“Bacon Bacon Bacon”: Food-Related Tweets and Sentiment in Metro Detroit

V. G. Vinod Vydiswaran,1 Daniel M. Romero,1 Xinyan Zhao,1 Deahan Yu,1 Iris Gomez-Lopez,2 Jin Xiu Lu,1 Bradley Iott,1 Ana Baylin,1 Philippa Clarke,1 Veronica Berrocal,1 Robert Goodspeed,1 Tiffany Veinot1

1 University of Michigan, Ann Arbor, MI
2 Miami University, Oxford, OH; gomezlin@miamioh.edu

Abstract

Initiatives to reduce neighborhood-based health disparities require access to meaningful, timely, and local information regarding health behavior and its determinants. In this paper, we examine the validity of Twitter as a source of information for analysis of dietary patterns and attitudes. We analyze the “healthiness” quotient of food-related tweets and sentiment regarding those tweets from metropolitan Detroit. Our findings demonstrate feasibility of using Twitter to understand neighborhood characteristics regarding food attitudes and potential use in studying neighborhood-based health disparities.

Introduction

The burden of negative health outcomes is, unfortunately, differential in the United States (US). Neighborhood-based disparities in dietary patterns and health behaviors influence health outcomes such as obesity, diabetes, kidney failure, and cardiovascular disease (Mozaffarian 2016). However, because people’s diets and attitudes are difficult to measure at scale, it has not yet been feasible to easily measure these phenomena at the local neighborhood level. In this paper, we ask whether social media data can help fill this gap, investigating whether social media can be used to assess dietary patterns and attitudes. In this work, (i) we present a general procedure to extract large-scale food-related content from social media from metropolitan Detroit, which can be extended to other localities; (ii) with input from a public health nutrition researcher, we develop a keyword-level “healthiness” score for common food words and aggregate those to measure the healthiness of any social media text about food; and (iii) we expand existing keyword-based sentiment analysis tools to include food-specific sentiment keywords.

Social media has been used for numerous public health applications, e.g., (Sarker et al. 2015; Kendall et al. 2011); this prior success suggests its promise for helping to fill gaps in the US population health information infrastructure. Further, in a food context, previous work has shown that tweets concerning food can be geo-located and correlated with aspects of the geographic context. For example, (Abbar, Mejova, and Weber 2015) examined food-related tweets at the level of the US state, showing that caloric value assigned to foods mentioned in tweets was associated with state-level obesity rates. Additionally, (Nguyen et al. 2017) conducted research at the zip code level in Salt Lake City, Utah, which identified higher obesity prevalence among people in a clinical sample of people who lived in areas in which tweets tended to refer to higher-calorie foods. However, calories are problematic as a measure of food-healthiness, in that they do not speak to the nutrient density of foods (Drewnowski 2005); moreover, since portion size is difficult to infer from tweets, caloric estimates are likely to be inaccurate.

Previous social media research concerning food within neighborhoods has focused primarily on attempts to infer health behavior concerning food. Furthermore, an increasing number of food-related interventions target environments which make it easier or more difficult to eat healthy or unhealthy food; for example, by focusing on improving the food supply in a neighborhood through initiatives such as farmer’s markets and community gardens (McCormack et al. 2010). Social media may provide valuable information concerning the location of behaviors and contexts surrounding the consumption of healthy and unhealthy food in a neighborhood. Therefore, we extend prior work by investigating tweet contents that reveal attitudes concerning food.

Finding food-related tweets

Twitter Data Sources

Twitter data were collected using three approaches. First, Twitter APIs were used to gather geo-tagged tweets from Metropolitan Detroit in southeast Michigan, USA. This includes the counties of Genesee, Lapeer, Lenawee, Livingston, Macomb, Monroe, Oakland, St. Clair, Washtenaw, and Wayne. Polygonal location-based queries and common English words (the hundred most frequent words such as the, is, etc.) were used to find tweets from the geographic area.

Second, the location query-based collection was enhanced using the Twitter Gardenhose stream. The Gardenhose stream comprises a 10% random sample of the entire Twitter collection. All geo-tagged tweets from the period 2013–15 within the ten-county area were identified and included.

Third, the collection was expanded using user timelines. Tweet authors were identified from those tweets gathered in the previous two methods, and their account timelines were crawled to collect all of their previous tweets, includ-
Table 1: Food keyword categories and examples

<table>
<thead>
<tr>
<th>Food word category</th>
<th>n</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definitely unhealthy</td>
<td>1,840</td>
<td>coke, bacon, candy</td>
</tr>
<tr>
<td>Unhealthy</td>
<td>1,065</td>
<td>fries, burrito, butter</td>
</tr>
<tr>
<td>Healthy</td>
<td>216</td>
<td>falafel, chicken breast, plain yogurt</td>
</tr>
<tr>
<td>Definitely healthy</td>
<td>807</td>
<td>apples, fish, carrot, peas</td>
</tr>
<tr>
<td>Total</td>
<td>3,928</td>
<td></td>
</tr>
</tbody>
</table>

Healthy and Unhealthy Food Vocabulary

To identify tweets that concern eating behavior and related attitudes, tweet content was mined for food-related terms. A vocabulary of food terms was compiled from multiple online sources. First, the comprehensive list of foods was collected from the United States Department of Agriculture (USDA) website, which included, for example, lists of fruits, vegetables, meats, and branded foods (Agricultural Research Service; United States Department of Agriculture; Nutrient Data Laboratory 2016). Second, we scraped Wikipedia pages that included lists of: branded products, raw and processed foods, international cuisines, cooking techniques, and restaurant chains in the United States. Finally, we included restaurant names from the list of the most popular fast food restaurants in the US from Business Insider (Business Insider 2015) and an industry publication (Nation’s Restaurant News 2012); and the list of fast food restaurants in the United States. Additionally, sentiment keywords were abstracted from Yelp reviews of 69 grocery stores in Metropolitan Detroit. This initial set contributed 81 positive and 101 negative new keywords.

Classifying relevant food-related tweets

The food terms vocabulary permitted identification of tweets that mention food words. However, such words are also used in non-food contexts, e.g. “I guess apple can put its apps where it wants”, to show endearment (e.g. honey, cupcake), or as metaphors (e.g. “If you were a fruit you’d be a clementine”, “beef with someone”). Hence, relevant food-related tweets had to be distinguished from non-relevant tweets containing food words. We designed a machine learning-based framework to identify relevant food-related tweets.

Creating labeled data set of tweets

To develop a supervised classification model to identify relevant food-related tweets, a labeled data set was constructed. First, 1,000 tweets were randomly selected from the set of filtered tweets with food keywords, and were labeled independently by four of the authors. These annotations helped create instructions to label food-related tweets. The annotation guidelines were used to train human annotators on the Amazon Mechanical Turk platform. Human intelligence tasks (HITs) were created to label 5,000 new tweets. Every tweet was labeled by at least four annotators, and was assigned the label agreed upon by at least three annotators. A group of five authors annotated 2,500 additional tweets. Tweets containing ambiguous food keywords (such as beef without a qualifier like “ground beef”) were subsequently removed, resulting in a total of 6,893 annotated tweets. 4,432 tweets (64.3% of the annotated set) were labeled as food-related tweets.

Training a food-related tweet classifier

Tweets were pre-processed to remove urls and callouts; hashtags and words were tokenized by remove punctuation marks. Unigram and bigram features were generated, and feature values were transformed using tf-idf normalization. Additional features were generated based on number of food words in the..
tweet, their categories, and counts of positive and negative sentiment words from the expanded sentiment vocabulary.

The labeled data set was split into train and test sets using an 80-20 split. Four classification models were trained, viz. (a) Support Vector Machine (SVM), (b) Random Forest, (c) multinomial naïve Bayes, and (d) logistic regression with stochastic gradient descent method for error minimization. In addition to these, the classification confidence estimate output by the SVM model was combined in a rule-based approach to build a Hybrid-SVM model. In this model, tweets were classified as relevant if (i) they had at least two food keywords or (ii) they had one food keyword and the SVM classification confidence was above a set threshold. The five-fold cross-validation approach, repeated ten times, was used to tune the best threshold ($\theta = 0.4852$). A similar approach was previously used by other researchers to train classifiers for short informal text messages (Thelwall et al. 2010).

Performance was measured using accuracy, precision, recall, and F1 averaged over the five cross-validation runs to decide the best classification model. The performance measures over the test data set are in Table 2. The Random Forest and SVM models performed equally well on F1, with the SVM model slightly better in accuracy. The Hybrid-SVM model performed the best on both accuracy and F1.

Filtering out tweets from non-layperson accounts: All tweets from organizations and well-known personalities were removed. Twitter user accounts in the data set were processed through Humanizr (McCorriston, Jurgens, and Ruths 2015), a pre-trained off-the-shelf tool for classifying Twitter accounts as individuals or organizations. Of the 78,794 accounts for which the profile information was available, Humanizr labeled 637 as Organizations. The labels were validated by one of the authors, and 194 accounts (30.5%) were re-classified as personal accounts. Additionally, tweets from verified accounts and online personal services, such as horoscopes and food service apps such as dinehere.us were removed. The Hybrid-SVM classifier was applied on all remaining tweets containing at least one food keyword.

“Healthiness” and Sentiment score for food tweets

Tweets identified as food-related were assigned a healthiness score. The healthiness score is computed in three parts: the healthy word score, the unhealthy word score, and the net healthiness score. The (un)healthy word score is computed as the number of (un)healthy food words in the tweet scaled by the level of (un)healthiness. The net healthiness score is computed as the difference of healthiness and unhealthiness word scores. For example, one of the tweets in our database,

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.737</td>
<td>0.806</td>
<td>0.753</td>
<td>0.770</td>
</tr>
<tr>
<td>Multinomial NB</td>
<td>0.676</td>
<td>0.652</td>
<td>0.993</td>
<td>0.787</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.738</td>
<td>0.737</td>
<td>0.878</td>
<td>0.801</td>
</tr>
<tr>
<td>SVM</td>
<td>0.758</td>
<td>0.790</td>
<td>0.813</td>
<td>0.801</td>
</tr>
<tr>
<td>Hybrid-SVM</td>
<td>0.805</td>
<td>0.806</td>
<td>0.918</td>
<td>0.858</td>
</tr>
</tbody>
</table>

Table 2: Best classification algorithms on test set.

Table 3: Most frequent keywords in food-related tweets.

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Top fifteen most frequent words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>want, need, great, like, good, thanks, love, best, special, lol, free, party, better, green, win</td>
</tr>
<tr>
<td>Negative</td>
<td>no, fuck, bad, stop, shit, brown, old, damn, shake, hate, bitch, sour, cut, mad, seriously</td>
</tr>
</tbody>
</table>

Table 4: Most frequent sentiment words in food tweets.

“Bacon bacon bacon”, mentions bacon thrice. Bacon is considered a “definitely unhealthy” food with a rating of $-2$. Since there are no healthy words in the tweet, the healthiness score for the tweet is 0, the unhealthiness score is $3 \times 2 = 6$, and the net healthiness score is $0 - 6 = -6$.

Sentiment analysis: We also measure the emotion contained in the tweets by counting the number of food-related sentiment words. All sentiment words in the tweet are identified using the sentiment vocabulary. For each tweet, positive and negative sentiment scores are computed by normalizing the number of positive (negative) sentiment words with the total number of words in the tweet.

Results

In all, 822,604 tweets authored by layperson accounts in Metropolitan Detroit were classified as food-related. There were 1,017,315 mentions of food words – 694,502 (68.3%) were unhealthy food words, while 322,813 (31.7%) were healthy food words. Table 3 lists the top ten most commonly used food words in each of the four healthiness categories.

In all, 550,560 sentiment and emotive words were found in the food-related tweets; 435,954 (79.2%) of which were positive sentiment words and 114,606 (20.8%) were negative sentiment words. Table 4 lists the top fifteen most commonly used positive sentiment words (that jointly account for 44.99% of all sentiment words) and negative sentiment words (that account for 9.89% of all sentiment words) in food-related tweets.

We analyzed the distribution of tweets with positive and negative sentiment words against healthy and unhealthy tweets. A positive (correspondingly, negative) sentiment tweet is defined as one that has at least one positive (negative) sentiment word. Similarly, a healthy (unhealthy) tweet is defined as one that has the net healthiness score of $> 0$ (correspondingly, $< 0$). Table 5 shows the $2 \times 2$ contingency matrix. $\chi^2$ statistic was 1,133.11 ($p = 2.1e-248$). The results show that the expression of positive sentiment is more com-
Table 5: Contingency matrix of distribution of tweets with positive and negative sentiment words against healthy and unhealthy tweets.

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Healthiness score &gt; 0</th>
<th>Healthiness score &lt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Healthy tweets</td>
<td>113,214 (79.3%)</td>
<td>29,533 (20.7%)</td>
</tr>
<tr>
<td>Unhealthy tweets</td>
<td>352,756</td>
<td>113,214 (79.3%)</td>
</tr>
</tbody>
</table>

Table 6: Positive and negative sentiment words in healthy and unhealthy tweets.

We conducted additional analysis of how sentiment was related to the healthiness and unhealthiness scores of a food-related tweet. We performed the non-parametric Kruskal-Wallis H test to determine if the medians of the two groups of tweets (those with positive sentiment words and those with negative sentiment words) were different. Table 6 summarizes the mean (and standard deviation), median (and inter-quartile range) and H-statistic for the two groups of tweets. For the healthiness score, the difference in the two groups was not found to be statistically significant (H-statistic of 3.087, p = 0.079). However, the difference in groups in unhealthiness score was found to be statistically significant (H-statistic of 2,726.743, p = 0.0).

Discussion and Conclusion

Our results showed that large proportions of tweets mentioned food-related behavior and locations where food and drink were consumed (e.g., Starbucks, Chipotle, etc.), from which certain aspects of diet can be inferred. With regard to attitudes, results showed stronger positive sentiment in relation to food, irrespective of healthiness score of the food.

Results of this paper suggest that social media data can provide a reliable signal for dietary patterns and food attitudes. Despite the noisy nature of the user generated text data, the limited fraction of geo-located tweets, and the likely bias from differences in users and non-users of social media, the results presented here are encouraging. These findings highlight the possibility of using social media as a signal of potential healthy and unhealthy food consumption, and related attitudes.

Our methodical approach to extract food related social media content at the census tract level, and to identify the healthiness of tweets and associated sentiment, can be applied to other social media sites besides Twitter, and to other geographical locations outside Metro Detroit. We believe that our results are an initial step towards a larger goal of utilizing social media as a tool for tracking a variety of health behaviors at large scale, but at localized geographical levels that allows us to understand differences between communities that, while at close proximity to each other, may exhibit significant variation along environmental, socioeconomic, demographic, and cultural dimensions.

References


Agricultural Research Service; United States Department of Agriculture; Nutrient Data Laboratory. 2016. Usda food composition and branded food products databases.


Tamean, K. 2009. 101+ descriptive words for foods. writing and illustrating.