we collected 79,848,992 million geotagged tweets from 603,363 unique Twitter users in the contiguous United States (including District of Columbia). The median number of tweets per user was 4. Job postings (identified through hashtags #hiring, #jobs, and #job) were removed from the final analytic sample of tweets because these were pervasive and not central to the neighborhood variables we constructed.

Spatial Join and Neighborhood Definition

Each geotagged tweet was assigned a corresponding census tract and zip code it falls within, based on the latitude and longitude coordinates of where the tweet was sent. This spatial join procedure was implemented in Python (version 2.7.12; Python Software Foundation), a popular programming language for spatial data processing [50]. Specifically, Python libraries were used to read shapefile format vector data (PyShp 1.1.4), build an R-tree index on the polygon data (Rtree 0.8.2), and perform a spatial join operation (Shapely 1.5.12 and Fiona 1.6.1). The R-Tree was used to build a spatial index [51] on the national census tract and zip code polygon data to speed

XSL•FO
RenderX

computation. Tweets that were not assigned a census tract or zip code location included those with destinations bordering the United States (ie, Mexico and Canada). We linked 99.8% of tweets with geocordinates to their respective 2010 census tract and zip code locations. The term *neighborhood* used in this paper refers to both zip codes and census tracts. We mapped tweets to these two geographic boundaries because they are among the most popular neighborhood definitions utilized by public health researchers [52-54].

Processing Tweets

Duplicate tweets (ie, tweets with the same tweet ID, <1%) were removed computationally. Although Twitter's API collects a random subset of 1% of publicly available tweets, users (especially spam accounts) who tweet often have potentially greater influence on variable values we construct. We examined outliers in our datasets (defined as the users whose tweets accounted for more than 1% of tweets in our dataset) and eliminated automated accounts and accounts for which the majority of tweets were advertisements. Processing and statistical analysis tasks were performed with Stata MP13 (StataCorp LP).

Construction of Neighborhood Variables From Twitter Data

From geotagged tweets, we derived variables that characterize happiness, food, and physical activity. Each tweet was divided into tokens using the Stanford tokenizer [55]. For processing of English text, tokens roughly correspond to words. We then built various algorithms utilizing tokens to create variables that characterize happiness and make references to food and physical activity. Below we describe in more detail our algorithms.

Sentiment Analysis

To conduct sentiment analysis, we utilized the Machine Learning for Language Toolkit (MALLET; AK McCallum, 2002), a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. We leveraged the Maximum Entropy text classifier in MALLET to classify tweets as happy and not happy [56]. In order to train our classifier, we obtained training sets from the following resources: Sentiment140 [57], Sanders Analytics [58], and Kaggle [59]. We trained our classifier to differentiate between happy and not happy sentiments. We then ran our classifier on our national Twitter data to compute a happy score (range 0-1) for each tweet, where higher happiness scores indicate more positive sentiment. MALLET estimates predicted probabilities that a tweet is happy based upon word-level features. The classifier uses search-based optimization to assign weights that maximize the likelihood of the training data. However, unlike Naïve Bayes, the Maximum Entropy classifier does not assume conditional independence among features.

To calibrate the generated happiness scores with human generated labels, two raters manually read a random subset of 1200 tweets and assigned a value of 1 to happy tweets and 0 to not happy tweets. The initial interrater reliability was 92%, and discordant values were reviewed until a 100% agreement between raters was reached. To decide on a cut point for MALLET scores at which we would classify tweets as happy, we computed accuracy levels at different cut points of MALLET scores (Multimedia Appendix 1). Increasing the MALLET score improves the accuracy against human annotations but also reduces the calculated prevalence of tweets deemed as happy. A MALLET score of 0.80 achieves the highest level of accuracy while still maintaining a prevalence of happy tweets of 19% (which approximates the prevalence obtained by human annotations). Area under the receiver operating characteristic curve is approximately 0.7 for all MALLET cut points between 60 and 85.

Food Analysis

We compiled a list of over 1430 popular food words from the US Department of Agriculture's National Nutrient Database [60]. Each food item was associated with a measure of caloric density, operationalized as calories per 100 grams. Fruits, vegetables, nuts, and lean proteins (ie, fish, chicken, and turkey) were labeled as healthy foods (340 food terms in total). Fried foods were not considered healthy foods. Our food list also contained popular national fast food restaurants such as McDonald's and Kentucky Fried Chicken (captured via 154 food terms including popular variations of restaurant names) to enable quantification of fast food references. From April 2015 to March 2016, we collected and processed 4,041,521 geotagged food tweets. In the food dataset, the median number of tweets per user was 12 tweets.

To analyze food culture, each tweet was examined for words or phrases matching those on our list. Each food item on our list was described by one or two words. Our text-matching algorithm first searched over a tweet for matches to two-word foods (eg, orange chicken). It then searched over the remaining words for matches to one-word food terms (eg, taco). We computed caloric density by summing up all the foods mentioned in the tweet. We also created a count of healthy food references and fast food restaurant references for each tweet. Moreover, we leveraged our sentiment analysis to assess sentiment toward food. Specifically, we tracked sentiment around healthy foods and fast food. These variables (any food references, healthy food references, fast food references, caloric density, and sentiment toward healthy foods and fast food) were then aggregated and summarized at the census tract and zip code level to create neighborhood indicators of food culture.

Physical Activity Analysis

We created a list of physical activities using published lists of physical activity terms gathered from physical activity questionnaires, compendia of physical activities, and popularly available fitness programs [61,62]. Our physical activity list had 376 different activities that incorporate gym-related exercise (eg, treadmill, weight lifting), sports (eg, baseball), recreation (eg, hiking, scuba diving) and household chores (eg, gardening). We excluded popular phrases that generally do not relate to physical activity such as "walk away" and "running late." Using metabolic equivalents associated with physical activities, we quantified the exercise intensity of each physical activity mention, scaled for a duration of 30 minutes and for a 155-pound individual [63], which approximates the weight of an average American adult [64,65].

XSL•FO
RenderX

Upon piloting our algorithm, we identified commonly used phrases or pop culture references that do not involve physical activity (eg, walking dead) which were manually coded and excluded. Moreover, in order to help reduce the possibility that the tweet was about watching rather than actually participating in the physical activity, we excluded the tweet if it contained any of the following terms: “watch,” “watching,” “watches,” “watched,” “attend,” “attending,” “attends,” and “attended.” In reviewing preliminary labeled physical activity data, we found that most tweets (over 90%) pertaining to team sports (eg, baseball, basketball, football, soccer) were about watching games rather than participating in them. Thus, for team sports, we required that the tweet include the words “play,” “playing,” or “played.”

Our algorithm created the following physical activity variables for each tweet: any physical activity mention, exercise intensity, and sentiment around physical activity. From April 2015 to March 2016, we collected 1,473,976 geotagged physical activity tweets. In the physical activity dataset, the median number of tweets per user was 5 tweets.

Quality Control Activities

A total of 5000 tweets have been manually labeled by two of the authors for quality control activities on food and physical activity. The authors manually labeled whether each tweet was food-related (2000), non–food-related (500), physical activity-related (2000), or non–physical activity-related (500). Excellent interrater reliability was achieved with greater than 90% agreement in all categories, and differences were discussed and resolved.

Among tweets our algorithm had labeled as food-related, 83% were labeled accurately when compared to labels generated by manual categorization. Among tweets our algorithm had labeled as non–food-related, 81% were labeled accurately (ie, both algorithm and human categorizers labeled the tweet as non–food-related). Overall, accuracy for food tweets was 83% and the *F* score was 0.86. It should be noted our algorithm could label a food-related tweet as non–food-related if the food reference was not in our food dictionary. Food items that are often associated with non-related food meaning, such as “perch,” have been excluded from our food dictionary. For tweets that had been mislabeled as food-related, commons reasons included food term used as a metaphor, in a pun, or for food advertisement.

Among tweets our algorithm had labeled as physical activity-related, 82% of them were labeled accurately when compared to labels generated by human categorizers. An accuracy of 97% was found among tweets labeled as non-physical activity-related by our algorithm. The *F* score was 0.90 and the overall accuracy was 85% for physical activity tweets. Typical errors in classification of physical activity tweets included the use of an idiom (eg, running late) or the tweet was about watching sports games rather than playing sports.

Additionally, we evaluated our algorithm on its ability to identify relevant food and physical activity terms within tweets. To do this, we examined a random subset of tweets that the algorithm had identified as positive for food (n=200) and physical activity (n=200). Here we focused on the accuracy of our algorithm to conduct string detection. We manually read the tweets to verify that manual annotations agreed with the terms detected. For food tweets, 87% of manual annotations matched all detected terms from the algorithm. Errors for nondetection of terms occurred when the tweet included a hashtag that had multiple food terms without spacing (eg, #chocolatebrownie) or when there were misspellings (eg, sandwhich) or when the food was not on the food list. String detection for physical activity-related terms was more accurate with 98% of manual annotations matching detected terms from the algorithm. Errors included the omission of certain terms from the dictionary (eg, cycling) and use of hashtags without spacing of terms (#runrunrun).

We further evaluated our sentiment analysis activities through Amazon Mechanical Turk (Mturk; Amazon.com Inc, Seattle, WA, USA), an online crowdsourcing marketplace [66]. We randomly selected 500 tweets with 50% labeled as happy and 50% as not happy by our algorithm. Then, we created 20 online surveys through random sorting, with each survey consisting of 25 tweets. We asked participants to rate the sentiment of each tweet. All 20 surveys were live on April 1, 2015. Each online survey closed itself when 15 responses had been reached; the last survey closed on April 5, 2015. For each completed survey, 25 cents ($0.25) was deposited into the participant’s Mturk account. A total of 32 participants completed 300 surveys (ie, 15 responses per survey, 20 surveys). Some participants completed multiple surveys rather than just one. Each tweet was then assigned a label of either happy or not happy based on the modal response from Mturkers (participants from Amazon Mturk). We found an accuracy of 69% for happy tweets and 80% for nonhappy tweets when compared to responses from Mturkers. The overall accuracy for sentiment was 78%, with an *F* score of 0.54.

We additionally compared performance of MALLET with two other sentiment analysis techniques: a popular bag-of-words technique involving the use of a 10,000 word list [67] and Sentiment140, a machine-learning classifier [68]. Among the 500 control tweets from our LabMT experiment, the bag-of-words algorithm had an accuracy of 73% (*F* score 0.55) and Sentiment140 was had an accuracy of 77% (*F* score 0.47).

Other Publicly Available Neighborhood Data

To examine how Twitter-derived neighborhood variables relate to more traditional neighborhood variables, we merged our social media dataset with the 2010 Census and 2014 American Community Survey data which comprised the following demographic, household, and economic characteristics: household size, median family income and percent of the following: 65 years and older age group, 10-24 years, male, African American, white, Hispanic, households with relatives (other than spouse and children), households with unmarried partner, single female-headed households, householder living alone, owner-occupied housing, college graduates, unemployed, less than a high school degree and families living in poverty. A census tract was urban if the geographic centroid of the tract was in an area with more than 2500 people; all other tracts are rural. A zip code was defined as urban if the majority (75% or

XSL•FO
RenderX

more) of its land area was characterized as urban (ie, containing at least 2500 people).

Data on business types at the zip code level were obtained from the 2013 US Census Bureau zip code business patterns accessed via American FactFinder [69]. The following North American Industry Classification System (NAICS) codes were utilized to categorize businesses: 722410 (drinking places [alcoholic beverages]; these places are also known as bars, taverns, night clubs and primarily serve alcohol and may have limited food services) and 722511 (full-service restaurants; these include, for instance, diners and steakhouses). Fast food was defined by the following NAICS codes: 722513 (limited-service restaurants; these include carryout restaurants, drive-in restaurants, and other fast food restaurants) and 722515 (snack and nonalcoholic beverage bars). We also tracked supermarkets and grocery stores (NAICS code 445110) and convenience stores (NAICS code 445120). To examine associations between Twitter physical activity mentions and presence of recreational facilities, we retrieved business data for the following types of establishments: fitness and recreational sports centers (NAICS code 713940), nature parks (NAICS code 712190), zoos and botanical gardens (NAICS code 712130), golf courses and country clubs (NAICS code 713910), skiing facilities (NAICS code 713920), and bowling centers (NAICS code 713950).

We obtained state-level health outcome data including age-adjusted death rates due to all-causes and homicides from 2013 National Vital Statistics Reports. Data in this report was based on information from all resident death certificates filed in the 50 states and the District of Columbia. Death certificates are generally completed by funeral directors, attending physicians, medical examiners, and coroners. Age-adjusted death rates expressed per 100,000 population were based on the 2000 US standard population. Causes of death statistics were classified by the International Classification of Diseases, Tenth Revision, and based on the underlying cause of death.

We obtained age-adjusted prevalences of health risk behaviors and chronic conditions of US adult residents for the 50 states from the 2013 Behavioral Risk Factor Surveillance System (BRFSS), the nation's premier system of health-related telephone surveys. The questionnaires were created by BRFSS state coordinators and Centers for Disease Control and Prevention staff. BRFSS data includes self-reported physical activity, self-rated health, body mass index (BMI, kg/m^2), and medical diagnoses of chronic conditions aggregated to the state level. Data from a national health survey suggests that BMI estimates derived from self-reported height and weight were lower than those are derived from measured height and weight, although BMI differences were generally less than 1.0 kg/m^2 across sex and age groups [70]. State-level BRFSS data is publicly available. Smaller area aggregations can require data use agreements. In addition to state-level BRFSS data, we also utilized restricted-access zip-code–level data from the 2009-2014 Utah BRFSS survey to examine zip-code–level health outcomes [71,72].

Regression Analyses

We implemented adjusted linear regression models to examine associations between area-level Twitter characteristics and other area-level characteristics (demographics, business characteristics, and health outcomes). To facilitate interpretation of findings for different variables, we standardized all variables to have a mean of zero and standard deviation of one. We investigated spatial autocorrelation and found that Moran's I was highest for census tract Twitter happiness (0.12) and less than 0.04 for other Twitter tract and zip code summaries. To account for spatial autocorrelation of area-level values in linear regression analyses, we adjusted standard errors for clustering of census tract and zip code values within a county. Statistical analyses were implemented with Stata MP13 (StataCorp LP) and ArcGIS Desktop version 10.1-10.3 (Esri).

Results

Table 1 displays descriptive statistics. Approximately 20% of tweets were happy. About 5.1% of tweets were about food and 1.8% were about physical activity. The mean and median caloric density of food references were 239 and 209 calories per 100 grams, respectively. Tweets about healthy food were happier than tweets about fast food (28.3% vs 14.5%; $P<.001$). The mean and median exercise intensity of physical activity mentions (assuming 30 minutes for a 155-pound person) were 199 and 130 calories, respectively.

Figure 1 presents the spatial distribution of happy tweets by census tract, highlighting variation across the United States. Multimedia Appendix 2 presents the spatial distribution of happy tweets by zip code. The proportion of happy tweets was highest in the following states: Montana, Tennessee, Utah, New Hampshire, Arkansas, Maine, Colorado, and New York (Multimedia Appendix 3). By contrast, the proportions of happy tweets were lowest for the following states: Louisiana, North Dakota, Oregon, Maryland, Texas, Delaware, West Virginia, and Ohio (Multimedia Appendix 3).

Table 2 presents the results of adjusted linear regression analyses examining the associations between population characteristics and Twitter-derived characteristics at the census tract level (percent of tweets that were happy, percent of tweets about healthy food, percent of tweets about fast food, and percent of tweets about physical activity). Census tract characteristics like percent African American (beta coefficient, B=−.11), greater household size (B=−.18), and economic disadvantage (B=−.19) were related to lower tract happiness. Economic disadvantage was negatively related to healthy food tweets (B=−.09), fast food tweets (B=−.09), and physical activity tweets (B=−.03). Urbanicity was strongly related to higher frequency of fast food tweets (B=.29). Greater household size was related to both lower healthy food tweets (B=−.11) and fast food tweets (B=−.07).

XSL•FO
RenderX

Figure 1. National distribution of happy tweets, by census tract. Geotagged tweets were spatially joined to their 2010 census tract locations and sentiment scores were computed. This color coded map presents the proportion of happy tweets in each census tract, with darker colors signifying higher proportions of happy tweets.

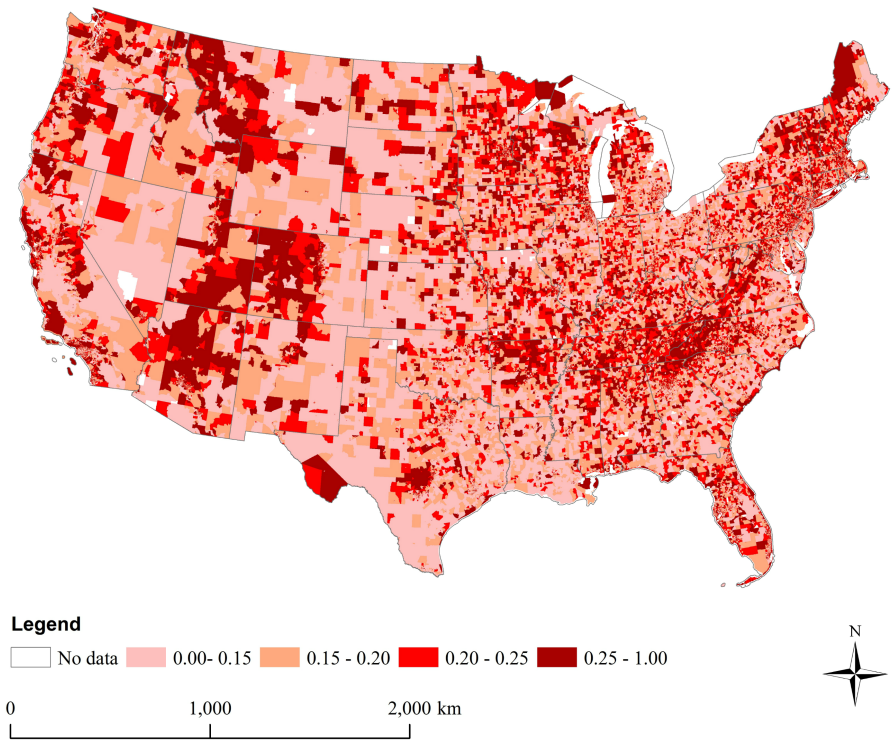


Table 1. Descriptive statistics of our national Twitter database, April 2015 to March 2016 (N=79,848,992).

	Mean (SD)
Happiness	
% Tweets that are happy	19.9 (6.7)
Food culture	
% Tweets about food	5.1 (22.0)
% Food tweets about healthy foods	15.9 (36.6)
% Food tweets about fast food	9.2 (29.0)
Caloric density of food tweets (per 100 grams)	238.5 (219.8)
% Food tweets that are happy	27.0 (44.4)
% Healthy food tweets that are happy	28.3 (45.0)
% Fast food tweets that are happy	14.5 (35.2)
Physical activity culture	
% Tweets about physical activity	1.8 (13.3)
Exercise intensity (per 30 minutes)	199.1 (117.5)
% Physical activity tweets that are happy	28.2 (45.0)

XSL•FO
RenderX

Table 2. Demographic and economic predictors of happy, food, and physical activity tweets from 70,515 census tracts (data source: 2010 US Census data).

Tract characteristics	% happy tweets Beta (95% CI)[a]	*P* value	% healthy food tweets Beta (95% CI)[a]	*P* value	% fast food tweets Beta (95% CI)[a]	*P* value	% physical activity tweets Beta (95% CI)[a]	*P* value
Urban (yes)	−.01 (−.04 to .03)	.79	.01 (−.02 to .03)	.54	.29 (.26 to .31)	<.001	−.02 (−.03 to −.01)	<.001
Population density	.06 (.03 to .08)	<.001	.04 (.02 to .07)	.001	−.03 (−.03 to −.02)	<.001	.00 (−.01 to .00)	.82
% 65 years and older	.02 (−.01 to .04)	.09	−.03 (−.04 to −.02)	<.001	−.03 (−.04 to −.01)	<.001	.02 (.02 to .03)	<.001
% 10-24 years	−.02 (−.04 to .00)	.01	−.05 (−.05 to −.04)	<.001	.00 (−.01 to .01)	.49	.00 (−.01 to .00)	.14
% Male	.04 (.03 to .06)	<.001	.01 (.00 to .02)	.21	−.05 (−.06 to −.04)	<.001	.01 (.01 to .02)	<.001
% African American	−.11 (−.14 to −.07)	<.001	−.03 (−.04 to −.01)	<.001	−.03 (−.04 to −.02)	<.001	−.01 (−.02 to −.01)	<.001
% Hispanic	−.04 (−.08 to .00)	.05	.02 (.01 to .03)	.00	.07 (.05 to .09)	<.001	.00 (.00 to .00)	.77
Household size	−.18 (−.20 to −.15)	<.001	−.11 (−.12 to −.09)	<.001	−.07 (−.09 to −.05)	<.001	−.01 (−.01 to −.01)	<.001
Economic disadvantage[b]	−.19 (−.21 to −.16)	<.001	−.09 (−.10 to −.08)	<.001	−.09 (−.10 to −.07)	<.001	−.03 (−.04 to −.03)	<.001

[a]Adjusted linear regression included all tract demographic and economic predictors simultaneously. Standard errors accounted for clustering at the county level.

[b]Economic disadvantage factor score derived from the following census tract characteristics: percent female-headed households, percent families living in poverty, unemployment rate, percent college graduates (reverse coded), and median family income (reverse coded).

Sensitivity analyses were performed to examine the relationship between population characteristics and happiness for a different unit of aggregation: zip code areas. Relationships seen at the census tract level were similar to those at the zip code level, although they were more muted at the zip code level (not shown). This may be the case because census tracts are designed to be relatively homogenous with regard to characteristics such as economic status and demographic characteristics [73].

Healthy foods (ie, vegetables, fruits, nuts, lean proteins) composed 15.9% of food tweets, while fast food restaurant mentions composed 9.2% of food tweets. The most popular foods include coffee, beer, pizza, wine, chicken, ice cream, and sushi (Figure 2). Popular healthy food terms included chicken, eggs, salad, turkey, and banana (Figure 3). Starbucks was the most popular fast food place mentioned (accounting for 46% of all fast food restaurant mentions), followed by Chipotle (9.2%), Taco Bell (5.4%), and Buffalo Wild Wings (5.2%). We additionally examined the relationship between food tweets and business characteristics. At the zip code level, greater numbers of fast food restaurants were associated with more fast food tweets (B=.15), and higher caloric density of food mentions (B=.08). Urban areas had tweets with higher caloric density (B=.08) and more fast food restaurant mentions (B=.16). Happy tweets were more prevalent in zip codes with higher numbers of businesses (B=.11) and full-service restaurants (B=.16). Higher numbers of fast food restaurant (B=−.16) and convenience stores (B=−.07) were related to fewer happy tweets (Table 3).

Additionally, relatively few physical activity terms (13 terms) accounted for 75% of physical activity tweets (Figure 4) although our data collection system was set up to collect tweets on 376 physical activity terms. The most popular terms included walking, dancing, and running. At the zip code level, greater numbers of fitness and recreational sports centers were related to higher exercise intensity (B=.05) and happier tweets (B=.07). Surprisingly, the presence of nature parks was not associated with physical activity mentions. Urbanicity was associated with lower frequency of physical activity tweets and happy tweets but higher exercise intensity (Table 4). In supplemental analyses, we examined information on number of miles covered during physical activity if that was mentioned in the tweet (n=36,291; median 3.1 miles). Even fewer tweets contained information on amount of time the person engaged in physical activity. Among 5823 tweets that mentioned hour(s) of physical activity, the median amount was 2 hours. Among 2402 tweets that only referred to minutes of physical activity, the median number of minutes was 20.

XSL•FO
RenderX

Table 3. Zip code and business characteristics as predictors of food tweets and happiness (data sources: 2013 zip code business patterns and 2010 US Census data).

Zip code characteristics	Average caloric density of food tweets n=21,756 Beta (95% CI)[a]	*P* value	% fast food tweets n=21,756 Beta (95% CI)[a]	*P* value	% happy tweets n=26,584 Beta (95% CI)[a]	*P* value
Urban (yes)	.08 (.05 to .11)	<.001	.16 (.12 to .20)	<.001	−.02 (−.06 to .02)	.29
Population density	.00 (.00 to .01)	.24	.00 (−.01 to .01)	.86	.01 (.00 to .03)	.18
Number of businesses	−.01 (−.02 to .01)	.34	.02 (.00 to .04)	.04	.11 (.08 to .15)	<.001
Businesses that sell alcohol	−.03 (−.04 to −.02)	<.001	−.04 (−.05 to −.04)	<.001	−.01 (−.02 to .00)	.02
Full service restaurants	−.04 (−.06 to −.02)	<.001	.01 (−.01 to .03)	.43	.16 (.13 to .20)	<.001
Fast food restaurants	.08 (.06 to .10)	<.001	.15 (.13 to .17)	<.001	−.16 (−.20 to −.12)	<.001
Grocery stores	.01 (.00 to .01)	.28	−.04 (−.05 to −.03)	<.001	−.02 (−.04 to .00)	.05
Convenience stores	.02 (.01 to .02)	<.001	−.03 (−.04 to −.02)	<.001	−.07 (−.08 to −.05)	<.001

[a]Adjusted linear regression included all zip code and business characteristics simultaneously. Standard errors accounted for clustering at the county level.

Table 4. Zip code and business characteristics as predictors of physical activity tweets and happiness (data sources: 2013 zip code business patterns and 2010 US Census data).

Zip code characteristics	% physical activity tweets n=26,839 Beta (95% CI)[a]	*P* value	Exercise intensity n=20,715 Beta (95% CI)[a]	*P* value	% happy tweets n=26,839 Beta (95% CI)[a]	*P* value
Urban (yes)	−.09 (−.11 to −.07)	<.001	.07 (.04 to .11)	<.001	−.08 (−.12 to −.04)	<.001
Population density	−.01 (−.02 to .00)	.01	−.01 (−.01 to .00)	.03	.01 (.00 to .02)	.08
Fitness/recreational centers	.01 (.00 to .02)	.003	.05 (.04 to .06)	<.001	.07 (.06 to .08)	<.001
Nature parks	.01 (.00 to .02)	.05	−.01 (−.01 to .00)	.21	.03 (.02 to .04)	<.001
Zoos/botanical gardens	.00 (.00 to .01)	.19	.00 (−.01 to .00)	.35	.02 (.01 to .03)	<.001
Golf/country clubs	.03 (.02 to .03)	<.001	−.05 (−.06 to −.04)	<.001	.03 (.02 to .04)	<.001
Skiing facilities	.04 (.04 to .05)	<.001	.02 (.02 to .03)	<.001	.03 (.02 to .03)	<.001
Bowling centers	−.01 (−.02 to −.01)	<.001	−.01 (−.02 to .00)	.01	−.02 (−.03 to −.01)	<.001

[a]Adjusted linear regression included all zip code and business characteristics simultaneously. Standard errors accounted for clustering at county level.

Table 5. Twitter happiness as a predictor of health outcomes in 232 zip codes in Utah (data source: Utah Behavioral Risk Factor Surveillance System [BRFSS] survey 2009-2014. BRFSS underwent design feature changes. Life dissatisfaction values were only available for 2009 and 2010. All other variables were averages from available data from 2011-2014).

Zip code health outcomes	Beta (95% CI)[a] n=232	*P* value
Life dissatisfaction	.01 (−.13 to .15)	.91
Self-rated health (higher score=worse health)	−.08 (−.21 to .05)	.21
Any past month physical activity/exercise	.13 (.00 to .26)	.05
Body mass index (kg/m^2)	−.13 (−.26 to −.01)	.04

[a]Separate linear regression models for each zip code health outcome.

XSL•FO
RenderX

Table 6. State level Twitter sentiment predictors of health outcomes (N=49 states in the contiguous United States plus District of Columbia. Data sources: 2013 National Vital Statistics Reports and 2013 Behavioral Risk Factor Surveillance System [BRFSS] survey on adults).

	Twitter predictor variables					
State-level adult health outcomes	Happiness Beta (95% CI)[a]	*P* value	Positive sentiment toward healthy foods Beta (95% CI)[a]	*P* value	Positive sentiment toward physical activity Beta (95% CI)[a]	*P* value
All-cause mortality per 100,000	−32.34 (−61.59 to −3.09)	.03	−23.51 (−40.54 to −6.48)	.01	−25.37 (−42.00 to −8.74)	.004
Homicide per 100,000	−1.02 (−1.98 to −.06)	.03	−.76 (−1.28 to −.25)	.01	−.75 (−1.28 to −.23)	.01
% With diabetes	−.58 (−1.05 to −.12)	.02	−.52 (−.78 to −.27)	<.001	−.41 (−.68 to −.14)	.004
% With obesity	−2.27 (−3.35 to −1.18)	<.001	−1.67 (−2.25 to −1.09)	<.001	−1.43 (−2.05 to −.80)	<.001
% Poor/fair self-rated health	−1.13 (−2.13 to −.13)	.03	−.77 (−1.36 to −.19)	.01	−.61 (−1.21 to −.02)	.05
% With high cholesterol	−.78 (−1.66 to .11)	.08	−.51 (−1.04 to .01)	.06	−.75 (−1.25 to −.26)	.003
% Physical inactivity	−2.46 (−4.80 to −.12)	.04	−2.32 (−3.61 to −1.03)	.001	−1.59 (−2.97 to −.22)	.02
% Current smoking	−1.47 (−2.68 to −.27)	.02	−1.20 (−1.88 to −.52)	.001	−1.14 (−1.82 to −.45)	.002

[a]Each cell in the table represents the coefficient estimate of the predictor variable (given by the column) on the state-level health outcome (given by the row). Adjusted linear regression models controlled for state-level demographics: median age, % non-Hispanic white, median household income.

Additionally, merging in health-related datasets, we examined associations between our Twitter-based variables and other measures of health and well-being. Utilizing data from the 2009-2014 BRFSS in Utah, we found that zip codes in Utah with higher Twitter happiness scores were associated with lower body mass index and higher physical activity (Table 5). However, Twitter happiness scores were not statistically significantly related to self-rated health or life satisfaction.

Greater state-level happiness, as indicated by tweets, was related to lower prevalence of obesity; a one standard deviation increase in happiness was associated with two percentage points lower prevalence in obesity. Greater positive sentiment for healthy foods was related to lower prevalence of diabetes and obesity and lower percent of the population who are physically inactive or current smokers (Table 6). Positive sentiment toward physical activity was related to lower obesity.

Table 7 presents adjusted regression results for additional Twitter-derived variables (percentage of food tweets about healthy foods, percentage of food tweets about fast food, and percentage of tweets about physical activity) and a select number of state health outcomes. Out of the three Twitter-derived variables, percentage of tweets about physical activity was the strongest and most consistent predictor; more online discussion about physical activity was related to lower all-cause mortality and lower prevalence of obesity and fair/poor self-rated health.

Table 7. State level Twitter food and physical activity characteristics as predictors of health outcomes (N=49 states in the contiguous United States plus District of Columbia. Data sources: 2013 National Vital Statistics Reports and 2013 Behavioral Risk Factor Surveillance System [BRFSS] survey on adults).

	State-level adult health outcomes					
Twitter predictors	All-cause mortality per 100,000 Beta (95% CI)[a]	*P* value	% with obesity Beta (95% CI)[a]	*P* value	% poor/fair self-rated health Beta (95% CI)[a]	*P* value
% Of food tweets about healthy food	11.74 (−6.48 to 29.96)	.20	−.09 (−.64 to .45)	.73	.11 (−.48 to .70)	.71
% Of food tweets about fast food	9.84 (−8.56 to 28.25)	.29	.68 (.13 to 1.23)	.02	.77 (.18 to 1.37)	.01
% Of tweets about physical activity	−28.17 (−46.68 to −9.65)	.004	−1.86 (−2.41 to −1.31)	<.001	−.89 (−1.49 to −.29)	.01

[a]Adjusted linear regression models were run separately for each state-level health outcome (column) and included all three predictors (row) simultaneously in addition to the following state-level control variables: median age, % non-Hispanic white, median household income. Beta coefficient represents a change in the outcome for every standard deviation change in the predictor (row variable).

XSL•FO
RenderX

Figure 2. Items in the top 50% of food tweets.

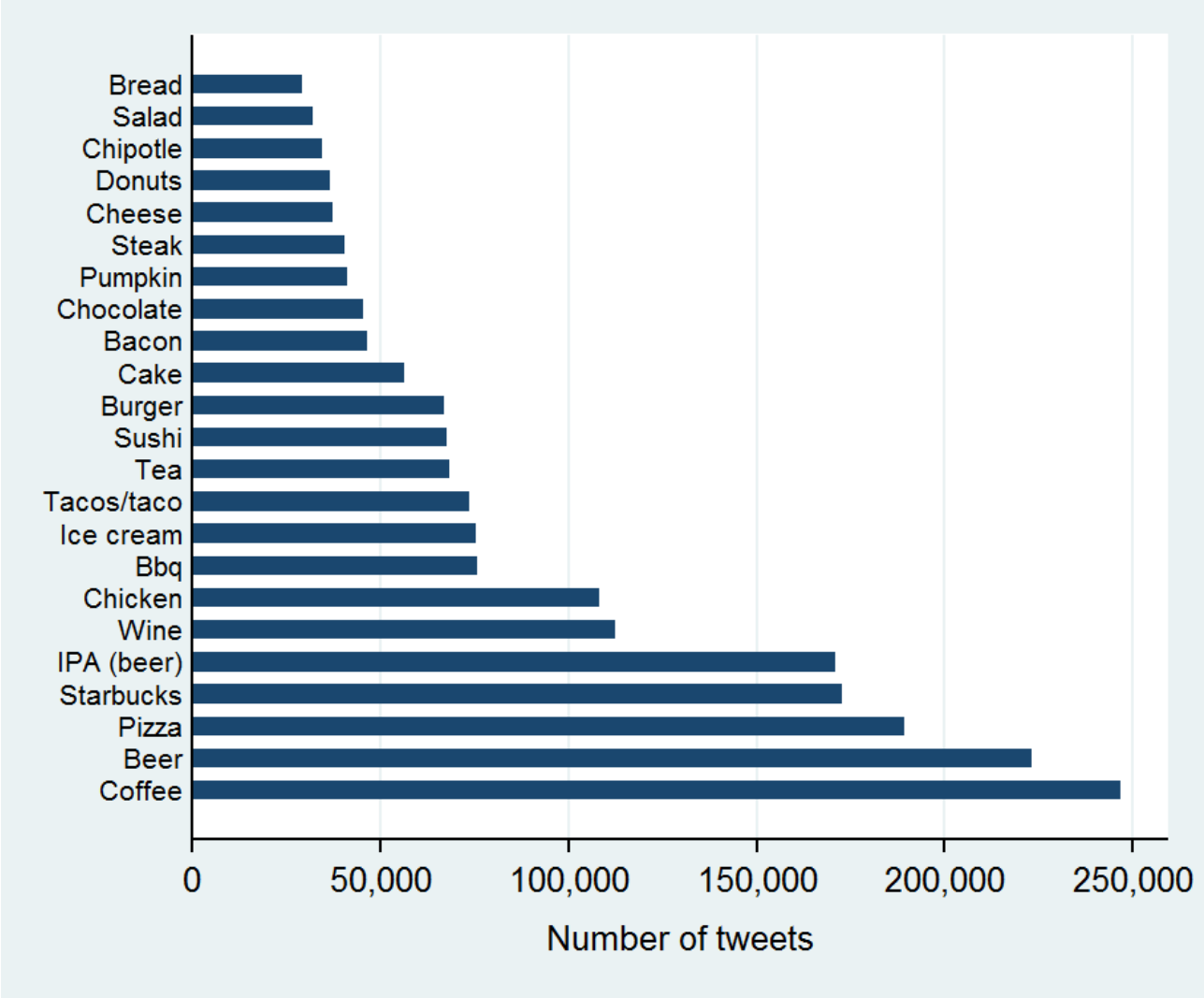
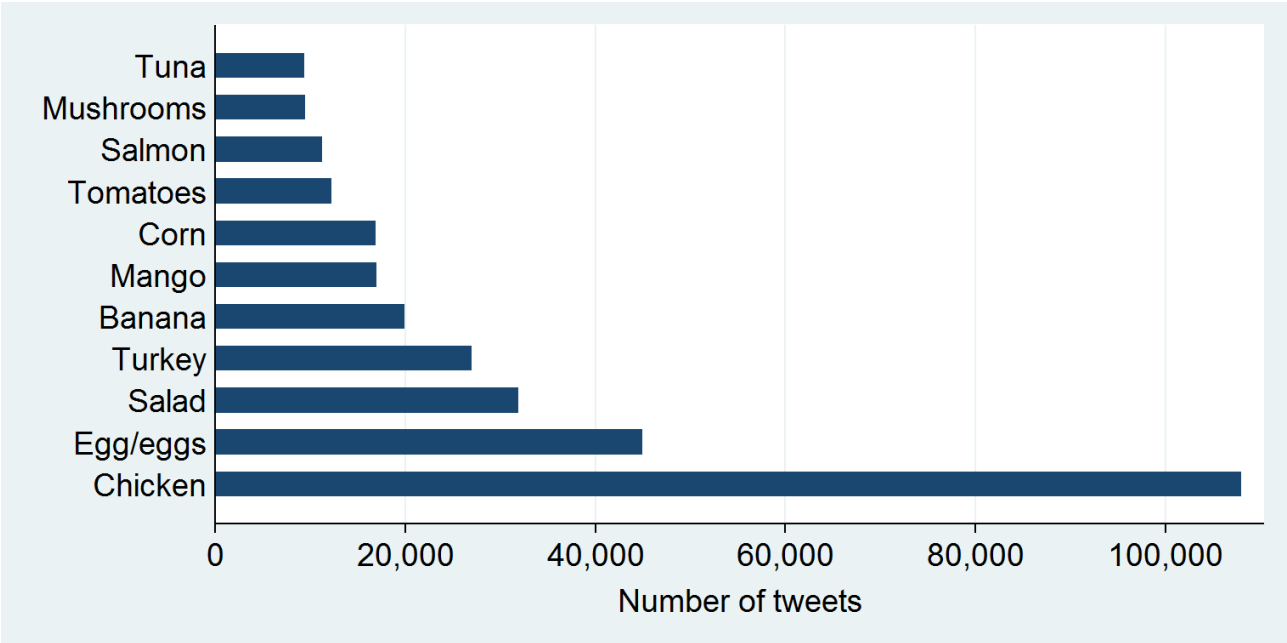
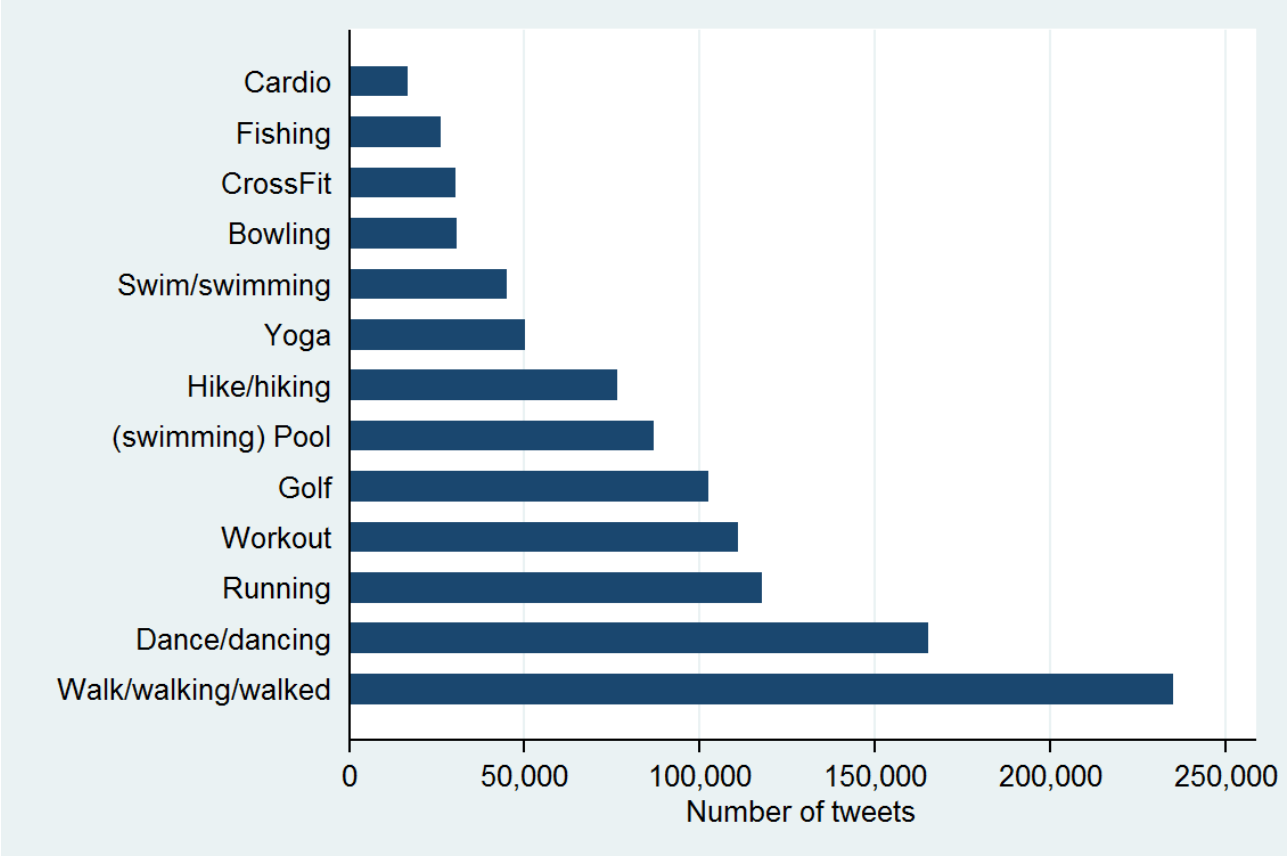


Figure 3. Items in the top 50% of healthy food tweets.



XSL•FO
RenderX

Figure 4. Items in the top 75% of physical activity tweets.



Discussion

Principal Findings

In this paper, we detail the building of a new national neighborhood data repository constructed from Twitter data which addresses a pressing need for neighborhood data that are available across large geographies and can be updated efficiently and cost-effectively. We demonstrate that simple machine learning algorithms for the construction of indicators for happiness, food, and physical activity can agree extremely well with manually generated labels. About one-fifth of tweets were identified as happy. There was substantial spatial variation in happiness across the United States. For instance, the proportion of tweets that were happy in Montana (the most happy state) was 10% greater than in Louisiana (the least happy state). Only a few terms are needed to capture the majority of tweets on food and physical activity. Economic disadvantage, urbanicity, and presence of fast food restaurants predicted lower area level happiness and lower frequency of healthy behavior mentions on Twitter. Moreover, we find that Twitter area-level characteristics are correlated with area-level health outcomes relating to health behaviors, chronic diseases, mortality, and self-rated health.

Study Findings in Context

Social media represents an important new data resource that is increasingly being harnessed for public health efforts such as surveillance of smoking behavior and sentiment toward tobacco products [74]. However, few studies are leveraging social media data for the investigation of local area characteristics. More commonly, studies utilizing social media data examine patterns at the city, county, or state level [67,75] rather than at finer levels of aggregation, which is necessary for understanding the potential impacts of neighborhood conditions.

Neighborhoods can impact health through a myriad of pathways. Disadvantaged neighborhoods may have fewer resources that support physical activity and healthy diets. Poor and minority neighborhoods have fewer large supermarkets (where healthy foods are more abundant and affordable) compared to wealthy and majority white neighborhoods. Studies have documented increased fruit and vegetable consumption with more supermarket availability [17]. Poor neighborhoods, which have been labeled food deserts, also tend to have more fast food restaurants, which can contribute to weight gain [6]. In this study, we found that higher numbers of fast food restaurants were associated with higher frequency of fast food mentions, lower frequency of healthy food mentions, and less positive sentiment about healthy foods on Twitter. Our results align with a recent study conducted analyzing Instagram posts, which found that posts originating from census tracts deemed as food deserts contained fewer mentions of fruits and vegetables compared to Instagram posts outside food deserts [76]. Additionally, neighborhoods may promote poor health through psychosocial pathways. Living in neighborhoods that are unclean, noisy, and violent can be psychologically harmful through over-activation of the stress response [77,78].

We found that economic disadvantage was related to lower frequency of happy tweets. Previous research by Mitchell and colleagues found that higher socioeconomic status was associated with higher Twitter happiness scores at the city level.

XSL•FO
RenderX

Moreover, they identified mild correlations (r=−0.34) between happiness and obesity rates for 190 metropolitan statistical areas [67] and that Twitter happiness scores were moderately correlated with other state-level indicators of well-being including shootings, the Peace index, America's Health Ranking, and the Gallup-Healthways Well-Being Index (correlations ranged between 0.51 and 0.64) [67].

Study Strengths and Limitations

In this paper, we describe the creation of a new neighborhood data repository constructed from Twitter data and merged with publicly available administrative datasets. However, this study is subject to several limitations. For instance, users of social media tend to be younger; in 2014, 37% of individuals aged 18 to 29 years old used Twitter compared to 12% of individuals aged 50 to 64 years and 10% among those 65 years and older. Nonetheless, adoption rates of social media have been steadily increasing [79]. Tweets also include information rarely found in other neighborhood sources. Twitter users are composed of individuals as well as groups of individuals, organizations, companies, and news outlets. Thus, compiling such information may allow for a more comprehensive examination of the social environment.

Moreover, we are only collecting a subset of publicly available tweets, and thus conclusions from our analytic sample may not generalize to the full population of tweets [80]. Our construction of neighborhood indicators from Twitter data necessitated that we restricted our data collection to geolocated tweets. We utilized Twitter's API which allows the retrieval of a maximum resulting volume of 1% of the total tweets at any given time point. Previous studies suggest that about 1% to 2% of tweets may contain global positioning system location information [81,82] and that use of Twitter's streaming API may obtain 40% to 90% of all geotagged tweets [81,82]. Tweets with location information may be different from those without. For example, tweets in which users share their locations may be more likely to contain public and social activities such as friends tweeting from a restaurant or an event. However, in sensitivity analyses with a subset of control tweets (n=138,152 tweets) collected from July 9 to July 14, 2015, we did not detect any statistically significant differences in happiness scores between tweets with and without geographic coordinates (not shown).

In creating our neighborhood indicators from Twitter data, we prioritized transparency and ease of implementation so that other researchers can replicate our algorithms. Our sentiment algorithm was trained to differentiate between happy and not happy sentiments (which encompasses neutral and sad sentiments). Thus, we were not able to specifically examine the prevalence of sad tweets, which may provide additionally useful information about the well-being of communities. In future work, we plan to target the identification of sadness. Our algorithms for food and physical activity implemented corpus-based classification with steps that are easily understandable. However, this technique does not take into account the entire context of sarcasm or humor in a tweet, challenges which still evade most natural language processing algorithms though some studies show promising results [83,84]. Our analysis of caloric density of food assumed calories per 100 grams. Most tweets do not specify the exact amount of food consumed, and thus our estimate is just an approximation.

Additionally, the content of tweets reflects the type of information that people feel comfortable reporting and may not represent the true spectrum of their feelings or their experiences. For instance, people may feel most comfortable presenting a neutral stance rather than voicing polarizing viewpoints. Certain foods (cupcakes) may get tweeted more often than others (celery). Additionally, we cannot be certain that the food that was tweeted was indeed consumed. Similarly, physical activity tweets may reflect a mixture of intentions, plans, and actual engagement in those physical activities. Also, exercise intensity for physical activities was assessed for 30 minutes of physical activity for an individual weighing 155 pounds, which can be an under- or overestimation depending on the type of activity and persons engaged in that activity.

Conclusions

The epidemic rise in obesity and related chronic diseases in recent decades signal the importance of structural forces and social processes, but the dearth of data on contextual factors limits the investigation of multilevel effects on health. Social media data can be uniquely harnessed to capture social and cultural processes with potential impacts on health [71,72,85-89]. For instance, public posts can be utilized to measure prevalent happiness which can impact health through emotional contagion and the interconnectedness between mental health and physical health. Additionally, public posts about health behaviors may help us understand the prevalence of those behaviors as well as local area social norms. We demonstrate that tweets can provide a means to assess prevalent sentiment and food behaviors and physical activity, which can inform health interventions and policies to meet the needs of different neighborhoods. In particular, as this study suggests, neighborhoods with social and economic disadvantage, high urbanicity, and those with more fast food restaurants may exhibit lower happiness and fewer healthy behaviors.

Acknowledgments

This work was supported by a National Institutes of Health grant (5K01ES025433) to Dr Nguyen. The research uses data from the Utah BRFSS survey, which is implemented by the Utah Department of Health in conjunction with the US Centers for Disease Control and Prevention. We thank Patsaporn Kanokvimankul for her assistance with locating some of the external health outcomes data for this paper. We thank Drs Jared B Hawkins and John S Brownstein for their assistance with quality control activities associated with the Twitter data.

XSL•FO
RenderX

Conflicts of Interest

None declared.

Multimedia Appendix 1

Varying MALLET cut points for happy tweets and comparisons with manually generated labels.

[PDF File (Adobe PDF File), 14KB-Multimedia Appendix 1]

Multimedia Appendix 2

National distribution of happy tweets, by zip code. Geotagged tweets were spatially joined to their 2010 zip code locations and sentiment scores were computed. This color coded map presents the proportion of happy tweets in each zip code area, with darker colors signifying higher proportions of happy tweets.

[JPG File, 6MB-Multimedia Appendix 2]

Multimedia Appendix 3

Proportion of happy tweets, by state.

[PDF File (Adobe PDF File), 18KB-Multimedia Appendix 3]

References

1. Social determinants of health. Washington, DC. US Department of Health and Human Services; Jan 30, 2016. URL: https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health [accessed 2016-09-28] [WebCite Cache ID 6ks1l76Df]
2. Marmot M. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet. Nov 08, 2008;372(9650):1661-1669.
3. Villanueva K. People living in hilly residential areas in metropolitan Perth have less diabetes: spurious association or important environmental determinant? Int J Health Geogr. 2013;12(1):1-11.
4. Schmidt NM, Lincoln AK, Nguyen QC, Acevedo-Garcia D, Osypuk TL. Examining mediators of housing mobility on adolescent asthma: results from a housing voucher experiment. Soc Sci Med. Apr 2014;107:136-144. [FREE Full text] [doi: 10.1016/j.socscimed.2014.02.020] [Medline: 24607675]
5. Nguyen QC, Rehkopf DH, Schmidt NM, Osypuk TL. Heterogeneous effects of housing vouchers on the mental health of US adolescents. Am J Public Health. Apr 2016;106(4):755-762. [doi: 10.2105/AJPH.2015.303006] [Medline: 26794179]
6. Morland K, Wing S, Diez RA, Poole C. Neighborhood characteristics associated with the location of food stores and food service places. Am J Prev Med. Jan 2002;22(1):23-29. [Medline: 11777675]
7. Stafford M, Cummins S, Ellaway A, Sacker A, Wiggins RD, Macintyre S. Pathways to obesity: identifying local, modifiable determinants of physical activity and diet. Soc Sci Med. Nov 2007;65(9):1882-1897. [doi: 10.1016/j.socscimed.2007.05.042] [Medline: 17640787]
8. Wang MC, Kim S, Gonzalez AA, MacLeod KE, Winkleby MA. Socioeconomic and food-related physical characteristics of the neighbourhood environment are associated with body mass index. J Epidemiol Community Health. Jun 2007;61(6):491-498. [FREE Full text] [doi: 10.1136/jech.2006.051680] [Medline: 17496257]
9. Inagami S, Cohen DA, Finch BK, Asch SM. You are where you shop: grocery store locations, weight, and neighborhoods. Am J Prev Med. Jul 2006;31(1):10-17. [doi: 10.1016/j.amepre.2006.03.019] [Medline: 16777537]
10. Christiansen KMH, Qureshi F, Schaible A, Park S, Gittelsohn J. Environmental factors that impact the eating behaviors of low-income African American adolescents in Baltimore City. J Nutr Educ Behav. 2013;45(6):652-660. [doi: 10.1016/j.jneb.2013.05.009] [Medline: 23916684]
11. Block JP, Scribner RA, DeSalvo KB. Fast food, race/ethnicity, and income: A geographic analysis. Am J Prev Med. 2004;27(3):211-217.
12. Roemmich JN, Epstein LH, Raja S, Yin L, Robinson J, Winiewicz D. Association of access to parks and recreational facilities with the physical activity of young children. Prev Med. Dec 2006;43(6):437-441. [doi: 10.1016/j.ypmed.2006.07.007] [Medline: 16928396]
13. Brownson RC, Hoehner CM, Day K, Forsyth A, Sallis JF. Measuring the built environment for physical activity: state of the science. Am J Prev Med. Apr 2009;36(4 Suppl):S99-S123. [FREE Full text] [doi: 10.1016/j.amepre.2009.01.005] [Medline: 19285216]
14. Mujahid MS, Diez Roux AV, Shen M, Gowda D, Sánchez B, Shea S, et al. Relation between neighborhood environments and obesity in the multi-ethnic study of atherosclerosis. Am J Epidemiol. Jun 1, 2008;167(11):1349-1357. [FREE Full text] [doi: 10.1093/aje/kwn047] [Medline: 18367469]

XSL•FO
RenderX

15. Yen IH, Kaplan GA. Poverty area residence and changes in physical activity level: evidence from the Alameda County Study. Am J Public Health. 1998;88(11):1709-1712.
16. Ross CE. Walking, exercising, and smoking: does neighborhood matter? Soc Sci Med. Jul 2000;51(2):265-274. [Medline: 10832573]
17. Morland K, Wing S, Diez RA. The contextual effect of the local food environment on residents' diets: the atherosclerosis risk in communities study. Am J Public Health. Nov 2002;92(11):1761-1767. [Medline: 12406805]
18. Black JL, Macinko J, Dixon LB, Fryer GE. Neighborhoods and obesity in New York City. Health Place. May 2010;16(3):489-499. [doi: 10.1016/j.healthplace.2009.12.007] [Medline: 20106710]
19. Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. Am J Public Health. Feb 1998;88(2):216-222. [Medline: 9491010]
20. Macintyre S, MacIver S, Sooman A. Area, class and health: Should we be focusing on places or people. J Soc Policy. 1993;22:213-233.
21. Duncan C, Jones K, Moon G. Context, composition and heterogeneity: using multilevel models in health research. Soc Sci Med. 1998;46:97-117.
22. Pearlin LI. The sociological study of stress. J Health Soc Behav. 1989;30(3):241-256.
23. Johns LE. Neighborhood social cohesion and posttraumatic stress disorder in a community-based sample: findings from the Detroit Neighborhood Health Study. Soc Psych Psych Epid. 2012;47(12):1899-1906.
24. Oswald AJ, Powdthavee N. Obesity, unhappiness, and the challenge of affluence: theory and evidence. Econ J. 2007;117:117.
25. Bray I, Gunnell D. Suicide rates, life satisfaction and happiness as markers for population mental health. Soc Psych Psych Epid. 2006;41(5):333-337.
26. Tella RD, MacCulloch RJ, Oswald AJ. The macroeconomics of happiness. Rev Econ Stat. 2003;85(4):809-827.
27. Blanchflower DG, Oswald AJ. Hypertension and happiness across nations. J Health Econ. 2008;27(2):218-233.
28. Dodds PS. Temporal patterns of happiness: information in a global social network: hedonometrics and Twitter. PLoS ONE. 2011;6(12):e26752.
29. Di Tella R, MacCulloch R. Gross national happiness as an answer to the Easterlin Paradox? J Devel Econ. 2008;86(1):22-42.
30. Bearman PS, Moody J. Suicide and friendships among American adolescents. Am J Public Health. 2004;94(1):89-95.
31. Larson R, Richards MH. Divergent Realities: The Emotional Lives of Mothers, Fathers, and Adolescents. New York, NY. Basic Books; 1994.
32. Fowler JH, Christiakis N. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study. Brit Med J. 2008;337:a2338.
33. Guan W, Kamo Y. Contextualizing depressive contagion: A multilevel network approach. Soc Ment Health. Dec 09, 2015. [doi: 10.1177/2156869315619657]
34. Pachucki MA, Jacques PF, Christakis NA. Social network concordance in food choice among spouses, friends, and siblings. Am J Public Health. Nov 2011;101(11):2170-2177. [FREE Full text] [doi: 10.2105/AJPH.2011.300282] [Medline: 21940920]
35. Keating NL, O'Malley AJ, Murabito JM, Smith KP, Christakis NA. Minimal social network effects evident in cancer screening behavior. Cancer. Jul 1, 2011;117(13):3045-3052. [FREE Full text] [doi: 10.1002/cncr.25849] [Medline: 21264828]
36. Rosenquist JN, Murabito J, Fowler JH, Christakis NA. The spread of alcohol consumption behavior in a large social network. Ann Intern Med. Apr 6, 2010;152(7):426-433. [FREE Full text] [doi: 10.7326/0003-4819-152-7-201004060-00007] [Medline: 20368648]
37. Mednick SC, Christakis NA, Fowler JH. The spread of sleep loss influences drug use in adolescent social networks. PLoS One. 2010;5(3):e9775. [FREE Full text] [doi: 10.1371/journal.pone.0009775] [Medline: 20333306]
38. National Archive of Criminal Justice. 2012. URL: http://www.icpsr.umich.edu/icpsrweb/PHDCN/ [accessed 2016-09-28] [WebCite Cache ID 6ks3vmrzJ]
39. Baltimore Neighborhood Indicators Alliance: Vital Signs 11. Sep 24, 2013. URL: http://bniajfi.org/wp-content/uploads/2014/04/VS-11-Intro.pdf [accessed 2016-09-28] [WebCite Cache ID 6ks49HpP9]
40. Peterson RD, Krivo LJ. National Neighborhood Crime Study. 2000. URL: http://www.icpsr.umich.edu/icpsrweb/RCMD/studies/27501 [accessed 2016-09-28] [WebCite Cache ID 6ks4ICpcz]
41. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annu Symp Proc. 2006.:244-248. [FREE Full text] [Medline: 17238340]
42. Eysenbach G. Infodemiology and infoveillance. Am J Prev Med. 2011;40(5):S154-S158.
43. Yepes AJ, Han B. Investigating public health surveillance using Twitter. ACL-IJCNLP. 2015;2015:164.
44. Nsoesie EO, Kluberg SA, Brownstein JS. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. Prev Med. Oct 2014;67:264-269. [FREE Full text] [doi: 10.1016/j.ypmed.2014.08.003] [Medline: 25124281]
45. Woo H, Cho Y, Shim E, Lee J, Lee C, Kim SH. Estimating influenza outbreaks using both search engine query data and social media data in South Korea. J Med Internet Res. 2016;18(7):e177. [FREE Full text] [doi: 10.2196/jmir.4955] [Medline: 27377323]
46. McIver DJ, Hawkins JB, Chunara R, Chatterjee AK, Bhandari A, Fitzgerald TP, et al. Characterizing sleep issues using Twitter. J Med Internet Res. 2015;17(6):e140. [FREE Full text] [doi: 10.2196/jmir.4476] [Medline: 26054530]

XSL•FO
RenderX

47. Nguyen QC. Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. Appl Geogr. 2016;73:77-88.
48. Yin Z, Fabbri D, Rosenbloom ST, Malin B. A scalable framework to detect personal health mentions on Twitter. J Med Internet Res. 2015;17(6):e138. [FREE Full text] [doi: 10.2196/jmir.4305] [Medline: 26048075]
49. Hawkins JB. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Safety. 2015.:4309.
50. Etherington TR. Teaching introductory GIS programming to geographers using an open source Python approach. J Geogr Higher Educ. 2016;40(1):117-130.
51. Guttman A. R-trees: a dynamic index structure for spatial searching. 1984. Presented at: 1984 ACM SIGMOD international conference on Management of Data; 1984;47-57; New York, NY.
52. Nguyen QC, Schmidt NM, Glymour MM, Rehkopf DH, Osypuk TL. Were the mental health benefits of a housing mobility intervention larger for adolescents in higher socioeconomic status families? Health Place. Sep 2013;23:79-88. [FREE Full text] [doi: 10.1016/j.healthplace.2013.05.002] [Medline: 23792412]
53. Larson NI, Story MT, Nelson MC. Neighborhood environments: disparities in access to healthy foods in the US. Am J Prev Med. Jan 2009;36(1):74-81. [doi: 10.1016/j.amepre.2008.09.025] [Medline: 18977112]
54. Roth C. Integrating population- and patient-level data for secondary use of electronic health records to study overweight and obesity. Stud Health Technol Inform. 2013.:192.
55. Stanford Natural Language Processing Group. Stanford tokenizer. 2015. URL: http://nlp.stanford.edu/software/tokenizer.shtml [accessed 2016-09-28] [WebCite Cache ID 6ks509JgN]
56. Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. 1999. Presented at: IJCAI-99 workshop on machine learning for information filtering; 1999; Stockholm, Sweden.
57. Sentiment140 for Academics. URL: https://sites.google.com/site/twittersentimenthelp/for-researchers [accessed 2016-08-16] [WebCite Cache ID 6joQzyTSS]
58. Twitter sentiment corpus. Sanders Analytics; 2011. URL: http://www.sananalytics.com/lab/twitter-sentiment/ [accessed 2016-08-16] [WebCite Cache ID 6joVh9V4R]
59. Kaggle in Class. 2011. URL: https://inclass.kaggle.com/c/si650winter11[WebCite Cache ID 6joWjx5fX]
60. National Nutrient Database. Washington, DC. United States Department of Agriculture; Feb 5, 2014. URL: http://ndb.nal.usda.gov/ndb/search/list?format=&count=&max=25&sort=&fg=&man=&lfacet=&qlookup=&offset=50 [accessed 2016-09-28]
61. Ainsworth BE. 2011 compendium of physical activities: a second update of codes and MET values. Med Sci Sport Exer. 2011;43(8):1575-1581.
62. Zhang N. Electronic word of mouth on Twitter about physical activity in the United States: exploratory infodemiology study. J Med Internet Res. 2013;15(11).
63. Kendall L. Descriptive analysis of physical activity conversations on Twitter. 2011. Presented at: CHI '11 Extended Abstracts on Human Factors in Computing Systems; 2011;1555-1560; Vancouver, Canada.
64. Body Measurements. National Center for Health Statistics, Centers for Disease Control and Prevention; Sep 2, 2012. URL: http://www.cdc.gov/nchs/fastats/body-measurements.htm [accessed 2016-09-28] [WebCite Cache ID 6ks62JHO7]
65. Harvard Health Publications. Calories burned in 30 minutes for people of three different weights. 2015. URL: http://www.health.harvard.edu/newsweek/Calories-burned-in-30-minutes-of-leisure-and-routine-activities.htm [accessed 2016-09-28] [WebCite Cache ID 6ks6AbOhg]
66. Snow R. Cheap and fast--but is it good? Evaluating non-expert annotations for natural language tasks. 2008. Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2008;254-263; Stroudsburg, PA.
67. Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. PLoS One. 2013;8(5):e64417. [FREE Full text] [doi: 10.1371/journal.pone.0064417] [Medline: 23734200]
68. Sentiment140 general information. URL: http://help.sentiment140.com/ [accessed 2016-09-29] [WebCite Cache ID 6kt10i1Ki]
69. ZIP Code Business Patterns. Washington, DC. US Census Bureau; 2015. URL: http://www.census.gov/newsroom/press-releases/2015/cb15-tps39.html [accessed 2016-09-29] [WebCite Cache ID 6kt1J9t9p]
70. Kuczmarski MF, Kuczmarski RJ, Najjar M. Effects of age on validity of self-reported height, weight, and body mass index: findings from the Third National Health and Nutrition Examination Survey, 1988-1994. J Am Diet Assoc. Jan 2001;101(1):28-34. [doi: 10.1016/S0002-8223(01)00008-6] [Medline: 11209581]
71. Behavioral Risk Factor Surveillance System Survey Data. Atlanta, GA. Centers for Disease Control and Prevention; 2013. URL: http://www.cdc.gov/brfss/ [accessed 2016-09-29] [WebCite Cache ID 6kt2xfrVk]
72. Utah Behavioral Risk Factor Surveillance System Survey Data. Office of Public Health Assessment, Utah Department of Health; 2014. URL: http://health.utah.gov/opha/OPHA_BRFSS.htm [accessed 2016-10-06] [WebCite Cache ID 6l4AUrogD]
73. Geographic terms and concepts: census tract. US Census Bureau; Jan 06, 2012. URL: https://www.census.gov/geo/reference/gtc/gtc_ct.html [accessed 2016-10-02] [WebCite Cache ID 6kxgOkozH]
74. Myslín M, Zhu S, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res. 2013;15(8):e174. [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]

XSL•FO
RenderX

75. Paul MJ, Derdze M. You are what you tweet: analyzing Twitter for public health. 2011. Presented at: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media; 2011; Barcelona, Spain.
76. De Choudry M, Sharma E, Kiciman E. Characterizing dietary choices, nutrition, and language in food deserts via social media. 2016. Presented at: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing; 2016;1157-1170; San Francisco, CA.
77. McEwen BS. Stress, adaptation, and disease: allostasis and allostatic load. Ann NY Academy Sci. 1998;840(1):33-44.
78. Seeman TE. Price of adaptation: allostatic load and its health consequences. Arch Intern Med. 1997;157(19):2259-2268.
79. Duggan M. Social Media Update 2014. Washington, DC. Pew Internet and American Life Project; 2015. URL: http://www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf [accessed 2016-09-29] [WebCite Cache ID 6kt2HVzYR]
80. Difference between sample and filter streaming API. Aug 06, 2016. URL: https://twittercommunity.com/t/diffence-between-sample-and-filter-streaming-api/15094 [accessed 2016-09-29] [WebCite Cache ID 6kt2LR5ve]
81. Burton SH. Right time, right place? Health communication on Twitter: value and accuracy of location information. J Internet Med Res. 2012;14(6).
82. Morstatter F. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. arXiv.5204v1 cs.SI. 2013.:1306.
83. Burfoot C, Baldwin TA. Automatic satire detection: Are you having a laugh? 2009. Presented at: Proceedings of the Association for Computational Linguistics-IJCNLP; 2009; Singapore.
84. Ptácek T, Habernal I, Hong J. Sarcasm Detection on Czech and English Twitter. In: COLING. 2014. Presented at: 25th International Conference on Computational Linguistics; August 23-29 2014;213-223; Dublin, Ireland.
85. Ali MM, Amialchuk A, Heiland FW. Weight-related behavior among adolescents: the role of peer effects. PLoS One. 2011;6(6):e21179. [FREE Full text] [doi: 10.1371/journal.pone.0021179] [Medline: 21731665]
86. Vartanian LR, Sokol N, Herman CP, Polivy J. Social models provide a norm of appropriate food intake for young women. PLoS One. 2013;8(11):e79268. [FREE Full text] [doi: 10.1371/journal.pone.0079268] [Medline: 24236117]
87. Cohen DA, Finch BK, Bower A, Sastry N. Collective efficacy and obesity: the potential influence of social factors on health. Soc Sci Med. Feb 2006;62(3):769-778. [doi: 10.1016/j.socscimed.2005.06.033] [Medline: 16039767]
88. Kim D. US state- and county-level social capital in relation to obesity and physical inactivity: A multilevel, multivariable analysis. Soc Sci Med. 2006;63(4):1045-1059.
89. Berkman L, Syme S. Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents. Am J Educ. 1979;190(2):186-204.

Abbreviations

API: application programming interface
BRFSS: Behavioral Risk Factor Surveillance System
MALLET: Machine Learning for Language Toolkit
Mturk: Mechanical Turk
NAICS: North American Industry Classification System

Edited by G Eysenbach; submitted 01.05.16; peer-reviewed by N Zhang, H Zhai, A Jimeno, A MacKinlay, C Seresinhe; comments to author 27.07.16; revised version received 29.08.16; accepted 15.09.16; published 17.10.16

Please cite as:
Nguyen QC, Li D, Meng HW, Kath S, Nsoesie E, Li F, Wen M
Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity
JMIR Public Health Surveill 2016;2(2):e158
URL: http://publichealth.jmir.org/2016/2/e158/
doi: 10.2196/publichealth.5869
PMID: 27751984

©Quynh C Nguyen, Dapeng Li, Hsien-Wen Meng, Suraj Kath, Elaine Nsoesie, Feifei Li, Ming Wen. Originally published in JMIR Public Health and Surveillance (http://publichealth.jmir.org), 17.10.2016. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Public Health and Surveillance, is properly cited. The complete bibliographic information, a link to the original publication on http://publichealth.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX