Obe-City: Predicting Obesity Using Yelp Reviews

CSE 6242: Final Report

Akanksha Bindal, Akhilesh Srikanth, Ishita Chordia, Kausar Mukadam, Pradyumna Tambwekar

I. Introduction:

The percentage of American food budget spent on eating out has grown steadily since the 1960s, resulting in a corresponding rise in adult obesity [1]. Today, almost one-third of the US population is considered obese, with an estimated medical cost of $147 billion [2], causing over 300,000 premature deaths per year in the United States alone [3]. Individual level variables such as age, gender, socio-economic level, smoking status, and race/ethnicity have been shown to be significant predictors of obesity but more work needs to be done looking at how environmental factors affect choice and obesity levels [2]. Since its inception, Yelp has been used by millions of restaurant reviewers every month. Individual eating habits are influenced by these reviews, making it an excellent method for analyzing eating habits of a population. Research on health and obesity facilitates policy making and campaigns, enabling governments to make informed decisions in tackling the rising problem of obesity. Healthiness is also an important factor that affects an individual's choice of restaurant. The restaurant health score generated through our project would enable these potential consumers to make informed choices, and also provide a metric for calculating obesity in areas where surveys have not been conducted, but crowdsourced reviews for restaurants are available.

II. Problem Definition and Survey:

Huang et al. [4], have discovered that there are sufficient indicators of health-related factors in user reviews on Yelp. This project is aimed at extracting health-related cues from reviews, and creating a health rating for each restaurant. The final goal is to correlate obesity levels in a county with the "healthiness" of restaurants based on Yelp reviews, by controlling for factors such as demographics, and generating a regression model that can be used to predict obesity in the future.

Ariyasriwatana et al. [5] found thirteen categories of health-related cues to action from Yelp reviews, including nutritional facts, portion sizes, ingredients and cooking methods. These cues have been primarily extracted by variations of Latent Dirichlet Allocation (LDA). Multi-grain LDA [6], bigram LDA [4] and hierarchical LDA [7] have been successful in extracting distinguishing sub-topics and interests from reviews, including different cuisines or review subtopics, such as décor or service. Additionally, bag of words [8], part-of-speech tagging [9] and paragraph vector models [10] have also been used for learning word and sentence representations.

Our work innovatively builds on previous work:

1. Previous research has not looked at multi-county online restaurant reviews to obtain a measure of healthiness.

2. We're using bigram-LDA to determine the healthiness score for each county. Other methods to analyze Yelp reviews have used multi-grain LDA or unigram-LDA.

3. Regression has not been used previously to determine the relationship between a reviews based healthiness-score and obesity.

III. Proposed Method

Intuition:

Eating habits of a population largely determine weight and health of individuals, hence impacting obesity rates of the area. Since it is an established fact that eating out is in fact unhealthier, determining the type of restaurants and food being served at these restaurants could be a good indicator of obesity. Since there is expected to be a lag between the establishment of restaurants in an area, and their effect on health, we have assumed a 3-year lag period in our analysis. While reviewing a restaurant, people comment about the quality and type of food served at the restaurant. By extracting these components of a review, we attempt to determine the health of a restaurant.

Description:

Data Collection:

The primary sources of data were: the CDC's 2013 Behavioral Risk Factor Surveillance System (BRFSS) obesity dataset [11], the Yelp 2016 Challenge Dataset [12], consisting of 2.7 million reviews from 2007 to 2016, US Census Data from 2013 [13], the Federal Communications Commission (FCC) geography dataset [14] and the MyNetDairy [15], which provides nutritional information.

CDC's BRFSS dataset is a monthly, nationwide telephone survey that collects information on health factors. CDC data is at the county level, so the FCC API is used to match latitude-longitude from Yelp data to county codes. To test the relationship between restaurant "healthiness" and obesity, we controlled for other factors such as age, gender, and race obtained through the Census' American Community Survey (ACS) (2013) [16]. MyNetDiary assigns a food score to over 453,000 foods and has been used to weigh the frequency counts of health related words in the reviews. Food Score is calculated using an equation, which uses twelve nutrients: total fat, saturated fat, cholesterol, sodium, total carb, fiber, sugar, protein, Vitamins A and C, calcium, and iron, derived from food ratings of nutrition experts using information found on the Nutrition Facts panel [17].

Several issues were faced during data cleaning and merging process: large data, poor documentation and data granularity. Several iterations were made to speed up the Python script, and intermediate outputs were saved from the API so that small errors wouldn't result in large recalculations. Variables extracted using the ACS API were validated against external data sources since documentation did a poor job in defining them. Finally, Yelp only released reviews for six American cities. Since obesity data at a zip code level was unavailable, and our Yelp dataset only contained 11 counties, the extent of our research was limited

Once merged, the dataset had 277,612 reviews. Rows with missing values and will less than 10 reviewed restaurants were removed. Statistics about the final dataset containing 261,898 reviews from 11 different counties are given in Table 1.

Table 1: Summary Statistics

	Number Restaurants	Number Reviews	Avg Adult Obesity(%)	Median Age
Arizona	**12470**	**108422**	**25.44**	**35.01**
Maricopa	12277	107762	25.40	35.00
Pinal	193	622	31.60	36.10
Illinois	**455**	**3846**	**25.10**	**29.10**
Champaign	455	3846	25.10	29.10
Nevada	**7586**	**115940**	**27.80**	**35.80**
Clark	7586	115940	27.80	35.80
North Carolina	**2492**	**13430**	**24.09**	**34.30**
Cabarrus	78	244	30.00	37.00
Gaston	33	86	33.30	39.00
Mecklenburg	2354	13022	23.90	34.20
Union	27	78	26.50	36.60
Pennsylvania	**1557**	**10770**	**26.70**	**41.10**
Allegheny	1557	10770	26.70	41.10
South Carolina	**67**	**200**	**29.50**	**37.57**
York	67	189	29.40	37.40
Wisconsin	**1336**	**9290**	**23.70**	**34.50**
Dane	1336	9289	23.70	34.50
Grand Total	**25963**	**261898**	**26.40**	**35.47**

Our Approach:

The Yelp data contained information about different types of local business, including restaurants, salons, shopping, etc. Since our focus is determining health for restaurants, all business reviews, except for restaurants, were filtered. Further, a restaurant review may consist of various topics or themes. For example, consider the review below:

*"Even though the line was very long for lunch at ****, I decided to give it a try today.*
*The food was amazing. I had the falafel sandwich, *** style. The falafel were perfectly cooked and the accompanying baba ghanoush, hummus, 4 Mediterranean salads, and tzatziki sauce rounded out the sandwich. I'll definitely be back this week!*
*Another plus: while waiting in line, *** greeted each guest warmly with a small cup of delicious soup. Great service! "*

For a health score, topics such as service, décor or crowd at a restaurant have no bearing on the result. Thus, our algorithm looks for features related to food, specifically words that convey some information about health. For separating different topics of a review, Latent Dirichlet Allocation (LDA) model was employed.

LDA is a generative probabilistic model that represents documents as random mixtures over latent 'topics' or themes, where each topic is characterized by a member set of words. Initially unigram LDA was used. But the context of words is not included when considering unigrams. For example, the word oily impacts a health score negatively, but if used in the context 'not oily', it reverses its impact, i.e. impacts positively. Thus, context of words is an important factor in determining health score. Therefore, bigram LDA was later used to determine features. By intuition, relevant topics i.e. topics that convey health-related information were selected from the words generated by LDA.

The feature set for regression was initially developed by scoring a restaurant based on frequency count of the above selected words. Negative health words like 'oily' were scored with a -1 for every occurrence, whereas positive health words such as 'fresh' were scored with a +1. But weighing all words equally yielded poor results.

A carefully chosen weighted combination of word frequencies yields a more optimal health score. Thus, a score for each word was determined, and the occurrences were weighed by these scores. For words that represented dishes or ingredients, such as French fries or fried chicken, the nutritional score developed by MyNetDiary, which was converted to a scale from -5 to 5, where 5 represents a healthy dish whereas -5 represents unhealthiness, was used as a scoring mechanism. For other words that are not included in this database, i.e. word like greasy, which are not ingredients or dishes, we used a scoring mechanism based on intuition (again on a scale of -5 to 5). To reduce bias, each item was scored by two team members, and the final score was an average value of these two scores.

Finally, the health scores generated for each review were averaged at the restaurant level to generate a health score. By aggregating these health scores at a county level, and combining it with the demographics of the area, a regression model was used to learn the obesity rate. We also determined the correlation value between the health metric generated by our algorithm and the obesity value for the county (as provided by the CDC) to calculate an obesity value. Since Yelp reviews were obtained at the city level and obesity data at the county level, we anticipated a bias in our predicted obesity values.

IV. Experiments/Evaluation

Research Questions:

Our project aims to answer the following questions:

Research Question 1: Is it possible to extract relevant health-related keywords from user-generated reviews in order to create an accurate health rating for each restaurant?

Research Question 2: What are the best methods to aggregate health-related keywords into health ratings?

Research Question 3: How predictive are these health ratings of obesity in a county?

Topic Modelling: (Latent Dirichlet Allocation)

Using Mallet (8), 50 'food-related" topics and top-10 words (unigram and bigram) in each topic were discovered. This generated a list of 1000 words, which were filtered manually. Using intuition, we selected 'health related' words from these topics, which were then used as our feature set.

Bag of Words:

The bag of words model outputs two results,

1. A vocabulary consisting of most common words/phrases that occur in the review set.

2. A count vector for each review displaying the count for the most common words for a review.

To determine whether yelp reviews had significant information about health, we ran a bag of words model for a set of words that we determined indicated healthiness. Testing for these words on reviews from 2013 (consisting of 344,443 reviews), we obtained the following word counts (Table 2.).

Table 3. Frequency counts for health-related words

Word	Frequency Count
kale	1056
starch	245
coke	1716
bacon	11300
fat	3709
quinoa	515
oil	6630
pepsi	334
shake	3681
carb	906
crispy	5
broccoli	1792

Word	Frequency Count
gluten	2937
cream	16531
salad	35642
spinach	3532
sugar	3633,
soup	18126,
portion	17750
health	2272
soda	3896
cake	12983
fresh	34326
salt	6359

The results show that many informative words like “fat”, “soda”, and “oil” frequently occur, which could indicate obesity, whereas words like “salad” and “fresh” could map to the other side of the obesity spectrum. One limitation here is that unigram bag-of-words considers individual words without looking at their neighboring words which could lead to the misinterpretation of context. Therefore, for our final algorithm, both unigram and bigram words were used to fix the lack of context problem.

A word cloud depicting some words that were used by our algorithm is given in Figure 1.

Figure 1. Word cloud depicting most commonly occurring health related words

Generation of Health Score:

The next step in the project was to calculate scores for each restaurant, based on the words obtained from the unigram and bigram LDA model. Before calculating these scores, each of these words needed to be assigned a weight in accordance with their healthiness. We took the help of a website called mynetdiary.com, which assigns a weight to each food related word based on its nutritional value. There were a few relevant non-food words which we came across in our model, like “freshly made” or “clogged arteries”, for which the website could not output a rating. For such words, we approximated a rating based on intuition.

These weights were then used to calculate the scores for each restaurant, by iterating through all the reviews for each restaurant and checking for the occurrence of these words in reviews. If the word is present, then the weight of that word was added to the score of the restaurant. Once all the scores were calculated, these scores were normalized so that the value is between zero and five.

Regression and Correlation:

Since the context is lost while considering unigram results, and the words for this section were not weighted (just a +1 and -1 was used to indicate positive and negative effect on obesity), the results for regression on unigram results were not substantial. This prompted us to explore other options, and we eventually used bigram LDA along with a scoring mechanism to generate a more accurate score.

To evaluate the predictive power of our health score, we ran a multivariate regression [18]. The dependent variable was the percentage of each county that is classified as obese (BMI greater than or equal to 31). The covariates included our health score, median age of the county, median income, the number of people in the population that have health insurance, and race.

We see that age and income are significant at 5% level, which is consistent with the literature on obesity [1]. Although it’s not significant, the average health score that we’ve created is negatively correlated with obesity. This coincides with our intuition that health scores of restaurants are correlated with obesity. Since in our scoring metric, 0 represents an unhealthy restaurant and 5 represents a healthy restaurant, we

can see from our correlation that as a health score increases (becomes healthier), the obesity decreases. Further, our health scores intuitively seem to make sense—fast food restaurants like McDonald's and Chick-Fil-A seem to have much unhealthier scores that sit-down restaurants. Since our dataset only contains 11 counties, the data for regression is limited and hence may not be an accurate indicator. Future research should focus on obtaining data for more counties or at a different level of granularity. Second, because we manually weighted the keywords using scores available on FoodNet.com, there may have been irregularities introduced into the data. Often FoodNet.com gave different scores for a given word and team members had to average the score and make a judgement call. The lack of a precise algorithm to determine the score for each keyword may have impacted the validity of the scores.

Table 3. Results from Regression Model

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	87.26655112	29.75762354	2.93257797	0.06087375
Average Health Score	-21.75916587	9.556893502	-2.2768032	0.10725049
Health Insurance	-7.28967E-07	5.26917E-07	-1.3834555	0.26049832
Median Age	0.662829078	0.182843499	3.62511701	0.03611716
Median Income	-0.000678672	0.000152154	-4.4604158	0.0209824
Percentage of Population: Black	1.918557317	4.962685228	0.38659662	0.72484545
Percentage of Population: Asian	-20.81827058	20.67405341	-1.0069758	0.38812782
Percentage of Population: Hispanic	14.77774548	5.698954351	2.59306262	0.08085943

V. Visualization:

Figure 2. Health scores for six cities from our dataset

The scores generated from our algorithm were visualized using the Google Maps API. By mapping restaurants according to their location, along with information about their star rating and health score

(generated above), we give the end user a medium to choose from restaurants across the city taking into account location, star rating and healthiness. The unhealthy restaurants are marked by red moderately unhealthy as orange, moderately healthy as yellow and healthy as green location icons.

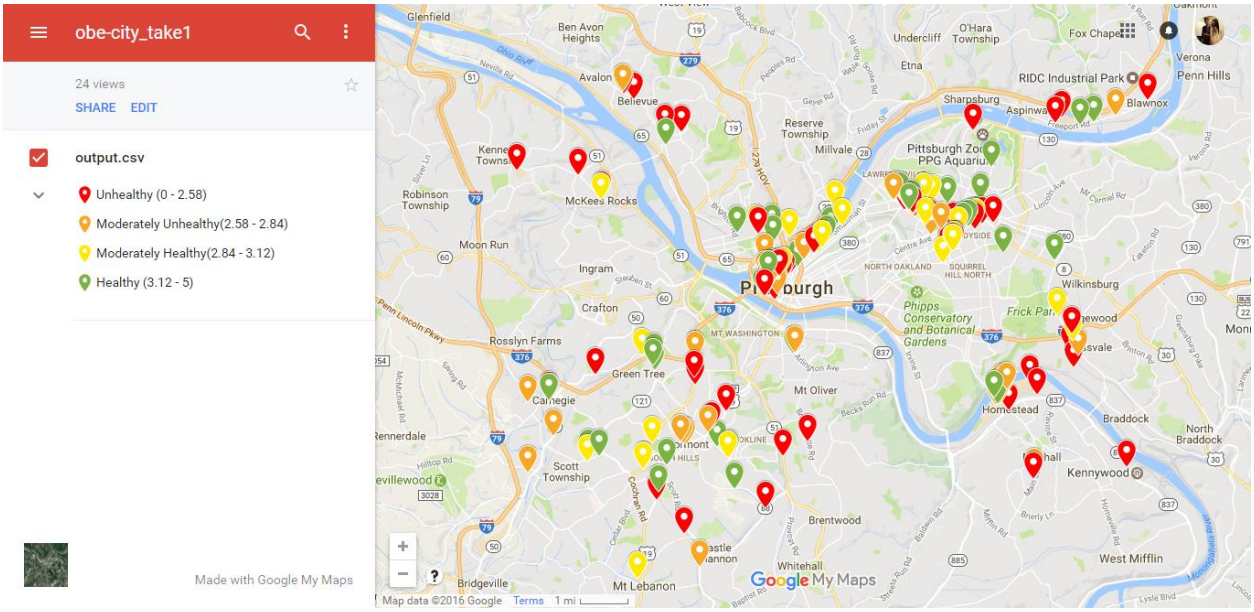

Figure 3. Health scores for a particular city (Pittsburgh)

VI. Conclusion/ Discussion

We can intuitively see that the health scores generated through our algorithm were correct, for example fast food restaurants were correctly categorized as unhealthy, and we also discovered a negative correlation between these scores and obesity, which was as expected. One drawback in our analysis is the lack of data points in our correlation and regression analysis (only 11 counties), which may have unfairly biased the results. Further research could include data about more number of counties, to accurately determine the correlation, and use other scoring metrics for words. Since we used scores generated by MyNetDiary alone, we may have introduced a bias. By aggregating health scores from two or more sites, a more accurate representation of weights can be generated, which may provide even better results.

VII. Distribution

All team members contributed similar amounts of effort towards the project.

VIII. References

[1] "Weight status and restaurant availability a multilevel analysis," 2008.

[2] "Centers for Disease Control and Prevention," [Online]. Available: https://www.cdc.gov/obesity/data/adult.html.

[3] "An economic analysis of adult obesity: results from the Behavioral Risk Factor Surveillance System," 2002.

[4] "Improving Restaurants by Extracting Subtopics from Yelp Reviews," *Yelp Dataset Challenge Winner*, 2013.

[5] "Categorizing Health-Related Cues to Action: Using Yelp Reviews of Restaurants in Hawaii," *New Review of Hypermedia and Multimedia*, 2014.

[6] "Modeling Online Reviews with Multi-Grain Topic Models," in *International World Wide Web Conference Committee*, 2008.

[7] "Summarizing Amazon Reviews using Hierarchical Clustering," 2009.

[8] "Exploring the Yelp Data Set: Extracting Useful Features with Text Mining and Exploring Regression Techniques for Count Data".

[9] "Mining Opinion Features in Customer Reviews," *American Association for Artificial Intelligence*, 2004.

[10] "Learning Sentence Vector Representations to Summarize Yelp Reviews," *Stanford*, 2015.

[11] [Online], "Centers for Disease Control and Prevention," [Online]. Available: http://www.cdc.gov/brfss/.

[12] [Online], "Yelp Dataset Challenge," [Online]. Available: https://www.yelp.com/dataset_challenge/dataset.

[13] [Online], "US Census Data," [Online]. Available: https://www.census.gov/programs-surveys/acs/guidance/comparing-acs-data/2013.html.

[14] [Online], "FCC Geography Dataset," [Online]. Available: https://www.fcc.gov/general/download-fcc-datasets.

[15] "My Net Diary," [Online]. Available: http://www.mynetdiary.com/foodSearch.do.

[16] [Online]. [Online]. Available: https://www.census.gov/programs-surveys/acs/.

[17] "Modeling Expert Opinionson Food Healthiness: A Nutrition Metric," *Journal of the American Dietetic Association*, 2009.

[18] "Introduction to statistical modelling: linear regression".