Characterizing Dietary Choices, Nutrition, and Language in Food Deserts via Social Media

Munmun De Choudhury
College of Computing
Georgia Tech
munmund@gatech.edu

Sanket Sharma
College of Computing
Georgia Tech
sanket@gatech.edu

Emre Kiciman
CLUES Group
Microsoft Research
emrek@microsoft.com

ABSTRACT
Social media has emerged as a promising source of data for public health. This paper examines how these platforms can provide empirical quantitative evidence for understanding dietary choices and nutritional challenges in "food deserts" — Census tracts characterized by poor access to healthy and affordable food. We present a study of 3 million food related posts shared on Instagram, and observe that content from food deserts indicate consumption of food high in fat, cholesterol and sugar; a rate higher by 5-17% compared to non-food desert areas. Further, a topic model analysis reveals the ingestion language of food deserts to bear distinct attributes. Finally, we investigate to what extent Instagram ingestion language is able to infer whether a tract is a food desert. We find that a predictive model that uses ingestion topics, socio-economic and food deprivation status attributes yields high accuracy (>80%) and improves over baseline methods by 6-14%. We discuss the role of social media in helping address inequalities in food access and health.

ACM Classification Keywords
H.4 Information Systems Applications: Miscellaneous

Author Keywords
Instagram; nutrition; social media; food; food desert; health

INTRODUCTION
"Food deserts" are urban neighborhoods or rural towns characterized by poor access to healthy and affordable food. These areas are known to be associated with poor diet and diet-related health outcomes, such as obesity, diabetes and cardiovascular disease. US Department of Agriculture's Economic Research Service (USDA) estimates that 23.5 million people in the US live in food deserts[1]. Because food deserts exist mostly in socio-economically disadvantaged areas, concerns have been raised beyond public health. Food deserts may contribute to social disparities, whereby area-level deprivation compounds individual disadvantage [37, 44, 20].

[1]http://apps.ams.usda.gov/fooddeserts/fooddeserts.aspx

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA
©2016 ACM. ISBN 978-1-4503-3592-8/16/02...$15.00
DOI: http://dx.doi.org/10.1145/2818048.2819956

Consequently, there has been an increasing interest in precisely identifying regions of the country likely to be food deserts, including recognizing their nutritional and dietary challenges. Most of these efforts rely on surveys, self-reported information, and sometimes anecdotal evidence from populations of limited size [28, 42]. It has been recognized that some of the current efforts lack rigorous research methods to achieve a scientific approach towards measuring people's access to food in these disadvantaged areas.

In order to curb these limitations, researchers have suggested relying on naturalistic observations as a way to gather better empirical evidence on the health inequalities agenda [30]. Adoption of social media such as Twitter and Facebook has been on the rise. In fact, a rich body of research has emerged which has identified content and language usage in these platforms to reflect individual's and population's milieu [9]. Among the many mundane details individuals are known to share on these platforms, ingestion and dining experiences constitute a unique category [13, 21]. Twitter, for instance, captures a number of minute details about our daily lives, including dietary and dining choices, and prior work has indicated it to be a viable resource that can be leveraged to study ingestion and public health related phenomena [1, 13]. Noteworthy is the prominent social media site Instagram, which has emerged as a popular platform for sharing food related content. In 2013, the Business Insider reported that food photos are a "phenomenon" on Instagram[2]. Due to the visual experience of Instagram, the platform serves as an attractive choice to individuals intending to share photos and videos of the food they are consuming anytime, anywhere.

In this paper, we examine the potential of social media as a "sensor" to capture people's dining experiences and the nutritional information of the food they are consuming, with a particular focus on areas characterized by food deprivation i.e., the food deserts in the US. We are specifically interested in examining the linguistic constructs that can characterize food deserts and the dietary choices there, as well as to what extent social media data may complement conventional means of identifying food deserts by the Census Bureau.

We address the following research questions in this paper:

RQ 1. How can we characterize the dietary choices of food deserts over geographical regions using social media data?

[2]http://www.businessinsider.com/instagram-food-photos-are-a-phenomenon-2013-1

RQ 2. What are the nutritional attributes of ingestion content shared on social media from different food deserts?

RQ 3. How can we model and identify linguistic constructs associated with ingestion content shared from food deserts on social media?

RQ 4. Can the linguistic constructs of ingestion content be utilized to infer food desert status of different areas?

To answer the above research questions, we first develop a statistical matching technique that allows comparison between Instagram food posts from USDA defined areas identified to be food deserts and otherwise. The method controls for geo-cultural and socio-economic differences across areas. Employing this method and based on a dataset of 3 million food posts from Instagram, we are able to empirically confirm and expand insights into several previously speculated characteristics of food consumption in food deserts.

First, we map food related posts from different areas to canonicalized food names and their USDA nutritional profiles[7]. We observe that posts from food deserts depict consumption of food higher in fat, sugar and cholesterol by 5-17% over the same measured in posts from "matching" (i.e., demographically similar) non-food desert areas. Further, a topic modeling approach reveals that food desert ingestion content is distinct from that shared in other similar areas by 8-17%. However, across different regions of the US, there are systematic differences in nutritional characteristics and ingestion topics shared in food deserts posts.

Finally, we propose a predictive model for inferring whether an area qualifies to be a food desert or not, utilizing gold standard labels provided by the USDA. We find that the topic distribution of the areas, when utilized in conjunction with socio-economic attributes and attributes used by the USDA to evaluate an area's food deprivation status, are able to predict the food desert status with high accuracy and precision (>80%). In fact our prediction model improves over a baseline method by 6-14% that uses socio-economic and food deprivation attributes alone.

Through this research, we provide one of the first large-scale empirical evidence into leveraging social media for studying food choices and dietary patterns in disadvantaged areas like food deserts. Our findings thus extend the growing body of literature in social computing on how cues obtained from online social platforms may help inform improved health and wellbeing of populations.

BACKGROUND AND PRIOR WORK

Food Deserts

Food deserts have been a major challenge to local, state, and federal governments in the United States, since they are characterized by socioeconomic inequalities in nutrition environments [42]. Although there is no universally accepted definition of food deserts [20], the United States Department of Agriculture's Economic Research Service (USDA) identifies them to be regions featuring large proportions of households with inadequate access to transportation, and a limited number of food retailers providing fresh produce and healthy groceries for affordable prices[1]. It is reported that bulk of the population in these regions lives more than 10 miles from any supermarket or supercenter. Further, in contrast to other areas, food deserts tend to have populations with lower income and education level, greater poverty rates, larger shares of people who are older and higher numbers of small grocers and convenience stores per capita [7].

Considerable research in this space has focused on assessing healthy food accessibility characteristics associated with these areas, such as income, vehicle availability, and access to public transportation [42]. Despite government efforts to identify nutrition deprivation in these areas, empirical data and observations on the nature of *actual* deprivation is somewhat lacking [7, 45]. Little is understood in terms of the types of food choices adopted by populations in these areas [20]. Some work indicates that fruit and vegetable consumption in these areas is low [28, 14]; others have found, qualitatively, that fat and cholesterol rich food is common [34].

Note that most of these studies on food deserts have been conducted in smaller geographical regions, spanning neighborhoods and states, limiting the scope and generalizability of the findings [8, 14, 2, 34]. Further, often the above information, particularly around identification of food deserts, is collected through surveys on small samples. Such survey data primarily focuses on (1) identifying the locations of different types of stores, and (2) assessing the availability of healthy food options in retail food businesses [28, 42]. Health or economic development officials also purchase proprietary retail data to map the precise locations of retail food providers in different areas throughout the state. Besides lacking statistical power, such methods are expensive and intrusive. They also suffer from the issue that the surveys happen only every few years — areas currently designated to be food deserts are based on 2000 and 2006 Census data[1]. Finally, due to the frequencies in which such survey data are collected, often there is a lag between capturing information about newly opened and recently closed food retail businesses. This has been found to hinder gathering accurate and up-to-date information regarding food access in food deserts [20].

Consequently, there is an identified need for devising complementary ways that can capture the nutritional profiles of populations in food deserts in a national scale. It is also recognized that discovering new types of data sources that can provide more fine-grained information towards identification of food deserts will be helpful [42]. In this paper, we examine the potential of leveraging self-reported information on diet shared on social media throughout the United States as a way to circumvent many of these issues. We also intend to provide a complementary perspective beyond what is known about the nutritional limitations of food deserts. We believe the ability to directly measure, via social media, ingestion related language, can enable more accurate identification of such areas, instead of inferring the same through low-income and low-food-access criteria[1].

Social Media and Food

There is a growing body of social media research focusing on identifying linguistic characteristics of content around food and dietary patterns [13]. West et al. and Wagner et al. leveraged recipes and their associated information shared on recipe websites to extract food names. They then used them as a proxy to derive consumption and dietary patterns of individuals [43, 41]. Aiming to characterize dietary patterns of different counties in the US and their public health characteristics, Abbar et al. [1] proposed a method to derive nutritional content in Twitter posts. They observed a variety of correlations between Twitter derived nutritional information and prevalence of obesity and diabetes in different counties in the US. Although this work does not focus on food deserts, the approach has motivated our investigation. Most relevantly, Sharma and De Choudhury [36] focused on using Instagram to study ingestion practices and nutritional patterns, and identified how the broader Instagram community responds to low and high calorific food. Additionally, in a recent work, Mejova et al. [21] also utilized the Instagram platform to identify obesity patterns. This two pieces of work inspired the choice of Instagram as the platform of investigation in this paper. We expand this existing emergent body of work by leveraging food related content on Instagram to understand food choices and nutritional characteristics in food deserts of the US.

Social Media, Health, and Well-being

Social media research has indicated that psychological states, health, and well-being status may be gleaned via analysis of language and online social interactional patterns. These include understanding conditions, health statistics [6] and symptoms related to diseases [27], influenza propagation [33], substance abuse [23, 24], mental health [11, 26, 16, 10, 39], insomnia [18] and others. Moreover, research has showed social media to illustrate several geographical attributes of populations. Twitter was used by Quercia et al. [31] to quantify sentiment across neighborhoods of different socioeconomic standing. In a work close to ours, topical characteristics of Twitter were used to find association with deprivation scores of areas [40]. Schwartz et al. [35] correlated life satisfaction score of counties with socioeconomic factors and Twitter language. This emergent body of work has established the viability of social media data to complement conventional measurements of population health and well-being, and in being able to provide a less intrusive and more scalable way to collect and characterize health data and related phenomena.

Our paper builds on this growing body of work by focusing on populations living in disadvantaged geographical areas like food deserts. We investigate how social media may be able to shed important insights into understanding their nutritional health as well as how to better characterize and identify them.

DATA

Social Media Data

We utilized data obtained from the popular social media platform Instagram. Instagram is a fast growing photo-sharing platform with an underlying social network. Users can share public or private photo posts, often tagging them with topical or other descriptive terms. Currently 26% online adults use Instagram and 53% of young adults ages 18-29 use the service as of 2014[3]. As indicated earlier, the platform is extensively used for sharing ingestion and dietary experiences of people [21] — ranging from bragging about eating healthy foods and adoption and maintenance of healthy lifestyles; confessions, preferences, culinary weaknesses, and cravings about high calorie intake; everyday dining outings, thoughts and feelings about food or "#foodporn"; food and eating related journaling; to providing and soliciting social recommendations about restaurants, cooking and recipes to one's social network (examples in Table 1).

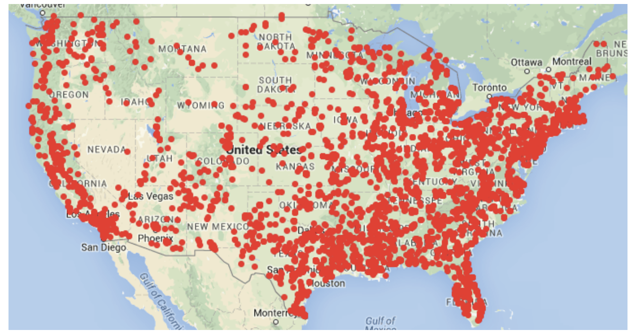

Figure 1: Food deserts visualized over the US map. Food deserts in the West and Mid-West regions appear sparse due to the larger size of the Census tracts.

For the purposes of data collection, we leveraged Instagram's official API[4] to obtain public posts and associated meta-data on ingestion activities. We first referred to the work of Sharma and De Choudhury [36] to obtain a list of 588 food-related words which were likely to be typically used as tags to describe food content on Instagram. Sharma and De Choudhury [36] compiled this list by deploying both automated and manual coding and filtering methods. They primarily relied on a popular online food vocabulary word list[5], also used in [13]. Examples of such food name tags include 'chocolate', 'oatmeal', 'kale', 'beef', 'hummus', 'tofu', 'thyme'. We enriched this list of food names by identifying the items which were fruits and vegetables – we used annotations from two nutritionists for the purpose, with an inter-rater agreement of $\kappa = .99$ — this classification was for RQ 1.

Based on this list of food tags (we will refer to them as canonical food names like [36]), we started a crawl of Instagram posts. The Instagram API does not allow us to query with both geolocation and tags, so we collected data using only the tags. For the purposes of this paper, we considered only English language public posts and associated meta-data that were tagged with at least one of the canonical food names as tags. We did not download the image or video themselves for our analysis, except for exemplary purposes. Our final dataset consisted of over 14 million posts from over 8 million Instagram users, which were shared in a timeframe between July 2013 and March 2015. Out of this set, 35.5% posts had geolocation tags. We did not observe any systematic difference in posting activity distribution between the set of geocoded and non-geocoded posts (based on an independent sample t-test).

[3]http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

[4]http://instagram.com/developer/

[5]http://www.enchantedlearning.com/wordlist/food.shtml

Post	Post tags	Canonical name(s)	Energy	Sugar	Fat	Chol.	Fiber	Protein
	butter, cakes, peanut, jelly, kellyloucakes, tea, decorating, cupcake	butter, cake, peanut, jelly, tea, cupcake	436.26	95.9	25.3	256	3.08	9.401
	healthyfood, meal, goodfood, foodgasm, carrots, vitamin, cucumber, veggies, foodisfuel, corns, lime, beetroot, nofilter, salad, potato, instafood, eatcleanmenu, apples, rich, fruits	carrot, cucumber, corn, lime, beetroot, salad, potato, apple	206.01	81.1	16.8	86	79	25.4
	strawberry, strudel, dessert, sweet_taste, jar, pastry	strawberry, strudel, pastry	322.91	21.3	80.2	114	10.2	4.202
	cajun, instagram, monday, foodporn, food, bestoftheday, instagramhub, oregano, cook, breakfast, instagood, instadaily, ig, bread, tomato, organic, iggers, egg, iphoneonly, iphonesia, morning, yum, iphone, fresh, spice	oregano, bread, tomato, egg	205.97	43.9	73.6	230	42.5	84.63
	food, dessert, lovelife, chocolate, cookie, delicious, tasty, raspberry, yummy, dough, dinner, pudding, treat, loveit, epic, pizza	chocolate, cookie, raspberry, pudding, dough, pizza	371.63	23.7	15.7	221	35.2	5.952

Table 1: Example Instagram posts with their tags, matching canonical food names, and their derived nutritional profile. Here energy is given in kcal, all other nutrients are in grams, except cholesterol (Chol.) which is in milligrams. Images are shown for exemplary purposes and were not included in our approach.

Extracting Nutritional Information

Since our goal is to characterize nutritional challenges in food deserts, we now present a two step approach to measure nutritional information of the Instagram posts. Our approach utilized the one in [36] that has been found to accurately describe the nutritional information in Instagram posts with 89% accuracy; similar approaches were also used in [1, 13].

Specifically, like [36] we referred to the official US Department of Agriculture (USDA) National Nutrient Database for Standard Reference database[6]. This resource provides precise nutritional values of over 30 nutrients for 8,618 food items, spanning calorific content, protein, fat, cholesterol, sugar, fiber etc. Further, food items in the USDA database are described in varying granularities and contain detailed illustrations of the ingredients and method of preparation, referred to as “food descriptors”. Note that the default nutritional information is reported based on per 100 grams of serving, which is the portion size of food we use to describe Instagram posts.

The method is briefly described as follows: We first developed a regular expression matching framework in which each tag in a given post was compared to the items in above described list of canonical food names. A second matching framework was developed to map the canonical food names corresponding to a post’s tags to the USDA food descriptors. This allowed us to associate a nutritional profile defined by the USDA to each post. Posts with no matches to USDA descriptors were disregarded. We pursued using the following six major nutritional information in our analysis — energy (kcal), protein (g), fat (g), cholesterol (mg), sugar (g) and fiber (g). For posts with more than one match with USDA food descriptors, we computed aggregate nutrient information based on the average across all matches. Finally, we were able to extract nutrient information in 93.5% posts in our Instagram dataset. Table 1 provides examples of posts with USDA derived nutritional information.

[6]http://ndb.nal.usda.gov/

Food Desert Data

In a parallel data collection task, from United States Census databases, we obtained cartographic information on 69,401 tracts throughout the US[7], of which 4484 tracts are officially identified to be food deserts by the USDA, per 2000 and 2006 Census data[8]. Census tracts are relatively permanent subdivisions of a county and usually have between 2,500 and 8,000 people. Census tracts do not cross county boundaries, and are designed, when established, to be homogeneous with respect to population characteristics, economic status, and living conditions. Tracts are the smallest granularity at which food deserts are defined by the USDA. Refer to Figure 1 for a map of the food deserts throughout the US. Additionally, for each Census tract (both food deserts and non-food deserts), we obtained the most recent (2014) socio-economic information based on the Federal Financial Institutions Examination Council (FFIEC) Online Census Data System[9]. A list of the socio-economic variables collected is given in Table 3.

Mapping Posts to Food Deserts

[7]http://www.census.gov/geo/maps-data/data/cbf/cbf_tracts.html
[8]http://www.ers.usda.gov/data/fooddesert
[9]http://www.ffiec.gov/census/Default.aspx

Region	States	FD Posts	FD Users	NFD Posts	NFD Users
North East	PA, NY, VT, NH, ME, MA, RI, CT, NJ, DE, MD, Washington DC	18,985	10,146	812,991	261,389
West	MT, WY, CO, UT, NV, CA, OR, WA, ID, AK, HI	30,253	17,456	1,099,633	361,281
Mid West	ND, SD, NE, KS, MO, IA, MN, WI, IL, IN, MI, OH	16,471	9,214	313,875	123,428
South West	AZ, NM, TX, OK	23,937	14,048	233,382	91,051
South East	AR, LA, TN, KY, WV, VA, NC, SC, GA, FL, AL, MS	27,994	15,408	495,889	188,102
Total		117,640	66,272	2,955,770	934,200

Table 2: Number of Instagram posts and users in food deserts and non-food deserts per geographic region.

We discuss our method of identifying the Census tract information associated with each Instagram post. Through this task, we aim to map Instagram posts to food desert and other tracts. Since our tag based crawl returned food posts from around the world and many of them did not have any geotags associated, we first filtered those with valid latitude-longitude information. We then utilized the Federal Communications Commission (FCC) API[10] to query the latitude-longitude pair of each post for possible mapping to one of the US Census tracts. The API query returns what is known as an FIPS code: a 15 character Census Bureau Census Block number (blocks are the smallest geographical units officially defined by the Census; several blocks make up a tract). The first 11 digits of this FIPS code uniquely identifies a tract. For lat-long coordinates outside US the API returns a null FIPS code.

This way we mapped over 3M posts to the 69,401 tracts: 117,640 posts in food deserts and 2,955,770 posts in non-food deserts. Table 2 gives statistics of the number of posts and unique users associated with each of the five official geographical regions of the US — North East (NE), West (W), Mid West (MW), South West (SW), and South East (SE).

METHODS

Generating Matched Samples

We note that all of our four research questions (RQ 1-4) involve comparison of food desert Instagram food content to that from the non-food desert tracts. However, given that our data is observational, we need to ensure that we control for confounding variables unrelated to food desert characterization of tracts that could impact ingestion language manifested on the platform. For instance, prior literature indicates that ingestion and related language are influenced by demographic attributes, e.g., income, race and ethnicity, geography and so on [22]. In fact, low income alone has been known to be associated with poor nutrition, irrespective of whether the population is from a food desert or not [37].

Statistics literature indicates that confounding or latent variable bias in the study of an outcome is ideally confronted through randomized experiments, where two separate grounds are carefully crafted: the "treatment group" (a population exposed to conditions hypothesized to affect the outcome) and the "control" (unexposed) group [46]. However for a study like ours, randomly assigning populations to live in food desert and non-food desert tracts presents obvious ethical and practical challenges. Regression modeling and propensity score matching are widely adopted methods in cases where randomized experimentation is not possible [46]. Note however that latent bias can still exist when unobserved variables affect treatment status or outcomes. Regression modeling can further be challenging when the dimensional space of variables is large, or when the variables are likely to have mutual interactions.

[10]http://www.fcc.gov/developers/census-block-conversions-api

Similarity Features	
% minority† population	population
% non-Hispanic whites	#households
median house age	median family income
owner occupied housing units	#families
distressed/underserved tract‡,§	

Table 3: Similarity features used for creating the matched samples. †Census defines minority to be anyone who is not non-Hispanic white. ‡ Binary variable. § The FFIEC[9] tates that a tract is considered distressed if it is in a county with one or more of the following: an unemployment rate of at least 1.5 times the national average; a poverty rate of 20 percent or higher; a population loss of 10 percent (or more) since the previous census; or a net migration loss of 5 percent (or more) during the five-year period preceding the most recent census. A tract is designated as underserved if it meets criteria for population size, density and dispersion that indicate the area's population is sufficiently small, thin and distant from the population center that the tract is likely to have difficulty financing the fixed costs of meeting essential community needs.

Hence in this paper we developed a matching methodology to "match" food deserts ("treatment group") to non-food deserts ("control group"), so that it would control for confounding variables — in our case these being socio-economic and geo-cultural attributes. Our method is motivated from recent work on utilizing matching and stratification to reduce latent variable bias in social media studies of health [12]. Our technique used the following steps:

- We constructed region-wise sets of food desert and non-food desert tracts (see Table 2 for the five regions used). Dividing into regions allowed us to understand patterns controlling for geo-cultural characteristics.
- Given a geographic region and the corresponding food desert and non-food desert tracts, we used the socio-economic variables described in Table 3 to compute distances between all pairs of the tracts. We used the Mahalanobis distance metric [3], which is suitable to detect similarity between multi-dimensional objects and is a generalization of the Euclidian distance metric. We weight all variables equally.
- For each food desert tract in the different geographic regions, we employed the k Nearest Neighbors algorithm to identify a matched sample of k most (socio-economically

and geo-culturally) *similar* non-food desert tracts. [11] For the purposes of this paper, we chose $k = 20$. [12]

- Based on caliper matching [32], we disregarded those food desert tracts which had fewer than 20 non-food desert tract matches at or above the 50% similarity threshold.

We note that per this method, a non-food desert tract may be matched to more than one food desert, since we select non-food deserts for matching with replacement. However since Census tracts are homogenous, we expect this choice to impact all food desert matches uniformly without bias.

The above method gave us 4365 matched samples from the 4484 food desert tracts, with each matched sample containing 20 non-food deserts. The ensuing comparisons of food deserts and non-food deserts (RQ 1-3) use the aggregate statistics across all of these matched non-food desert samples corresponding to each food desert tract. The matched samples would allow us to determine the *expected* food choice, nutrition and ingestion language for populations in non-food deserts. They would thereafter let us examine to what extent these patterns are distinct from that in food desert tracts with similar socio-economic and geo-cultural characteristics.

Modeling Ingestion Language

Next we discuss our method of modeling ingestion language of food desert and matching non-food desert Instagram posts. We employ topic modeling for the purpose (Latent Dirichlet Allocation [15]), a method that has been commonly employed to analyze health related social media data [27], as well as to cluster food related social media posts [13]. We expect LDA to identify topics around patterns in diet, language, and lifestyle, thereby allowing us to go beyond specific food names in our data.

For the combined set of posts spanning all food deserts and their matched non-food desert tracts, we obtain topics by running the online version of LDA given in Python Gensim library. We use the default hyper-parameter settings; 100 topics were found to work well in initial experiments. Thereafter we compute the posterior probability of each topic in each post belonging to food deserts and their matching non-food deserts. Finally, we obtain mean probability of the topics from the ratio of the probability of topics across all posts in a tract to the total number of unique users in that tract (a method similar to [6]).

We propose two measures to compare topic distribution of food desert tracts with that of matching other tracts: (1) *Jensen-Shannon (JS) divergence* [3]; and (2) *Topical content difference*. The latter measure is given by the mean normalized difference between the frequencies of tags belonging to each topic in food deserts and the frequencies of tags belonging to the same topic in non-food desert tracts: $1/K\{\sum_{i=1}^{K}\{1/N_i \sum_{j=1}^{N_i}(|n_{ij}(f) - n_{ij}(o)|/n_{ij}(o))\}\}$, where $n_{ij}(f)$ is the frequency of tag j in topic i for food desert tracts, $n_{ij}(o)$ is the frequency for the same tag in the same topic in matching other tracts, N_i is the number of tags in topic i and K is the total number of topics.

Next we devised an iterative greedy strategy to identity the subset of topics which would *distinctively* characterize food desert tracts versus others, motivated from [10]:

(1) We sort the topic probability distribution of each food desert tract in a region from the highest to least. We iteratively eliminate topics from the sorted distribution, starting with the one with maximum probability. Correspondingly, we eliminate the same topic from the mean topic probability distribution of the tract's matched non food desert samples.

(2) In each iteration, we compute the JS divergence between the food desert and its matching non-food deserts, using the topic distribution over the uneliminated topics. We terminate the elimination task when the JS divergence between the topic distributions of the food desert and its matched sample non-food deserts is minimum (~0). The set of eliminated topics is then considered to be the topics that distinctly characterize ingestion language in food deserts, since they maximize the divergence between the topic distributions of the food deserts and their corresponding matching tracts.

Predicting Food Desert Status

To what extent can ingestion related language on social media, as captured through topic distribution, predict the food desert status of any given Census tract? Corresponding to this RQ 4, we propose the prediction task to be a binary classification task — our goal is to predict the USDA defined food desert status of a tract. We develop three different classification models with different sets of features, corresponding to each of the five regions. Our first model, referred to as **S + F**, uses respectively the socio-economic attributes of tracts (Table 3) and the USDA identified attributes of food deprivation (Table 4). In a second model **T** we include as features the LDA topics derived from the tracts' ingestion language. The final model **S + F + T** combines the socio-economic, food deprivation status and LDA topics as features. On each of these classification models, we apply Principal Component Analysis (PCA) [3] to reduce the high dimensionality of our feature space (the **S + F +T** model has 117 features), reduce the effect of correlated features, handle sparsity, as well as to assign appropriate weights to the most predictive features.

Finally, we use a binary Support Vector Machine (SVM) classifier (with a linear kernel) [3] to infer food desert status in a region, as SVMs are adept at handling large dimensionality of and arbitrary relationships in data.

RESULTS

RQ 1: Dietary Choices in Food Deserts

Corresponding to our first research question, we begin by investigating the dietary choices in food desert tracts versus their matching non-food desert counterparts. In Table 5 we report the presence of different canonical food names as tags in Instagram posts shared from the food deserts and their

[11] While some studies use paired matches (i.e., k=1) we choose to oversample our matches to reduce the variance in our matched comparisons, though we do note the trade-off of an increase in bias due to comparing to additional neighbors that are a slightly poorer fit than the nearest neighbor.

[12] The value of k was chosen via a model selection procedure [29], in which for different values of k between 1 and 50, we built one model each, estimated the log likelihood of each model, and then computed the Bayesian Information Criterion (BIC) for them. $k = 20$ was the model for which the BIC was minimized , giving the optimal k.

Food Deprivation/Desert Status Features	
% below poverty line	
% low access, low income people	urban/rural‡
% low access 0-17yrs	% low access 65+yrs
% group quarters§ population	vehicle access‡
% low access housing units	

Table 4: Food deprivation/desert status features used by the USDA to identify a tract to be a food desert (or not). ‡ Binary variable. § Census[9] defines Group Quarters (GQ) to be "places where people live or stay, in a group living arrangement, which is owned or managed by an entity or organization providing housing and/or services for the residents".

More FD	LLR	Eq. Freq	LLR	Less FD	LLR
Mid West (MW)					
hamburger	1.5687	soup	0.0493	bean	-1.1864
hotdog	1.1679	rice	0.0485	turkey	-0.2152
brisket	0.7562	pudding	0.0339	spinach	-0.1881
meat	0.4914	cake	0.0285	kale	-0.1773
pork	0.2316	pawpaw	0.0037	cucumber	-0.1019
West (W)					
pie	0.7927	sauerkraut	0.0420	quinoa	-0.4935
beef	0.3108	coffee	0.0375	apple	-0.4397
sausage	0.2417	cherry	0.0240	chicken	-0.3528
potato	0.1490	steak	-0.0166	crab	-0.2682
corn	0.1269	chocolate	-0.0614	blackberry	-0.1770
North East (NE)					
pork	0.4346	bread	0.1208	bagel	-0.3666
mayonnaise	0.2646	pizza	0.0910	kale	-0.2354
cookie	0.2144	lox	0.0536	hummus	-0.1667
pasta	0.2100	chowder	0.0404	soup	-0.1257
milkshake	0.2096	cheesecake	0.0347	mushroom	-0.0470
South East (SE)					
bacon	0.2980	okra	0.0319	collardgreen	-0.3122
potato	0.2047	biscuit	0.0198	orange	-0.3080
brisket	0.1302	cajun	0.0005	peach	-0.2221
grits	0.1229	chicken	-0.0001	bean	-0.2150
sweetpotato	0.0705	taco	-0.0484	pecan	-0.1163
South West (SW)					
barbecue	0.9506	salsa	-0.0092	tomato	-0.8385
meat	0.4501	guacamole	-0.0401	asparagus	-0.2866
pork	0.2778	taco	-0.0401	banana	-0.2771
burrito	0.0684	tamale	-0.0917	lemon	-0.1765
rice	0.0387	jalapeno	-0.1101	pepper	-0.1602

Table 5: Prevalence of different canonical food names with their log likelihood ratios in FDs and non-FDs. The Log-likelihood ratios were computed only for the canonical food names which appear at least five times in either FD posts or non-FD posts.

matched non-food deserts. We calculate the log likelihood ratios (LLR) of each of the canonical names. It is given as the natural logarithm of the ratio between their normalized frequency of occurrence in each food desert of a region, and that in the matching non-food deserts corresponding to each food desert. Here normalized frequency of a canonical name in a food desert is given by the total frequency of occurrence of the term divided by the number of unique users in the tract. We obtain an aggregate value of the LLR of each canonical name by taking its mean across all food deserts and their matches in a region.

Note Table 5 has three categories of canonical food names and their LLRs for each region — (1) leftmost column are the food names with the most positive LLR, i.e., they appear more frequently in food deserts compared to non-food deserts; (2) rightmost column comprises the food names that are more frequent in non-food deserts, i.e., they have the most negative LLRs; and finally, (3) the middle column indicates those food names whose LLRs are close to zero, i.e., they are almost equally likely in both food and non-food deserts.

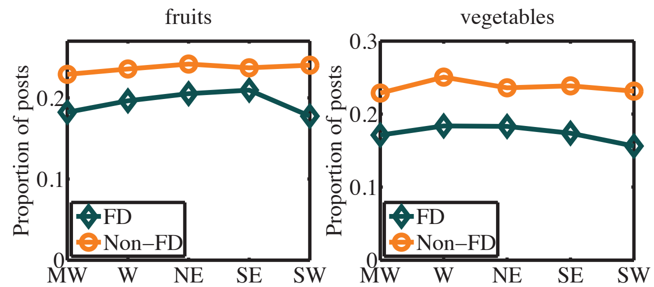

Figure 2: Proportion of posts containing at least one fruit or vegetable canonical name shared in food deserts (FD) and matching non-food deserts (non-FD). Results are shown corresponding to the five US geographical regions.

We find that the food names in the middle column of Table 5 (equally frequent) capture the cultural food habits and dietary preferences of populations in the different regions of the US. For MW, "soup", "pudding", "cake" and "pawpaw" are known to be central to the local cuisine [13]. In the case of the western states, we find that "sauerkraut" and "steak" are prominent — food items widely consumed in the Great Plains [4]. Further, "coffee" is popular in the Pacific Northwestern states (Washington), and "cherries" are grown in Oregon. Instagram posts from the northeastern states indicate what the region's cuisine is known for — "pizza" (New York), "chowder" (New England cuisine), "cheesecake" (Pennsylvania). We observe predominance of Tex-Mex and Mexican food names in posts from the southwestern region: "salsa", "guacamole", "taco", "jalapeno". Finally, the characteristic Southern cuisine manifests itself in the posts from the SE states — "okra", "biscuit", "cajun", "chicken". Broadly, the prevalence of these food names in both food deserts and non-food deserts indicates that ingestion related content on Instagram captures the dietary preferences in different parts of the US. These observations also align with observations in recent work on utilizing social media for identifying dietary patterns [13].

Food items more extensively mentioned in Instagram posts from food deserts reveal distinctively that high calorific food is common. For MW, "hamburger", "hotdog", "meat" characterize the food deserts, whereas "spinach", "kale", "cucumber" are predominant in the non-food deserts of the same region. Similarly, "sausage", "potato", "corn" appear more in posts from food deserts of the western states. In non-food deserts from the same region, we observe mentions of "apple", "crab" and "blackberry" — all of which are characteristic food items from the region [1]. Northeastern food desert posts mention "mayonnaise" and "milkshake", while those from the non-food deserts mention "kale", "soup", "mushroom" etc. Finally, food desert posts from SW and SE, while capturing the cultural dietary choices of the regions, tend to primarily focus on items on the high calorific side of the nutritional spectrum, including "meat", "burrito", "rice" (SW); "sweetpotato", "brisket", "grits" (SE).

Fruit and Vegetable Consumption. Next, we specifically examine the extent of consumption of fruits and vegetables in the food deserts of the different regions, versus their matching non-food deserts of the same region. This analysis is motivated from prior work [28, 14] which have argued that the consumption of these items in food deserts is limited. In Figure 2 we report the proportion of Instagram posts that contain at least one canonical food name tag labeled as a fruit or a vegetable in food desert and other tracts. Food desert posts, across all regions, are less frequently tagged with fruit and vegetable names — this difference is significant (fruit: $F = -4.4; p < 10^{-3}$; vegetable: $F = -6.1; p < 10^{-4}$) based on the Clifford, Richardson, and Hemon [5] (CRH test), a method that corrects traditional p-value calculation by taking into account spatial autocorrelation in data. Summarily, while from Table 5 we observed cultural influences in *which* food names are common in different regions, the differences between fruit and vegetable mentions in food and non-food deserts show noted variation.

RQ 2: Nutritional Profiles of Food Deserts

Corresponding to RQ 2, we investigate whether, controlling for socio-economic and geo-cultural factors, food deserts' dietary choices are less nutritional. For the purpose, we compare the values of inferred energy (kcal), sugar, fat, protein, fiber and cholesterol between food deserts and their matched non-food desert post samples (Figure 3). In the figure, we additionally show the overall consumption of these nutrients in posts across *all* food deserts and matching other tracts.

We observe that for all regions, Instagram manifested consumption in food deserts is characterized by higher calorific content, high sugar, fat and cholesterol, however low protein and fiber food. On examining whether these differences are statistically valid, from Table 6 we find that across all regions, sugar, fat and cholesterol intake manifested in food desert posts is significantly higher relative to their matching non-food desert posts, based on Clifford, Richardson, and Hemon [5] (CRH) tests. A deeper investigation of the intake of various nutrients in different regions, however, shows notable differences between Instagram content from food deserts and non-food deserts.

The nutritional differences are most evident for the West (W) and South West (SW) regions. For the former, everything except protein shows significance, whereas for the latter, everything except calorific content (energy) does. In fact, statistics indicate that the Great Plains and the Rocky Mountains regions have one of the highest prevalence of food deserts [4]. Mid West (MW) exhibits the next most distinct differences, with sugar being the only nutrient without statistically significant differences across the two cohorts. South East (SE) is the region where the least number of nutrients show statistically significant differences between the two groups. Note that this finding may seem counter-intuitive at first, given that the SE region has one of the highest percentage of food desert tracts (see Figure 1) and that the Center for Disease Control's Behavioral Risk Factor Surveillance System (BRFSS) 2012 survey[13] reports the SE region to have the highest prevalence of diabetes. However, statistics from the President's Council

[13] http://kff.org/other/state-indicator/adults-with-diabetes/

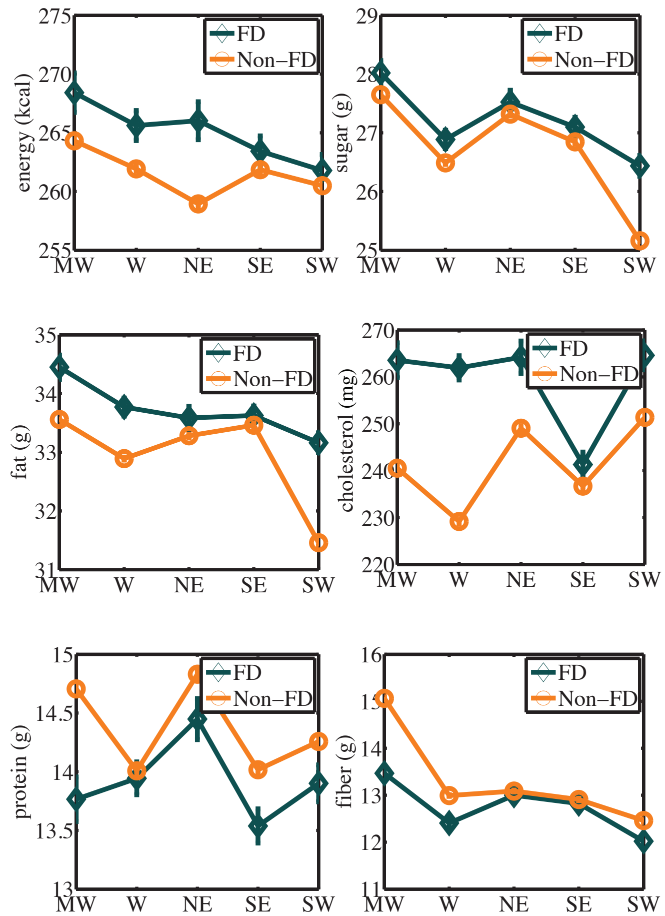

Figure 3: Nutritional measurements (means) in food deserts (FD) and non-food deserts (non-ED) corresponding to the five US geographical regions.

of Fitness, Sports, and Nutrition would indicate that SE, in general, has relatively high intake of calorific food, as well as food rich in fat and cholesterol. Broadly, these differences across regions show that ingestion and eating habits are often influenced by geo-cultural attributes. Naturally, the manner in which the presence of food deserts impacts public health and nutrition needs to take into account the particular context of geography and culture that are likely to drive food consumption of specific populations.

Summarily, the findings provide evidence to the previously speculated but not empirically validated hypothesis [7] that the nutritional differences that characterize food deserts against other tracts is the high sugar, fat and cholesterol consumption, and against popular intuition, they are not distinct in terms of the amount of calories consumed.

RQ 3: Linguistic Signatures of Food Deserts

Per RQ 3, we begin by reporting the distinctions between ingestion language in food deserts of different regions versus their matching non-food deserts. Table 7 gives the mean

	MW		W		NE		SE		SW		All	
	F	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*	*F*	*p*
Energy	3.101	*	3.766	***	5.897	***	1.514	–	1.194	–	1.218	–
Sugar	1.799	–	2.687	*	1.102	–	1.684	–	7.558	***	5.09	***
Fat	3.074	*	4.309	***	1.167	–	0.747	–	7.224	***	5.643	***
Cholesterol	4.447	***	9.692	***	3.42	**	1.259	–	2.979	*	5.285	***
Protein	-7.641	***	-0.72	–	-3.31	**	-5.587	***	-3.758	***	-1.584	–
Fiber	-12.105	***	-6.292	***	-0.829	–	-0.932	–	-4.101	***	-1.493	–

Table 6: Statistical significance comparison between nutritional attributes of food deserts and non-food deserts. Independent sample *t*-tests with Clifford, Richardson, Hemon correction for spatially autocorrelated data were used, along with Bonferroni correction ($\alpha/6$) to correct for familywise error rate ($\alpha = 0.05, 0.01, 0.001$).

	MW	W	NE	SE	SW
JS div.	0.32	0.61	0.24	0.13	0.53
	(±.05)	(±.11)	(±.09)	(±.03)	(±.07)
TC Diff.	12.7%	16.5%	10.3%	7.6%	14.6%
	(±2.4)	(±2.9)	(±1.5)	(±1.4)	(±2.5)
F	4.26**	8.39***	3.99**	1.84*	6.58***
TE	24%	38%	17%	14%	29%

Table 7: Mean (and std. dev.) of JS divergence and Topical content difference (TC diff.) between topic distributions of food deserts and non-food deserts. Significance results are based on a modified *t*-test that uses the Clifford, Richardson, Hemon correction for spatially autocorrelated data. Bonferroini correction ($\alpha/5$) was used to control for familywise error rate ($\alpha = 0.05, 0.01, 0.001$). The last row corresponds to the proportion of topics eliminated (TE) through our greedy iterative strategy, before the topic distribution of food deserts approximates that of the matching non-food deserts.

Jensen Shannon (JS) divergence and topical content difference between LDA derived topics from all food deserts corresponding to a particular region and the matching non-food deserts in the same region. There are consistent differences across regions. The distinction of ingestion topics is most significant for West (W) followed by South West (SW); for South East (SE) it is the least. These findings bolster our observations in RQ 2, that there are not only varied nutritional differences between food deserts and non-food deserts in different regions, but even the associated ingestion language follows similar variation.

Next we present results of the iterative greedy topic elimination task on posts from food deserts and their matching non-food deserts in each region. We find that *different* number of topics are eliminated in different regions (Table 7), before the topic distribution of food deserts approaches that of the non-food deserts. For instance, 38 topics (out of 100) are eliminated for food desert posts in West (W) compared 14 for South East (SE). Note that the former is the region with highest JS divergence between ingestion topics in food deserts and non-food deserts, while the latter with the least. Similarly, 24 topics are eliminated for MW food desert posts, 17 for NE and 29 for SW. In essence, *more* topics are eliminated for regions where ingestion topics of food deserts is significantly distinct from that of the matching non-food desert tracts.

Nevertheless, it is important to note here that the specific topics that are eliminated corresponding to each region, are not consistently the same topics. To test if the topics eliminated across regions are significantly distinct, we performed Kruskal Wallis one-way analysis of variance test on the tag frequenct distributions associated with the eliminated topics in the five regions — the outcome was statistically significant ($F(df = 17, 991) = 6.35; p < 10^{-6}$).

What are the characteristics of these eliminated topics in each region? For this, we refer to the topic descriptions: ten most frequent tags in two eliminated topics per region with the highest posterior probability (Table 8). We observe that in general the eliminated topics describe consumption of less nutritional or high calorific food ("bacon", "pizza", "icecream", "burger", "fries", "macandcheese", "hotdogs"). Eliminated topics also capture attributes of lifestyles and casual expressions associated with such dietary patterns ("foodporn", "allyoucaneat", "friedfood", "burgerporn", "foodbeast", "fatty"). Finally, as also observed in the results of RQ 1, there are specific high calorific food item mentions in the eliminated topics that are characteristic of the regions ("burrito" in SW; "bbq", "brisket", "poboy" in SE).

Together, we conclude that ingestion language, as captured by these eliminated topics in each region, reflect the dietary habits in their corresponding food deserts.

RQ 4: Predicting Food Desert Status

For our final research question, we examine the potential of using the topic distribution, socio-economic attributes of tracts, as well as USDA defined food deprivation features of a tract to infer their food desert status. For the purposes of classification, we identified a random sample of 10% tracts as our test set, and performed k-fold cross validation ($k = 10$) on the remaining 90% tracts for parameter tuning purposes. The results in Table 9 are based on classifying the test set.

The results of the classification task using our full model **S + F + T** are given in Table 9 (see ROC curves in Figure 4). We find that the prediction performance of **S + F + T** varies based on the region. Particularly, we observe highest accuracy (88.3% and 84.1%) and F1 score (.86 and .82) for those regions where the JS divergence between the topic distributions of the food deserts and non-food deserts were high (W and SW — Table 7). Hence we conjecture that in these regions, the ingestion topics of the tracts bear valuable 'signatures' relating to their food desert status. Conversely, perfor-

	Topic	Representative tags
MW	id=157	pork, shrimp, porkshoulder, creamcheese, foodporn, chicken, capers, food, crinklefries, bagel
	id=3	burgers, cheeseburger, yumminess, chicago, burger, burgerporn, holistic, foodbeast, mediumrare, seasoning
SW	id=34	bacon, brunch, breakfast, omelet, atx, austin, cheese, pastrami, egg, caterer
	id=159	instayum, pizza, pepperoni, mexicanfood, fatty, burritto, cheatmeal, riceandbeans, meal, bake
NE	id=36	dessert, nyc, chocolate, icecream, foodporn, brooklyn, sherbet, dessertporn, food, cooking
	id=123	egg, roasted, poached, fried, culinary, balsamic, buttery, cheese, nyc, dinner
W	id=64	risotto, toffee, addicted, candy, teriyaki, muffin, pancakes, wildrice, allyoucaneat, hotdogs
	id=103	mushroom, sausage, burger, fries, sundayfunday, frenchfries, nomnom, hashbrown, selfie, toast
SE	id=156	brisket, bbq, ribs, macandcheese, cream, pulledpork oil, foodporn, grill, thc
	id=149	chili, crust, nutmeg, southern, catfish, veganism, friedfood, bratwurst, icancook, poboy

Table 8: Two eliminated topics with highest posterior probabilities associated with the food deserts in each of the five regions of the US. Corresponding to each of the two topics per region, we indicate the topic id and also show the 10 most representative (highest frequency) tags for each of them (each row).

mance is relatively lower in the case of the SE and NE regions since their topic distributions were closer for food deserts and non-food deserts (Table 7). The nutrient levels in the posts from food deserts and non-food deserts in this region also showed less difference (Figure 3).

Baseline Comparison

We now compare the performance of this binary classifier to the baseline model **S + F** which uses only the socio-economic and the food deprivation features per region, and the model **T** that uses the topics alone. Overall **S + F** performs notably worse compared to our model **S + F + T** (ref. Table 9 and Figure 4). Mean accuracy of this baseline is less by 11%, precision by 11%, recall by 13%, and F1 by 12%. Model **T** performs the worst of the three models (not shown for brevity). We find that for this model, the mean accuracy across regions is 61.46%, and the mean F1 is .64. We conjecture that since this model disregards any of the larger population and food access characteristics, social media ingestion topics on their own are not adequate in assessing food desert status of tracts.

The extent to which adding Instagram ingestion topics to socio-economic and food deprivation features improves prediction differs significantly depending on the region. The **S + F + T** model performs very well over **S + F** for W and NE because these regions are the most urban (per Census 2010, 89.8% and 85% population in these regions are urban[14]) compared to the other regions. Further in our data, prevalence of Instagram use in populations in W is .54% and in NE is .48%[15]. Both these proportions are higher compared to other regions (Wilcoxon rank sum test, $p < .001$). That is, we conjecture that tracts in these regions have a larger and more diverse population using Instagram, including more individuals who may be residing in food desert tracts. This is likely to contribute to better prediction performance in these regions.

However, we note that the **S + F + T** model gives only modest improvements over **S + F** for the SE region. We explain this finding per our observations from Table 7, which shows there are little differences in food desert and non-food desert topical content in SE. Moreover, the proportional representation of population in our data for SE is .28%%, which is lower than other regions[14].

[14]http://www.census.gov/newsroom/releases/archives/2010_census/cb12-50.html

[15]http://www.census.gov/popclock/data_tables.php?component=growth

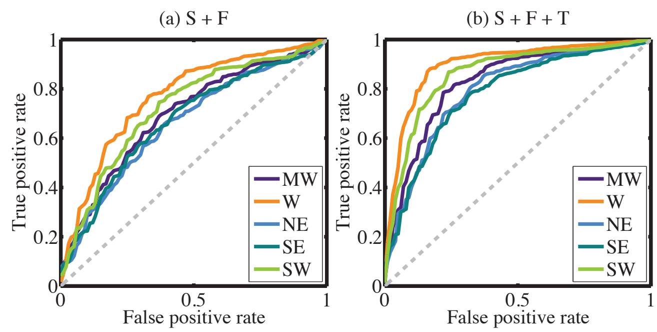

Figure 4: Receiver operating characteristic (ROC) curves for predicting food desert status in different regions. We compare across the **S + F** (left) and the final **S + F + T** (right) models.

Summarily, these findings indicate that the ingestion topics of Instagram posts, along with the information on socio-economic attributes and food deprivation status helps us better infer (with sufficiently high accuracy (80%) and F1 score (.78)) whether a tract is likely to be a food desert.

	Accuracy			F1 score		
	S+F	**S+F+T**	Δ	**S+F**	**S+F+T**	Δ
MW	69.92	79.705	14.27	0.643	0.745	15.36
W	74.36	88.316	18.99	0.717	0.866	20.23
NE	63.28	77.067	21.84	0.602	0.748	24.39
SE	63.32	68.444	8.78	0.594	0.637	7.77
SW	72.73	84.142	15.37	0.735	0.827	12.63

Table 9: Predicting food desert status of Census tracts. **S + F** is the model that uses socioeconomic and food deprivation attributes; **S + F + T** uses these attributes along with the topic distribution features of the tracts. Δ indicates normalized percentage change in accuracy or F1 score.

Error Analysis

In order to investigate deeper into how LDA ingestion topics can improve identification of food desert tracts than possible via socio-economic and food deprivation attributes alone, we discuss two case examples below.

For the first, we identified all tracts in a region predicted to be food deserts by **S + F + T**. Then corresponding to the same region we identified those tracts which were most similar to them in terms of the socio-economic and food deprivation attributes, but **S + F** identified them to be non-food deserts. We discuss one such "similar" tract pair: Census tract 113 in western Washington state's Thurston county and tract #7 in the Marathon county of central Wisconsin. Neither of these tracts are predicted to be food deserts by **S + F** and the predictions correspond to USDA's ground truth labels. However on using the **S + F +T** model, while the latter is predicted to not be a food desert, the former is. We conjecture differences in Instagram ingestion topics to be accountable for this difference. For instance, the WA tract's ingestion topics include "takeout", "cream", "spaghetti", "tvdinnerlife" which indicate low nutritional food consumption; however for the WI tract no such observations can be made ("riceandbeans", "dinner", "honeycancook"). It is known that a number of areas in south-western WA which were formerly industrial towns and have historically grown a variety of food, have undergone considerable economic decline of late, including avenues which may allow access to affordable or healthy food [25]. Since Instagram's ingestion topics utilize self-reported information on food, they likely capture these changes better than socio-economic or food deprivation attributes: the latter are known to reflect changes slowly.

We conduct a reverse exercise for the second case example: identifying tracts in a region predicted to be non-food deserts by **S + F + T**, and "similar" tracts in the same region which **S + F** indicated to be food deserts. We discuss the following pair: Census tract 52, an urban geographical area to the east of downtown Atlanta in Georgia, and tract number 24 in the Montgomery county in south-eastern Alabama. Both these tracts were observed to be highly similar in socio-economic composition and their food access status. Although the model **S + F** identifies the former to be a food desert that aligns with its true label (from USDA), **S + F + T** predicts otherwise. Like before, we ascribe the observed discrepancy in inference to differences in Instagram derived ingestion topics between the two tracts. The GA tract's topics mention healthy eating lifestyles ("smoothie", "organic", "farmtotable", "baking") perhaps attributable to its rapid gentrification, growth in real estate and influx of small and big businesses in recent years [17]. We were not able to find evidence of such changes in the case of the AL tract. The impact of gentrification on food are likely more readily observable via social media than through demographic composition or food access statistics of a population.

DISCUSSION

Health Implications

Direct Measurement of Nutrition Quality. With the rising trajectory of national health issues, such as the incidence of obesity and the growing prevalence of diabetes and other related diseases, the concept of healthy food availability has become increasingly important in public policy [42]. Our findings provide the first empirical insights into the potential of leveraging social media to understand nutritional limitations in areas challenged by healthy and affordable food access: food deserts. In this manner, our work attempts to overcome challenges of prior work [7] by *directly measuring* the quality of food in different geographical locations: Most studies [22] on food deserts typically have not directly measured the quality of food available (e.g., nutritional adequacy), rather use access to a conventional supermarket as an indicator of quality. We find that several of our findings on the dietary choices and nutritional challenges of food deserts align with qualitative evidence on the same [20]. For instance, we find that counter-intuitively, calorie intake of the food posted by people in food deserts is not significantly higher than that in other locations, however fat, cholesterol, and sugar intake in food, as indicated by Instagram content, is notably high.

Geographic Nutritional Differences and Granularity. Our results also show that *not all* regions of the US are equally nutritionally challenged in their food deserts. Furthermore, while broad nutritional limitations are consistent in food deserts throughout the country, specific nutrients might be inadequate in specific regions. For instance, cholesterol intake manifested in Instagram content is high in Mid West and West, while protein content in diet is low in the South East. These nuances gleaned from social media may provide fresh insights and complementary information to health planners and policy makers geared toward improving food safety among disadvantaged residents in different parts of the country. Further, we note that national surveys on nutrition are often not powered at the tract level, making the use of social media data at this granularity, as demonstrated by our approach, particularly valuable.

Role of Food Access. Our work proposes a novel approach to study dietary and nutritional characteristics of food deserts via social media and comparing them with geo-culturally and socio-economically similar non-food desert tracts (RQ 1, 2). Interestingly, we found that controlling for these characteristics, nutritional and dietary differences still exist across food deserts and other tracts. Consequently, our findings align with prior work on food deserts that found that the unique aspects of lowered accessibility to healthy food is often the factor behind for poor dietary choices in food deserts, rather than socio-economic deprivation alone [7, 42].

Implications for Social Computing Research

Population-scale Phenomena with Social Media. As we discussed before, measuring population-scale attributes from observational social media data has been recognized as a notable methodological challenge [46], since typically used statistical models like regression may incur omitted variable bias. In this paper, we have proposed a matching methodology to control for such biases (geo-cultural and socio-economic attributes) while measuring the effect of food desert characteristics on nutrition of populations. We believe this kind of methodology is generalizable and can be applied to study social media derived population characteristics in a variety of different contexts and settings, especially ones in which randomized experimental design to control for confounding variable effect may not be practicable.

While in this paper we specifically focused on Instagram, our methods involving detection of nutritional levels from text and identifying linguistic constructs in ingestion related content are generalizable, and may be applied to text data derived from other social media platforms like Twitter or Facebook.

Improving Identification, Surveillance of Food Deserts. Our predictive model showed that topics associated with ingestion content may help detect the food desert status of different areas (RQ 4). In fact, these topics do *indeed* contain valuable cues that help us predict food desert status with better accuracy and precision than is possible by using the socio-economic and USDA defined food deprivation features alone.

Our methods and findings may help health agencies to identify food deprivation areas in a more empirically-driven manner; also to track the nutritional status of different locations less intrusively and more periodically. Current state-of-the-art efforts are plagued by access to appropriate data, the span or scale of the data, and the time gaps in which the data are collected [30]. Since we leverage naturalistic data shared *publicly* on social media by millions of individuals and over a long period of time, we believe our approach of food deprivation status inference can complement well existing survey-driven methods of identifying challenges in food deserts.

Additionally, our prediction model may be useful in providing surveillance for areas that may be at risk of becoming food deserts. It can aid policymakers in formulating policies suited to the specific needs of populations in these disadvantaged circumstances or to monitor dietary habits after policy changes are enforced in food desert areas. Our findings can also help public health officials develop hypotheses to study further the mechanisms by which food deserts arise. Finally, as observed in our error analysis, introducing social media ingestion topics into food desert identification task may be able to capture recent changes in dietary habits and food availability, characteristics which might not always be reflected immediately in Census or USDA defined attributes.

Food and Language. Broadly, our results also showed that there are unique ingestion related linguistic signatures in Instagram posts in different regions (RQ 3). We find that social media could act as a new data sensor in food and nutrition research — that there are socio-geo-cultural dimensions to food [43, 41, 1]. We believe our findings provide more thorough understanding of the links between food and language.

Limitations and Future Work

Nutrition Inference. There are notable limitations to the nutrition inference method we employed in our data. While similar methods have been explored in prior work with success [1, 13, 36], we acknowledge that the USDA database likely does not include all possible food names, especially processed foods, gourmet recipes, or specific restaurant dishes. We also normalized consumption across all posts and users in our data at the 100g serving size level; of course, we suspect there would be individual-centric differences across the actual amount of food consumed. Inferring the proportion of food consumed based on tags or images is a challenging problem as observed in recent work [38]. We also did not remove brand-accounts or celebrities from our data who may post about food and ingestion; we did not expect such accounts to impact our analysis. However, since we employed a statistical matching procedure to compare consumption in food desert and non-food desert areas, we expect the impact of these limitations to be low.

Social Construction of Diet. It is important to note here that there is a social component to one's dietary habits and choices. Being the social platform it is, Instagram content is likely biased by people's personal and cultural perceptions and decisions regarding *what* type of ingestion activities are appropriate, desirable or interesting to be shared publicly, in ways that would enact and preserve the "image" they intend to portray on the platform. In essence, Instagram content are more appropriately described as identity statements, instead of actual behavior. More generally, self-reported information on social media platforms may be biased by social comparison or self-presentation concerns. Hence, we acknowledge that the data we study here is not a perfectly true reflection of what people are eating, but rather one distorted by cultural values, personal identity and social habits.

Geographic Context. Next, it is important to bear in mind that our data and ensuing analysis are able to measure self-reported food consumption in different geographic areas, without specific claims whether these individuals are actual residents of the area or not. In fact, individuals may move around throughout the course of a day, or may travel to another place for the short or long-term. Such mobility patterns are often difficult to measure from social media data directly without explicit self-reported information, and hence we do not claim that our methods actually capture what residents of a geographical area are ingesting.

Platform Choice and Generalizability. Finally, we acknowledge the limitations posed by the use of the social media platform Instagram. Studying population-scale phenomenon via social media has its known shortcomings [19]. In the case of Instagram, bias may exist in the demographic population who use the platform. The Pew Internet survey indicate that women, Hispanics, African-Americans, young adults and urban/suburban residents are more likely users of the platform[3]. Moreover, Instagram population bias may also be non-uniform across the country, and it is possible different types of food related content appear differentially in different locations. In fact along these lines, one potential criticism of the use of Instagram to identify nutritional characteristics in food deserts could be that the user base on the platform is non-representative for this particular question. It is known that people residing in food deserts are economically challenged [7], and statistics indicate Instagram to be prevalent among more affluent communities[3]. Our results are indeed affected by this non-representativeness issue to some extent. We observed that the food desert status predictive model performs better in regions with larger urban populations and with greater representative population in our Instagram dataset.

One can argue that these challenges may bias estimates of absolute consumption of different food items in food deserts and other locations. However we rationalize that since our methodology involves comparing consumption of food between food desert and non-food desert tracts, the biases would impact, to a lesser extent, the *relative* measurement of differences in diet and nutrition in the two categories of tracts. Further, we note that we presented a careful matching methodology that would counter most of the bias effects. After controlling for geo-cultural and socio-economic variables, we find that the nutritional attributes of food deserts as measured via Instagram posts are notably distinct from those in non-food deserts. Hence we are confident that the relative differences on which we base our major findings on, are less affected by our choice of the Instagram platform.

Visual Content of Images. In this paper, we did not leverage the content of the images themselves shared on Instagram. Although our findings show that for images that have textual tags associated with them we can extract their nutritional profiles with sufficient confidence, in future work we would like to examine how these tags may be boosted with visual features of Instagram images.

CONCLUSION

We investigated how ingestion and food related content extracted from social media, specifically Instagram, may lend valuable empirical insights into food and nutritional choices in areas challenged by healthy food access: food deserts. We proposed a matching methodology to control for socioeconomic and geo-cultural differences in food deserts and non-food deserts, so that we can accurately identify the characteristic dietary choices in each category. In a corpus of over 3 million Instagram posts, we found that ingestion content in food deserts was associated with low fruit and vegetable consumption and higher levels of fat, cholesterol and sugar. Further, distinctive linguistic markers, extracted through a topic model, corresponded to the ingestion content of food deserts and other areas. In fact, we were able to predict USDA defined food deprivation status of a tract by utilizing these ingestion topics together with baseline demographic variables. Our results bear implications in how longitudinal inferences of nutritional and food deprivation status of areas derived from social media may be useful in improved detection of food deserts and thereby helping reduce inequalities in health.

REFERENCES

1. Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You Tweet What You Eat: Studying Food Consumption Through Twitter. In *Proc. CHI*.
2. Philippe Apparicio, Marie-Soleil Cloutier, and Richard Shearmur. 2007. The case of Montreal's missing food deserts: evaluation of accessibility to food supermarkets. *Intl. journal of health geographics* 6, 1 (2007), 4.
3. Christopher M Bishop and others. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.
4. Troy C Blanchard and Todd L Matthews. 2007. Retail Concentration, Food Deserts, and Food-Disadvantaged Communities in Rural America. *Remaking the North American food system: Strategies for sustainability* (2007), 201.
5. Peter Clifford, Sylvia Richardson, and Denis Hémon. 1989. Assessing the significance of the correlation between two spatial processes. *Biometrics* (1989), 123–134.
6. Aron Culotta. 2014. Estimating county health statistics with Twitter. In *Proc. CHI*. ACM, 1335–1344.
7. Steven Cummins. 2002. Food deserts. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society* (2002).
8. Steven Cummins and Sally Macintyre. 1999. The location of food stores in urban areas: a case study in Glasgow. *British Food Journal* 101, 7 (1999), 545–553.
9. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Major life changes and behavioral markers in social media: case of childbirth. In *Proc. CSCW*. ACM, 1431–1442.
10. Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. 2014. Characterizing and Predicting Postpartum Depression from Facebook Data. In *Proc. CSCW*. ACM.
11. Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proc. ICWSM*.
12. Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using Matched Samples to Estimate the Effects of Exercise on Mental Health from Twitter. In *Proc. AAAI*.
13. Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, and Dane Bell. 2014. Analyzing the language of food on social media. In *Proc. IEEE Big Data*. 778–783.
14. Deja Hendrickson, Chery Smith, and Nicole Eikenberry. 2006. Fruit and vegetable access in four low-income food deserts communities in Minnesota. *Agriculture and Human Values* 23, 3 (2006), 371–383.
15. Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online learning for latent dirichlet allocation. In *Proc. NIPS*. 856–864.
16. Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. 2014. Social structure and depression in TrevorSpace. In *Proc. CSCW*. 615–625.
17. Dan Immergluck. 2009. Large redevelopment initiatives, housing values and gentrification: the case of the Atlanta Beltline. *Urban Studies* 46, 8 (2009), 1723–1745.
18. Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. 2012. I can't get no sleep: discussing# insomnia on twitter. In *Proc. CHI*. 1501–1510.
19. David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* (2014).
20. Sally Macintyre. 2007. Deprivation amplification revisited; or, is it always true that poorer places have poorer access to resources for healthy diets and physical activity? *International Journal of Behavioral Nutrition and Physical Activity* 4, 1 (2007), 32.
21. Yelena Mejova, Hamed Haddadi, Anastasios Noulas, and Ingmar Weber. 2015. # FoodPorn: Obesity Patterns in Culinary Interactions. In *Proc. Digital Health*.
22. Latetia V Moore and Ana V Diez Roux. 2006. Associations of neighborhood characteristics with the location and type of food stores. *American journal of public health* 96, 2 (2006), 325–331.
23. Megan A Moreno, Dimitri A Christakis, Katie G Egan, Libby N Brockman, and Tara Becker. 2011. Associations between displayed alcohol references on Facebook and problem drinking among college students.

Archives of pediatrics & adolescent medicine (2011), Archpediatrics–2011.

24. Elizabeth L Murnane and Scott Counts. 2014. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proc. CHI*. ACM, 1345–1354.
25. Kit Oldham, Peter Blecha, and HistoryLink (Firm). 2011. *The Story of the Port of Seattle: Rising Tides and Tailwinds: 1911-2011*. University of Washington Press.
26. Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-depressed Users in Twitter. In *Proc. ICWSM*.
27. Michael J Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proc. ICWSM*.
28. Tim Pearson, Jean Russell, Michael J Campbell, and Margo E Barker. 2005. Do food deserts influence fruit and vegetable consumption?A cross-sectional study. *Appetite* 45, 2 (2005), 195–197.
29. Dan Pelleg, Andrew W Moore, and others. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters.. In *ICML*. 727–734.
30. Mark Petticrew, Steven Cummins, Catherine Ferrell, Anne Findlay, Cassie Higgins, Caroline Hoy, Adrian Kearns, and Leigh Sparks. 2005. Natural experiments: an underused tool for public health? *Public health* 119, 9 (2005), 751–757.
31. Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2012. Tracking gross community happiness from tweets. In *Proc. CSCW*. ACM, 965–968.
32. Paul R Rosenbaum and Donald B Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 1 (1985), 33–38.
33. Adam Sadilek, Henry A Kautz, and Vincent Silenzio. 2012. Modeling Spread of Disease from Social Interactions. In *Proc. ICWSM*.
34. Kai A Schafft, Eric B Jensen, and C Clare Hinrichs. 2009. Food deserts and overweight schoolchildren: Evidence from Pennsylvania. *Rural Sociology* 74, 2 (2009), 153–177.
35. Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, and Lukasz et al. Dziurzynski. 2013. Characterizing Geographic Variation in Well-Being Using Tweets. In *Proc. ICWSM*.
36. Sanket Sharma and Munmun De Choudhury. 2015. Detecting and Characterizing Nutritional Information of Food and Ingestion Content in Instagram. In *Proc. WWW Companion*.
37. George Davey Smith and Eric Brunner. 1997. Socio-economic differentials in health: the role of nutrition. *Proc. nutrition society* 56, 1A (1997), 75–90.
38. Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D Abowd. 2015. Inferring Meal Eating Activities in Real World Settings from Ambient Sounds: A Feasibility Study. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 427–431.
39. Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. In *Proc. CHI*. 3187–3196.
40. Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. 2015. Measuring Urban Deprivation from User Generated Content. In *Proc. CSCW*.
41. Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2014. Spatial and temporal patterns of online food preferences. In *Proc. WWW Companion*. 553–554.
42. Renee E Walker, Christopher R Keane, and Jessica G Burke. 2010. Disparities and access to healthy food in the United States: a review of food deserts literature. *Health & place* 16, 5 (2010), 876–884.
43. Robert West, Ryen W White, and Eric Horvitz. 2013. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proc. WWW*. 1399–1410.
44. Amanda Whelan, Neil Wrigley, Daniel Warm, and Elizabeth Cannings. 2002. Life in a'food desert'. *Urban Studies* 39, 11 (2002), 2083–2100.
45. Martin White, Jane Bunting, Liz Williams, Simon Raybould, Ashley Adamson, and John Mathers. 2004. Do food deserts exist? A multi-level, geographical analysis of the relationship between retail food access, socio-economic position and dietary intake. *Food Standards Authority, London* (2004).
46. Christopher Winship and Stephen L Morgan. 1999. The estimation of causal effects from observational data. *Annual review of sociology* (1999), 659–706.