Applied Geography 54 (2014) 189–197

Contents lists available at ScienceDirect

Applied Geography

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

Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US

Michael J. Widener [a], [*], Wenwen Li [b], [1]

[a] Department of Geography, University of Cincinnati, Cincinnati, OH, USA
[b] GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA

Keywords:
Food deserts
Twitter
Nutrition
Spatial analysis
Public health

ABSTRACT

Mining the social media outlet Twitter for geolocated messages provides a rich database of information on people's thoughts and sentiments about myriad topics, like public health. Examining this spatial data has been particularly useful to researchers interested in monitoring and mapping disease outbreaks, like influenza. However, very little has been done to utilize this massive resource to examine other public health issues. This paper uses an advanced data-mining framework with a novel use of social media data retrieval and sentiment analysis to understand how geolocated tweets can be used to explore the prevalence of healthy and unhealthy food across the contiguous United States. Additionally, tweets are associated with spatial data provided by the US Department of Agriculture (USDA) of low-income, low-access census tracts (e.g. food deserts), to examine whether tweets about unhealthy foods are more common in these disadvantaged areas. Results show that these disadvantaged census tracts tend to have both a lower proportion of tweets about healthy foods with a positive sentiment, and a higher proportion of unhealthy tweets in general. These findings substantiate the methods used by the USDA to identify regions that are at risk of having low access to healthy foods.

1. Introduction

Twitter provides a massive amount of spatiotemporal information about individuals broadcasting their opinions, moods, and activities. These data have been utilized in a number of diverse ways to understand a range of social phenomena. For example, Vieweg, Hughes, Starbird, and Palen (2010) analyze two natural hazard events to understand if tweets contribute to increasing the situational awareness of nearby residents. Herdağdelen, Zuo, Gard-Murray, and Bar-Yam (2012) map social, political, and geographic attributes of tweeters sharing online news articles to explore the relationship between individual identities and collective group dynamics. Others have explored everything from the diffusion of political unrest (Howard et al., 2011) to tracking popular social trends (Naaman, Becker, & Gravano, 2011).

One important, and largely untapped use of Twitter's massive data feed is to utilize tweets as a means for understanding trends in public health. Of the health-related research that has used these data, many focus on tracking and understanding the diffusion of various diseases, e.g. influenza and cholera (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; American Society for Microbiology 2011; Culotta, 2010; Ehrenberg, 2012). One reason for this is the relative ease with which things like flu symptoms can be traced in tweets and linked to reports from the Center for Disease Control. However, less research has explored other, non-diffusion related public health-related uses (Ghosh & Guha, 2013).

Given the vast amount of available Twitter data, and researchers' increasing ability to handle "Big Data," it is important to explore the potential of using these data in a way that informs broader public health problems that go beyond disease diffusion. The research presented in this paper studies the spatial patterns and sentiment of food-related tweets to gain perspective on trends of food consumption across the continental United States. Additionally, a statistical analysis is conducted to explore whether spatial variations in food-related tweets correspond to the locations of US Department of Agriculture (USDA)-classified food desert census tracts. In other words, the data are analyzed to gauge whether the content and sentiment of food tweets corroborate the implication that residents in these food desert tracts consume less healthy foods. More explicitly, the goals of this paper are to:

[*] Corresponding author. Department of Geography, University of Cincinnati, 401E Braunstein Hall, Cincinnati, OH 45221-0131, USA. Tel.: +1 513 556 4829.

E-mail addresses: michael.widener@uc.edu (M.J. Widener), wenwen@asu.edu (W. Li).

[1] Tel.: +1 480 727 5987.

http://dx.doi.org/10.1016/j.apgeog.2014.07.017
0143-6228/© 2014 Elsevier Ltd. All rights reserved.

1. introduce a new framework for exploring health-related social media data that employs sentiment analysis at a large scale based on the acquired big spatial data,
2. analyze the overall spatial distribution and sentiment of tweets on healthy and unhealthy foods, and
3. explore the relationship between the locations of tweets on healthy and unhealthy food and USDA-designated food desert census tracts.

To the best of the authors' knowledge, this is the first study to attempt such analyses, and will contribute to the growing public health literature on understanding the geography of healthy food consumption and accessibility, through the use of a novel combination of big spatial data and Geographic Information Science-oriented concepts.

The rest of the paper is organized as follows. Section 2 provides a review of Twitter-based Geographic Information Systems (GIS) research, previous work in public health using Twitter, and relevant work on healthy diets and food deserts. Next, Section 3 provides a detailed overview of our methods and data. Results from our analyses are presented and discussed in depth in Section 4. Finally, conclusions, limitations, and future directions are offered in Section 5.

2. Literature review

While social media technology has only recently become ubiquitous, researchers have been quick to realize that the massive amount of information posted online can provide potentially valuable information about myriad topics.

2.1. Twitter data's role in research

GIScience research utilizing social media data, such as Twitter, can be classified into two categories: general exploratory data analysis and applied science research using Twitter as a complimentary source of data. GIS researchers are particularly interested in studying the location awareness and the social-economic characteristics of tweets (Alampay, 2006; Cheng, Caverlee, & Lee, 2010; Li, Goodchild, & Xu, 2013; Soule, Shell, & Kleen, 2003; Xu, Wong, & Yang, 2013; Zhao & Rosson, 2009). While the previously mentioned analyses improve the understanding of Twitter data, they are more focused on the locational property rather than content of the tweets. Due to its characteristics of real-time, large-scale and quick-propagation, Twitter data has attracted attentions from applied scientists to facilitate the knowledge discovery process in a wide variety of fields. For example, Twitter, along with other crowd sourcing GIS methods are identified as a useful data source to improve geospatial support for disaster management (Goodchild & Glennon, 2010). Similar examples include work by Zook, Graham, Shelton, and Gorman (2010) and Kumar, Barbier, Ali Abbasi, and Liu (2011) who both look at the utility of crowd sourcing to aid in disaster relief. In another application, Cranshaw, Schwartz, Hong, and Sadeh (2012), delineate people's "livehood" (a dynamic organizational urban structure of lived spaces) by analyzing behaviors of social media users.

Twitter-based research in the field of public health is similarly nascent. In comparison to the exploratory analyses described above, public health studies employing social media data combine content analysis, location analysis and domain knowledge to improve applied science research. For example, and as alluded to in the introduction, researchers have studied the utility of Twitter data to detect the outbreak of seasonal influenza, which in some cases is demonstrated to be more effective than more traditional surveillance methods (Dugas et al., 2012; Lee, Agrawal, & Choudhary, 2013; Signorini, Segre, & Polgreen, 2011).

However, public health work beyond monitoring vector-borne diseases remains lacking. Examples of papers exploring other aspects of public health include Scanfield et al.'s (2010) work using Twitter to better understand the use and misuse of antibiotic medications. Two other especially relevant papers that utilize Twitter data to identify important public health topics and understand their spatial patterns have been recently published and are presented here. Ghosh and Guha (2013) focus on obesity and fast food locations, and demonstrate the potential for using Twitter as a data source for a variety of public health applications, beyond infection disease monitoring. A second innovative paper examines the relationship between the locations of "healthful" and "unhealthful" tweets and the location of food vendors that sell healthy or unhealthy foods in the city of Columbus, Ohio (Chen & Yang, 2014). This work suggests that Twitter data can be linked to environmental factors to understand behavioral choices. The research presented here expands on both of these papers by implementing an automated sentiment analysis of a large national dataset of tweets, and comparing their locations to the government-maintained USDA food accessibility mapper.

2.2. "Food deserts" and public health

One of the most pressing public health issues in the United States is the inequity in access to healthy foods. The issue receives ample attention in both popular (Barclay, 2013) and academic publications (McKinnon, Reedy, Morrissette, Lytle, & Yaroch, 2009; Wrigley, 2002). Regions that lack sufficient access to healthy food stores are commonly referred to as "food deserts," implying the area is deserted of vendors that regularly offer affordable, fresh, and healthy foods (Shaw, 2006). A large literature has developed over the past half decade, seeking to understand how to best characterize these environments (Farber, Morang, & Widener, 2014; Larson, Story, & Nelson, 2009; McKinnon et al., 2009; Walker, Keane, & Burke, 2010). However, there is a growing body of work that critiques the notion that food deserts can be so simply described (Cummins & Macintyre, 2002; Shaw, 2006). Widener, Metcalf, and Bar-Yam (2011), Widener, Farber, Neutens, and Horner (2013) and Burgoine and Monsivais (2013) demonstrate the spatio-temporal movements of food vendors (e.g. farmers' markets opening and closing) and urban residents (e.g. changing spatial distributions due to commuting) can alter the level of access populations have to healthy food vendors. An and Sturm (2012) directly examine the relationship between the food environment and diet among children, aged 5–17, in California. The authors fail to find a robust link between healthier eating and residing in food deserts, but assert that further study is needed.

Beyond mapping the locations of tweets about healthy and unhealthy food in the continental United States, the research presented in this paper attempts to extend An and Sturm's work by studying the relationship of healthy eating and location at a national scale. While Twitter data has limitations, it does provide a novel means for exploring food consumption in a way that was previously unavailable, even a decade ago. Ultimately, this big data approach seeks to add to the food accessibility conversation by furthering the understanding of the spatial distribution of diets and access across the US.

3. Data and methods

In this paper, we are interested in the spatial patterns of food tweets. We collect tweets from Twitter's streaming application programming interface (API) from 6/26/2013 to 7/22/2013. The streaming API provides a near-real time sample of all tweets that match our submitted query about healthy and unhealthy foods, of which we store only those tweets that are georeferenced (2.5% of the total tweets). Finally, only tweets that occur within the continental US are kept and analyzed to map what regions more frequently mention healthy or unhealthy foods, and whether their sentiment is positive or negative. The entirety of this process is thoroughly described in the next subsection.

3.1. Collecting Twitter data

Fig. 1 demonstrates the software architecture for data retrieval, and the analysis framework that fetches and analyzes real-time Twitter data. Two modules are developed to complete the data acquisition task: a retrieval module and an analysis module. The retrieval module is responsible for establishing a connection with the Twitter web server through its streaming API. The "public stream" is used for collecting data, as the focus of this retrieval task is public data flow containing tweets that mention healthy and unhealthy food, rather than a data stream from a specific user or website. Different from receiving data through Twitter's search API, which provides a REST (Representational State Transfer) interface, the method utilized here requires the establishment of a persistent HTTP (HyperText Transfer Protocol) connection for receiving streaming tweet data. To satisfy this requirement, the retrieval module first opens a socket connection and then performs HTTP post requests in which the queried tweets are identified. An error handling mechanism is adopted to check if the communication channel remains open every time when a new HTTP request is sent. If the number of failed attempts is more than a predefine value, the program automatically stops.

Once the raw tweets are returned from the server, they are serialized and stored in a database. In order to make use of these data in a timely manner, a separate analysis module simultaneously reads the data as they are flushed into the database. The parsing module first deserializes the raw tweet data into a text based JSON (JavaScript Object Notation) object and then extracts desired information (e.g. user ID, user profile, tweet text, location and tweet time) from the object. Before entering them into the database for further analysis, two rules are applied to filter out noisy tweets, which contain food keywords but actually refer to something different due to the polyseme issue. The first rule is defined to remove tweets mentioning specific companies with food names, such as Apple and Blackberry. Company profiles are compiled and are composed of a set of keywords that are related to the company (Yerva, Miklós, & Aberer, 2010). For instance, when "iPhone", "iPad", "iPod", "keynote", "iTunes", etc are mentioned, the tweet is likely about the computer company Apple Inc. If a tweet contains a company's name, and at the same time contains any keyword that appears in the company's preconstructed profile, it will be considered irrelevant to food, and disregarded. The second rule is applied for more general cases. A taxonomy analysis is performed on each tweet and a list of themes will be returned. If none of the themes are related to the category "Food and Drink" in the taxonomy (Alchemy, 2014), this tweet will be filtered out. For instance, although a food name "orange" appeared in tweet "I'm at Orange County Convention Center West Concourse (Orlando, FL)," this tweet is identified as noise and disregarded, because the tweet is categorized as having a travel theme by the taxonomy analysis. By applying the above rules for disambiguation, many irrelevant tweets were removed, increasing the number of tweets actually about food saved in the database.

Next, an analysis module is called to gain perspective on users' preferences about food through sentiment analysis. The next section describes the process in detail.

3.2. Sentiment analysis

Sentiment analysis of social media, which computationally extracts the implied opinions, or "sentiment", of text from online

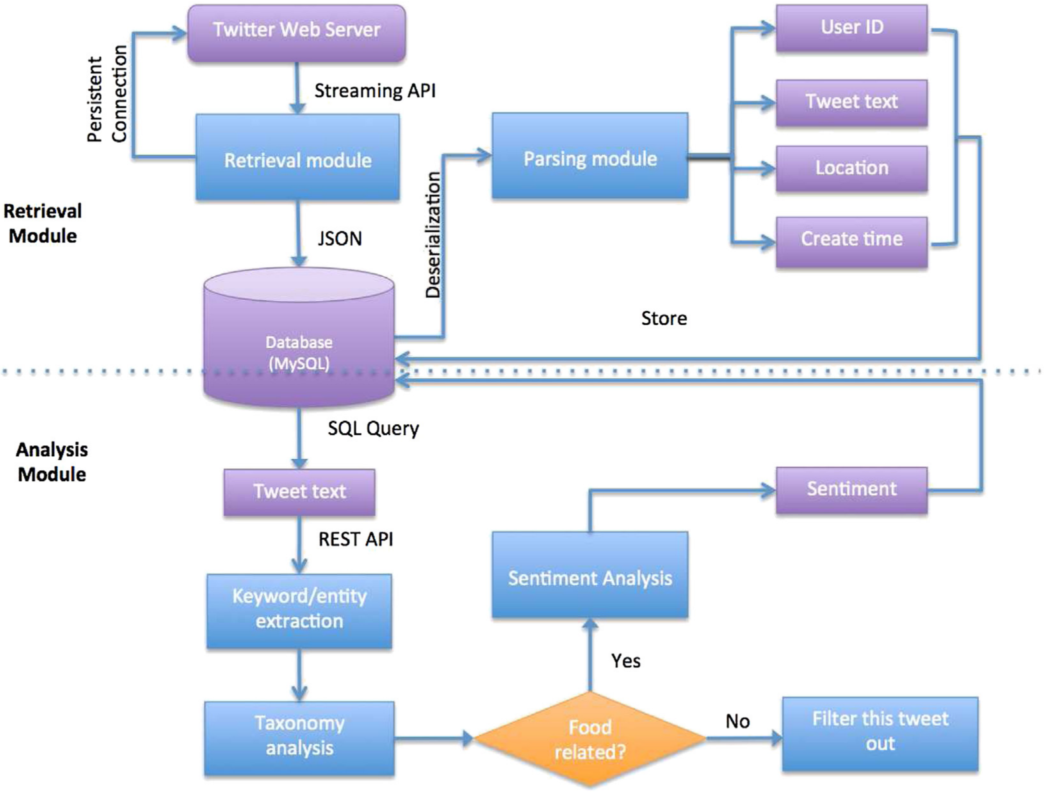

Fig. 1. A tweet data retrieval and analysis framework. Purple modules represent static modules and blue modules represent software programs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
List of foods searched for in tweets.

Healthy foods			Unhealthy foods		
Fruits	Vegetables		General	Fast food	
apple juice	acorn squash	hubbard squash	bacon	Arby's	Panda Express
apples	artichokes	iceberg (head) lettuce	cakes	Baskin-Robbins	Panera Bread
apricots	asparagus	kale	cheese	Bojangles'	Papa John's
bananas	avocado	kidney beans	cookies	Boston Market	Papa Murphy's
blueberries	bean sprouts	lentils	donuts	Burger King	Pizza Hut
cantaloupe	beets	mesclun	energy drinks	Captain D's	Popeyes Louisiana Kitchen
cherries	black beans	mushrooms	fruit drinks	Carl's Jr.	Qdoba Mexican Grill
fruit cocktail	black-eyed peas	mustard greens	hot dogs	Checkers/Rally's	Quiznos
grape juice	black-eyed peas (dry)	navy beans	ice cream	Chick-fil-A	Sbarro
grapefruit	bok choy	okra	pastries	Chipotle Mexican Grill	Sonic Drive-In
grapefruit juice	broccoli	onions	pizza	Church's Chicken	Starbucks
grapes	brussels sprouts	pinto beans	ribs	CiCi's Pizza	Steak 'n Shake
honeydew	butternut squash	plantains	sausages	Culver's	Subway
kiwi fruit	cabbage	potatoes	soda, pop	Dairy Queen	Taco Bell
lemons	carrots	pumpkin	sports drinks	Del Taco	Tim Hortons
limes	cassava	red peppers		Domino's Pizza	Wendy's
mangoes	cauliflower	romaine lettuce		Dunkin' Donuts	Whataburger
nectarines	celery	soy beans		Einstein Bros. Bagels	White Castle
orange juice	collard greens	spinach		El Pollo Loco	Zaxby's
oranges	corn	split peas		Five Guys Burgers & Fries	
papaya	cowpeas	sweet potatoes		Hardee's	
peaches	cucumbers	taro		In-N-Out Burger	
pears	dark green leafy lettuce	tomato juice		Jack in the Box	
pineapple	eggplant	tomatoes		Jason's Deli	
plums	field peas	turnip greens		Jimmy John's	
prunes	garbanzo beans (chickpeas)	turnips		KFC	
raisins	green bananas	water chestnuts		Krispy Kreme	
raspberries	green beans	watercress		Krystal	
strawberries	green lima beans	wax beans		Little Caesars	
tangerines	green peas	white beans		Long John Silver's	
watermelon	green peppers	zucchini		McDonald's	

users, offers an opportunity for researchers to make new discoveries by associating the implied meaning of statements with physical location at a broad scale. Sentiment analysis can be classified into three general categories, differentiated based on the manner in which text is being analyzed: sentiment classification, feature-based sentiment analysis and sentiment analysis of comparative sentences. Sentiment classification aims at assigning a text block (of any size) a single positive or negative score (Aue & Gamon, 2005; Devitt & Ahmad, 2007; Wan, 2008). Feature-based sentiment analysis is used to further extract the objects within a text block, and judge the positive or negative opinion of each object (Bethard, Yu, Thornton, Hatzivassiloglou, & Jurafsky, 2004; Choi, Cardie, Riloff, & Patwardhan, 2005; Kim & Hovy, 2004). Finally, a sentiment analysis of comparative sentences is used to understand a user's opinion of objects being compared in a single sentence, such as “KFC's chicken tastes better than McDonald's”.

Analyzing the sentiment of social media text is a challenging task due to the sophistication of natural language and the typical short length and irregular structure of user-generated content (Saif, He, & Alani, 2012). To improve the accuracy of sentiment analysis results, we adopted a feature-based sentiment analysis by obtaining sentiment of the actual food entity mentioned in a tweet rather than getting an overall score of the whole tweet. Given a tweet or data from other social media, sentiment analysis can be defined by the below quadruple:

$$< op_{ijk}, u_i, o_j, t_k > ,$$

where op_{ijk} is the opinion from a user u_i on some object o_j at time t_k. The variable op_{ijk} usually has a continuous range of $[-1,1]$, where a value of 1 indicates a strong positive sentiment, a value of 0 is neutral, and a value of -1 is a tweet expressing a strong negative sentiment.

While there are multiple popular tools for sentiment analysis, including Alchemy (2013), Zemanta (2013), and OpenCalais (2013), Alchemy API is chosen in our work because it has been validated through earlier studies to yield more accurate sentiment classification (Meehan, Lunney, Curran, & McCaughey, 2013; Saif et al., 2012). Alchemy API introduces a combined use of linguistic analysis, which considers a sentence's composition, and statistic analysis, which handles noisy content (e.g. misspellings). The actual sentiment analysis is a classification process, built upon learned patterns used to make predictions about the text's intended sentiment. The sentiment analysis process includes the following steps:

(1) Since opinion words and phrases are always the prominent feature used for sentiment classification, the first step is to extract words that provide a good indication about subjectivity or opinions. Usually these words are adjectives and adverbs. An adjacent noun of these opinion indicators is also extracted to provide sufficient context to determine a sentence's opinion orientation.
(2) A supervised classification is conducted to estimate the positive or negative orientation of the extracted words identified in step 1. To ensure classification accuracy, a large collection of training data from the web was crawled, as well as top social media platforms. Over 200 billion words were extracted for pattern detection. The orientation of the sentiment is measured by the co-existence of a phrase with perfect positive terms, such as “excellent”, and with negative terms, such as “poor”. By matching the classified patterns in the tweet content, the opinion orientation of entities in each tweet can be identified.

The results of sentiment analysis have range of $[-1,1]$. The higher the value is, the more positive a tweet is. A value of 1 means

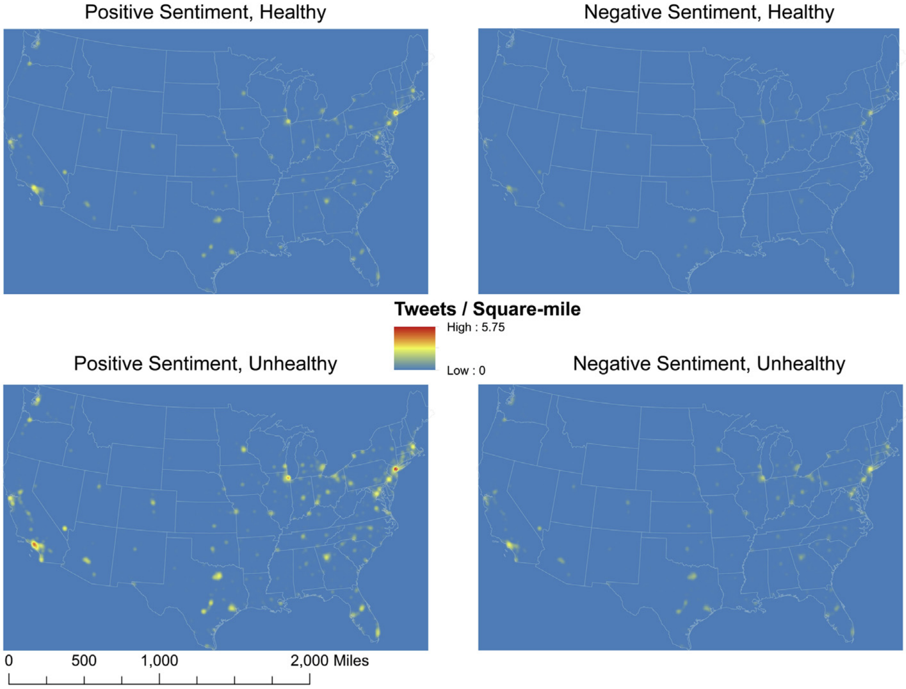
Positive Sentiment, Healthy
Negative Sentiment, Healthy
Tweets / Square-mile
High : 5.75
Low : 0
Positive Sentiment, Unhealthy
Negative Sentiment, Unhealthy
0
500
1,000
2,000 Miles

Fig. 2. Kernel density surfaces (tweets/square-mile) of four subgroups of tweets.

a person holds very positive opinion about an object (in our case, the object is the food mention). A value of −1 means that a person is extremely negative about an object, and 0 means this is a neutral tweet.

3.3. *What foods are healthy?*

In order to evaluate which food is healthy and whether people like it or not by tweeting it, the first step involves defining a list of healthy and unhealthy foods. www.choosemyplate.gov is used as a starting point for populating this list. This website catalogues commonly eaten foods from the five food categories considered to be apart of a healthy diet (fruits, vegetables, grains, protein foods, and dairy), as well as foods classified as "empty calories." For this study's healthy food list, we only consider fruits and vegetables for three reasons. First, produce is commonly recommended to make up about half of a person's healthy diet (Control, 2013). Second (and related to the first reason), research has found a strong relationship between the consumption of fruits and vegetables and decreased risk of a number of chronic diseases (Bazzano et al., 2002; He et al., 2004; Higdon, Delage, Williams, & Dashwood, 2007; Hung et al., 2004). Third, grains, protein foods, and dairy are healthy when consumed in moderation, and can have negative health effects if consumed in larger quantities (Drewnowski, Kurth, Holden-Wiltse, & Saari, 1992). Therefore, the lists of fruits and vegetables are utilized as a proxy. For the unhealthy food list, we take the empty calories catalogue from www.choosemyplate.gov, and supplement it with the names of fast food restaurants listed in QSR's Top 50 quick-service restaurants (QSR, 2012). Both lists are presented in Table 1.

3.4. *Preparing data for analysis*

As previously mentioned, of all tweets with relevant keywords from Table 1 available via the Twitter's streaming API, 2.5% are available with location information and are stored for further analysis. This process occurs from 6/26/2013 to 7/22/2013, at which point a total of 500,000 geolocated tweets are recorded. The data are then further examined, with all tweets outside of the contiguous 48 United States eliminated, resulting in a total of 148,533 tweets. The spatial distribution of tweets closely resembles the population distribution of the United States with most occurring in major metropolitan regions throughout the East Coast, South, Midwest, and West Coast. Once collected, as described in detail in Section 3.1, a textual analysis is executed in order to gauge whether a tweet is about healthy or unhealthy foods, as defined by Table 1, and if the sentiment of the tweet is positive or negative. This results in the derivation of four subgroups, where sentiment values of exactly zero are not included in any subgroup: Healthy/Positive (HP), Healthy/Negative (HN), Unhealthy/Positive (UP), and Unhealthy/Negative (UN). There are a total of 128,914 tweets in the four subgroups.

Next, in order to control for the large volume of tweets happening in major cities, a population density surface is constructed. A population-weighted kernel density estimate (KDE) is computed for the contiguous United States using tract centroid and population data from the 2010 U.S. Census, with a search radius of 20 miles and 1 square mile output cells with units of people/square mile. The 20 mile search radius is chosen so that metropolitan regions are contiguously represented and the output cells of 1 square mile are chosen because they provide a relatively high resolution representation of the population, while not causing an unreasonable computational burden. Using the point data from the four tweet subgroups (HP, HN, UP, and UN), an additional four tweet KDE surfaces are computed using the same search radius and cell size values, resulting in units of tweets/square mile (Fig. 2). Finally the four tweet density surfaces are divided by the population density surface to compute four raster surface layers with units of tweets/people. The resulting surfaces are shown in Fig. 3.

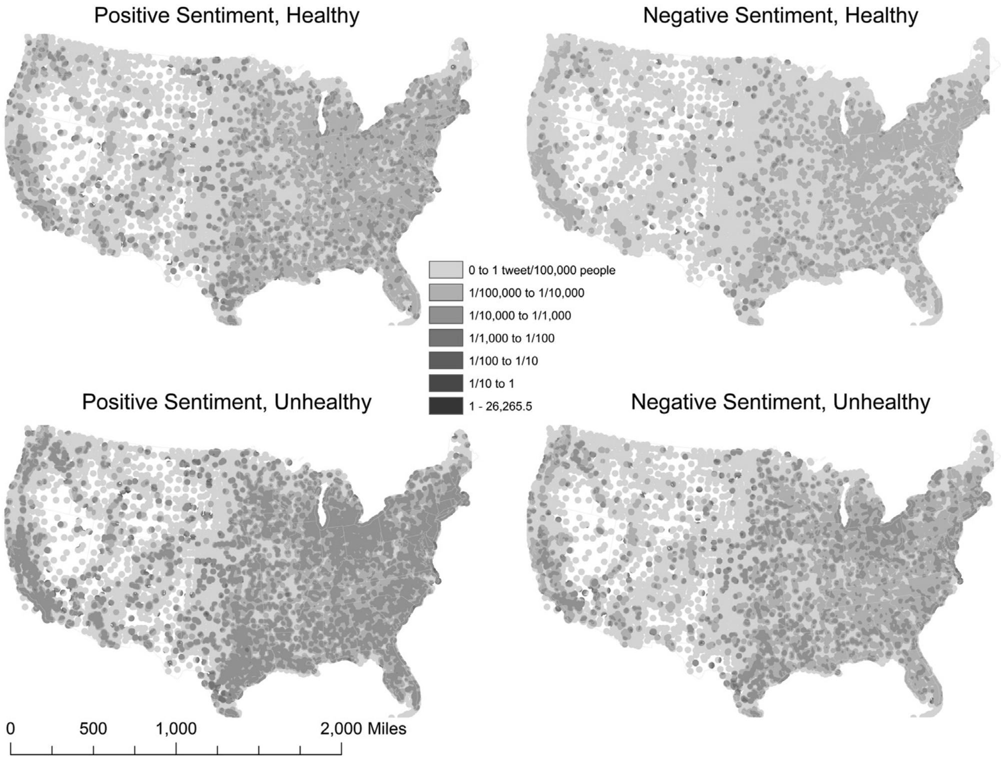
Positive Sentiment, Healthy
Negative Sentiment, Healthy
0 to 1 tweet/100,000 people
1/100,000 to 1/10,000
1/10,000 to 1/1,000
1/1,000 to 1/100
1/100 to 1/10
1/10 to 1
1 - 26,265.5
Positive Sentiment, Unhealthy
Negative Sentiment, Unhealthy
0
500
1,000
2,000 Miles

Fig. 3. Kernel density surfaces of four subgroups of tweets, controlling for the spatial distribution of population.

4. Results and analysis

Before exploring the relationships between content and sentiment of tweets and their locations in US Department of Agriculture designated low access, low income (LILA) census tracts (also known as food deserts), it is important to provide an overview of the spatial distribution of tweets across the study area. Subsection 4.1 presents such an overview of the previously described tweet density surfaces, as well as a number of derived maps. This is followed by statistical analyses aimed at understanding whether food tweets located in regions with low access to healthy foods are predominantly healthy or unhealthy in their content. Finally, a series of statistical models that relate tweet content to census tract characteristics, like LILA status, are presented to further understand the link between tract-level demographics and Twitter behavior.

4.1. Overview of spatial distribution of Twitter data

Adjusting for population, some interesting geographic patterns are identifiable. In particular, even after adjusting for population, large metropolitan regions are still easily identifiable. This corresponds with the findings of Smith and Brenner (2012), who note that urban and suburban residents are significantly more likely to use Twitter than are residents of more rural regions. However there are notable differences in the intensity of the tweets per person rates across the four categories and space. For example, the map displays the subtle spatial variations in the rate of tweeting a message with a negative sentiment about healthy food and the rate of tweeting a message with a negative sentiment about unhealthy food in the state of Michigan.

It is important to note that extreme peaks in the rate of tweeting per person are likely due to sensitivity to low population counts. For example, the highest rates (the dark grey and black regions) occur in sparsely populated areas and are an artifact of the method of calculation (described in Section 3.3). It is also possible that these peaks are related to temporal anomalies in tweeting that correspond to events focused on food, like the week long National Cherry Festival in Traverse City, Michigan (population of approximately 15,000), that attracts over 500,000 attendees to celebrate the cherry harvest. This event occurred during our data collection period, which may account for the fact that Traverse City, Michigan had a rate of 0.0003 tweets/person, which is greater than rates of approximately 0.0001 found in cities like Memphis, Louisville, Jacksonville, and Buffalo, which all have metro-region populations of around 1,000,000.

Despite this, the descriptions of per capita food tweets for more populated areas are likely more robust, as they are not subject to wild fluctuations in the population. The Northeastern, Midwestern, and West Coast megalopolises, along with other large urban regions (e.g. the Dallas-Houston-San Antonio triangle in Texas), have relatively high food tweet to person ratios.

4.2. Analysis of tweet locations and low income, low access census tracts

One of the major goals of this research is to understand whether tweets about food can serve as a means of public health surveillance at a national scale by analyzing differences in content and sentiment in food desert and non-food desert census tracts. To explore this, a polygon layer with data pertinent to residents' abilities to access food at the tract level is obtained from the USDA's Economic Research Service's Food Access Research Atlas, formerly known as the Food Desert Locator (USDA ERS, 2013). For this paper, census tracts considered to be “low income, low access” are of interest, where:

- the tract is considered to be low access if 33% or at least 500 residents are 0.5 miles away from supermarkets in urban regions, and 10 miles away from supermarkets in rural regions,

- and the tract is considered to be low income if the tract's poverty rate is 20% or greater, or the median family income is less than or equal to 80% of the state-wide median income, or it is in a metropolitan area and has a median family income less than or equal to 80% of the metro-area's median family income (USDA ERS Documentation, 2013).

The location of these LILA tracts is displayed in Fig. 4 (USDA ERS, 2013).

First, it is of interest to simply understand where various types of tweets are located. Table 2 displays a cross tabulation of tweets in the previously described four subgroups versus their presence in a LILA census tract, where a “1” indicates that the tweet occurred in a LILA tract and “0” means the tweet occurred in a non-LILA tract. Generally, the volume of tweets that occur in LILA tracts is much lower than those tweets occurring in non-LILA tracts. This makes sense, as there are 63,389 non-LILA tracts and only 8894 LILA tracts. However, comparing the column-wise proportions reveals noteworthy results. For example, 73.6% of tweets are about unhealthy food in LILA tracts, while 72.7% of tweets in non-LILA tracts are about unhealthy food. While the difference of 0.9% is small, it is found to be statistically significant via the difference of proportions test at a $p < 0.05$ level. Likewise, there is a significant difference (at $p < 0.05$) between the 16.7% of HP tweets in LILA tracts and the 17.9% of HP tweets in non-LILA tracts.

The differences in the lower prevalence of HP tweets and higher prevalence of unhealthy tweets in LILA tracts lends credence to the supposition that there are forces driving residents of low income neighborhoods with low access to healthy food stores, like supermarkets, to maintain less healthy diets. In this regard, the analysis of the Twitter dataset has provided useful information.

More nuanced models are needed to understand the impact of being located in a LILA tract, controlling for other local demographic factors. To explore these issues, a number of logistic regression models are constructed and tested. The models' results are presented in Table 3, with the binary variable of “healthy tweet content” as the dependent variable, where the dependent variable is equal to one if a tweet contained a keyword considered to reference healthy food (Table 1). It is important to note that these models use all of the approximately 149,000 tweets. Unlike the contingency tables, tweets with a sentiment value of zero are included in this analysis. Independent variables include the sentiment of the tweet (ranging from -1 to 1), and a number of tract-level demographic variables like total population (continuous integer), proportion of males (continuous value from 0 to 1), median age (continuous integer), and proportion of black residents (continuous value from 0 to 1). These variables are all relevant to dietary intake (Forshee & Storey, 2006). Additionally, independent binary variables indicating the tweet is located in a tract considered to be both low income and low access, just low access, or just low income are included in Models 1, 2, and 3, respectively.

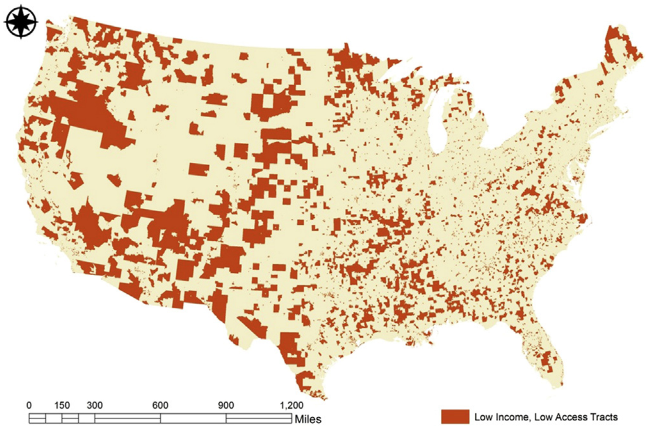

Fig. 4. Location of low income, low access tracts.

Table 2
Contingency table of subgroups versus LILA tract status.

	LILA tracts			
	0	1	Row total	
Healthy, negative	12,442	1678	14,120	N
	0.881	0.119	0.10	N/Row total
	0.095	0.097		N/Col total
	0.084	0.011		N/Table total
Healthy, positive	23,465	2901	26,366	
	0.890	0.110	0.18	
	0.179	0.167		
	0.158	0.020		
Unhealthy, negative	37,858	5356	43,214	
	0.876	0.124	0.29	
	0.289	0.308		
	0.255	0.036		
Unhealthy, positive	57,393	7440	64,833	
	0.885	0.115	0.44	
	0.438	0.428		
	0.386	0.050		
Column total	131,158	17,375	148,533	
	0.883	0.117		

Coefficients (as adjusted odds ratios) and significance levels for the three models are presented in Table 3. The AIC values (a relative goodness of fit statistic) of the models are all similar, with Model 2 having a marginally better fit. The direction and magnitude of the odds ratios are consistent across all three models, indicating that there are some general trends that can be gathered from this analysis. All three models indicate that the odds of a tweet being about healthy food increase with an increase in the sentiment score, suggesting that the more positive the tweet is, the more likely that tweet contains healthy content. For all three models, a tweet being located in a LILA, low access to supermarket, or low income tract (from Models 1, 2, and 3 respectively) result in a decrease in the odds that the tweet will contain healthy content, holding all else equal. The low access tract variable (Model 2) has the greatest decrease in odds, followed by the LILA tract variable, with the low income tract variable not being significant. Variables describing the demographic characteristics of tracts show that areas with a larger median age and more male residents have increased odds of a tweet containing healthy content, while the proportion of black residents is not significant. Finally, the tract population from the 2010 census,

Table 3
Logistic regression models exploring the relationship of tweet content with tract-level variables.

	Dependent variable: healthy tweet content		
	All coefficients presented as adjusted odds ratios		
	Model 1	Model 2	Model 3
Sentiment (-1 to 1)	1.210***	1.210***	1.210***
LILA tract	0.960*	–	–
Low access tract	–	0.930***	–
Low income tract	–	–	0.994
Median age	1.003**	1.004***	1.003**
Proportion male	3.020***	2.870***	3.040***
Proportion black	1.050	1.040	1.040
Population (2010)	1.000***	1.000*	1.000***
AIC	173,130.62	173,118.16	173,135.42

* = $p < 0.05$,** = $p < 0.01$,*** = $p < 0.001$.

which was included to help control for urban/rural differences, does not affect the odds of the models' outcomes. Ultimately, the three exploratory logistic regression models reinforce the finding from the contingency table that tweets in tracts where there may be spatial or economic accessibility issues are less likely to contain healthy content than tracts with higher incomes and/or better access.

5. Conclusion

The results presented in Tables 2 and 3 suggest that healthy food might be more prevalent in regions not considered food deserts by the USDA. This finding is important because it provides some justification for the aggregate level methods that identify variably sized regions as food deserts, as is done in the USDA's food access research atlas. Additionally, the use of geolocated tweets allows researchers to generate maps of various subcategories of social media messages, as is done in Fig. 3, for spatial analyses. Certain anomalies, like the previously discussed Traverse City case, can be further inspected to determine whether their contributions should be discounted.

Findings from this research imply that data from social media services, like Twitter (which is easily available through their API), can be used for public health purposes that range beyond that of simple diffusion mechanics. In fact, such services provide myriad opportunities for “taking the pulse” of what and where a topic of interest is being discussed. Monitoring the prevalence of a topic over space allows researchers to compare that topic with other data compiled using different means. Specifically, in this paper, the content and sentiment of geolocated food tweets are compared to a dataset that identifies regions with low access to healthy foods. The idea that census tracts with low access to supermarkets and low income residents are at risk for maintaining less nutritious diets is reinforced by the data showing a higher proportion of tweets with unhealthy content in food desert tracts, and a negative correlation between a tweet being about healthy content and located in a food desert tract.

Of course, there are important limitations that must be considered when conducting any analysis using social media data. First, the use of a service like Twitter is elective. Because of this, it is unlikely that the data represent the complete population of interest. This is reinforced by findings from Smith and Brenner (2012), showing that only 15% of internet using adults use Twitter. Additionally, young adults, African Americans, urban/suburban residents, and mobile users use Twitter at higher rates (Smith & Brenner, 2012). Therefore, any analysis using social media data should not overstate claims about representing all residents in a study area. For example, and relevant to this paper, internet usage is negatively correlated with poverty and age (Zickuhr, 2013), two groups most vulnerable to having low economic access to healthy foods. Similarly, the higher urban and suburban usage may mean the regression models presented in Table 2 are more representative of food deserts in non-rural regions. While services like Twitter do allow researchers to collect and analyze spatial data with at a scale that was previously not easily achievable, the uncertain demographic make up of users poses problems when attempting to establish irrefutable conclusions. This is a problem that affects all research utilizing this type of data. Correcting for these population biases could cause different or non-significant model outcomes. Thus, it is important to note that the results from this paper are derived from a new and large, but inevitably flawed, dataset. So, while it is difficult to make conclusions about broader populations, the results seen here do provide some new insights into the prevalence and sentiment of food tweets by the population of Twitter users.

A second limitation is related to the reliability of the content and sentiment analysis methods presented in Section 3. While automated analyses of content and sentiment have improved, they are still imperfect. Limitations with the sentiment analysis method include the inability to pick up nuanced or ambiguous meanings, like sarcasm. This is a particular challenge that affects all areas of research involving the interpretation of online or digital text (Davis, Bolding, Hart, Sherr, & Elford, 2004; Stapleton, 2005).

Related to the second limitation, a third limitation is that Twitter data is sensitive to when it is collected. For example, by collecting our data during the middle of the summer, there may be more mentions of farmers' markets and food festivals, such as the Cherry Festival in Traverse City mentioned in the results section, than there might be during winter months.

A fourth limitation specific to this study has to do with the food list used and displayed in Table 1. While we assume fast food restaurants are a reasonable analog for less-than-healthy food, it is possible to order healthier options at these types of restaurants. Future work should explore methods for determining a list of terms that can be used with social media analysis and related to actual food consumption.

A final limit to note is that this method may not be applicable to certain public health issues. While food is a relatively common and easy thing to tweet about, other topics might be more personal, for example, sexual health. Along the same lines, people may be more apt to tweet about food when they go out to a restaurant for dinner, where they could be eating less healthy foods than they might otherwise.

Despite these limitations, the findings of this work point to a convincing relationship between increased prevalence of unhealthy foods in food deserts at a national scale. Such a finding lends credence to the measures used by the USDA to locate regions at risk of having poor spatial and economic access to healthy foods. Of course, local studies with carefully designed surveys will provide a more detailed and reliable means of analyzing healthy and unhealthy food consumption, but ultimately, the methods presented here provide a novel way for utilizing social media data across a large geographic space to understand the landscape of public health.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2011). Predicting flu trends using Twitter data. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 702–707). IEEE.

Alampay, E. A. (2006). Analysing socio-demographic differences in the access & use of ICTs in the Philippines using the capability approach. *The Electronic Journal of Information Systems in Developing Countries, 27*.

Alchemy. (2013). *Alchemy API*. http://www.alchemyapi.com/. Last accessed at 02.01.14.

Alchemy. (2014). *Alchemy AI's IAB++ taxonomy*. http://www.alchemyapi.com/sites/default/files/taxonomyCategories.zip. Last accessed at 20.07.14.

American Society for Microbiology. (2011). *Genomics and social network analysis team up to solve disease outbreaks* [online]. ScienceDaily http://www.sciencedaily.com/releases/2011/05/110522141549.htm Accessed 01.07.13.

An, R., & Sturm, R. (2012). School and residential neighborhood food environment and diet among California Youth. *American Journal of Preventive Medicine, 42*(2), 129–135.

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: a case study. In *Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria*.

Barclay, E. (2013). *Can star power make New Orleans' food deserts bloom* [online]. Available from http://www.npr.org/blogs/thesalt/2013/05/15/183992818/can-star-power-make-new-orleans-food-deserts-bloom Accessed 03.07.13.

Bazzano, L., He, J., Ogden, L., Loria, C., Vupputuri, S., Myers, L., et al. (2002). Fruit and vegetable intake and risk of cardiovascular disease in US adults: the first National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *American Journal of Clinical Nutrition, 76*(1), 93.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., & Jurafsky, D. (2004). Automatic extraction of opinion propositions and their holders. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text, Stanford, CA*.

Burgoine, T., & Monsivais, P. (2013). Characterising food environment exposure at home, at work, and along commuting journeys using data on adults in the UK. *International Journal of Behavioral Nutrition and Physical Activity, 10*(1), 85.

Chen, X., & Yang, X. (2014). Does food environment influence food choices? A geographical analysis through "tweets". *Applied Geography, 51*(0), 82–89.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759–768).

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, Canada*.

Control, C. F. D. (2013). *Choose my plate* [online]. Available from http://www.choosemyplate.gov/ Accessed 07.01.13.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. M. (2012). The livehoods project: utilizing social Media to understand the dynamics of a city. In *ICWSM*.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics* (pp. 115–122).

Cummins, S., & Macintyre, S. (2002). "Food deserts"—evidence and assumption in health policy making. *BMJ: British Medical Journal, 325*(7361), 436.

Davis, M., Bolding, G., Hart, G., Sherr, L., & Elford, J. (2004). Reflecting on the experience of interviewing online: perspectives from the Internet and HIV study in London. *AIDS Care, 16*(8), 944–952.

Devitt, A., & Ahmad, K. (2007). Sentiment analysis in financial news: a cohesion based approach. In *Proceedings of the Association for Computational Linguistics (ACL), Prague, Czech Republic* (pp. 984–991).

Drewnowski, A., Kurth, C., Holden-Wiltse, J., & Saari, J. (1992). Food preferences in human obesity: carbohydrates versus fats. *Appetite, 18*(3), 207–221.

Dugas, A. F., Hsieh, Y.-H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., et al. (2012). Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases, 54*(4), 463–469.

Ehrenberg, R. (2012). Twitter kept up with Haiti cholera outbreak social media can track disease spread even in poorest countries. *ScienceNews, 181*(4), 16.

Farber, S., Morang, M. Z., & Widener, M. J. (2014). Temporal variability in transit-based accessibility to supermarkets. *Applied Geography, 53*, 149–159.

Forshee, R. A., & Storey, M. L. (2006). Demographics, not beverage consumption, is associated with diet quality. *International Journal of Food Sciences and Nutrition, 57*(7–8), 494–511.

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.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth, 3*(3), 231–241.

He, K., Hu, F., Colditz, G., Manson, J., Willett, W., & Liu, S. (2004). Changes in intake of fruits and vegetables in relation to risk of obesity and weight gain among middle-aged women. *International Journal of Obesity, 28*(12), 1569–1574.

Herdağdelen, A., Zuo, W., Gard-Murray, A., & Bar-Yam, Y. (2012). *An Exploration of Social Identity: The geography and politics of news-sharing communities in Twitter* (arXiv:1202.4393).

Higdon, J., Delage, B., Williams, D., & Dashwood, R. (2007). Cruciferous vegetables and human cancer risk: epidemiologic evidence and mechanistic basis. *Pharmacological Research, 55*(3), 224–236.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M., Mari, W., & Mazaid, M. (2011). *Opening closed regimes: What was the role of social media during the Arab Spring?* (Vol. 2013). Seattle: Project on Information Technology & Political Islam. [online] Available from http://pitpi.org/index.php/2011/09/11/opening-closed-regimes-what-was-the-role-of-social-media-during-the-arab-spring/ Accessed 27.05.13.

Hung, H. C., Joshipura, K. J., Jiang, R., Hu, F. B., Hunter, D., Smith-Warner, S. A., et al. (2004). Fruit and vegetable intake and risk of major chronic disease. *Journal of the National Cancer Institute, 96*(21), 1577–1584.

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the International Conference on Computational Linguistics (COLING), Geneva, Switzerland*.

Kumar, S., Barbier, G., Ali Abbasi, M., & Liu, H. (2011). TweetTracker: an analysis tool for humanitarian and disaster relief, demo. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM-11), July 17–21. Barcelona, Spain*.

Larson, N. I., Story, M. T., & Nelson, M. C. (2009). Neighborhood environments: disparities in access to healthy foods in the US. *American Journal of Preventive Medicine, 36*(1), 74–81.

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-time disease surveillance using Twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1474–1477).

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science, 40*(2), 61–77.

McKinnon, R. A., Reedy, J., Morrissette, M. A., Lytle, L. A., & Yaroch, A. L. (2009). Measures of the Food Environment: a Compilation of the Literature, 1990–2007. *American Journal of Preventive Medicine, 36*(4), 124–133.

Meehan, K., Lunney, T., Curran, K., & McCaughey, A. (2013). Context-aware intelligent recommendation system for tourism. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on* (pp. 328–331). IEEE.

Naaman, M., Becker, H., & Gravano, L. (2011). Hip and trendy: characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology, 62*(5), 902–918.

OpenCalais, A. P. I. (2013). http://www.opencalais.com. Last accessed 11.02.14.

QSR. (2012). *The QSR 50* [online]. Available from http://www.qsrmagazine.com/reports/qsr50-2012-top-50-chart Accessed 07.01.13.

Saif, H., He, Y., & Alani, H. (2012). *Semantic sentiment analysis of Twitter*. The Semantic Web–ISWC 2012 (pp. 508–524). Springer Berlin Heidelberg.

Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control, 38*(3), 182–188.

Shaw, H. (2006). Food deserts: towards the development of a classification. *Geografiska Annaler: Series B, Human Geography, 88*(2), 231–247.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS One, 6*(5), e19467.

Smith, A., & Brenner, J. (2012). *Twitter use 2012*. Washington D.C.: Pew Research Center.

Soule, L. C., Shell, L. W., & Kleen, B. A. (2003). Exploring Internet addiction: demographic characteristics and stereotypes of heavy Internet users. *Journal of Computer Information Systems, 44*(1), 64–73.

Stapleton, P. (2005). Evaluating web-sources: Internet literacy and L2 academic writing. *ELT Journal, 59*(2), 135–143.

USDA ERS. (2013). *Food access research atlas* [online]. Available from http://www.ers.usda.gov/data-products/food-access-research-atlas/go-to-the-atlas.aspx Accessed 30.05.13.

USDA ERS Documentation. (2013). *Food access research atlas* [online]. Available from http://www.ers.usda.gov/data-products/food-access-research-atlas/documentation.aspx Accessed 13.09.13.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 1079–1088). ACM.

Walker, R. E., Keane, C. R., & Burke, J. G. (2010). Disparities and access to healthy food in the United States: a review of food deserts literature. *Health & Place, 16*(5), 876–884.

Wan, X. (2008). Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of EMNLP08, Honolulu, HI* (pp. 553–561).

Widener, M. J., Farber, S., Neutens, T., & Horner, M. W. (2013). Using urban commuting data to calculate a spatiotemporal accessibility measure for food environment studies. *Health & Place, 21*, 1–9.

Widener, M. J., Metcalf, S. S., & Bar-Yam, Y. (2011). Dynamic urban food environments: a temporal analysis of access to healthy foods. *American Journal of Preventive Medicine, 41*(4), 439–441.

Wrigley, N. (2002). 'Food deserts' in British cities: policy context and research priorities. *Urban Studies, 39*(11), 2029–2040.

Xu, C., Wong, D. W., & Yang, C. (2013). Evaluating the "geographical awareness" of individuals: an exploratory analysis of Twitter data. *Cartography and Geographic Information Science, 40*(2), 103–115.

Yerva, S. R., Miklós, Z., & Aberer, K. (2010). It was easy, when apples and blackberries were only fruits. In *The 3rd Web People Search Workshop, Padua, Italy, No. EPFL-WORKING-151616*.

Zemanta, A. P. I. (2013). http://www.zemanta.com. Last accessed 11.02.14.

Zhao, D., & Rosson, M. B. (2009). How and why people Twitter: the role that microblogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work* (pp. 243–252).

Zickuhr, K. (2013). *Who's not online and why*. Washington D.C.: Pew Research Center.

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical & Health Policy, 2*(2), 7–33.