Improving the Health and Safety of Dining Establishments in Butler County

By: Magda Odrowaz, Drew Nanda, Riley Nguyen & Matt Smuda

I. Abstract

Millions of Americans are affected by foodborne gastrointestinal illness each year, which poses a direct threat to public health, and disproportionately the immunocompromised and elderly populations. Foodborne illness occurs when pathogens are not properly deactivated before, during, or after food preparation causing illness among those who consume the contaminated food. Safe food handling practices are the best way to prevent disease and ensure health and safety at public dining establishments. The Butler County Health department regularly inspects restaurants and bars to ensure they adhere to safe food handling practices, as well as followup on inspections that were failed or when complaints are received. This project provides an exploratory analysis of public health inspection records as well as a supervised learning model that predicts restaurant risk levels based on non critical health inspection violation frequency. Such an analysis results in a word frequency word clouds, topic modeling, classification matrix evaluation, and feature importance delineation for what attributes most affect health and safety of dining establishments.

II. Introduction

Food-borne illnesses are a serious public health problem that affect millions of Americans each year. According to the CDC, “known foodborne disease agents are estimated to cause approximately 9.4 million illnesses each year in the United States” (CDC, 2018). The severity of these cases can range from mild to life-threatening, however those that are immunocompromised can be at a much higher risk of hospitalization and even death. Safe food handling practices are the best way to prevent gastrointestinal illness from foodborne pathogens. It is required by law for local health departments to regularly inspect dining establishments, although these regular inspections cannot take on the full responsibility of ensuring public health and safety at all times, everywhere. Health inspections are conducted on a standardized, semi-annual basis, as well as in response to complaints or concerns filed by members of the community. The goal of this analytics project is to develop and deploy a model that could inform the Butler County Health Department as to which eateries are at higher risk of unsafe food handling practices, and which establishments require more frequent or immediate health inspections. Such a model can promote the efficiency and productivity of the Butler County Health Department, but most importantly it can promote the health and wellness of the community.

This analysis will focus on dining establishments and bars across Butler county in Ohio. Within these geographical constraints, data about health inspection reports as well as reviews for these respective establishments will be collected from Yelp. In order to form recommendations for the Butler County Health Department, topic modeling, keyword and sentiment analysis, as well as a supervised classification model will be deployed. These analytical tools are expected to reveal associations between Yelp review sentiments and health code violations at respective

establishments, as well as predict which specific eateries or bars may be at a higher risk of unsafe food handling practices thus needing more inspections or intervention from the health department. A Random Forest Classification model was deployed to predict restaurant risk status defined as “good” or “bad”. This classification defines a “good” establishment as one that has less than six non-critical health violations in the given time span of available health inspection data (about three years), and a “bad” establishment as one that has more than six non critical violations in the same given period of time. Topic modeling and the development of word clouds with cleaned string type text data suggest that when it comes to leaving reviews on Yelp, food, service, and the taste of food are some of the most frequent topics expressed. Additionally, neutral sentiment scores for restaurant written reviews appear to be most associated with a restaurant being classified as “good” or “bad”.

The solution outlined in this report provides an exploratory and predictive analysis of how the public perception can be associated with dining establishment safety and health criteria. While more research is needed to determine how exactly the sentiments of written reviews of eateries play a role in their cleanliness and sanitation, this project provides a framework for guiding the discovery of human sentiments and their link to health and wellness.

III. Data Collection, Preparation and Analysis

The data collected for analysis in this project are derived from two sources. The first source is gleaned from the Butler County Health Department website, where facility information and health inspection records are published for public review. The facility information page contains restaurant and eatery information such as the phone number and address, while the health inspection records contain information on critical and non-critical inspection counts, as well as the inspection type. The second data source collected for this analysis was collected from Yelp. String type Yelp reviews containing information about the restaurant name, phone number, address, and date of review, along with the actual string text review and review rating were collected.

Health inspection data from the Butler County Health Department website were collected via web scraping techniques. Web scraping resulted in the collection of valuable information such as facility name, address, the phone number, critical violation counts, non critical violation counts, as well as descriptions of critical and non critical violations. Packages in the Python programming language such as Beautiful Soup (Richardson, 2007) and the Selenium WebDriver API were used to gather this data. The resulting dataset contained 1300 rows of inspection records for various eateries in Butler county, which includes restaurants, school cafeterias, vending machines, etc. Out of these 1300 eateries, we then filtered down to actual restaurant names to fit in our scope.

Web scraping was similarly performed for the gathering of Yelp review data. In order to obtain reviews, the Beautiful Soup (Richardson, 2007) package as well as the Selenium WebDriver API were again used to obtain around 47,000 Yelp reviews for all eateries in Butler

county. The process of web scraping for Yelp required more computing power and time, as after a long scraping period, the Yelp website started to block users from getting more information. Many iterations had to be done to collect the amount of reviews mentioned. Also, there was a problem of mis-matched restaurant names between the Yelp and Butler County data, which resulted in very few matches in the initial merge. To resolve this problem, we had to perform conversion for Butler County's restaurant names through running a script to search for the names on Yelp and replacing the original names format with the names from Yelp. A similar process was also done with the address in Butler County's restaurant. After this conversion, we had consistent name and address formats for the two datasets. Then, the reviews were merged to each individual restaurant based on restaurant name and address to account for chain restaurants that had more than one location in Butler county.

Web scraping from these two data sources resulted in one merged dataset that contained facility name, inspection type, inspection date, critical violation count, non critical violation count, Yelp ratings, Yelp reviews for respective establishments, as well as the date of the Yelp review. Sentiment analysis was then integrated into the merged dataset using the NLTK Sentiment Intensity Analyzer (Hutto & Gilbert, 2014) to obtain neutral, positive, negative, and compound sentiment scores for Yelp reviews. Descriptive analysis of the dataset with sentiment scores revealed that after merging on the facility name and address, 185 restaurants resulted with respective Yelp reviews. Non critical violation counts were far more frequent for eateries than critical violations, therefore the count of non critical violations would be the most informative metric for developing a response variable. The median non critical violation count was 6, which as discussed above would serve as the benchmark for informing the public and the Butler County Health Department of "good" or "bad" eateries, or eateries that pose more risk to public health and safety than others. A histogram of non critical violation frequencies revealed that the majority of restaurants in Butler county had less than or equal to 5 non critical violations, with very few restaurants ranging to the frequency of 50 non critical violations.

Wordcloud development that explored keywords in Yelp reviews revealed that a majority of the public leaving Yelp reviews is not as focused on the sanitation and health attributes of eateries, but rather the service and quality of food at these eateries. Some keywords that were observed at the highest frequency among Yelp reviews that were cleaned to rid of unrelated food and restaurant terms were "good", "time", "ordered", "delicious", "best", and "loved". Topic modeling, which is an unsupervised machine learning algorithm, revealed a similar finding where customers leaving Yelp reviews often left reviews based on topics of "food", "good", and "place". There are not a lot of negative keywords associated with the restaurants, so we decided to use every review to have more data for our model to learn, rather than focusing on just the negative ones.

Figure 1. Word cloud from Yelp reviews.

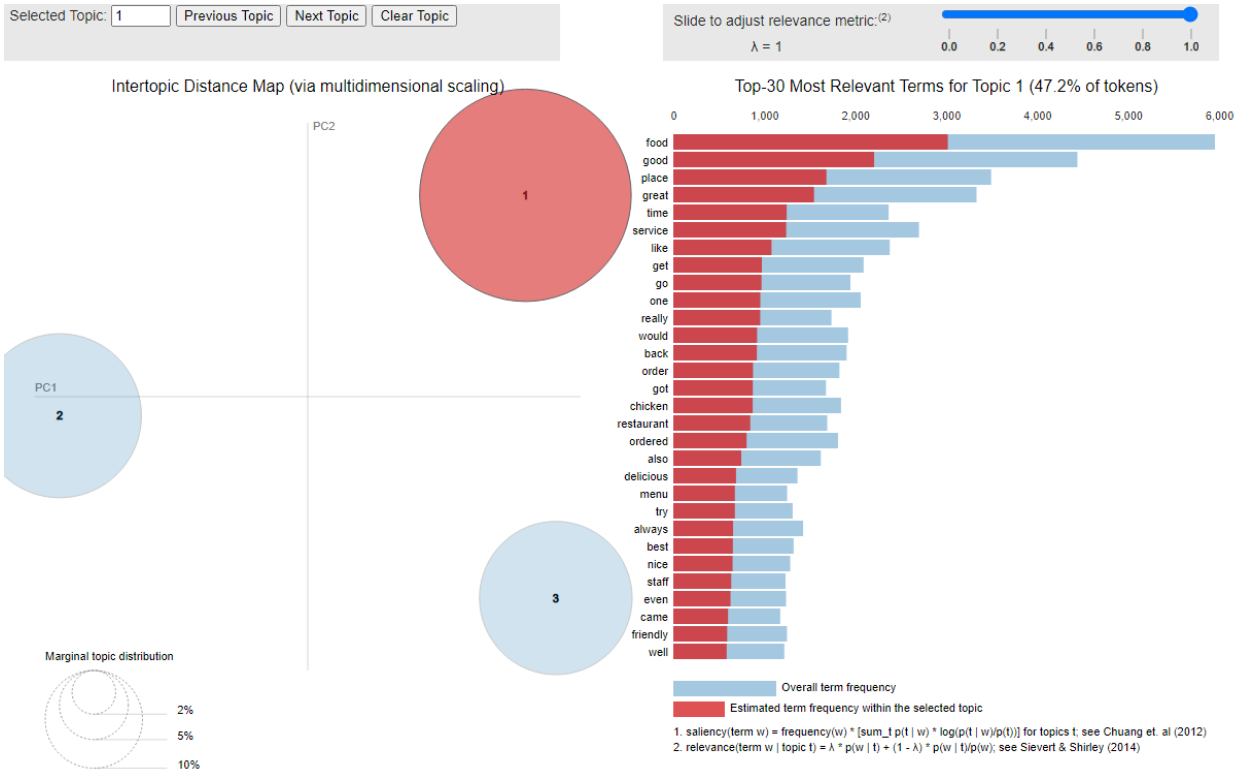

Figure 2. Topic modeling for Yelp reviews.

Other than text mining and topic modeling, we also performed some further exploratory research through data visualization to see if there is a geographical pattern among the non-critical violations - the attribute that we used as benchmark for our response variable. Apart from our

model, this exploratory analysis can also be used to monitored different areas and see if there is any particular area with more violations than usual that needs more attention. The figure below is a snapshot from the map focusing in the Oxford area, which does not have too many non-critical violations.

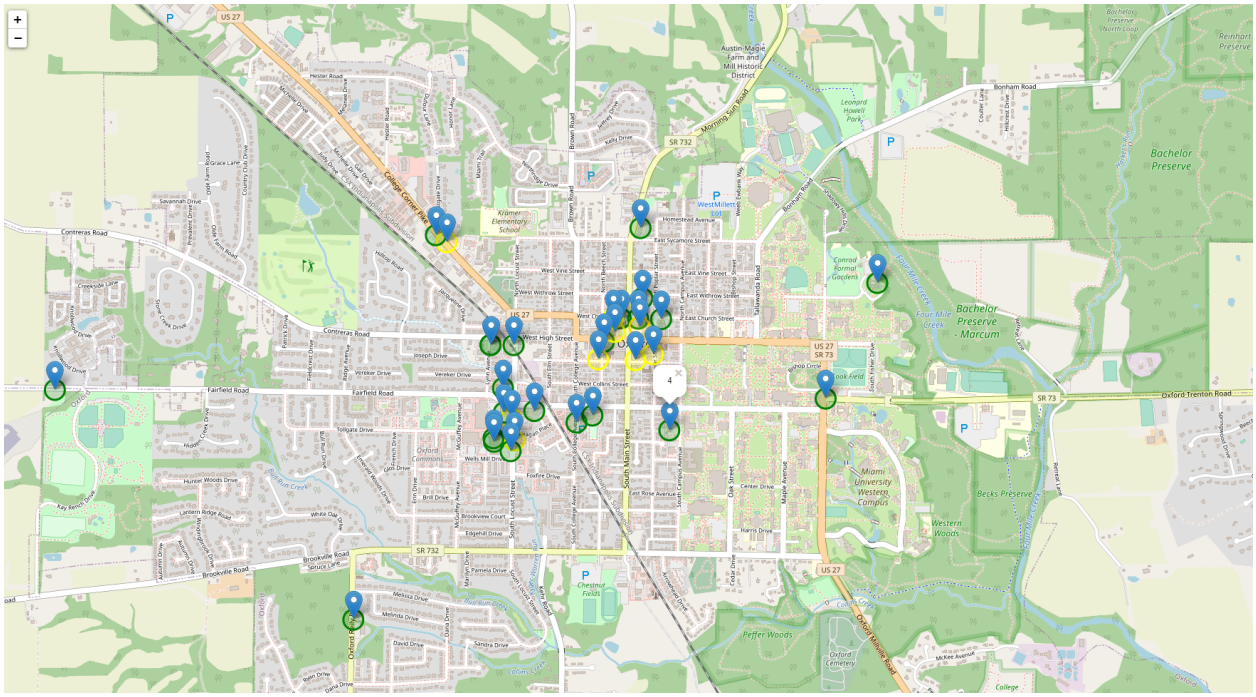

Figure 3. Map of restaurants and their non-critical violations.

IV. Model Building and Evaluation

The second portion of this analysis was the deployment of a supervised learning model that could guide the Butler County Health Department and the public at large in predicting "risky" eateries to a certain extent of reliability. The first step in preparing this machine learning model is to design the predictor variables and the response variable. As mentioned previously, the response variable for the model deployed was "good" or "bad" classification of a restaurant. A "good" classification would suggest that an establishment was safe to eat in terms of food safety attributes, and at low risk or low need of health inspections. Predictor variables for the deployment of the model were chosen based on value in terms of the intended solution as well as predictive performance. The predictor variables included for the final deployment of the model were risk rating, or the standardized risk level of eateries as classified by the Butler County Health Department, total number of Yelp reviews, Yelp review score out of five stars, the city the restaurant was located in, the string type Yelp review, as well as the positive, negative, neutral, and compound sentiment scores of Yelp reviews.

Many different supervised learning classification models were tested and deployed in the preliminary exploration of this data, although the most accurate model with the best predictive ability in terms of classifying a restaurant as "good" or "bad" was a Random Forest Classifier

model (Ho, 1995). The dataset used for analysis was split into 70% training data and 30% test data in order to evaluate predictive performance. The Random Forest Classifier was fit to the training dataset, and tuned using grid search which allows many parameters to be run simultaneously within a Random Forest Classifier to choose the best overall parameters. Such a grid search to obtain the best parameters for the Random Forest Classifier resulted in the best parameters to be n_estimators = 50, min_samples_split = 8, min_samples_leaf = 6, max_features = log2, max_depth = 50, bootstrap = False. In order to obtain the predictive performance of this model, the same parameters were fitted on the test dataset to determine how accurately this model would perform when given data it had never experienced before.

After fitting the model to the test data, a classification matrix was returned. The overall accuracy of the model was 0.5541, which means that the model was able to correctly predict about 55.41% of restaurants as "good" or "bad" in comparison to their actual classification. Furthermore, the precision of a "bad" classification was 0.59, which means the model did somewhat of a good job at correctly identifying bad restaurants - which is our main goal for the model. The sensitivity for a "bad" classification was also around 0.6, which means the model was able to identify 60% of all the actual bad restaurants in our test dataset.

Further evaluation of the Random Forest Classifier model in terms of feature importances revealed that some predictor variables had more influence in terms of risk classification than others. Neutral sentiment analysis of Yelp reviews appears to have a high likelihood of predicting restaurant risk level as well as positive sentiment scores, negative sentiment scores, and the number of Yelp reviews written for a particular establishment.

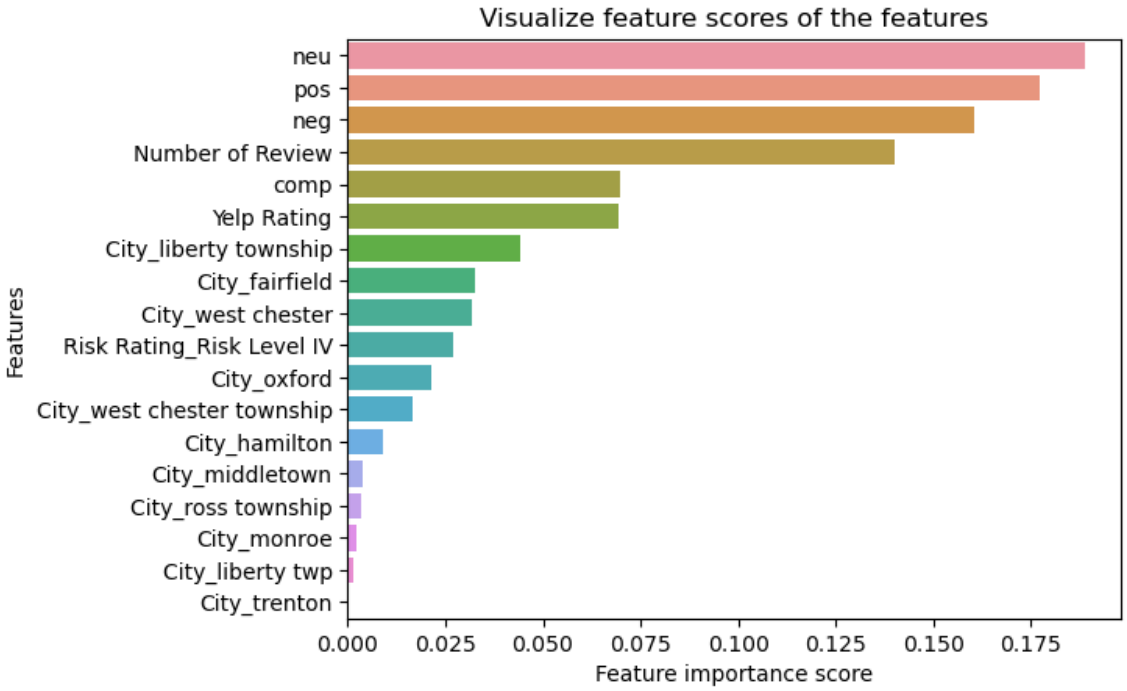

Figure 4. Feature importance from Random Rainforest Model.

V. Conclusion

The supervised learning model and the exploratory analysis of the Butler County health inspection data as well as the Yelp reviews data are a great starting point when looking at the potential association between customer sentiment and the health and safety of dining establishments. The Random Forest Classifier provides predictive power and delineates the attributes most associated with risky classification of eateries in Butler county. Such a delineation of attributes that may affect the food handling practices at dining establishments can be a valuable tool for the Butler County Health Department in their efforts to maintain public health and safety. The results from feature importance selection in the classification model suggest that the Butler County Health Department may want to focus their efforts on reading neutral sentiment analysis reviews to guide their inspection frequency necessity on an establishment by establishment basis. Additionally, it would be valuable to isolate and explore the neutral sentiment Yelp reviews with more advanced text cleaning packages to uncover a pattern or effect that certain words or phrases may have in predicting risky dining establishments. The feature importance of geographic locations as related to health risk level can be useful in guiding consumers and the Butler county health department as to what geographic areas in the county need more frequent inspection than others. For example, Liberty Township, West Chester, and Oxford are the top three most important locations to look at in terms of establishment risk level to eat at. This information can be used to direct efforts of health inspections to specific areas of the county where “good” and “bad” classifications of establishments tend to be correctly identified and maintained. Perhaps these locations have a lot of customers at dining establishments who write many Yelp reviews that correlate with non critical violations counts.

The analysis of Yelp reviews suggests that the scope of most customer Yelp reviews tends to focus on areas of customer service, taste of food, and the location of restaurants as opposed to attributes that are associated with sanitation, health, and safety. Perhaps the issue in uncovering a pattern between Yelp reviews and restaurant health and safety practices is the nature of the data, where we may conclude that the people of Butler county put much more importance on the service and taste of food at dining establishments than they do on sanitation attributes. Moving forward, the model can be rerun using data from a more populated area to see if this is the case.

Overall, the frequency of historical critical and non critical violation counts may be the most indicative factors associated with restaurant risk level as classified by “good” or “bad”. It makes logical sense to think that eateries with a lot of critical and/or non critical violation counts would pose more of a sanitation and health risk than eateries who do not have many critical or non critical violation counts. Perhaps a general shift in focus to looking at the eateries that have more than 10, 20, or 30 non critical violations would be more indicative of uncovering a valuable pattern that predicts restaurant or eatery risk level.

VI. Limitations & Future Studies

One major limitation with this project is that the dataset collected contains a limited amount of observations. The final dataset contained 185 restaurants located in the Butler county area. This is due to the fact that data was collected from Butler County, rather than a more highly populated area with the potential for more observations of data. For future analysis, data collection of health inspection records as well as Yelp reviews for a more highly populated area might be more effective at providing a balanced dataset. For example, targeting areas with more populated and developed cities in Ohio like Cincinnati, Columbus, or Cleveland, or even looking at some of the biggest cities in the United States, like Los Angeles or New York, could be more beneficial for this type of analysis.

A limitation within the collected dataset was that there were far more Yelp reviews that netted positive sentiment scores compared to Yelp reviews that netted negative sentiment scores. When looking through Yelp to see the possible reasoning as to why there were so few negative sentiment reviews compared to positive sentiment reviews, we noticed that there are a number of reviews of certain eateries that have been removed from the site for violating Yelp's Terms and Service agreement. We believe that they were removed because they contained language that Yelp felt was inappropriate, and most likely, these reviews would have netted very negative sentiment scores. The loss of these likely negative sentiment reviews are one reason as to why we were able to collect less negative sentiment reviews.

Another limitation we discovered was that when looking at our sentiment analysis results, a few Yelp reviews did not net super negative sentiment scores because they contained words like "good" for example, even though they were actually pretty negative reviews. The phrasing of certain reviews, despite giving a bad rating, did not necessarily use words that would be associated with a high negative sentiment score. In terms of future similar studies, it may be beneficial to find other sources of customer reviews like Google reviews to gather more observations for sentiment scores and modeling purposes.

A similar study was performed in San Francisco in 2016 where in a dataset of 440 restaurants, a predictive model was deployed using keyword analysis in Yelp reviews to not only detect high-risk restaurants, but also identify specific health code violations that would be made (Schomberg et al., 2016). The aforementioned model performed with an accuracy of 78%, and even saw similar performance measures when the model was deployed in New York. Different variations of the Yelp review dataset performed highly in terms of accuracy and predictive power. An analysis of complete Yelp reviews data in San Francisco revealed an AUC value of 0.98, which suggests that Yelp review analysis in larger cities is an extremely effective way of managing food safety risk (Schomberg et al., 2016). The success achieved in the aforementioned study makes it a very confident assumption that an expanded dataset of Yelp reviews of a larger city could provide an extremely useful model with high accuracy to health departments that can be an innovative guide to health inspections.

References

Allen, T., Walshe, K., Proudlove, N., & Sutton, M. (2019, December 26). *Using quality indicators to predict inspection ratings: Cross-sectional study of General Practices in England*. The British journal of general practice : the journal of the Royal College of General Practitioners. Retrieved December 8, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6917361/

Centers for Disease Control and Prevention. (2018, November 5). *Burden of foodborne illness: Findings*. Centers for Disease Control and Prevention. Retrieved December 8, 2022, from https://www.cdc.gov/foodborneburden/2011-foodborne-estimates.html

Centers for Disease Control and Prevention. (2018, July 26). *Surveillance for Foodborne Disease Outbreaks - United States, 2009–2015*. Centers for Disease Control and Prevention. Retrieved December 8, 2022, from https://www.cdc.gov/mmwr/volumes/67/ss/ss6710a1.html

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Richardson, L. (2007). Beautiful soup documentation. *April*.

Schomberg, J. P., Haimson, O. L., Hayes, G. R., & Anton-Culver, H. (2016, March 29). *Supplementing public health inspection via social media*. PloS one. Retrieved December 8, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4811425/

Will yelp remove a false or defamatory review? Yelp. (n.d.). Retrieved December 8, 2022, from https://www.yelp-support.com/article/Will-Yelp-remove-a-false-or-defamatory-review?l=en_US

Yelp. (n.d.). Retrieved December 8, 2022, from https://www.yelp.com/