Applied Geography 73 (2016) 77–88

Contents lists available at ScienceDirect

Applied Geography

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

Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity

Quynh C. Nguyen [a, *], Suraj Kath [b], Hsien-Wen Meng [a], Dapeng Li [c], Ken R. Smith [d], James A. VanDerslice [e], Ming Wen [f], Feifei Li [b]

[a] Department of Health Promotion and Education, College of Health, University of Utah, Salt Lake City, UT, USA
[b] School of Computing, University of Utah, USA
[c] Department of Geography, University of Utah, USA
[d] Department of Family and Consumer Studies, University of Utah, USA
[e] Division of Public Health, Department of Family and Preventive Medicine, School of Medicine, University of Utah, USA
[f] Department of Sociology, University of Utah, USA

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
17 June 2016
Accepted 19 June 2016

Keywords:
Twitter messaging
Neighborhood
Happiness
Physical activity
Diet
Food

ABSTRACT

Objectives: Using publicly available, geotagged Twitter data, we created neighborhood indicators for happiness, food and physical activity for three large counties: Salt Lake, San Francisco and New York.
Methods: We utilize 2.8 million tweets collected between February–August 2015 in our analysis. Geocoordinates of where tweets were sent allow us to spatially join them to 2010 census tract locations. We implemented quality control checks and tested associations between Twitter-derived variables and sociodemographic characteristics.
Results: For a random subset of tweets, manually labeled tweets and algorithm labeled tweets had excellent levels of agreement: 73% for happiness; 83% for food, and 85% for physical activity. Happy tweets, healthy food references, and physical activity references were less frequent in census tracts with greater economic disadvantage and higher proportions of racial/ethnic minorities and youths.
Conclusions: Social media can be leveraged to provide greater understanding of the well-being and health behaviors of communities—information that has been previously difficult and expensive to obtain consistently across geographies. More open access neighborhood data can enable better design of programs and policies addressing social determinants of health.

1. Introduction

The literature examining neighborhood effects on health has flourished in the last decade (Diez Roux, 2001). Extant research has provided evidence on associations between the neighborhood environment and mortality risk (Eames, Ben-Shlomom, & Marmot, 1993; Morris, Blane, & White, 1996; Townsend, Phillimore, & Beattie, 1988; Tyroler et al., 1993; Waitzman & Smith, 1998a, 1998b; Wing, Barnett, Casper, & Tyroler, 1992), life expectancy (Clarke et al., 2010), mental health (Truong & Ma, 2006), self-rated health (Wen, Browning, & Cagney, 2003), obesity, (Black, Macinko, Dixon, & Fryer, 2010; Heinrich et al., 2008; Mujahid et al., 2008; Smith et al., 2008), and diabetes (Grigsby-Toussaint et al., 2010; Lysy et al., 2013)—even after adjusting for individual characteristics. Poor access to healthy food (Christiansen, Qureshi, Schaible, Park, & Gittelsohn, 2013; Inagami, Cohen, & Finch, 2006; Morland, Wing, Diez Roux, & Poole, 2002; Morland, Wing, Roux, 2002; Stafford, 2007; Wang, Kim, Gonzalez, MacLeod, & Winkleby, 2007), fast food chains (Block, Scribner, & DeSalvo, 2004), the lack of recreational facilities (Brownson, Hoehner, Day, Forsyth, & Sallis, 2009; Roemmich et al., 2006), and higher crime rates (Mujahid et al., 2008; Stafford, 2007) all correlate with higher obesity rates. Community happiness levels also have been inversely related to obesity as well as other outcomes including hypertension, suicide, and life expectancy (Blanchflower & Oswald, 2008; Bray & Gunnell, 2006; Di Tella & MacCulloch, 2008; Dodds, Harris, Kloumann, Bliss, & Danforth, 2011; Oswald & Powdthavee, 2007; Tella, MacCulloch, & Oswald, 2003). Adverse neighborhood conditions concentrate in poor, minority neighborhoods (Black et al., 2010; Diez-Roux,

* Corresponding author. Department of Health Promotion and Education, University of Utah, 1901 E South Campus Drive, Annex B 2124, Salt Lake City, UT 84112, USA.
E-mail address: quynh.nguyen@health.utah.edu (Q.C. Nguyen).

http://dx.doi.org/10.1016/j.apgeog.2016.06.003
0143-6228/

1998; Duncan, Jones, & Moon, 1998; Macintyre, Maciver, & Sooman, 1993), thereby increasing health disparities. Furthermore, the epidemic rise in obesity and related chronic diseases in recent decades signal the importance of structural forces and social processes.

Nonetheless, the dearth of data on contextual factors limits the investigation of multilevel effects on health. Certain places (National Archive of Criminal Justice; Baltimore Neighborhood Indicators Alliance – The Jacob France Institute) have extensive neighborhood data collected on them, but they are the exception rather than the rule, and it is difficult to make comparisons across geographies because available measures vary greatly across them. Also patterns seen in specific places may not apply to other places. For instance, estimates and patterns seen in urban areas may not apply to rural areas. Neighborhood data collection is expensive and time consuming, and then only available for certain places or time periods and become outdated quickly (Peterson and Krivo, 2000). Moreover, while comparable neighborhood data across large areas are highly lacking, the neighborhood data we do have are typically data on compositional characteristics (e.g., percent females) and features of the built environment (e.g., number of grocery stores and health care clinics). These data do not capture the social environment, or an individual's interactions with that environment.

Social processes and networks can affect health via a myriad of mechanisms, such as 1) the maintenance of norms around healthy behaviors via informal social control; 2) the stimulation of new interests such as a new sport or exercise; 3) political advocacy for access to neighborhood amenities and protection against stressors and toxic agents; 4) emotional support; and 5) the dispersal of knowledge about health promotion practices (Ali, Amialchuk, & Heiland, 2011; Berkman & Syme, 1979; Cohen, Finch, Bower, & Sastry, 2006; Kim, Subramanian, Gortmaker, & Kawachi, 2006; Vartanian, Sokol, Herman, & Polivy, 2013). According to Social Learning Theory, learning takes place in a social context (Bandura, 1977). Behaviors are adopted by observing how the behavior is performed by others, attitudes around that behavior, and outcomes associated with that behavior. Empirically, the adoption of specific health behaviors related to food consumption, health screening, smoking, alcohol consumption, drug use, and sleep has been observed to disperse through social networks (Keating, O'Malley, Murabito, Smith, & Christakis, 2011; Mednick, Christakis, & Fowler, 2010; Pachucki, Jacques, & Christakis, 2011; Rosenquist, Murabito, Fowler, & Christakis, 2010; Roy, 2004; Smith & Christakis, 2008). Similarly, evidence suggests that emotional states such as mood (Kramer, Guillory, & Hancock, 2014), happiness (Fowler & Christakis, 2008), depression (Rosenquist, Fowler, & Christakis, 2011), and suicidality (Bearman, & Moody, 2004) can spread through social networks. The measurement of area-level happiness and subjective-well-being is a new and expanding research endeavor (Gallup-Healthways, 2013; Gill, French, Gergle, & Oberlander, 2008; Helliwell, Layard, & Sachs, 2012; Kramer, 2010; Quercia, Ellis, Capraz, & Crowcroft, 2012). For instance, in 2012, the United Nations began its annual release of a World Happiness Report (Helliwell et al., 2012). Social media may influence individuals' health behaviors but may also be a way to characterize prevalent community characteristics and patterns of behaviors.

1.1. Study aims

Given the vast literature documenting the influence of social networks on individual health behaviors and health outcomes, we believe that social media data represent an important new data resource for neighborhood researchers. Thus, using publicly available, geotagged Twitter data, we construct novel indicators of neighborhood happiness levels, healthiness of food, and physical activity. We conduct quality control activities and perform validation analysis comparing Twitter-derived neighborhood indicators to demographic and economic characteristics of the corresponding census tract. In order to test our computer algorithm for constructing neighborhood indicators, we selected three counties that display diversity in regards to geographical location, landscape, housing market, cultural characteristics, and demographic characteristics (e.g., racial/ethnic composition, age distribution, and household size). The three counties are the following: Salt Lake County, San Francisco County, and New York County.

2. Methods

2.1. Social media data collection

From February–August 2015, we utilized Twitter's Streaming Application Programming Interface (API) to continuously collect a random 1% subset of publicly available tweets with latitudes and longitude coordinates. We present in-depth analyses and findings for three counties in the United States: Salt Lake County (367,204 tweets); San Francisco County (same as San Francisco city; 653,670 tweets); and New York County (1,828,026 tweets).

2.2. Spatial join

We linked 99.8% of tweets with available GPS coordinates to their respective 2010 census tract locations. We used Python and relevant GIS libraries (Shapely and Fiona) to accomplish this task. An R-Tree was used to build a spatial index (Guttman, 1984) on census tract polygon data. R-tree indexing allows for faster spatial searches on the data because the R-tree groups data into bounding rectangles and narrows the search space. A query that does not intersect the bounding rectangle cannot intersect any of its component parts. Utilizing the latitude and longitude coordinates of where tweets were sent, spatial joins were performed on tweets to identify the corresponding census tracts. Tweets that were not assigned a census location included those with destinations bordering the United States (i.e., Mexico and Canada).

2.3. Processing tweets

We processed tweets to create variables that measure sentiment, food, and physical activity. To accomplish this task, we utilized a bag-of-words algorithm which creates a simplifying representation of tweets that disregards grammar and word order, but has the capacity to track the frequency of terms or components of tweets, and then performs computations on those components and terms. Several steps were conducted that first included dividing each tweet into tokens (Stanford Natural Language Processing Group). A tokenizer divides text into a sequence of tokens, which roughly correspond to "words." We are using a tokenizer particularly suitable for processing English text called the PTBTokenizer (aka the Stanford Tokenizer). The PTBTokenizer is an efficient, fast, and deterministic tokenizer. It can tokenize text at a rate of about 1,000,000 tokens per second on a standard personal computer. Utilizing heuristics, it can usually differentiate when single quotes are part of words and when periods do and do not imply sentence boundaries. After we obtain the tokens (i.e., individual words) from a tweet, we then search each word in our word dictionary to get its corresponding happiness score for sentiment analysis. Currently, we are ignoring words which are not present in our word dictionary. Using this algorithm, sentiment scores can be assigned to approximately 80–85% of tweets across geographies.

Below we describe in more detail our algorithms for constructing each variable.

2.4. Sentiment analysis

To implement sentiment analysis, we utilized the Language Assessment by Mechanical Turk (LabMT) word list, compiled by obtaining the most frequently occurring words in each of the following four text sources: Google Books (English), music lyrics, the New York Times and Twitter. About 10,000 of these individual words have been scored by users of Amazon's Mechanical Turk service on a scale of 1 (sad) to 9 (happy), resulting in a measure of average happiness for each given word. We extended the LabMT list by adding 158 commonly used emoticons – pictorial representation of a feeling or facial expression expressed through a combination of numbers, letters, and punctuation marks. For instance,:-) represents a smiley face. Happiness values for the emoticons were assigned according to the word meaning of the emoticons.

Average happiness for each tweet was computed by taking the average of happiness values of words composing the tweet, excluding neutral words (e.g., "the," "is") (Mitchell, Frank, Harris, Dodds, & Danforth, 2013). The happiness value for a census tract was then computed as the average of sentiment values of all tweets in that census tract. Our algorithm tracked and accounted for intensifiers (such as the use of exclamation marks and all-capitalized words) in computing sentiment. For tweets with happiness scores above 6, we added 1 for the use of one exclamation mark (7.2% of tweets) and 2 for use of 2 exclamation marks (4.7% of tweets). For tweets with happiness scores less than 4, 1 and 2 was subtracted, respectively. For tweets with at least two words in which all the letters were capitalized (5.7% of tweets), 1 was added to the happiness score of tweets with scores originally above 6 and subtracted from the happiness scores of tweets with original scores below 4. If three or more words were capitalized, 2 was added if the happiness score was originally above 6 and 2 was subtracted if the happiness score was below 4. We did not modify tweets with sentiment scores between 4 and 6 (i.e., neutral tweets) because the directionality of the emphasis was ambiguous. We required that at least two words be capitalized in order to avoid extra weight being given to tweets that utilized capitalized state initials (e.g., CA) or denotations of time (e.g., "AM" or "PM"). Tweets can have use of exclamation marks and at least two all-caps words (1.4% of tweets).

For quality control, we evaluated our sentiment analysis against the ratings of human coders (Snow, O'Connor, Jurafsky & Ng, 2008). We created 20 online surveys consisting of 25 publicly available tweets each (500 tweets in total). The 500 tweets were randomly selected within sentiment groups such that 50% of tweets were happy, 25% were sad and 25% were neutral. We randomly sorted this master list of tweets and created 20 online surveys from the sorted list. Each online survey received 15 responses. Upon completion of the survey (i.e., participant rated all 25 tweets), 25 cents ($0.25) was deposited into the participant's Amazon Mturk account per online survey completed. The study was approved by the University of Utah Institutional Review Board. A total of 32 people participated in our study (participants generally did several batches of tweets rather than just one batch). Study time was four days (April 1 – April 5, 2015). We utilized the human ratings to calibrate the scaling of our sentiment analysis. The bag of words technique utilized by Mitchell and colleagues (Mitchell et al., 2013) categorized happy, sad and neutral tweets as having the following sentiment scores: $\geq$6, $\leq$4, and 4–6. However, after calibration to human ratings such that the proportion of happy, sad and neutral tweets matched human categorization, sentiment scores were categorized as such $\geq$6.6 (30%), $\leq$4.9 (33%), and 4.9–6.6 (37%). The kappa statistics for agreement between human vs. algorithm-derived sentiment ratings was 73%, if we dichotomize sentiment as happy vs. not happy at the modified cut points. Additionally, sensitivity analyses with a different set of weights for capitalizations and exclamation marks (e.g., 0.5 and 1.0, 1.0 and 2.0, and 2.0 and 3.0) resulted in qualitatively similar agreement between human coders and the algorithm (70–71%) among tweets with intensifiers.

2.5. Food analysis

We compiled a list of over 1300 popular food words from the U.S. Department of Agriculture's National Nutrient Database (United States Department of Agriculture). Each food item was associated with a measure of caloric density, operationalized as calories per 100 g. Fruits, vegetables, nuts, and lean proteins (e.g., fish, chicken, and turkey) were additionally labeled as "healthy foods." We excluded fried foods from our count of healthy foods. Our food list also contained 48 popular national fast food restaurants such as McDonald's and Kentucky Fried Chicken to enable quantification of fast food references.

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 computer algorithm first searched over a tweet for matches to two-word foods (e.g., orange chicken). It then searched over the remaining words for matches to one-word food terms (e.g., 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.

We attempted to calculate happiness scores for all tweets, including those involving food, using the algorithm described in our sentiment analysis section. For food tweets, separate variables were calculated for the average sentiment of tweets involving healthy foods and fast food. That is, if a food tweet mentioned a fast food restaurant, its happiness score was included in our calculation of the average sentiment around fast food. Alternatively, if a food tweet mentioned a healthy food, it was included in our calculation of the average sentiment around healthy foods. These variables (i.e., any food references, healthy food references, fast food references, caloric density, and sentiment towards healthy foods and fast food) were then aggregated and summarized at the census-tract level to create neighborhood indicators of food culture.

Quality control activities were implemented to determine whether tweets identified by our algorithm as food-related would be rated as pertaining to food by humans. Two of the co-authors rated a random assortment of 2500 tweets. Initial inter-rater reliability was 92% and discordant values were reviewed until 100% agreement between raters was reached. For tweets that our algorithm had labeled as non-food related, 81% of those labels were correct according to human ratings. For tweets that our algorithm labeled as food related, 83% of those labels were correct. The kappa statistics for agreement between labels produced by our algorithm and human coders was 83%. Our algorithm missed food references if the food term was not included in our food list. We excluded food items with many non-related food meanings such as "perch." For tweets that had been mislabeled as food-related, common reasons included the following: food term utilized in the tweet had a non-food related meaning (e.g., pho fighters); the term was utilized in a food pun or metaphor; or the tweet was a food advertisement.

2.6. Physical activity/recreation 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 (Ainsworth et al., 2011; Zhang et al., 2013). Our physical activity list has 258 different activities that incorporate gym-related exercise (e.g., treadmill, weight lifting), sports (e.g., baseball), recreation (e.g., hiking, scuba diving) and household chores (e.g., gardening). Using Metabolic Equivalents (METs) associated with physical activities, we quantified the caloric expenditure of each physical activity mention, scaled for a duration of 30 min and for a 155 pound individual (Kendall, Hartzler, Klasnja, & Pratt, 2011). Upon piloting our algorithm, we identified commonly utilized phrases or pop culture references that do not involve physical activity (i.e., walking dead, running late, yoga pants) 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 (i.e., baseball, basketball, football, soccer) were about watching games rather than participating in them. Thus, for team sports, we also required that the tweet include the word play, playing, or played. Our algorithm created the following physical activity variables for each tweet: indicator variable for any physical activity mention, indicator variable for gym-related vs. other types of physical activity, sum of caloric expenditure of mentioned activities, and sentiment around physical activity.

Quality control activities were implemented to determine accuracy of tweets that had been labeled by our algorithm as being related to physical activity. Two annotators labeled a random assortment of 2500 tweets with an initial kappa agreement of 95%. Again discordant values were reviewed until 100% inter-rater agreement was reached. For tweets that the computer algorithm had labeled as not related to physical activity, 97% of those labels were correct according to human ratings. For tweets that our algorithm labeled as related to physical activity, 82% of those labels were correct. Kappa agreement between labels generated by our algorithm and human coders was 85%. Common classification errors included the following: 1) tweet was about watching a professional sport rather than participating in it and 2) tweet utilized a term that took on a non-physical activity related meaning (e.g., running a fever).

2.7. Census tract characteristics

To examine how Twitter-derived neighborhood variables relate to more traditional neighborhood variables, we merged our social media dataset with the 2010 Census and 2013 American Community Survey data which comprised the following demographic, household and economic characteristics: percent 65 years+; percent 10–24 years; percent male; percent African American;

Table 1
Descriptive statistics of tract-level characteristics.

	San Francisco County N = 196)	Salt Lake County (n = 212)	New York County (n = 288)
	Mean (Standard deviation)	Mean (Standard deviation)	Mean (Standard deviation)
Happiness scores (continuous)	6.4 (0.2)	6.2 (0.2)	6.5 (0.2)
Categorized happiness scores			
Percent happy	46.2 (9.0)	37.9 (6.4)	52.3 (8.6)
Percent neutral	43.4 (6.0)	48.1 (4.3)	39.0 (5.8)
Percent sad	10.4 (4.8)	14.0 (3.7)	8.8 (5.0)
Food culture			
Percent of tweets that have food references	6.6 (4.1)	3.1 (1.8)	6.6 (3.8)
Percent of tweets that have healthy food references	1.1 (0.8)	0.5 (0.4)	1.1 (0.7)
Percent of tweets that have fast food restaurant references	0.4 (0.8)	0.3 (0.4)	0.6 (1.0)
Calories density of food mentions (calories per 100 g)	260.3 (71.3)	261.2 (48.9)	249.6 (56.0)
Sentiment of healthy foods	6.6 (0.4)	6.5 (0.5)	6.7 (0.4)
Sentiment of fast food	6.3 (0.8)	6.4 (0.7)	6.5 (0.5)
Physical activity culture			
Percent of tweets with physical activity references	1.7 (1.4)	1.7 (1.3)	1.7 (1.1)
Caloric expenditure of physical activity references (calories per 30 min)	218.5 (47.2)	216.8 (38.5)	218.6 (46.0)
Physical activity is gym-related	17.0 (17.1)	13.3 (12.7)	23.5 (16.4)
Sentiment of physical activity	6.5 (0.4)	6.4 (0.3)	6.6 (0.3)
Census 2010 and ACS 2013 Data			
Population characteristics			
Percent 65 years+	13.8 (7.0)	9.3 (4.8)	13.1 (6.8)
Percent 10–24 years	14.8 (7.5)	22.5 (3.7)	17.6 (9.0)
Percent male	51.0 (5.7)	50.4 (3.5)	47.9 (5.7)
Percent African American	6.6 (9.6)	1.7 (1.5)	17.1 (21.5)
Percent white	50.2 (22.2)	81.5 (12.1)	56.9 (27.0)
Percent Hispanic	14.8 (11.8)	16.5 (13.7)	23.4 (23.1)
Percent relatives (besides spouse and children) living in households	8.0 (6.5)	7.0 (3.3)	5.7 (5.3)
Percent unmarried partner living in households	4.2 (2.5)	2.1 (1.2)	3.5 (1.7)
Household characteristics			
Household size	2.4 (0.7)	3.0 (0.7)	2.0 (0.5)
Percent single female-headed households	9.5 (7.6)	5.9 (2.5)	13.0 (12.1)
Percent householder living alone	36.1 (15.9)	21.7 (13.0)	44.9 (12.6)
Economic characteristics			
Percent owner-occupied housing	37.7 (23.0)	67.7 (21.9)	21.7 (18.4)
Median family income	98893.2 (49650.0)	64267.6 (24449.0)	117429.9 (76118.1)
Percent college graduates	52.1 (21.5)	20.6 (9.5)	58.3 (25.7)
Unemployment rate	8.7 (4.6)	7.7 (3.6)	9.4 (8.3)
Percent less than high school graduate	8.6 (12.1)	15.2 (10.9)	11.4 (14.3)
Percent families living in poverty	8.6 (8.3)	9.9 (8.2)	13.0 (12.9)

All variables are standardized to have a mean of 0 and standard deviation of 1.

percent white; percent Hispanic; percent in households with relatives (other than spouse and children); percent in households with unmarried partner; household size; percent single female-headed households; percent householder living alone; percent owner-occupied housing; median family income; percent college graduates; percent unemployed; percent with less than a high school degree; percent families living in poverty (Table 1).

2.8. Analytic approach

We grouped tweets by census tract and created indicators of happiness, food, and physical activity for each census tract. We then merged our tract-level social media database with census data (see above). To facilitate interpretation of findings for different variables, we standardized all variables to have a mean of zero and standard deviation of one. We implemented linear regression models to examine associations between Twitter-derived neighborhood variables and census tract characteristics for three demonstrative counties: Salt Lake, San Francisco, and New York. In addition, we mapped the distribution of happiness scores across census tracts in the three counties. We ranked food and physical activity terms by popularity, and reported the most popular terms.

3. Results

Across the three counties, tweets were more likely to be neutral or positive rather than negative in sentiment (Table 1). Prevalence of happy tweets was highest in New York, followed by San Francisco and Salt Lake counties. About 3.1–6.6% of the tweets were food-related. Of these food tweets, about 16–17% mentioned healthy foods and 6–10% mentioned fast food restaurants. The mean (standard deviation) caloric density of food references was 250–261 calories (per 100 g). Food tweets that did not include fast food were slightly happier than those with fast food references ($p < 0.01$). About 1.7% of tweets were about physical activity, and most physical activity references were not gym-related (e.g., skiing, hiking)—although higher proportions of gym-related activities were reported in New York than San Francisco and Salt Lake counties. The mean caloric expenditure of physical activity references was 217–219 calories (assuming duration of 30 min for a 155 pound person). Positive sentiment around physical activity was higher in New York compared to Salt Lake County ($p < 0.001$).

The demographic and economic composition of residents differed in the three counties. More youth were present in Salt Lake County than the other two (Table 1). Salt Lake County had fewer racial/ethnic minorities and larger household sizes. Median family incomes and percent college graduates were much higher in New York and San Francisco than in Salt Lake County. New York had a higher poverty rate than Salt Lake or San Francisco.

While possible happiness scores can range from 1 to 9, at the census tract level, most locations had scores that varied between 6.0 and 7.0 (Figs. 1–3). The majority of areas in Salt Lake County had

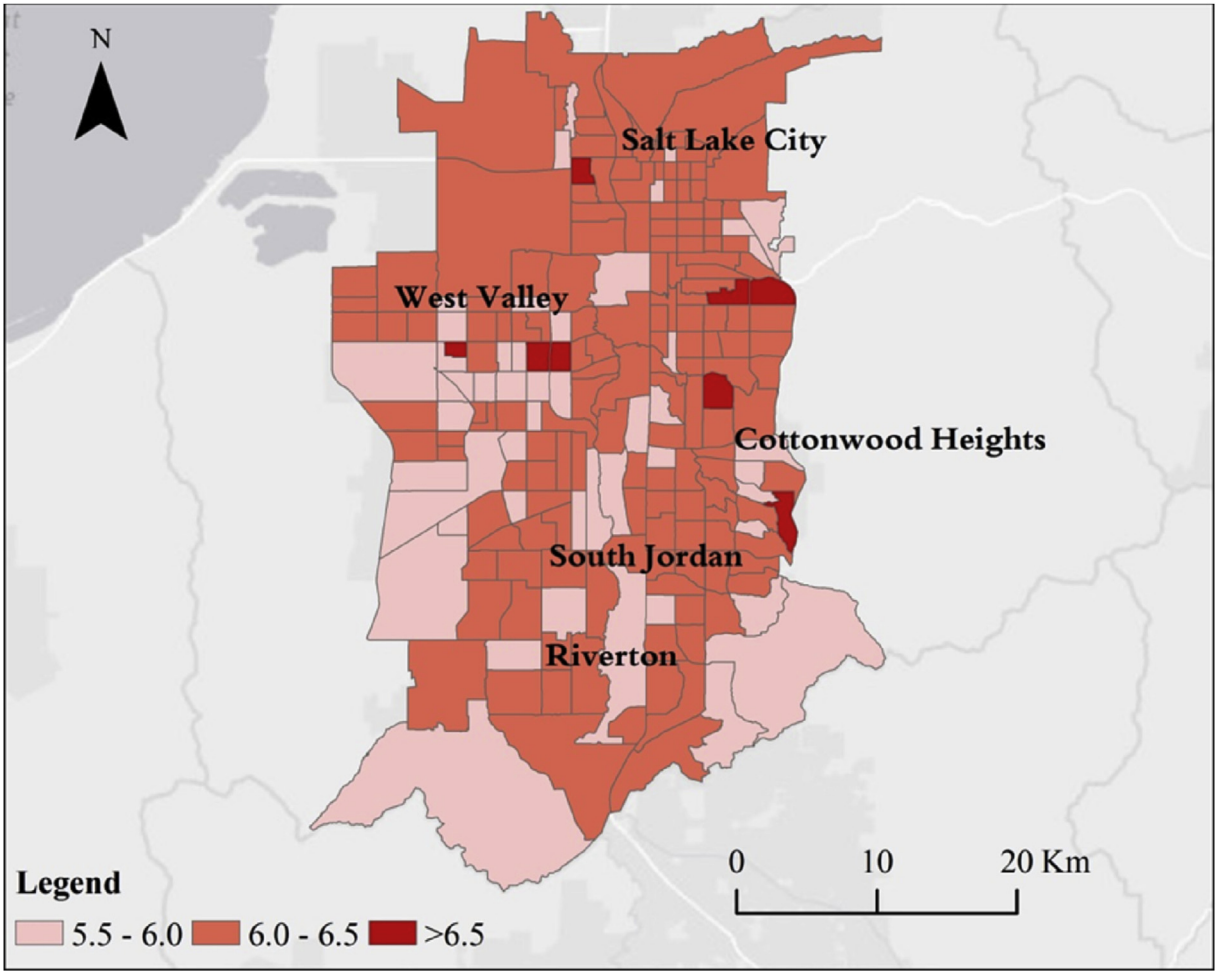

Fig. 1. Geography of happiness scores across census tracts in Salt Lake County.

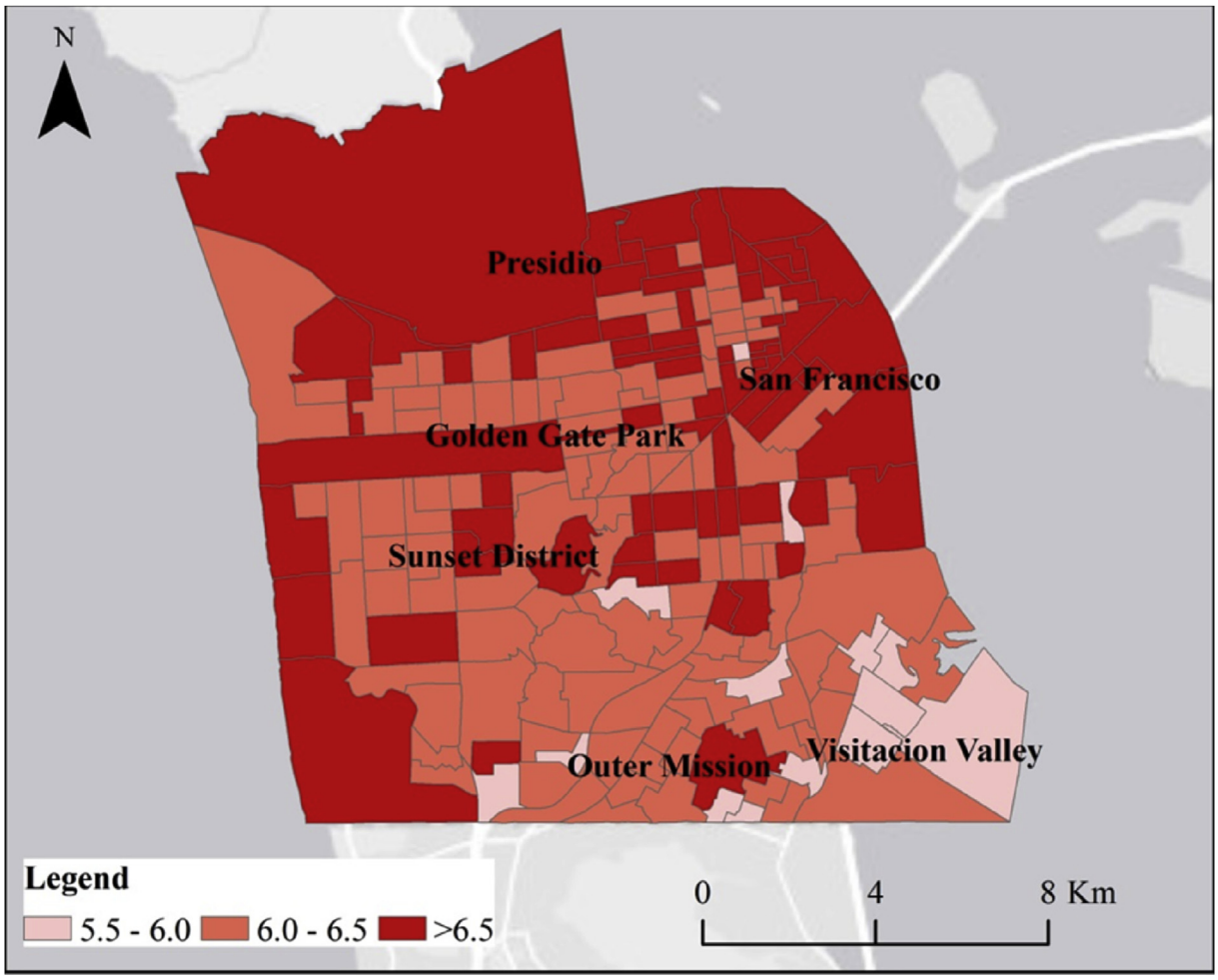

Fig. 2. Geography of happiness scores across census tracts in San Francisco County.

sentiment scores between 6.0 and 6.5, although lower sentiment scores were more prevalent on the west and south sides (Fig. 1). For San Francisco, happiness scores were highest near areas bordering the bay or ocean (e.g., Presidio, Embarcadero) as well as areas near Golden Gate Park and the Mission District (Fig. 2). Of the three counties, the proportion of tweets characterized as happy was highest in New York County and hotspots of happiness were found near Central Park, Meatpacking District, Greenwich Village, Tribeca, Soho and also Wards Island and upper areas like Inwood (Fig. 3).

We conducted crude analyses in which each demographic and economic characteristic was used as a regressor for standardized tract level happiness. Higher proportions of whites and college graduates were positively associated with higher tract happiness levels (eTable 1). Conversely, higher proportion of young people aged 10–24 years, Hispanics, single female-headed households, and higher unemployment rate, poverty rate, and household sizes were associated with lower happiness.

We additionally implemented adjusted analyses controlling for all predictors simultaneously. However, due to high collinearity of census tract characteristics, we reduced the number of variables included in the adjusted models and factor-analyzed four highly correlated characteristics (i.e., percent female-headed households, percent families living in poverty, unemployment rate, and median family income) to create an economic disadvantage factor score. In San Francisco, percent African American and greater household size were related to lower happiness scores (Table 2). In Salt Lake County, higher percentages of elderly individuals and males were associated with higher happiness scores while greater household size, percent Hispanic, and economic disadvantage were related to lower happiness scores—although the latter two characteristics bordered statistical significance (Table 2). In New York City, proportion of young people aged 10–24 years was linked to lower happiness scores. Sensitivity analyses using dichotomized measures of sentiment (i.e., happy vs. not happy; sad vs. not sad) resulted in finding the same statistically significant associations with census tract characteristics as compared to analyses utilizing continuous happiness scores (not shown).

While our food list consisted of over 1300 food terms, 50% of food tweets could be characterized by approximately 25 of the most popular food terms (Table 3). Across the three counties, top food terms included coffee, beer, tea, wine, pizza, burger, ice cream, chocolate, cake, chicken, sushi, and salad among others. In crude analyses, healthy food references increased with higher proportions of unmarried cohabitating adults, householder living alone, and college graduates (eTable 2). There is also some suggestion that healthy foods references increase with higher proportion of whites. Healthy food references were less frequent in tracts with higher proportions of individuals aged 10–24 years, larger household sizes, and more economic disadvantage.

In adjusted analyses, increasing household size and greater proportions of individuals aged 10–24 were associated with less frequent healthy food references (Table 4). In San Francisco, we also

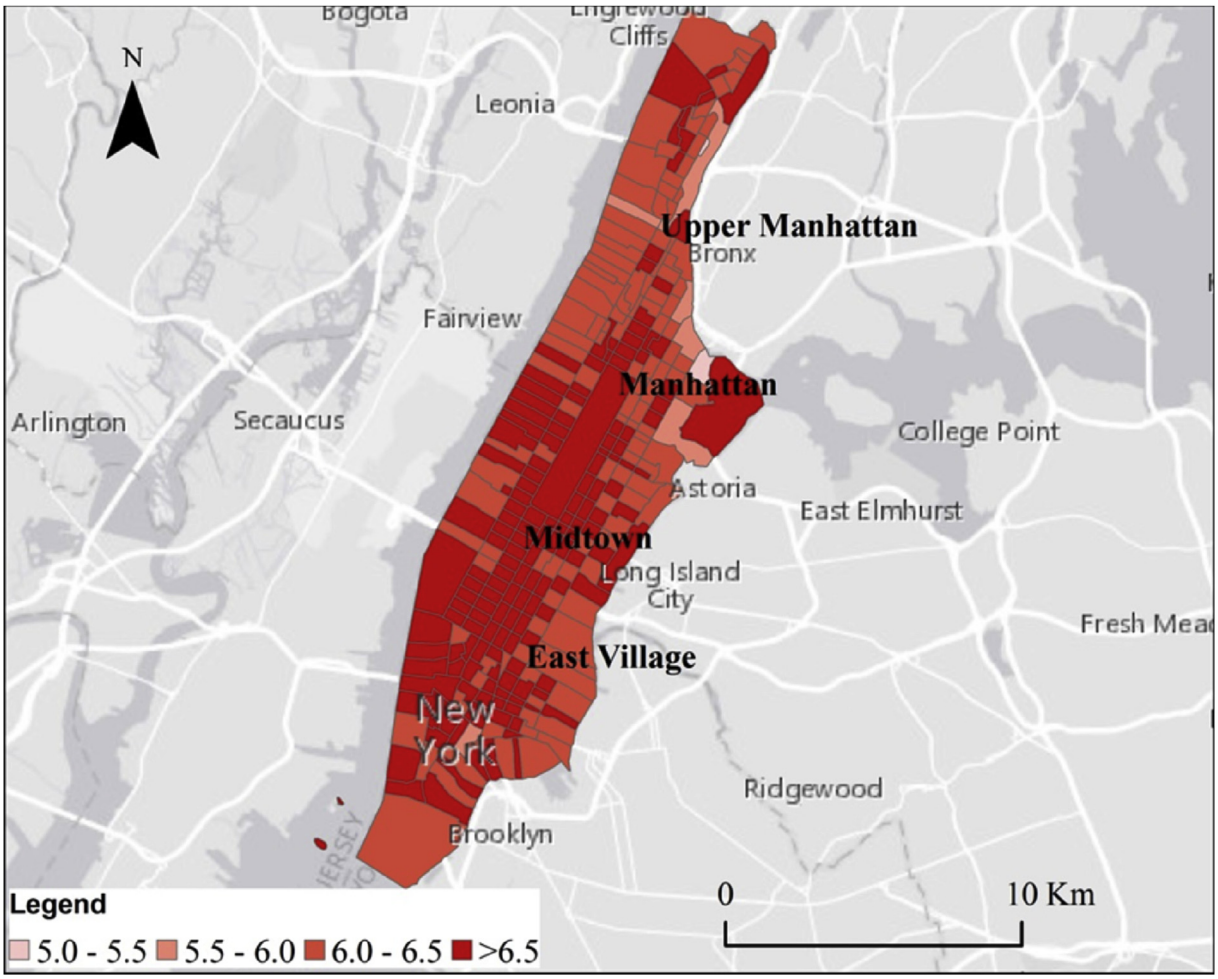

Fig. 3. Geography of happiness scores across census tracts in New York County.

saw that greater percent of elderly, males, and African Americans were related to fewer healthy food references (Table 4). No socio-demographic characteristics statistically significantly predicted sentiment around healthy foods (not shown). In New York County, greater household size (B = −0.43; p = 0.02) and economic disadvantage (B-0.45; p = 0.04) predicted less positive sentiment around fast food. Additionally, in New York (B = −0.55; p < 0.002) economic disadvantage was linked to lower frequency of fast food tweets. However, in Salt Lake County, economic disadvantage was marginally related to more fast food tweets (B = 0.24; p = 0.09).

Although our physical activities list comprised more than 250 terms, across the three counties, 75% of physical activity references could be characterized by about 10 of the most popular physical activity terms. Top physical activity terms included the following:

Table 2
Census tract demographic and economic predictors of happiness scores.

Census tract characteristics	San Francisco County	Salt Lake County	New York County
	Beta (95% CI)	Beta (95% CI)	Beta (95% CI)
Population characteristics			
Percent 65 years+	0.00 (−0.12, 0.13)	0.18 (0.01, 0.35)*	−0.09 (−0.22, 0.03)
Percent 10–24 years	−0.08 (−0.23, 0.06)	0.02 (−0.13, 0.17)	−0.22 (−0.35, −0.09)*
Percent male	0.03 (−0.11, 0.17)	0.26 (0.05, 0.48)*	−0.04 (−0.21, 0.13)
Percent African American	−0.58 (−0.88, −0.28)*	0.17 (−0.04, 0.38)	−0.07 (−0.21, 0.07)
Percent Hispanic	−0.01 (−0.26, 0.23)	−0.24 (−0.49, 0.01)#	−0.11 (−0.30, 0.08)
Household size	−0.44 (−0.57, −0.30)*	−0.26 (−0.47, −0.05)*	−0.03 (−0.33, 0.27)
Economic disadvantage[a]	0.10 (−0.11, 0.32)	−0.20 (−0.45, 0.04)#	−0.20 (−0.50, 0.09)
Adjusted r-square	0.37	0.28	0.23

N = 196 census tracts in San Francisco County, N = 212 tracts in Salt Lake County, and N = 288 tracts in New York County.
All variables are standardized to have a mean of 0 and standard deviation of 1.
Adjusted linear regression models includes all predictors simultaneously, and were run separately for each county.
*p < 0.05 #p < 0.10.
[a] Economic disadvantage factor score derived from the following census tract characteristics: percent female-headed households, percent families living in poverty, unemployment rate, and median family income (reverse coded).

Table 3
Most popular food items, in descending order of popularity.

San Francisco County	Salt Lake County	New York County
Coffee	Coffee	Coffee
Beer	Beer	Starbucks
Starbucks	Pizza	Beer
Wine	Starbucks	Pizza
Tea	Sushi	Wine
Pizza	Chicken	Ice cream
Ice cream	Ice cream	Burger
Sushi	Chocolate	Cheese
Chocolate	Cake	Bbq
Burger	Cookies	Chicken
Chicken	Wine	Tea
Cheese	Cheesecake	Chocolate
Bacon	Burger	Sushi
Crab	Tea	Salad
Bbq	Tacos	Cake
Salad	Salad	Lobster
Ramen	Eggs	Milk
Shrimp	Fries	Ramen
Burrito	Whiskey	Steak
Milk	Bbq	Tacos
Tacos	Cheese	Bacon
Steak	Burrito	Chipotle
Cake	Taco bell	Egg
Egg	Donut	Shrimp
Pork	Pie	
	Milk	
	Chips	
	Cheese	
	Egg	
	McDonalds	
	Pancakes	
	Rice	
	Donuts	

These food items compose 50% of all food references.

walking, running, dance, golf, hiking, skiing, yoga, and workout (Table 5). In crude analyses, predictors of tweets involving physical activity were similar between San Francisco and Salt Lake counties – with percent aged 10–24 years, percent Hispanics, and percent in households with relatives associated with lower frequency of physical activity tweets. In San Francisco and Salt Lake counties, percent college graduates was associated with higher frequency of physical activity tweets. In New York County higher median income was associated with higher frequency of physical activity tweets (eTable 3). In adjusted models for physical activity, higher

Table 5
Most popular physical activity terms, in descending order of popularity.

San Francisco County	Salt Lake County	New York County
Walk/walking	Walk/walking	Walk/walking
Dance/dancing	Dance/dancing	Dance
Running	Hike/hiking	Running
Yoga	Running	Yoga
Hike/hiking	Work out/workout	Workout
Workout	Golf	(Swimming) pool
(Swimming) pool	Skiing	Dancing
Golf	(Swimming) pool	Golf
Swimming	Yoga	
	Swim	
	Bowling	

These terms compose 75% of all physical activity references.

proportions of individuals aged 65 + were associated with higher frequency of physical activity tweets in San Francisco. In Salt Lake and New York counties, economic disadvantage was related to lower frequency of physical activity tweets (Table 6). Larger household size was associated with lower positive sentiment around physical activity in San Francisco (B = −0.21, p = 0.02) and Salt Lake (B = −0.24; p = 0.05) counties. We also examined predictors of caloric expenditure of activities mentioned in tweets; higher percent of males (New York: B = −0.27; p = 0.01; San Francisco: B = −0.15; p = 0.10) and economic disadvantage (New York: B = −0.49; p = 0.005; San Francisco: B = −0.23; p = 0.10) were related to lower caloric expenditure. Additionally, in New York greater household size (B = 0.60; p = 0.001) was related to higher caloric expenditure.

4. Discussion

Social media is a massive data resource that is beginning to be leveraged for health research such as the prediction of flu (Centers for Disease Control and Prevention), predicting the onset of depression (De Choudhury, Gamon, Counts, & Horvitz, 2013), modeling outbreaks (Evans, Fast, & Markuzon, 2013), characterization of emergency response (Lamb, Paul, & Dredze, 2012), investigating spatial patterns in obesity-related tweets and their proximity to McDonalds (Ghosh & Guha, 2013), and identifying associations between healthful tweets and presence of green retailers (Chen & Yang, 2014). In this study, across three different counties in the United States, we generally find that happy tweets,

Table 4
Census tract predictors of tweets with healthy food references.

Census tract characteristics	San Francisco County	Salt Lake County	New York County
	Beta (95% CI)	Beta (95% CI)	Beta (95% CI)
Population characteristics			
Percent 65 years+	−0.24 (−0.39, −0.09)*	−0.15 (−0.35, 0.05)	−0.13 (−0.28, 0.01)#
Percent 10–24 years	−0.24 (−0.42, −0.07)*	−0.05 (−0.22, 0.12)	−0.15 (−0.30, 0.00)*
Percent male	−0.20 (−0.37, −0.03)*	−0.06 (−0.30, 0.19)	−0.08 (−0.27, 0.11)
Percent African American	−0.44 (−0.80, −0.08)*	−0.06 (−0.31, 0.18)	−0.01 (−0.17, 0.14)
Percent Hispanic	0.22 (−0.07, 0.51)	−0.01 (−0.29, 0.28)	−0.16 (−0.38, 0.06)
Household size	−0.25 (−0.41, −0.09)*	−0.35 (−0.59, −0.11)*	−0.10 (−0.44, 0.24)
Economic disadvantage[a]	0.14 (−0.12, 0.40)	0.00 (−0.28, 0.28)	−0.01 (−0.35, 0.33)
Adjusted r-square	0.14	0.05	0.08

N = 196 census tracts in San Francisco County, N = 212 tracts in Salt Lake County, and N = 288 tracts in New York County.
All variables are standardized to have a mean of 0 and standard deviation of 1.
Adjusted linear regression models includes all predictors simultaneously, and were run separately for each county.
*p < 0.05 #p < 0.10.
[a] Economic disadvantage factor score derived from the following census tract characteristics: percent female-headed households, percent families living in poverty, unemployment rate, and median family income (reverse coded).

Table 6
Census tract predictors of tweets with physical activity references.

Census tract characteristics	San Francisco County	Salt Lake County	New York County
	Beta (95% CI)	Beta (95% CI)	Beta (95% CI)
Population characteristics			
Percent 65 years+	0.25 (0.10, 0.41)*	−0.05 (−0.24, 0.14)	−0.01 (−0.16, 0.14)
Percent 10–24 years	−0.02 (−0.21, 0.16)	−0.06 (−0.22, 0.11)	−0.15 (−0.30, 0.01)#
Percent male	0.07 (−0.10, 0.25)	0.06 (−0.18, 0.30)	−0.12 (−0.32, 0.07)
Percent black	−0.05 (−0.43, 0.32)	0.07 (−0.17, 0.31)	0.05 (−0.11, 0.21)
Percent Hispanic	−0.06 (−0.37, 0.25)	0.08 (−0.20, 0.36)	0.01 (−0.22, 0.23)
Household size	−0.03 (−0.20, 0.14)	−0.20 (−0.44, 0.04)#	0.37 (0.01, 0.72)*
Economic disadvantage[a]	−0.03 (−0.29, 0.24)	−0.40 (−0.68, −0.12)*	−0.29 (−0.64, 0.06)#
Adjusted r-square	0.05	0.06	0.02

N = 196 census tracts in San Francisco County, N = 212 tracts in Salt Lake County, and N = 288 tracts in New York County.
All variables are standardized to have a mean of 0 and standard deviation of 1.
Adjusted linear regression models includes all predictors simultaneously, and were run separately for each county.
*p < 0.05 #p < 0.10.
[a] Economic disadvantage factor score derived from the following census tract characteristics: percent female-headed households, percent families living in poverty, unemployment rate, and median family income (reverse coded).

healthy food references, and physical activity references were less frequent in census tracts with greater economic disadvantage and higher proportions of racial/ethnic minorities and youths. However, we did not find consistent relationships between economic disadvantage and fast food tweets.

Nascent work suggests that social media interactions may serve to share knowledge, provide support, and influence behaviors and mental states including happiness (Bliss, Kloumann, Harris, Danforth, & Dodds, 2012; Coulson & Knibb, 2007; Coulson, Buchanan, & Aubeeluck, 2007). Our algorithms demonstrated levels of accuracy comparable to other natural language processing techniques (Pak & Paroubek, 2010; Verma et al., 2011). Nonetheless, while sentiment analysis is an emerging endeavor in computer science and natural language processing, very few studies have been conducted on the classification of food and physical activity references using social media. Additionally, our algorithms are unique in creating indicators at the neighborhood level. The vast majority of Twitter analyses are conducted at the city, county or state levels. We plan to make our database, *HashtagHealth*, available to other public health researchers upon one full year of data collection.

Our sentiment analysis was adapted from methodology developed by Dodds and colleagues (Dodds et al., 2011). and further tested by Mitchell and colleagues (Mitchell et al., 2013). Our sentiment analysis extended the methodology by accounting for emoticons and use of capitalizations and exclamation marks in estimating the sentiment of a tweet. Mitchell and colleagues found that higher socioeconomic status was associated with higher happiness scores at the city level. Moreover, they found a mild ($r = -0.34$) correlations between happiness and obesity rates for 190 Metropolitan Statistical Areas (MSAs) (Mitchell et al., 2013) and that happiness scores 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) (Mitchell et al., 2013). Widener and Li conducted Twitter analysis of food references and found that disadvantaged areas had fewer positive references for fruits and vegetables (Widener & Li, 2014). Ghosh and Guha found a positive correlation between obesity prevention themed tweets and number of policies related to obesity, nutrition and physical activity at the state level (Ghosh & Guha, 2013). They also found a strong positive correlation between tweets about high calorie foods/obesity and locations of McDonalds (Ghosh & Guha, 2013). Zhang and colleagues manually coded a random assortment of 30,000 tweets about physical activity, finding that most were neutral in sentiment and only 9.0% offered social support (Zhang et al., 2013).

4.1. Study strengths and limitations

Our Twitter database leverages publicly exhibited text to construct indicators of the social environment previously not available to neighborhood effects researchers. This project is made possible through collaborations between public health, social sciences, and computer science. Using social media for neighborhood research purposes has substantial strengths including the public nature of tweets, the regularity and pervasiveness of tweets, which enables easy updating of constructed indicators, and participation from increasing segments of the population. Although we utilized Twitter as our main source of social media data, Twitter users with multiple social media accounts such as Facebook, Instagram and Vine will often link their accounts to display their posts on multiple platforms. We found tweets often incorporated links to Facebook, Instagram and Vine posts.

Nonetheless, social media as a data resource is not without limitations which include an over-representation of young individuals. While more users of social media tend to be younger, adoption rates have been steadily increasing over the years for all age groups and as of 2014, usage of social networking sites ranged from 89% among 18–29 year olds to 49% among those 65 years and older (Duggan, Ellison, Lampe, Lenhart, & Madden, 2015). Moreover, arguably a person can be affected by social media even if he or she is not an active user, for instance, through discussions by friends and colleagues or headlines in the news that now regularly feature social media content. 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.

Nevertheless, social media researchers can only harvest and utilize information that people are willing to share. This information includes their activities, intentions, and opinions. Also, our construction of neighborhood indicators from Twitter data necessitated that we restrict our data collection to geotagged tweets (i.e., tweets in which users enabled location on their mobile phones). Previous studies suggest that about 1–2% of tweets may contain GPS location information, and that use of Twitter's Streaming API may obtain 90% of all geotagged tweets (in comparison to Twitter's Firehouse which returns all tweets) (Burton, Tanner, Giraud-Carrier,

West, & Barnes, 2012; Morstatter, Pfeffer, Liu, & Carley, 2013). Tweets with location information may differ from those who do not. For instance, our body of tweets may have a higher proportion of public and social activities such as friends tweeting from a restaurant or an event.

Moreover, tweets were spatially joined to the location of where they were sent, not the user profile location (which is a "text entry" of home city, state, or country). Thus, tweets aggregated to a census tract can be composed of those written by people visiting the area and those living there. We believe that including data from visitors is a strength of the study design because people can be mobile throughout the course of the day and regardless of their residency (Luo, Cao, Mulligan, & Li, 2015), people can directly influence the social environment in which they are in physical proximity. Additionally some places such as New York City's Times Square and Fisherman's Wharf are characterized by the large influx of visitors. A potential concern is that the demographic or economic characteristics of visitors may differ from the compositional characteristics of residents. However, people may tend to visit areas that are aligned with their own background characteristics (i.e., assortative sorting) as suggested by the literature on neighborhood segregation (Keels et al. 2005; Cutler et al., 2008). Moreover, the data is naturally weighted towards residents or people who spend greater amounts of time in a census tract (given that we are collecting a random subset of tweets and people who spend more time in a census tract are more likely to have their tweet included in our dataset). Nonetheless, the influence of a tweet can, of course, extend much further than the actual physical location of where the tweet was sent.

An additional complexity is that the location of where the tweet was sent may differ from where the sentiment, event or activity originated. For instance, frustrations about work could be expressed later in the day via a tweet at a bar. Thus having only access to the geographic information of where the tweet sent may mean that we are obtaining only partial information of the all possible geographies described in the tweet. This potential "migration bias" may reduce the strength of observed associations.

Moreover, taking census tracts as the neighborhood unit can be problematic given the variation in population density and physical size. However, census tracts are useful neighborhood approximations given that they are designed to be relatively homogenous with respect to sociodemographic composition and living conditions. Additionally, publicly available administrative data is available at the census tract level. Aggregating tweets to the census tract level also ameliorates concerns about privacy because individual-level tweets and any identifying information are not made available. Our neighborhood dataset contains only summaries of group-level characteristics. In addition, geotagged data were voluntarily contributed by Twitter users via an opt-in mechanism on the Twitter mobile app.

In creating our neighborhood indicators from Twitter data, we prioritized transparency and ease of implementation so that other researchers can replicate our database or refine it for their specific purposes. Our algorithms implemented corpus-based classification, with steps that are easily understandable. Furthermore, this algorithm is efficient in terms of computer processing speed. However, this technique does not take into account the entire context of a tweet and does not account for sarcasm or humor, challenges which still evade the majority of natural language processing algorithms. Our analysis of caloric density of food assumes calories per 100 g. Most tweets do not specify the exact quantity of food consumed and thus our estimate is just an approximation. Also, caloric expenditure for physical activities was assumed for 30 min of physical activity for an individual weighing 155 pounds, which can be an under- or over-estimate depending on the type of activity and persons engaged in that activity.

4.2. Conclusion

There is a growing interest in social determinants of health and how place-based factors affect health. However, one tremendous impediment to such research endeavors includes the lack of open access neighborhood data beyond census demographic and economic characteristics. Moreover, when neighborhood data is available it is inconsistent across geographies and time, thus making neighborhood comparisons difficult, expensive, and time-consuming. Widespread usage of the internet and open recording of many transactions has led to the availability of massive amounts of data that enable understanding of previously hidden micro-level interactions. Social media data uniquely allows us to capture characteristics of the social environment that are understudied. For instance, public posts allow us to potentially measure prevailing sentiment among the people in an area. Additionally public posts about food and physical activity can help us understand local area social norms and the potential impacts of social modeling of health behaviors.

In this study, we describe algorithms and quality control results for measuring neighborhood characteristics from publicly available, geotagged Twitter data. We demonstrate that tweets can provide a means to assess prevalent sentiment and popularity of types of foods and physical activities in communities, which can enable more efficient targeting of programs and policies to meet the needs of different neighborhoods. In particular, as this study suggests, neighborhoods with social and economic disadvantage as well as those with high percentages of youth may have greater need for health promoting resources.

Acknowledgements

This work was supported by the following grants: Dr. Quynh Nguyen was PI on NIH grant 5K01ES025433; Dr. Feifei Li was supported by NSF grants 1443046 and 1251019, and in part by NSFC grant 61428204 and a Google research award. The funding sources did not have any involvement in the study design; collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.apgeog.2016.06.003.

References

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett, D. R., Tudor-Locke, C., et al. (2011). Compendium of physical activities: A second update of codes and MET values. *Medicine & Science in Sports & Exercise, 43*(8), 1575–1581.

Ali, M. M., Amialchuk, A., & Heiland, F. W. (2011). Weight-Related Behavior among Adolescents: The Role of Peer Effects. *PLoS ONE, 6*(6), e21179.

Baltimore Neighborhood Indicators Alliance – The Jacob France Institute. Vital Signs 11 Reports. http://bniajfi.org/vs11_report Accessed 24.09.13.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, New Jersey: Prentice-Hall.

Bearman, P. S., & Moody, J. (2004). Suicide and friendships among American adolescents. *American Journal of Public Health, 94*(1), 89–95.

Berkman, L., & Syme, S. (1979). Social networks, host resistance, and mortality: A nine-year follow-up study of alameda county residents. *American Journal of Education, 190*(2), 186–204.

Black, J. L., Macinko, J., Dixon, L. B., & Fryer, J. G. E. (2010). Neighborhoods and obesity in New York city. *Health & Place, 16*(3), 489–499.

Blanchflower, D. G., & Oswald, A. J. (2008). Hypertension and happiness across nations. *Journal of Health Economics, 27*(2), 218–233.

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., & Dodds, P. S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science, 3*(5), 388–397.

Block, J. P., Scribner, R. A., & DeSalvo, K. B. (2004). Fast food, race/ethnicity, and income: A geographic analysis. *American Journal of Preventive Medicine, 27*(3), 211–217.
Bray, I., & Gunnell, D. (2006). Suicide rates, life satisfaction and happiness as markers for population mental health. *Social Psychiatry and Psychiatric Epidemiology, 41*(5), 333–337.
Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical Activity: State of the science. *American Journal of Preventive Medicine, 36*(Suppl. 4), S99–S123. e12.
Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., & Barnes, M. D. (2012). Right time, right place" health communication on twitter: Value and accuracy of location information. *Journal of Medical Internet Research, 14*(6).
Centers for Disease Control and Prevention. CDC Competition Encourages Use of Social Media to Predict Flu. http://www.cdc.gov/flu/news/predict-flu-challenge.htm Accessed 02.02.14.
Chen, X., & Yang, X. (2014). Does food environment influence food choices? A geographical analysis through "tweets". *Applied Geography, 51*, 82–89.
Christiansen, K. M. H., Qureshi, F., Schaible, A., Park, S., & Gittelsohn, J. (2013). Environmental factors that impact the eating behaviors of low-income african american adolescents in Baltimore city. *Journal of Nutrition Education and Behavior, 45*(6), 652–660.
Clarke, C. A., Miller, T., Chang, E. T., Yin, D., Cockburn, M., & Gomez, S. L. (2010). Racial and social class gradients in life expectancy in contemporary California. *Social Science & Medicine, 70*(9), 1373–1380.
Cohen, D. A., Finch, B. K., Bower, A., & Sastry, N. (2006). Collective efficacy and obesity: The potential influence of social factors on health. *Social Science & Medicine, 62*(3), 769–778.
Coulson, N. S., Buchanan, H., & Aubeeluck, A. (2007). Social support in cyberspace: A content analysis of communication within a Huntington's disease online support group. *Patient Education and Counseling, 68*(2), 173–178.
Coulson, N. S., & Knibb, R. C. (2007). Coping with food Allergy: Exploring the role of the online support group. *CyberPsychology & Behavior, 10*(1), 145–148.
Cutler, D. M., Glaeser, E. L., & Vigdor, J. L. (2008). When are ghettos bad? Lessons from immigrant segregation in the United States. *Journal of Urban Economics, 63*(3), 759–774.
De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the 7th international AAAI conference on weblogs and social media boston, MA* (p. 2).
Di Tella, R., & MacCulloch, R. (2008). Gross national happiness as an answer to the Easterlin Paradox? *Journal of Development Economics, 86*(1), 22–42.
Diez Roux, A. V. (2001). Investigating neighborhood and area effects on health. *American Journal of Public Health, 91*(11), 1783–1789.
Diez-Roux, A. (1998). Bringing context back into epidemiology: Variables and fallacies in multi-level analysis. *American Journal of Public Health, 88*, 216–222.
Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social Network: Hedonometrics and twitter. *PLoS ONE, 6*(12), e26752.
Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). In Pew Research Center (Ed.), *Social media update 2014*. Available at: http://www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf.
Duncan, C., Jones, K., & Moon, G. (1998). Context, composition and heterogeneity: Using multilevel models in health research. *Social Science & Medicine, 46*, 97–117.
Eames, M., Ben-Shlomom, Y., & Marmot, M. G. (1993). Social deprivation and premature mortality: Regional comparison across england. *BMJ, 307*, 1097–1102.
EnchantedLearning.com. Food and Eating Vocabulary Word List. http://www.allaboutspace.com/wordlist/food.shtml Accessed 26.02.14.
Evans, J., Fast, S., & Markuzon, N. (2013). Modeling the social response to a disease outbreak. In A. Greenberg, W. Kennedy, & N. Bos (Eds.), *Social computing, behavioral-cultural modeling and prediction. Lecture notes in computer science* (Vol. 7812, pp. 154–163). Springer Berlin Heidelberg.
Fowler, J. H., & Christakis, N. (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study. *British Medical Journal, 337*, a2338.
Gallup-Healthways. State of American Well-being: 2013 State Rankings and Analysis. http://info.healthways.com/wbi2013 Accessed 11.01.14.
Ghosh, D., & Guha, R. (2013). What are we 'tweeting' about obesity? mapping tweets with topic modeling and geographic information system. *Cartography and Geographic Information Science, 40*(2), 90–102.
Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The language of emotion in short blog texts. In *Proceedings of the 2008 ACM conference on computer supported cooperative work* (pp. 299–302). San Diego, CA, USA: ACM.
Grigsby-Toussaint, D. S., Lipton, R., Chavez, N., Handler, A., Johnson, T. P., & Kubo, J. (2010). Neighborhood Socioeconomic Change and Diabetes Risk: Findings from the chicago childhood diabetes registry. *Diabetes Care, 33*(5), 1065–1068.
Guttman, A. (1984). R-Trees: A dynamic index structure for spatial searching. *Proceedings of the 1984 ACM SIGMOD international conference on Management of Data, 14*(2), 47–57.
Heinrich, K. M., Lee, R. E., Regan, G. R., Reese-Smith, J. Y., Howard, H. H., Haddock, C. K., et al. (2008). How does the built environment relate to body mass index and obesity prevalence among public housing residents? *American Journal of Health Promotion, 22*(3), 187–194.
Helliwell, J. F., Layard, R., & Sachs, J. D. (2012). *World happiness report 2013*. UN Sustainable Development Solutions Network.
Inagami, S., Cohen, D. A., & Finch, B. K. (2006). You are where you shop: Grocery store locations, weight, and neighborhoods. *American Journal of Preventive Medicine, 31*, 10–17.
Keating, N. L., O'Malley, A. J., Murabito, J. M., Smith, K. P., & Christakis, N. A. (2011). Minimal social network effects evident in cancer screening behavior. *Cancer, 117*, 3045–3052.
Keels, M., Duncan, G. J., Deluca, S., Mendenhall, R., & Rosenbaum, J. (2005). Fifteen years later: Can residential mobility programs provide a long-term escape from neighborhood segregation, crime, and poverty. *Demography, 42*(1), 51–73.
Kendall, L., Hartzler, A., Klasnja, P., & Pratt, W. (2011). *Descriptive analysis of physical activity conversations on twitter. CHI '11 extended abstracts on human factors in computing systems* (pp. 1555–1560). Vancouver, BC: Canada. ACM.
Kim, D., Subramanian, S. V., Gortmaker, S. L., & Kawachi, I. (2006). Us state- and county-level social capital in relation to obesity and physical inactivity: A multilevel, multivariable analysis. *Social Science & Medicine, 63*(4), 1045–1059.
Kramer, A. D. I. (2010). An unobtrusive behavioral model of "gross national happiness". In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 287–290). Atlanta, Georgia, USA: ACM.
Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788–8790.
Lamb, A., Paul, M. J., & Dredze, M. (2012). In *Investigating twitter as a source for studying behavioral responses to epidemics. AAAI fall Symposium: Information retrieval and knowledge discovery in biomedical text. Arlington, VA* (pp. 81–83).
Luo, F., Cao, G., Mulligan, K., & Li, X. (2015). Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case study of Chicago. *Applied Geography, 70*, 11–25.
Lysy, Z., Booth, G. L., Shah, B. R., Austin, P. C., Luo, J., & Lipscombe, L. L. (2013). The impact of income on the incidence of diabetes: A population-based study. *Diabetes Research and Clinical Practice, 99*(3), 372–379.
Macintyre, S., Maciver, S., & Sooman, A. (1993). Area, class and health: Should we be focusing on places or people. *Journal of Social Policy, 22*, 213–233.
Mednick, S. C., Christakis, N. A., & Fowler, J. H. (2010). The spread of sleep loss influences drug use in adolescent social networks. *PLoS One*, e9775.
Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of Happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE, 8*(5), e64417.
Morland, K., Wing, S., Diez Roux, A., & Poole, C. (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American Journal of Preventive Medicine, 22*(1), 23–29.
Morland, K, Wing, S., & Roux, A. D. (2002). The contextual effect of the local food environment on residents' diets: the atherosclerosis risk in communities study. *American Journal of Public Health, 92*(11), 1761–1768.
Morris, J. N., Blane, D. B., & White, I. R. (1996). Levels of mortality, education, and social conditions in the 107 local education authority areas of England. *Journal of Epidemiology and Community Health, 50*, 15–17.
Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). *Is the sample good enough? comparing data from Twitter's streaming API with Twitter's firehose*. arXiv: 1306.5204v1 [cs.SI].
Mujahid, M. S., Roux, A. V. D., Shen, M., Gowda, D., Sánchez, B., Shea, S., et al. (2008). Relation between neighborhood environments and obesity in the multi-ethnic study of atherosclerosis. *American Journal of Epidemiology, 167*(11), 1349–1357.
National Archive of Criminal Justice. Project on Human Development in Chicago Neighborhoods http://www.icpsr.umich.edu/icpsrweb/PHDCN/ Accessed 24.09.12.
Oswald, A. J., & Powdthavee, N. (2007). Obesity, unhappiness, and the challenge of affluence: Theory and evidence. *Economic Journal, 117*, 117. F441–54.
Pachucki, M. A., Jacques, P. F., & Christakis, N. A. (2011). Social network concordance in food choice among spouses, friends, and siblings. *American Journal of Public Health, 101*, 2170–2177.
Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh conference on International Language Resources and Evaluation, 10*, 1320–1326.
Peterson, R. D., & Krivo, L. J. (2000). *National Neighborhood Crime Study (NNCS)*. http://www.icpsr.umich.edu/icpsrweb/RCMD/studies/27501 Accessed 25.09.13.
Quercia, D., Ellis, J., Capraz, L., & Crowcroft, J. (2012). *Tracking "gross community happiness" from tweets. Association for computing machinery annual conference on computer supported cooperative work seattle, Washington* (pp. 965–968).
Roemmich, J. N., Epstein, L. H., Raja, S., Yin, L., Robinson, J., & Winiewicz, D. (2006). Association of access to parks and recreational facilities with the physical activity of young children. *Preventive Medicine, 43*(6), 437–441.
Rosenquist, J. N., Fowler, J. H., & Christakis, N. A. (2011). Social network determinants of depression. *Molecular Psychiatry, 16*(3), 273–281.
Rosenquist, J. N., Murabito, J., Fowler, J. H., & Christakis, N. A. (2010). The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine, 152*, 426–433.
Roy, J. P. (2004). Socioeconomic status and health: A neurobiological perspective. *Medical Hypotheses, 62*, 222–227.
Smith, K. R., Brown, B. B., Yamada, I., Kowaleski-Jones, L., Zick, C. D., & Fan, J. X. (2008). Walkability and body mass index: Density, design, and new diversity measures. *American Journal of Preventive Medicine, 35*(3), 237–244.
Smith, K. P., & Christakis, N. A. (2008). Social networks and health. *Annual Review of Sociology, 34*, 405–429.
Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast-but is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–263.

Stafford, M. (2007). Pathways to obesity: Identifying local, modifiable determinants of physical activity and diet. *Social Science & Medicine, 65*, 1882–1897.
Stanford Natural Language Processing Group. Stanford Tokenizer. http://nlp.stanford.edu/software/tokenizer.shtml.
Tella, R. D., MacCulloch, R. J., & Oswald, A. J. (2003). The macroeconomics of happiness. *The Review of Economics and Statistics, 85*(4), 809–827.
Townsend, P., Phillimore, P., & Beattie, A. (1988). *Health and deprivation. Inequality and the north*. London, England: Routledge.
Truong, K. D., & Ma, S. (2006). A systematic review of relations between neighborhoods and mental health. *Journal of Mental Health Policy and Economics, 9*(3), 137–154.
Tyroler, H. A., Wing, S. B., & Knowles, M. G. (1993). Increasing inequality in coronary heart disease mortality in relation to educational achievement: Profile of places of residence, United States, 1962–87. *Annals of Epidemiology, 3*(Suppl), S51–S54.
United States Department of Agriculture. National Nutrient Database. http://ndb.nal.usda.gov/ndb/search/list?format=&count=&max=25&sort=&fg=&man=&lfacet=&qlookup=&offset=50 Accessed 05.02.14.
Vartanian, L. R., Sokol, N., Herman, C. P., & Polivy, J. (2013). Social models provide a norm of appropriate food intake for young women. *PLoS ONE, 8*(11), e79268.
Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., et al. (2011). Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *Proceedings of the fifth international AAAI conference on weblogs and social media. Barcelona, Spain* (pp. 385–392).
Waitzman, N. J., & Smith, K. R. (1998a). Phantom of the area: Poverty-area residence and mortality in the United States. *American Journal of Public Health, 88*(6), 973–976.
Waitzman, N. J., & Smith, K. R. (1998b). Separate but lethal: The effects of economic segregation on mortality in metropolitan America. *Milbank Quarterly, 76*(3), 341–373.
Wang, M. C., Kim, S., Gonzalez, A. A., MacLeod, K. E., & Winkleby, M. A. (2007). Socioeconomic and food-related physical characteristics of the neighborhood environment are associated with body mass index. *Journal of Epidemiology Community Health, 61*(6), 491–498.
Wen, M., Browning, C. R., & Cagney, K. A. (2003). Poverty, affluence, and income inequality: Neighborhood economic structure and its implications for health. *Social Science & Medicine, 57*, 843–860.
Widener, M. J., & Li, W. (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. *Applied Geography, 54*, 189–197.
Wing, S., Barnett, E., Casper, M., & Tyroler, H. A. (1992). Geographic and socioeconomic variation in the onset of decline of coronary heart disease mortality in white women. *American Journal of Public Health, 82*, 204–209.
Zhang, N., Campo, S., Janz, K. F., Eckler, P., Yang, J., Snetselaar, L. G., et al. (2013). Electronic word of mouth on twitter about physical activity in the United States: Exploratory infodemiology study. *Journal of Medical Internet Research, 15*(11), e261.