Health Inspections and Online Reviews

Carson Bruno

Summer 2024

1 Problem Statement

Restaurant and market health inspections are usually performed one to three times a year, depending on the type of facility.[1] With potential for year long gaps between inspections, the latest score does not necessarily represent the current state of the establishments compliance with health regulations. Due to limited resources, it is not viable for departments to have monthly inspections for all restaurants when there may only be a few restaurants caught in violation. Regardless, restaurant patrons should be aware of updated potential threats to their health. A potential way to provide this up to date information comes from online reviews.

Many food establishments have reviews updated daily containing content about the service and state of the restaurant. If these reviews could be leveraged to predict the current inspection score, then patrons can make more informed decisions. Additionally, regulators could initiate an inspection for those establishments that are flagged as in violation since the last time.

Previous studies have examined this problem and found success in using yelp reviews to predict establishments considered unhygienic.[2] However, subsequent studies have questioned the validity of these results as they ignore class imbalances and do not randomly sample establishments for analysis.[3] Currently, scheduling for health inspections is determined primarily by the type of establishment (restaurant, market, etc.) and previous compliance with regulations.[4] This study examines if publicly available review information can help identify potentially unhygienic establishments and determine if they would fail an inspection if one was performed. All passing and failing inspections will be considered in this approach as opposed to previous work that only considers inspections with no violations versus those with extreme risks.[2] Unlike previous studies that utilized yelp review data, this analysis looks at information from google which has not been previously examined in this context.

2 Data Source

For this analysis, restaurants and food markets in Los Angeles and Chicago are examined. Health inspection data for Los Angeles is sourced from data.gov and includes each unique inspection, associated business information, and the assigned score ranging from 0-100, as well as a letter grade.[5]. For Chicago, health inspection data is pulled from the Chicago Data Portal containing inspections back to Jan 1, 2009 and is updated daily.[6]

This includes similar inspection information to the California dataset, however scoring for inspections is done on a pass/fail basis rather than a letter grade or numerical score.

Review data for Los Angeles and Chicago establishments is obtained from the Google Local Data datasets compiled by the McAuley Lab at UC San Diego.[7] These datasets includes information on different establishments around California and can be used to link to those with available health inspection scores. This source includes two separate datasets, business metadata and reviews. The business dataset includes detailed information about the establishments including name, address, average star rating, and total number of reviews. The reviews dataset provides individual reviews including both textual content and star ratings. This is reduced to a 10-core representation where each business and user are limited to 10 reviews each.

3 Methodology

In order to examine if online reviews can improve upon predictions of health inspection outcomes, binary classification tasks can be employed. The objective in this analysis is to predict whether an establishment will pass or fail a health inspection based on readily available business information (past inspection outcome, type of business) and if this prediction is improved by including review information. Several models are employed for this classification task including Support vector machine (SVM), logistic regression, Random Forest, and XGBoost. These are chosen based on previous studies that found success with these models for the task.

3.1 Data Preprocessing

The business metadata datasets for Los Angeles and Chicago are pulled from the same source and thus can be combined into one business dataset. The inspections dataset for the two cities are, however, from different sources and vary in their grading systems. Los Angeles county has a letter based inspection grading system (A,B,C), in order to treat this problem as a binary classification, inspections with a score of A and B are coded as passing, and C grades are coded as failing. The Chicago dataset includes 5 different categories for inspection outcomes; Pass, Pass with conditions, Fail, Out of Business, and not located. Pass with conditions is considered as simply passing and out of business/not located is dropped from the dataset. The two inspections datasets can then be combined into one.

To find review data associated with each inspection, the business and inspections dataset are first linked to find the associated google maps id. Linking these yields 10,706 unique inspections. The reviews dataset is linked to this by the google maps id. In order to avoid data leakage, reviews are only considered that occurred between inspection dates. For the first inspection scores available, all available reviews beforehand are included.

The review data is preprocessed before being included in the model, this includes removing punctuation,lower-casing, and removing stop words('the','and',etc.). Additionally, words are reduced to their base form via lemmatization. Compared to stemming which simply eliminates suffixes, lemmatization reduces words to their base form.[8] This is useful for review data as it reduces the overall vocabulary and will ensure consistency in language across reviews allowing better model performance.

3.2 Feature Extraction

In order to incorporate the textual data from reviews in the classification model, term frequency-inverse document frequency is implemented. The term frequency metric measures the frequency of a term in a document, while inverse document frequency measures the rarity of a term across all documents. The combination of these metrics results in a TF-IDF score for all words, with higher scores indicating words that appear frequently in a specific document but are rare across the whole corpus.[9] This is helpful in highlighting words that would be indicative of a certain inspection outcome. For example, we may expect to see the words "dirty" or "bugs" have high TF-IDF scores when associated with a failing inspection. For this analysis a single document is considered as a combination of all of the reviews after the previous inspection and before the current one being used for prediction. In order to capture short phrases, bigrams are included here. Additionally,the number of tf-idf features is limited to 5000 to reduce dimensionality.

While the tf-idf scores are a useful way to represent textual information, they fall short in capturing the intention of the original message, in order to gain this back, sentiment scores are calculated for the combined reviews in an inspection period. The VADER sentiment analyzer is used in this analysis due to it's ability to properly detect sentiment in informal review type data.[10]

In addition to textual features, numerical review metrics are also considered. These include the total number of reviews, and the average star rating (1-5). To prevent data leakage, only reviews posted prior to a specific inspection are considered. This mirrors the information that would be available at the time should the model be tested to predict a future inspection outcome.

In order to build a model for comparison with information that the health inspectors currently have, two features will be considered, the previous inspection outcome and the category of the establishment. Google business metadata includes a list of associated categories for each business ('restaurant','market','bar',etc.). Since there are a large number of different categories, this list is joined into a string for each business. Similarly to the review data, TF-IDF scores are computed on this string of categories to evaluate the importance of each in predicting inspection results.

Feature	Description
review text(TF-IDF)	TF-IDF scores for combined reviews within a set inspection period (unigrams and bigrams included, limited to 5000 features)
Sentiment score	Sentiment score for combined reviews within and inspection period.
Number of reviews	Total number of reviews before current inspection date
Average rating	Mean star rating of reviews before current inspection date.
Category (TF-IDF)	TF-IDF score for category of business being inspected as listed on Google (unigrams and bigrams included, limited to 5000 features)
Previous inspection outcome	Result of the previous inspection (0:Pass, 1:Fail)

Table 1: Features used in classification models.

**Inspection period is the time between the previous inspection up until and including the current inspection. For those with no previous inspections the time period is everything before the current date.*

4 Evaluation

Before performing classification, it's important to note that the dataset suffers from an extreme class imbalance (*Figure 1*). There are significantly less instances of businesses that fail their health inspections. This was combated in two ways. For the original proposal, this project only focused on businesses in Los Angeles, but found that only 35 instances of failing inspections were available with review information, more data was then included from Chicago to counteract this. While a large imbalance still remains, SMOTE is used prior to running models to oversample the minority class. This works by choosing a minority class point, randomly selecting one of k-nearest neighbors, and creating a synthetic example between these two points in the feature space.[11]

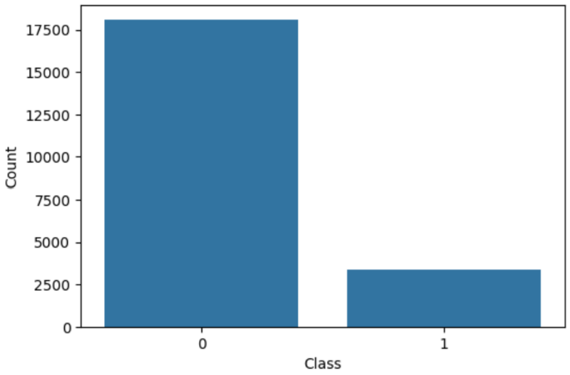
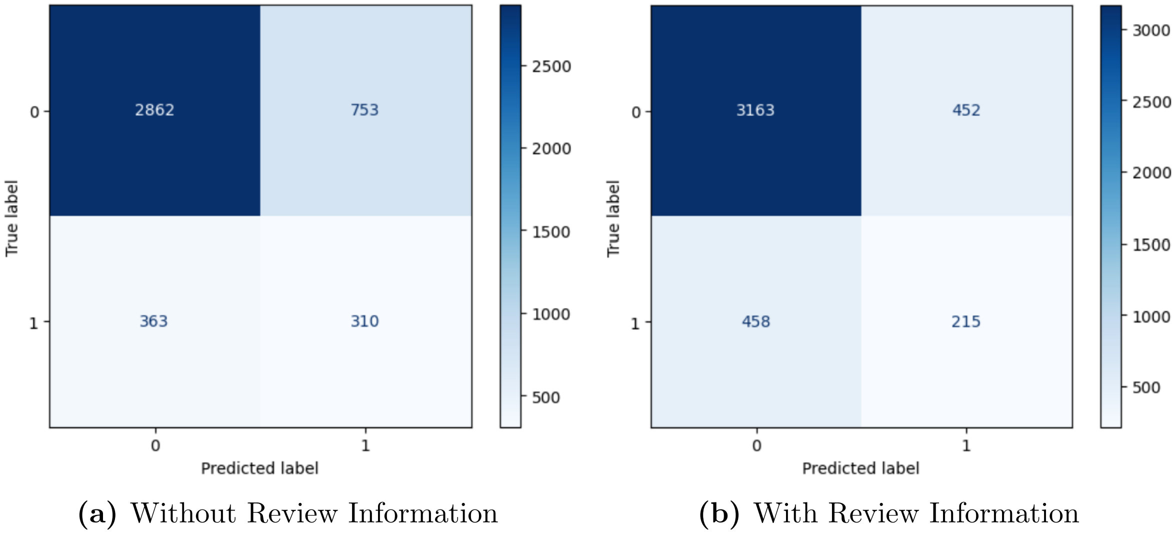


Figure 1: class distribution

The performance of the classification models with and without review information is summarized in Table 2. Out of these, XGBoost and Random Forest achieve the highest performance both with and without the review information. For comparison between the different feature sets, Random Forest was selected over XGBoost due to it's quick performance. While including the review information does increase the overall model performance from an F1 score of 0.75 to 0.79, Table 3 highlights a lack of improvement for class 1 (failing inspections). The model struggles in classifying failing inspections with an F1 Score of 0.32 across feature sets.

Model	Review Info?	Accuracy	Precision	Recall	F1 Score
SVM(linear)	No	0.59	0.79	0.59	0.64
SVM(linear)	Yes	0.66	0.78	0.66	0.71
Logistic Regression	No	0.60	0.79	0.60	0.66
Logistic Regression	Yes	0.68	0.79	0.68	0.72
Random Forest	No	0.73	0.78	0.73	0.75
Random Forest	Yes	0.80	0.79	0.80	0.79
XGBoost	No	0.69	0.79	0.69	0.73
XGBoost	Yes	0.80	0.78	0.80	0.79

Table 2: Classification performance with and without review information. Precision, recall, and F1-score are weighted averages. Models were evaluated using 5-fold cross-validation.



(a) Without Review Information

(b) With Review Information

Figure 2: Random Forest Confusion Matrix with and without review information

Features	Precision	Recall	F1 Score
No Reviews	0.34	0.31	0.32
With Reviews	0.32	0.32	0.32

Table 3: Class 1 metrics for Random Forest model with and without review features.

5 Results

With the goal of using text information to enrich classification of inspection outcomes and alert consumers and health officials of restaurants that would fail an inspection, the models presented here ultimately fall short. Interestingly, the base model without review information has poor performance in predicting failing inspections despite the fact that it includes information that is currently used for determining the frequency of inspections. This indicates that the information that is useful for scheduling when inspections happen is not necessarily useful in determining the outcome of said inspection. The inability of the inclusion of review data to improve prediction performance for failing inspections suggests that reviewers are most likely not picking up on, or documenting, aspects of an establishment that are negatively affecting the inspection score. Reviews most likely focus on service aspects which do not contribute to hygiene conditions. Reviews that do include mentions of hygiene issues are likely too infrequent to hold any real predictive power for the model.

6 Limitations

There are several limitations to this analysis that could contribute to the poor predictive power, the biggest of which being the class imbalance. While techniques such as SMOTE were employed to remedy this issue, they can not entirely account for the lack of information on the minority class. In future studies, inspection and review information could be brought into the model from additional cities to provide a more balanced dataset.

This study also does not take into account the expectation discomfirmation bias that is potentially present in reviews. This bias explains that reviewers are more likely to leave reviews when expectations are not met, meaning users that already expect poor conditions at a restaurant will likely not provide additional reviews on the matter.[12] Including metrics to account for this could potentially improve the predictive power of textual review data.

7 Future Studies

As mentioned, previous studies on this topic have found some success in using textual review data to improve model performance in predicting failing inspections. While this report disproves the notion, there are some potential extensions to this analysis that could yield better results. A more advanced feature extraction for text data, for example LDA for topic extraction could improve textual predictive power.[13] Additionally, including violation specific information in the base model could help with better predictions for failing inspections. The inspections datasets include information on what specific violations lead to each inspection score as well as additional inspector notes which could be good signals for how future inspections for the business will be scored.

8 References

1. County A. Retail Food Inspection Guide for Permanent Food Facilities — Los Angeles County Department of Public Health - Environmental Health. Lacounty.gov. Published 2025. Accessed June 30, 2025. http://www.publichealth.lacounty.gov/eh/inspection/retail-food-inspection-guide.htm

2. Kang J, Kuznetsova P, Luca M, Choi Y. Association for Computational Linguistics; 2013:1443-1448. https://aclanthology.org/D13-1150.pdf

3. Altenburger KM, Ho DE. Is Yelp Actually Cleaning Up the Restaurant Industry? A Re-Analysis on the Relative Usefulness of Consumer Reviews. Published online May 13, 2019. doi:https://doi.org/10.1145/3308558.3313683

4. County A. Retail Food Inspection Guide for Permanent Food Facilities — Los Angeles County Department of Public Health - Environmental Health. Lacounty.gov. Published 2025. http://www.publichealth.lacounty.gov/eh/inspection/retail-food-inspection-guide.htm

5. Restaurant and Market Health Inspections. Data.gov. Published November 30, 2020. https://catalog.data.gov/dataset/restaurant-and-market-health-inspections

6. City of Chicago. Food Inspections. Cityofchicago.org. Published August 8, 2011. https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/about_

7. UCSD CSE Research Project. Google Local review data 2018. Ucsd.edu. Published 2018. https://mcauleylab.ucsd.edu/public_datasets/gdrive/googlelocal/

8. IBM. Stemming and Lemmatization. Ibm.com. Published December 10, 2023. https://www.ibm.com/think/topics/stemming-lemmatization

9. GeeksforGeeks. Understanding TFIDF (Term FrequencyInverse Document Frequency). GeeksforGeeks. Published January 20, 2021. https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/

10. GeeksforGeeks. Sentiment Analysis using VADER Using Python. GeeksforGeeks. Published January 23, 2019. https://www.geeksforgeeks.org/python/python-sentiment-analysis-using-vader/

11. Brownlee J. SMOTE for Imbalanced Classification with Python. Machine Learning Mastery. Published January 16, 2020. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

12. Siering M. Leveraging online review platforms to support public policy: Predicting restaurant health violations based on online reviews. Decision Support Systems. 2021;143:113474. doi:https://doi.org/10.1016/j.dss.2020.113474

13. Wong S, Chinaei H, Rudzicz F. Predicting Health Inspection Results from Online Restaurant Reviews. https://arxiv.org/pdf/1603.05673