Predicting Restaurant Health Inspection Penalty Score from Yelp Reviews

Shashank Uppoor
University of California, San Diego
A53097623
La Jolla, California
suppoor@ucsd.edu

Shreyas Pathre Balakrishna
University of California, San Diego
A53096479
La Jolla, California
spathreb@ucsd.edu

ABSTRACT

This project presents an approach for governments to leverage social media information to make health inspections of restaurants more efficient. We have developed a model to predict restaurant's hygiene conditions given its Yelp's consumer reviews.

Keywords

Health inspection; Yelp reviews; Text mining

1. INTRODUCTION

According to the Centers for Disease Control and Prevention, about one in six Americans(48 million people) get sick, 128,000 are hospitalized, and 3,000 die of food-borne diseases. Food-borne illness is caused by consuming bacteria, viruses, or toxins in food. It is spread when people eat contaminated or undercooked meat, poultry, shellfish, fish, or other foods, or by drinking contaminated water. Public health inspection records aid customers to stay away from restaurants which have poor health scores. Some cities make it mandatory to display health inspection results at their premises. Studies have shown that this decreases profits earned, thereby motivating the establishments to improve their sanitary practices.As in many cities, health inspection in Boston is completely at random. This results in the entire process being inefficient, ensuring eateries with poor sanitation get away while a lot of effort is being spent on the ones that follow the rules closely.

Every year millions of people post reviews on Yelp regarding their experience at these restaurants, which have potential to serve as indicators of sanitary conditions. Our effort is to look for lexical cues in reviews and analyze past health records to arrive at ways to predict health and sanitation conditions of the restaurants. This helps the government agencies to target most of their resources on businesses that are likely to commit higher number of violations, effectively improving the efficiency of the overall exercise.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

We assume that the reviews have been written as soon as visiting the restaurants.

2. DATASET

We have used the dataset from a competition run on Drivendata.org titled 'Keeping it Fresh: Predict Restaurant Inspections'. The data consists of three parts,

2.1 All Historical Violations

All of the historical violations between April, 2006 and June, 2015 for Boston Restaurants. This consist of 34879 data points including test,train and validation data. Each data point consists of Date of inspection, restaurant ID, Number of minor, major and severe violations. A sample health inspection report is included in the Appendix for reference.[1]

2.2 Restaurant ID Mapping

Matches restaurant ID in the violations data to business ID in the Yelp data.

2.3 Yelp Restaurant Reviews

This consists of business details, customer reviews, customer tips, user details for all restaurants in the city of Boston which are inspected between April, 2006 and June, 2015. There are about 1868 restaurants, 235213 reviews, 24424 tips and 71122 users.

3. EXPLORATORY ANALYSIS

3.1 Penalty Scores

We define Penalty Score as follows,

$$PenaltyScore = minor_v + major_v + severe_v \quad (1)$$

Where $minor_v$, $major_v$ and $severe_v$ are minor, major and severe violations respectively.

3.2 Correlation Measures

We have used two widely used statistical parameters, ie Pearson product moment correlation and Spearman's rank correlation to explore the nature of the relationship between features and penalty scores, major, minor and severe violations.[3]

3.2.1 Pearson product moment correlation

This is a measure for evaluating linear correlation between two variables. The coefficient ranges from -1 to +1,

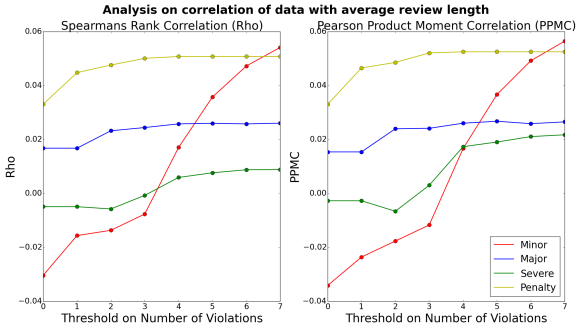

Figure 1: Correlation analysis for average review length

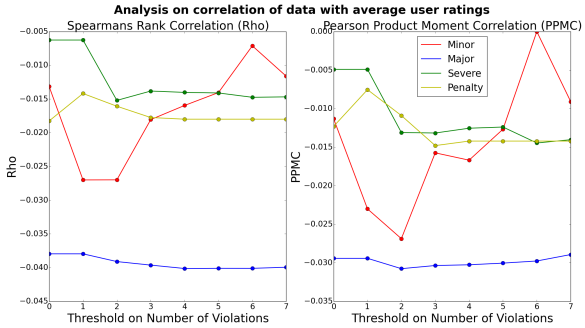

Figure 2: Correlation analysis for average user rating

where '-1' implies perfect negative linear correlation
where '0' implies no linear correlation
where '+1' implies perfect positive linear correlation

3.2.2 *Spearman's rank correlation*

This non parametric measure of statistical dependence assesses how well the relationship between two variables can be described by a monotonic function.The coefficient ranges from -1 to +1,
where '-1' implies perfect negative monotonic relationship
where '0' implies no monotonic relationship
where '+1' implies perfect positive monotonic relationship

3.2.3 *Statistical significance - 'p' Value*

'p' value is a function which is used to test a statistical hypothesis. For analyzing correlation with above measures, we have used a significance level 0f 5%.

3.3 Correlation Analysis and observations

Data is split into eight bins, based on their minor, major, severe violations and penalty score. Correlation Analysis is done on each bin separately to capture the granular changes in the correlation coefficients.

3.3.1 *Average Review Length*

Using Spearman's rank correlation we found that, Minor violations tend to have higher correlation with review length, when number of violations is higher. Major and severe violations also have positive correlation with Average review length.
From this we can infer that, People tend to write longer reviews when they see multiple minor violations.
Pearson's Coefficient doesn't satisfy p value test($p > 0.05$). Hence no conclusion regarding linearity could be obtained.
Fig:1 provides the analysis results.

3.3.2 *Average User Rating*

Spearman's coefficient gives negative correlation.
Restaurants with high ratings usually have less violations. We can infer that people tend to give higher ratings for hygienic restaurants compared to that of unhygienic restaurants. Pearson method also provides negative correlation coefficient which are statistically significant ($p<0.05$)
Outcome of the analysis can be found in Fig:2

3.3.3 *Review Count and Negative Review Count(Count of ratings which are less than three)*

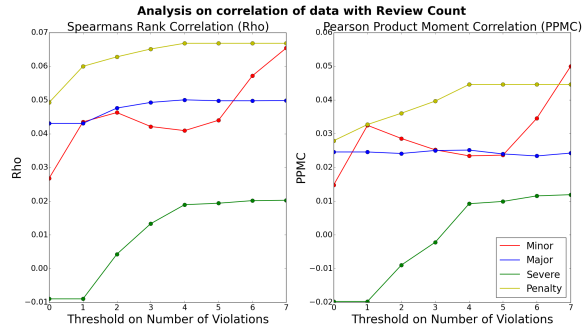

Figure 3: Correlation analysis for Review Count

In Fig:3 we see that minor and major violations have good positive correlation, whereas coefficient for severe violation is comparatively small (although it is positive). Possible explanation is people usually notice non critical violations like 'Leaky pipes', 'Unhygienic toilets', 'improper garbage disposal' etc. But severe violations usually happen in kitchen where customers are not allowed to enter. We found that negative review count gives better correlation compared to review count(Fig:4). From this we can infer that reviewers are vocal about unhygienic practices in their reviews and rate them poorly.Pearson's coefficient suggests that this relationship is linear.

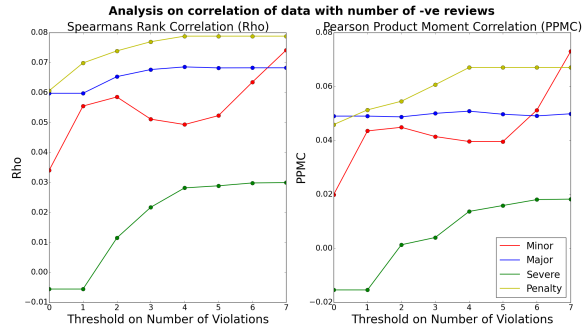

Figure 4: Correlation analysis for negative Review Count

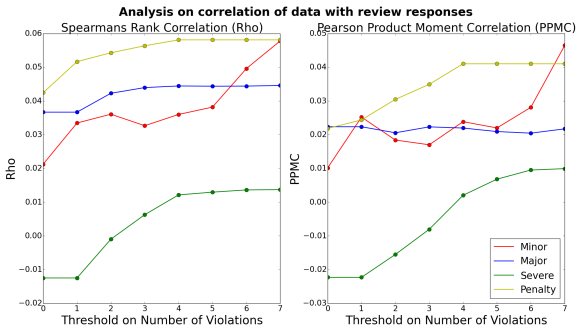

Figure 5: Correlation analysis for Review Responses

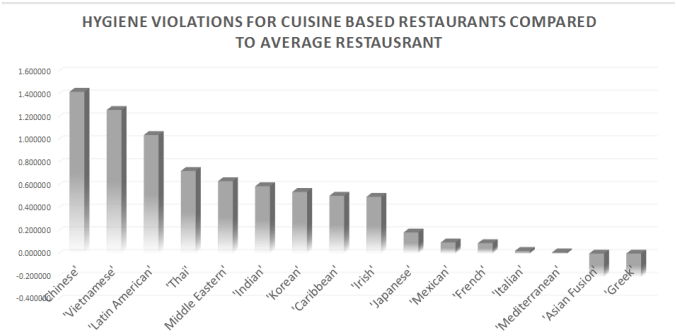

Figure 6: Hygiene Violations and Cuisines

3.3.4 Review Responses

Other reviewers on yelp usually acknowledge the reviews which highlight the violations committed by the restaurants.

4. PREDICTIVE TASK

Input Variable : Each data-point (input) has Date of inspection, restaurant ID, Number of minor, major and severe violations.
Target Variable : For our predictive task the target variable is penalty score.
Penalty score is defined in Eq:1.
The training set comprises of 30000 data points. Out of the remaining 4879 data points, 2000 are chosen at random to form validation set where we tune 'alpha', the regularization parameter of Ridge regression.
We have used Mean Absolute Error (MAE) to evaluate our models. Mean Absolute Error Equation is

$$MAE = \frac{\sum_{x \in X} |\bar{x} - x|}{N} \quad (2)$$

5. FEATURES

5.1 Average user rating

As seen from exploratory analysis, this feature is negatively correlated with penalty score.

5.2 Review Count

As seen from exploratory analysis, this feature is positively correlated with penalty score.

5.3 Average review length

Here, the average is taken over all reviews for the restaurant before the date of inspection. It is positively correlated with penalty score.

5.4 Review response

Penalty is positively correlated with number of review responses indicating that the particular review was 'useful', 'funny' and 'cool'.

5.5 Average of previous scores and previous penalty scores

Restaurants which have a history of violations tend to commit similar number of violations in future inspections. Average of previous penalty scores and the previous penalty score are used to model this behaviour.

5.6 Does the restaurant serve alcohol?

This is a binary feature. The restaurants which serve liquor tend to be unsanitary.

5.7 Is the business a fast food joint?

This is a binary feature. Fast food joints practice better hygiene practices although the co-efficient corresponding to this is very small.

5.8 Cuisine binary vector

One binary feature is used for every cuisine. We see that the Asian themed restaurants like Chinese, Vietnamese, Thai, Korean, Indian usually commit higher number of violations compared to others. This can be attributed to the cooking procedure which usually involves using bare hands to handle ingredients. Restaurants like Mexican, Italian, Mediterranean and French commit lesser number of violations compared to Asian restaurants. Greek and Asian-Fusion restaurants maintain very good hygienic practices. Fig6 depicts Hygiene violations for various cuisines.

5.9 Text mining of reviews and tips

5.9.1 Pre-processing

We created a 'bag of words model' of the Yelp 'reviews' and 'tips' pertaining to restaurant in every data-point, written before the date of inspection.
We ignore case and punctuation here. 1000 most frequently used words were used in the feature vector. Every entry of the feature vector is the number of occurrences of a particular word in that restaurant's list of reviews and tips.

6. MODEL

6.1 Baseline

Here we predict global average of penalty for all data points.

6.2 Ridge Regression

We ran Ridge regression on the the dataset, with regularization factor 3450. We tuned the regularization factor using the validation set. The ridge coefficients minimize a penalized residual sum of squares

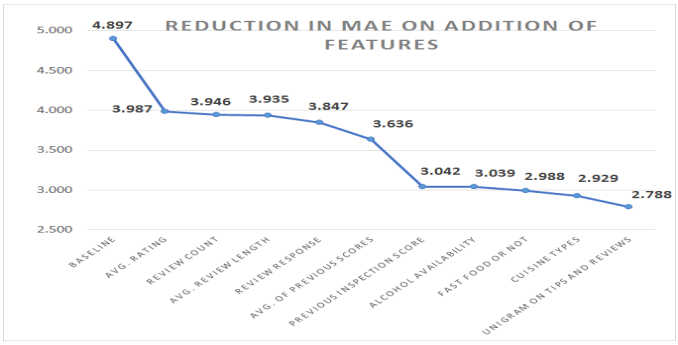

Figure 7: Summary of results

Table 1: Summary

Sl.No.	Features added cumulatively	Mean Absolute Error
1	Baseline	4.897
2	Average Rating	3.987
3	Review Count	3.946
4	Average Review Length	3.935
5	Review Response	3.847
6	Average of previous scores	3.636
7	Previous Inspection Scores	3.042
8	Alcohol Availability	3.039
8	Fast Food or not	2.988
8	Cuisine types	2.929
9	Text Mining:Unigrams	2.788

$$min_w ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2 \quad (3)$$

Where X is the matrix of features, w is the parameter vector and α is the regularizer.Sklearn was used to implement this.

7. RESULTS AND CONCLUSIONS

7.1 Summary of Results

Our model offers 43.07% improvement over Baseline. The result is summarized in Fig:7 and Table:1

7.2 Text Mining

Fig:8 and Fig:9 depict the Wordclouds for hygienic and unhygienic restaurants. The size of the word denotes the value of the Ridge regression coefficient, for that word. So, larger the word, more influence it has in determining the penalty.

7.2.1 Insightful lexical cues

See Table:2 and Table:3 for representative lexical cues for each class

Hygiene

Reviewers tend to use overwhelmingly negative words to describe unsanitary conditions. When the conditions are good they tend to focus on other aspects of restaurants, such as food, ambience etc.

Service and Atmosphere

Restaurants that use fancy ingredients such as veal, wine etc tend to be cleaner. Also establishments which are expen-

Figure 8: Hygienic

Figure 9: Unhygienic

Table 2: Lexical Cues and Examples - Hygienic (clean)

Hygiene	fresh, homemade, pleasant
Cuisines	Italian, homemade, vegetarian, Greek, Japanese
Healthy/Fancier Ingredients	wine, green, chowder, pasta, dumplings, calamari, blueberry, ricotta, veal, pastrami, creme, gelato, traditional, pistachio, poutine, cheese, grilled, dessert, tiramisu
Sentiment	awesome, tasty, expensive, reservations, authentic, pricey, mmm, gem, care, ambiance, perfectly, pleasant, smile, green, valet
whom and where	today, summer, morning, birthday, cafe, date, weekend, noon, downtown, lady, boyfriend, refreshing
Drinks	wine, cocktails, cafe, vanilla, honey, shakes, mocha

Table 3: Lexical Cues and Examples - Unhygienic (dirty)

Hygiene	bathroom, ill, raw, bathrooms,gross, sticky
Cuisines	Chinese, Irish, Mexican, Indian, American, Korean
Basic Ingredients	ribs, lobster, beef, rice, fish, egg, shrimp, seafood, clam, crab, pho, salty, naan
Sentiment	dont, bad, die, disappointed, hell, sucks, poor,beware, hate, cheap
Drinks	beer, ale, margaritas, guinness, cider, soda, liquor

sive(words such as valet, fancy etc) and where prior reservation is required are also cleaner.

Whom and When

If people talk about their date/boyfriend things seem to go well.

Way food is described

Basic ingredients such as soy, eggs, seafood correspond to unclean restaurants. Establishments serving beer tend to be unclean.

8. LITERATURE SURVEY

Some of the past studies on social media analysis for public health monitoring include work by Dredze on 'Social Media as a Sensor of Air Quality and Public Response in China' , Nicholas Generous on 'Global Disease Monitoring and Forecasting with Wikipedia' and Philippe Barboza on 'Evaluation of Epidemic Intelligence Systems Integrated in the Early Alerting and Reporting Project for the Detection of A/H5N1 Influenza Events'. Krieck et al. explored augmenting the traditional notification channels about a disease outbreak with Twitter data. Researchers have tried to analyze the overall trend of disease outbreak by analyzing social media(Culotta ;Lampos et al.; Chunara et al.). There has also been research on study of seasonal trends in mental health disorders such as depression across the globe(Golder and Macy).

Our work mainly draws its inspiration from 'Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews' by Jun Seok Kang et al. 2013)[2]. Here, online reviews of restaurants based in Seattle are used to predict hygiene inspection records. Prediction strategy used here was liblinear's Support Vector regression. We have used online reviews of eateries based in Boston. We have used Sklearn's implementation of Ridge regression to perform predictions. Some of the things we borrowed from the paper include using Spearman's coefficient to estimate correlation, bag of words approach to analyze lexical cues etc. In addition to Spearman's co-efficient we use Pearson co efficient to guess the nature of relationship between features and output variable. Some of the additional features we have used, with respect to the work described above are user tips, business categories, user compliments etc.

9. FUTURE WORK

We could use mixture of unigrams and bigrams to extract meaningful features from review text to further improve prediction. Also, we can predict Minor, Major and Severe Penalties separately. In order to accomplish the above, we need more data points.

10. ACKNOWLEDGMENTS

We would like to thank Prof McAuley and the Teaching assistants for all the guidance and support.

APPENDIX

A. REFERENCES

[1] Sample Health Inspection Report `http://goo.gl/v78PTn`

[2] Jun Seok Kang. *'Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews'*

[3] Pearson and Spearman's Coefficient definition www.statisticssolutions.com/correlation-pearson-kendall-spearman/