Obe-City

Predicting Obesity Using Yelp Reviews

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

INTRODUCTION

Obesity is a rising health problem, causing over 300,000 premature deaths in the US alone. The effect of environmental variables on obesity has not been researched to a large extent. We explore the correlation of one such environmental variable, the restaurants in an area, with obesity. Using data from the Yelp Dataset Challenge, we assign a 'health score' to each restaurant, and aggregate these scores at the county level to correlate with obesity.

DATA

The primary downloaded sources of data used were: the CDC's 2013 Behavioral Risk Factor Surveillance System obesity dataset, the Yelp 2016 Challenge Dataset, consisting of 2.7 million reviews from 2007 to 2016, US Census Data from 2013, the Federal Communications Commission geography dataset and MyNetDairy, which provides nutritional information about ingredients and foods. Since we expect a lag between restaurants and their effect on obesity, a 3 year lag was introduced in our dataset. We used reviews from 2006-2010 to determine health scores, and validated it using obesity data from 2013. After data cleaning (removing reviews about other businesses and restaurants with less than 10 reviews), the final dataset consisted of 261,898 reviews.

PROPOSED METHOD

A restaurant review consist of various topics. For example, a reviewer may talk about the decor, service, ambiance, and food quality at a restaurant. Therefore, we used topic modelling to extract relevant topics, and using intuition selected food and health related words. The frequency of these features (or words) in a review, weighted according to a nutritional score, determines the health score for a restaurant. Initially, a non-weighted combination was used, but since that yielded poor results, for our final implementation, we weighted words by the score obtained through MyNetDiary, which assigns a food score (normalized from -5 to 5), based on the nutritional content. These health scores were then aggregated at the county level. By controlling for race, age, gender and other demographics, we obtained a correlation between the health score generated and obesity.

RESULTS

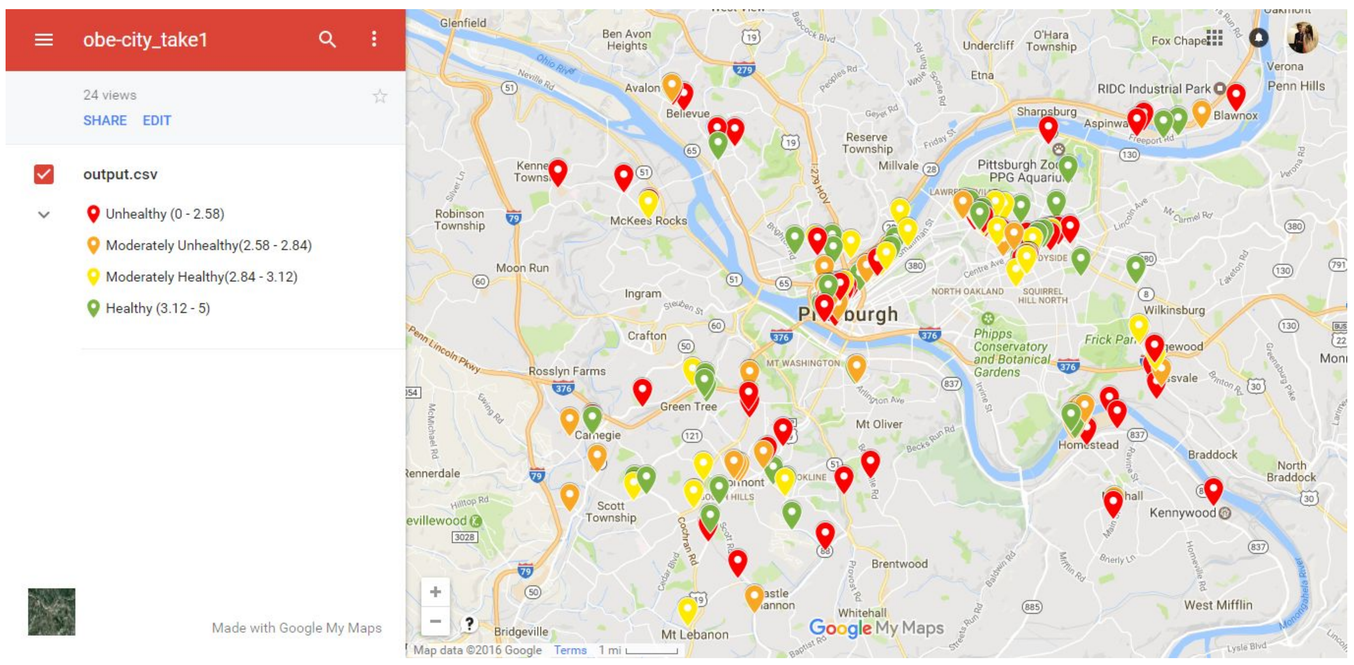
Figure 1: Word cloud depicting features extracted using LDA

The topics and words obtained using Latent Dirichlet Allocation show the presence of health related information in reviews, which can be seen from the word cloud.

Figure 2: Correlation between generated health scores and obesity

Our correlation shows that the scores we generated were negatively correlated with obesity, which is in accordance with the initial assumption we made. 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.

	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



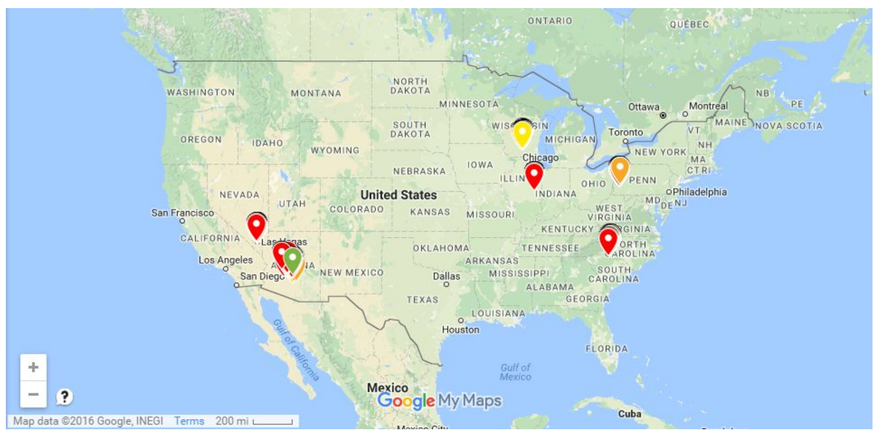


Figure 4: Visualization depicting scores for a particular city

Figure 3: Visualization depicting scores for a particular city

We visualized the scores generated through our algorithm on the map of the city, using location data to provide an interactive interface to the end user.

CONCLUSION AND 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.