Using Yelp data to predict health violations of restaurants

Mansee Jadhav
University of California- Los Angeles
Masters in Computer Science
UCLA id- 204567818
(mansee8@cs.ucla.edu)

Mauli Shah
University of California- Los Angeles
Masters in Computer Science
UCLA id- 004567942
(maulishah91@cs.ucla.edu)

Avneet Oberoi
University of California- Los Angeles
Masters in Computer Science
UCLA id- 104572014
(avneet@cs.ucla.edu)

(Dated: May 30, 2016)

I. ABSTRACT

Yelp is a widely used application which connects people with local businesses by gathering a rich corpus of data about customers experiences at those businesses via reviews, tips, check-ins and business attributes. This paper offers an approach for governments to harness the information contained in social media in order to make public inspections and disclosure more efficient. As a case study, we turn to restaurant hygiene inspections which are done for restaurants throughout the United States and in most of the world and are a frequently cited example of public inspections and disclosure. We use data from Yelp restaurant reviews to narrow the search for health code violations that predict violations and future health score that will be assigned to a business at their next health inspection and thus help public health inspectors do their job better. This was a competition held by Yelp, in association with Kaggle last year. We present the first empirical study that shows the viability of statistical models - Linear regression, Random Forest, MARS, Neural Network and KNN, that learn the mapping between textual information of restaurant reviews and the hygiene inspection records from the Department of Public Health.

II. MOTIVATION

The Center for Disease Control(CDC) has reported that more than 48 million Americans fall prey to food poisoning and estimate that 75% of those are due to food consumed at delis and restaurants. Currently, inspectors are sent to restaurants in a random fashion. Since cities only have a limited number of health inspectors, quite often their time is wasted on spot checks at clean, rule-abiding restaurants. This also means that sometimes restaurants with poor health and safety records are discovered too late. Thus by predicting health scores of restaurants, we can direct health inspectors to unhealthy establishments which will lead to fewer people falling sick due to food-borne illnesses. Also by linking the health score to the establishment on Yelp, thereby making it publicly available, best practices will improve across the eatery industry.

III. INTRODUCTION

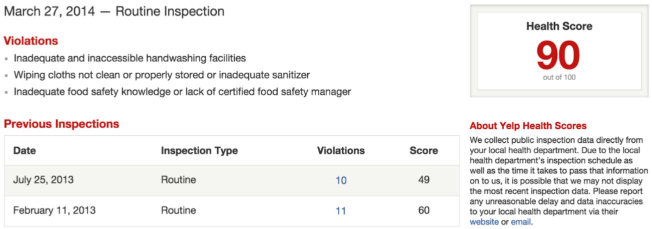
Public health inspection records help customers to be wary of restaurants that have violated health codes. In some counties and cities, e.g., LA, NYC, it is required by restaurants to post their inspection grades at their premises, which have shown to affect the revenue of the business substantially (e.g., Jin and Leslie (2005), Henson et al. (2006)), thereby motivating restaurants to improve their sanitary practice. Other studies have reported correlation between the frequency of unannounced inspections per year, and the average violation scores, confirming the regulatory role of inspections in improving the hygiene quality of the restaurants and decreasing food-borne illness risk.

However, one practical challenge in the current inspection system is that the department of health has only limited resources to dispatch inspectors, leaving out a large number of restaurants with unknown hygiene grades. We postulate that online reviews written by the very citizens who have visited those restaurants can serve as a proxy for predicting the likely outcome of the health inspection of any given restaurant. Such a prediction model can complement the current inspection system by enlightening the department of health to make a more informed decision when allocating inspectors, and by guiding customers when choosing restaurants.

The following models were developed - Linear regression, Random Forest, MARS, Neural Network and KNN, to help predict health violation level V1(Severity level 1), V2(Severity level 2), V3(Severity level 3) for restaurants, thus discriminating severe offenders from places with no violation, and provides insights into salient cues in reviews that are indicative of the restaurants sanitary conditions. Our study suggests that public disclosure policy can be improved by mining public opinions from social media to target inspections and to provide alternative forms of disclosure to customers.

IV. DATASET DESCRIPTION

Using Yelps data for restaurants, food and nightlife businesses in Boston as well as past history of health inspections, we **predict the future health violation level that will be assigned to a business at their next health inspection**. Yelp partnered with the City of San Francisco and City of New York to develop the Local Inspector Value-Entry Specification (LIVES). **Local Inspector Value-Entry Specification (LIVES)** is an open data standard which allows municipalities to publish restaurant inspection information to Yelp. The dataset provided consists of information about the Business, Users, reviews, inspection history, check in information and violations.



March 27, 2014 — Routine Inspection

Violations

- Inadequate and inaccessible handwashing facilities
- Wiping cloths not clean or properly stored or inadequate sanitizer
- Inadequate food safety knowledge or lack of certified food safety manager

Health Score

90

out of 100

Previous Inspections

Date	Inspection Type	Violations	Score
July 25, 2013	Routine	10	49
February 11, 2013	Routine	11	60

About Yelp Health Scores

We collect public inspection data directly from your local health department. Due to the local health department's inspection schedule as well as the time it takes to pass that information on to us, it is possible that we may not display the most recent inspection data. Please report any unreasonable delay and data inaccuracies to your local health department via their website or email.

FIG. 1. LIVES provided information to Yelp

Since the City of Boston records health violations at three different levels *, **, and ***, which can be thought of as ”minor”, ”major”, and ”severe” violations, the aim is to predict the count of each of these levels of violations for an inspection of a particular restaurant, given the history of violations of the restaurant. The dataset can be downloaded from reference link [7].

business_id	name
address	city
state	postal-code
latitude	longitude
contact-details	

TABLE I. Business.csv

votes (funny,useful,cool)	user_id
review_id	stars
date	text
time	

TABLE II. reviews.csv

yelping_since	votes (funny,useful,cool)
review_count	name
user_id	friends
fans	average_star
type	compliments

TABLE III. users.csv

check in info	type	business_id

TABLE IV. checkin info.csv

votes (funny,useful,cool)	user_id
business_id	review_id
stars	date
text	type

TABLE V. All violations.csv

business_id	date
score	result
description	type

TABLE VI. inspection.csv

user_id	text
business_id	likes
date	type

TABLE VII. tips.csv

V. PROJECT WORKFLOW

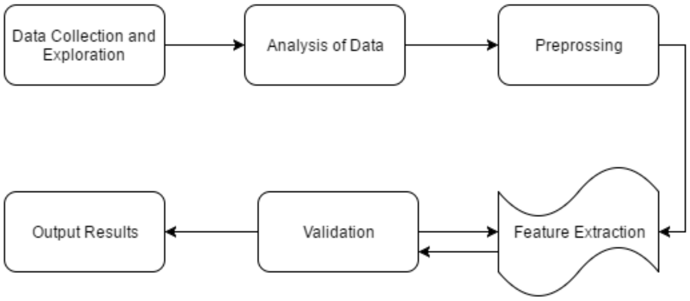


FIG. 3. Workflow

Beginning with the Data set collection and exploration, we started to analyze our data. Many of the unneccessary columns were deleted and new features based on their importance were added. Prepossessing the data included elimination of unwanted attributes/features, checking for missing values in data-set,data formatting and merging data-sets. Finally the results were validated after training the model to get best results. More of this is explained in section: Phases of the Project.

VI. LANGUAGE AND PACKAGES USED:

Language: Python, R.

Packages: openNLP, quanteda, jsonlite, randomForest, caret, e1071, gbm, stringr, kernlab, earth, nnet, nltk and pandas.

VII. PHASES OF THE PROJECT

1. *Loading the dataset*

We filtered out unwanted columns like *business.address/ city / state/ postal-code* by dropping them. Dataframes were created for each of the datasets. Based on primary key- foreign key concept we created associations between each of the loaded dataframes.

2. *Deciding on Features*

We created a dictionary where the keys are Yelp IDs and the values are Business IDs, so that we can use this to figure out which restaurant reviews match which hygiene inspections. Each review from the Reviews.csv was parsed, stemmed and lemmitized and the stop words were removed. We performed data analysis and visualizations in form of scatterplots and correlation matrices to eliminate completely uncorrelated features.

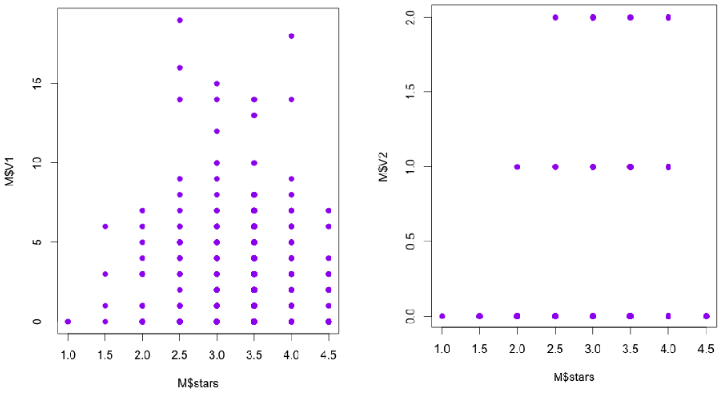


FIG. 4.

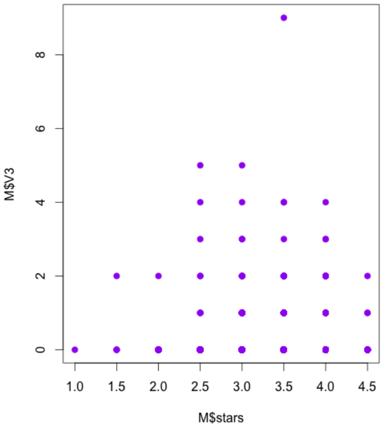


FIG. 5. Violations Vs Star Rating of Restaurants

As seen from the above plot of Violations Vs Ratings of the restaurants, we can see that violations (v1, V2, V3) are highly dependent on ratings of the restaurants. Restaurants with high ratings usually have less violations.

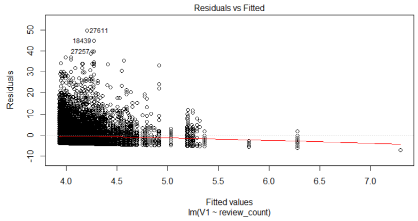


FIG. 6. Violations Vs Review Count

Also from the plots of Review-counts and Parking status, we can see that Parking is no where related to Hygiene Violations of restaurants, while review count plays

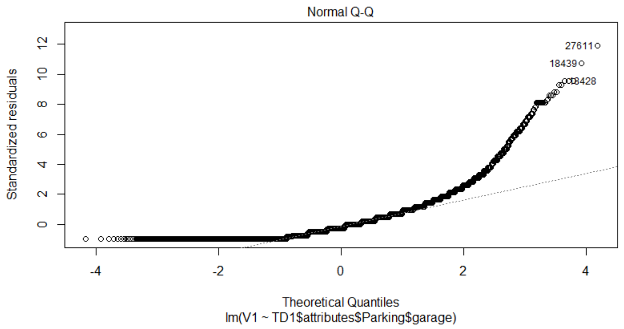


FIG. 7. Violations Vs Parking

significant role in predicting the health violations.

Following features were found to prominently improve the performance of the model:

- Number of review count for each of the restaurant.+
- Business Rating.+
- Difference between actual and average rating.+
- Average of star count received by the restaurant. +
- Length of the review text.+
- Number of characters in the review.+
- Number of Categories (Cool/Funny/Useful) of the review.+

In addition to above features, we also added features created by text processing. Text Processing was done in *Python*. We considered only those reviews which were prior to the most recent *inspection date* of particular restaurant, as we did not want any reviews after the considered inspection date to influence our next violation prediction. This assumption greatly improved the model prediction.

- Inspection date. +

Review text was stemmed, stop words were filtered out and lemmatized. *Tf-idf* vector matrix was created for review-text and then we extracted top 1500 words.

- Top 1500 words from review text based on tf-idf matrix. +

Thus a matrix is created with rows as the restaurant ID and columns as the 1500 words. The value gives the tf-idf weightage of the word for that restaurant. This can then used as a feature for the models created.

Each restaurant is known for the cuisine style it offers. Considering this fact, we maintained a dictionary of all cuisines(160) and maintained a one hot encoding cuisine vector for each of the restaurant.

- 1-hot encoded cuisines (160 features) vectors. +

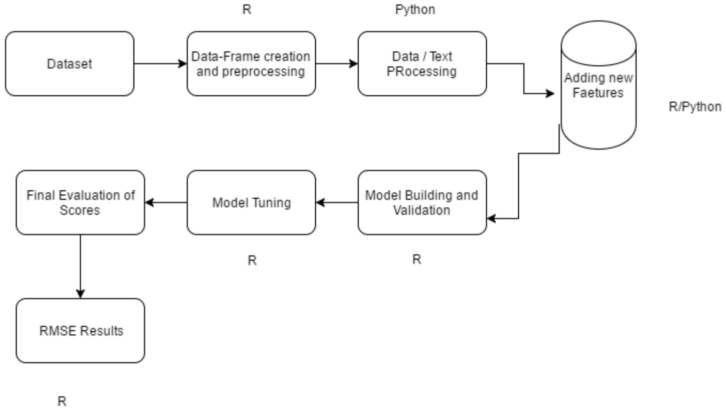

R
Python
Dataset
Data-Frame creation and preprocessing
Data / Text PRocessing
Adding new Faetures
R/Python
Final Evaluation of Scores
Model Tuning
Model Building and Validation
R
R
RMSE Results
R

FIG. 8. Data Flow Diagram

Data Flow diagram gives a glimpse of the languages used for features extraction.

3. ***Model Predictions***

VIII. MODEL PREDICTION AND RESULTS

The evaluation metric chosen for this competition was a weighted RMSE. So the aim was to reduce this metric. We tried and tested our extracted features on many regression models like linear regression, KNN, SVM, trees, Random forest etc. Of which GBM gave best prediction results. Due to large dataset,consisting of about 170 predictors, running each models takes 2 - 8 hours with cross validation. Hence fine tuning of model was difficult. It also takes 2 hours to create data-frames for model fitting. Due to the large size of reviews data, text processing was also very time consuming.Model tuning also required extensive research. However ***GBM*** has till now given the best prediction results.

IX. ANALYSIS

Since the City of Boston records health violations at three different levels *, **, and ***, which can be thought of as ”minor”, ”major”, and ”severe” violations, the aim is to predict the count of each of these levels of violations for an inspection of a particular restaurant.

Weighted RMSE Metric

$$= \sqrt{\frac{1}{N}\sum_{i=0}^{N}(log(y_i.W+1) - log(\hat{y_i}.W+1))^2}$$

where:
W: Weights assigned based on violation.
y_i : Actual output violation score.

$\hat{y_i}$: Estimated output violation score.

First we calculate a single score for the predicted and actual by multiplying each by vector of weights followed by log of this scores. Finally, we calculate the root-mean-squared error of the differences between the log scores. Weights are [1,2,5] for [*,**,***] respectively.
After evaluating each of the above models the score obtained are as follows:

Prediction Model	Evaluation Score
GBM	1.1178
Random Forest	1.119
MARS	1.121903
Neural Networks	1.1188
KNN	1.12382

As seen from the above table GBM (Generalized Boosted Model) gave the best results. Since the competiton is no longer active, we did not have a provided test file to run the scores against, so split our training data into test and training data. The winner of the contest had a very impressive RMSE score of ***0.4257***. With our score as per the public leaderboard, assuming similar results on the test file provided, we would rank amongst the ***top 30 out of 53*** entries to the contest

X. CONCLUSION

We have reported an empirical study demonstrating the promise of review analysis for predicting health violations score V1/V2/V3, introducing a task that has potentially significant societal benefits, while being relevant to much research in NLP for opinion analysis based on customer reviews.Infectious diseases spread through food or beverages are a common, distressing, and sometimes life-threatening problem for millions of people in the United States and around the world. The Centers for Disease Control and Prevention (CDC) estimates that each year in the United States, 1 in 6 Americans (or 48 million people) gets sick, 128,000 are hospitalized, and 3,000 die of foodborne diseases. Using Yelps data for restaurants, food and nightlife businesses as well as past history of health inspections to predict the future health score that will be assigned to business (hotels) at their next inspection, can be of great help to health inspectors to narrow down there inspection schedule and be more focused on more health violated restaurants.Our developed GBM based prediction model gives best prediction.

XI. RELATED WORK

There have been several recent studies that probe the viability of public health surveillance by measuring relevant textual signals in social media, in particular, microblogs (e.g., Aramaki et al. (2011), Sadilek et al. (2012b), Sadilek et al. (2012a), Sadilek et al. (2013), Lamb et al. (2013), Dredze et al. (2013), von Etter et al. (2010)). Our work joins this line of research but differs in two distinct ways. First, most prior work aims to monitor a specific illness, e.g., influenza or food-poisoning by paying attention to a relatively small set of keywords that are directly relevant to the corresponding sickness. In contrast, we examine all words people use in online reviews, review length, cuisine style, inspection dates and draw insights on correlating terms and concepts that may not seem immediately relevant to the hygiene status of restaurants, but nonetheless are predictive of the outcome of the inspections. Second, our work is the first to examine online reviews in the context of improving public policy, suggesting additional source of information for public policy makers to pay attention to.

We expect that previous studies for aspect-based sentiment analysis (e.g., Titov and McDonald (2008), Brody and Elhadad (2010), Wang et al. (2010)) would be a fruitful venue for further investigation.

XII. FUTURE WORK

- Improving RMSE value by adding and trying out more features.
- We think some more features can be squeezed out of the review text as well.
- We also would like to use some aspect of user's authenticity in determining how important his/her review should be weighted as.
- We would also like to try out more models and to better fine tune the models.

XIII. REFERENCES

[1] Jun Seok Kang Polina Kuznetsova
*http ://homes.cs.washington.edu/ yejin/Papers/emnlp*13_*hygiene.pdf*

[2] Local Inspector ValueEntry Specification (LIVES) *http : //www.yelp.com/healthscores*

[3] Kaggle Competition
https://www.yelpblog.com/2015/06/data-science-challenge-predict-restaurant-health-scores-with-yelp-data.

[4] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 15681576, Edinburgh, Scotland, UK., July. Association for Computational

Linguistics.

[5] Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 10, pages 804812, Stroudsburg, PA, USA. Association for Computational Linguistics.

[6] Tomislav Hengl, a, , Gerard B.M. Heuvelinkb, , Alfred Steina,
A generic framework for spatial prediction of soil variables based on regression-kriging
http://ftp.auckland.ac.nz/software/CRAN/src/contrib/Descriptions/gbm.html

[7] Dataset download link: https://www.drivendata.org/competitions/5/page/33/