Exploring Online Food Review Platform's Representation of Health Inspection

A Regression Analysis of Restaurants in Boston Using Big Data Tools

Zhehan (Andrew) Shi and Shubo (Gabriel) Xu
Processing Big Data for Analytics Applications
New York University
May 8, 2021

Abstract

To investigate the comprehensiveness of online crowdsourced review platform in reflecting restaurants' qualities, especially their cleanliness, we surveyed the restaurants in the city of Boston by acquiring data from Yelp and the Health Division of the Department of Inspectional Services. We utilized Big Data tools to clean and profile the datasets obtained from these two sources, merged them by matching restaurants through address, name, and geographical coordinates, and conducted regression analysis on the resulting dataset. The obtained results led to an unexpected weak correlation between a restaurant's performance on crowdsourced online review platforms, either in terms of rating (measured in star ratings) or popularity (measure in numbers of reviews) and its officially-recorded health inspections.

Introduction

We studied how representative a restaurant's user-facing profile on a crowdsourced online review platform was of its health and sanitary condition. It was paramount that users could obtain an accurate and well-represented impression of a restaurant's cleanliness from the information listed on the platform.

It was an implicit assumption that a score provided a holistic review of a food establishment. The reviews should consider its health and sanitary condition; therefore, we conducted the study to verify the aforementioned assumption. Our work provided insight into the representativeness of health and sanitary condition in the information.

Past research[1][2] utilized data from online reviews to assess a restaurant's cleanliness; however, most of these works either aimed at policymakers (e.g., arranging health inspection; disease tracing) or used linguistic data to predict health inspection outcomes. Consequently, there was room left for exploration in taking advantage of the full potential of the numerical information concerning health violations.

Furthermore, normal users were neither willing nor capable to afford time and effort combing through the data, forcing them to resort to ratings and popularity. The platform's existing functionality in allowing users to sort the results by popularity and rating also hugely affected their decision-making process; therefore, it was indispensable to ascertain the association of numerical information with the outcomes of health inspection. Our work provided a unique insight into the representativeness of numerical data, which are ratings and the number of reviews, to assess a food establishment's health and sanitary condition.

Since the data from a crowdsourced online review platform and data on restaurants' official health violations were stored and managed separately, with different schemas and identifiers, it was unclear how representative the former reflected the latter underutilized and overlooked general consumers and public alike. The two datasets often used slightly different names, addresses, and geographical coordinates for the same food establishments. Using Big Data tools, we implemented a merge of the two data sources using a partial string of name, address, and geographical coordinates of restaurants. Doing so ensured the accuracy and consistency of our results. The merge was crucial. Merging two, otherwise entirely unlinkable, datasets hugely enhanced our ability to analyze the data; therefore, the merge itself is our contribution to studies in this area.The workflow could be understood better through the illustration in Figure 1.

Regression analysis revealed a weak correlation between a restaurant's health inspection results and its respective ratings on the online reviews. The correlation between online popularity, measured by the number of reviews and health inspection results, was even weaker by another magnitude and had a negative slope. The negative correlation could mean that customers were more inclined to leave reviews after a negative experience than they would otherwise (expectation disconfirmation bias). The very weak correlation, an almost negligible r-squared value, brought to light the representativeness of existing online information. A logistic regression was also performed on the relationship between health inspection outcomes and restaurant openness. The area under receiver operating characteristic

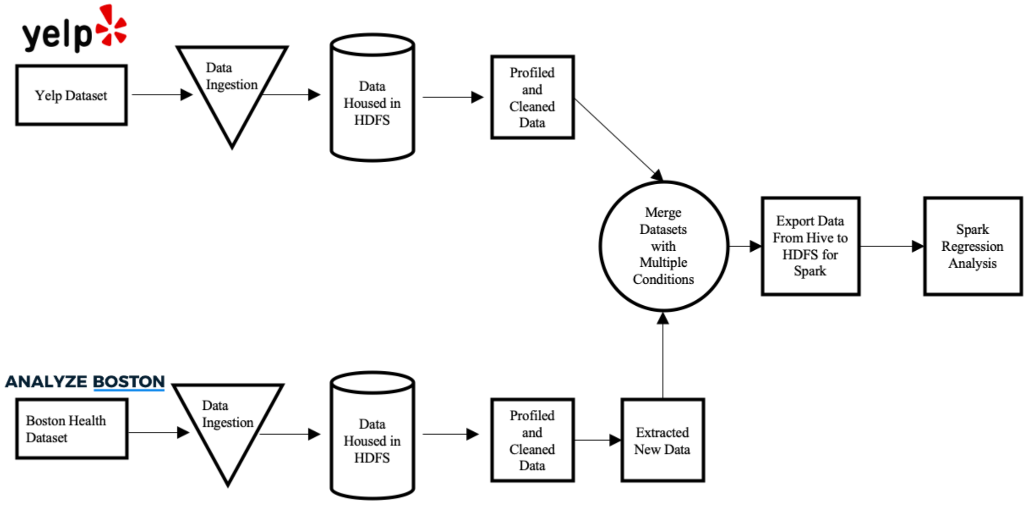
yelp
Yelp Dataset
Data Ingestion
Data Housed in HDFS
Profiled and Cleaned Data
Merge Datasets with Multiple Conditions
Export Data From Hive to HDFS for Spark
Spark Regression Analysis
ANALYZE BOSTON
Boston Health Dataset
Data Ingestion
Data Housed in HDFS
Profiled and Cleaned Data
Extracted New Data

Figure 1: Design Diagram

(ROC) shows that the health inspection outcomes were not predictive of restaurant openness in the future.

This insight had three forms of impact: 1. Make users aware of the risk of food safety if they only conduct a quick search on Yelp. 2. Inspire users to pay more attention to health conditions when leaving reviews in the future. 3. Push Yelp to better integrate the aspect of health conditions to their profiling of restaurants. The insight 1 and 2 create a positive feedback loop by reminding users of the significance of leaving comments on restaurant sanitary conditions for others.

Motivation

With the prosperity of online review platforms, a remarkable amount of data was being generated daily and provided a solid basis for data-oriented analysis. Significant efforts were devoted to this subject from the intersection of industry and academia. There was notable progress made in data extraction, the profiling of data, and data analysis. However, most of these works were conducted as tools to advance public policy efforts. For example, a typical use case of the data is to help the Health Department determine which restaurants to inspect, given a limited resource [1] or disease tracing purposes[2]. Considerably less attention was paid to prioritize the interests of individual consumers — who form the basis of the online review platform.

Our work was intended to fill this gap. We assumed that in contrast to governmental departments or policymakers who took a macro and holistic approach to the information available, individual consumers or online review platform users simply looking for the next meal were much less willing to research beyond the most fundamental criteria of rating and popularity. The observations further confirmed this assumption that most online review platforms offer users the option to sort specific restaurants by rating or popularity. By analyzing the comprehensiveness of these most common aspects of restaurants checked by individual users, we provided data-backed insights on how much/little they reflect a restaurant's cleanliness, which we believe is of the utmost importance to every consumer. Individual users who are now conscious of the limits of a restaurant's existing online profile would be likely in the future to leave reviews that push such boundaries. A positive feedback loop was therefore formed.

Related Work

As mentioned earlier in the paper, a considerable amount of research was conducted to investigate the potential use of online crowd-based data to evaluate the cleanliness of restaurants. In particular, Michael Siering explored the role of online reviews in helping health departments decide how to strategize health inspections given constrained resources[1]. In his work, Siering described a high degree of data utilization that extracts information, including ”restaurant visitor behavior, linguistic features, and review texts.” Among a composed list of classifiers, Siering used statistical tools to show that ”restaurant popularity, linguistic review information, and textual information help identify restaurant health violations” — the main result of his research.

Acknowledging the validity of Siering's work, we noted that our project had a different targeted group of beneficiaries — everyday consumers, instead of health department officials. Although this distinction in ideology and consequently data profiling led us to almost the contrary of Siering's qualitative result, we were inspired by his work to look for additional ways to integrate the data generated by everyday “citizens of the internet” and inspections by government officials. It was also refreshing to note that, in addition to the main result of his research, Siering also introduced the concept of expectation disconfirmation bias in the scenario of online reviews, which he defined as a pattern where ”users are more likely to contribute online reviews when their expectations are not met.” This concept could help explain some analytical results we have obtained later on where health score was negatively correlated with online popularity.

Unlike Siering's and our works, which both relied on data gained directly from online reviews, Sadilek et al. took a more holistic and costly approach to evaluate the health score of restaurants that draw data from online social media posts —– a much large pool of data in which the relevant information was significantly more scarce[2]. Their work centered around a program that combed through the entire Twitter database and targeted explicit and implicit mentions of food poisoning, which they then used to start a trace that would ideally result in the responsible restaurants. Sadilek's work and choice of data source helped us decide to use data from yelp, which was much more concentrated and incurred a much lower cost of analysis. This decision was made acknowledging that Sadilek et al.'s work was conducted nearly a decade ago, after which the size of Twitter's database had increased by magnitudes.

In the trade-off between accuracy and comprehensiveness, we emphasized the former as one of our main objectives is to evaluate the quality of consumer-facing online profiles of restaurants. Although we differed much in methodology, their work inspired us to profile data based on geographical coordinates, which proved to be an essential step in our work and its potential contributions. Discussions of the accuracy of their program also propelled us to look into calibration of the accuracy thresholds of the location coordinates in response to scenarios that contained a high concentration of restaurants, including food courts and Chinatown.

Data Sets

1. Yelp Dataset

The precleaned Yelp Dataset was in JSON format. The original dataset was provided by Yelp itself through a simple application on its site. There were multiple files provided and the whole JSON dataset was 11.42 GB; however, the portion we needed was only 124.4 MB in size and was last modified on January 28, 2021. It has 160,585 businesses from 8 metropolitan areas. From 14 fields we selected 9 of them for our cleaned Yelp dataset, and we chose Boston as our target city. The field *is_open* used 1 to mean open and 0 closed. See Table 1 for the schema with selected fields described above.

Table 1: **yelp_business**

Column Name	**Data Type**
business_id	string
name	string
address	string
city	string
stars	decimal(2,1)
review_count	int
is_open	int
latitude	string
longitude	string

2. Boston Food Establishment Inspections Dataset

The Boston Health Authority was extremely responsible and updated their table on a daily basis. The data for our purpose was downloaded from the platform on April 21, 2021. It was 249.9 MB large. It has 27 fields in total. We selected 7 of them for our data analysis. The field *property_id* indicated the property Id numbers for food establishments. The field *result* indicated the result of inspection.

Table 2: **boston_clean**

Column Name	**Data Type**
name	string
address	string
city	string
result	string
latitude	string
longitude	string

Analytic Stages

1. Data Ingestion

1.1 Yelp Dataset

For the Yelp Dataset in JSON format, we created an external table named **json_tab** and loaded the JSON file into a single column called *col1*. We chose *string* as the data type for latitude and longitude to make it more facile to use *SUBSTRING()* later. All the other tables adopted this data type idea for latitude and longitude.

There were 160, 585 rows in the table, **json_tab**, after data ingestion.

1.2 Boston Dataset

For the Boston Health Inspection dataset in CSV format, we also created an external table named **boston_raw** and loaded the CSV file into different columns, including *businessname*, *dbaname*, *legalowner*, *namelast*, *namefirst*, *licenseno*, *issdttm*, *expdttm*, *licstatus*, *licensecat*, *descript*, *result*, *resultdttm*, *violation*, *viollevel*, *violdesc*, *violdttm*, *violstatus*, *statusdate*, *comments*, *address*, *city*, *state*, *zip*, *property_id*, *latitude*, *longitude*. We also used Hive built-in function, *tblproperties("skip.header.line.count"="1")* to skip the headers.

There were 1, 851, 492 rows in the table, **boston_raw**, after data ingestion.

2. Data Cleaning and Profiling

2.0 Failed Attempts

We tried to clean the data using MapReduce; however, this approach failed us. It was significantly harder to parse JSON data using MapReduce. Furthermore, even for the Health Data, the data from many cities were either impossible to obtain or formatted in an inaccessible way. Originally, we would like to do a data analysis nationwide. The bottleneck for our project to scale to such extent was local governments. They collected data with different formats and metrics. It was imperative that they adopted a more unified approach in doing health inspections.

2.1 Yelp Dataset

For Yelp Dataset, we created a new table called **yelp_business** using the built-in function *get_json_object*. We extracted all the needed fields with appropriate data type listed in table 1 from temporary table **json_tab** and filtered out the redundant data.

There were 160, 585 rows in the table, **yelp_business**, after data cleaning.

2.2 Boston Dataset

For Boston Dataset, we extracted new data from table 2, **boston_clean**. During cleaning, we used *SUBSTRING()* function to filter out the punctuation marks for latitude and longitude. Afterwards, we transformed the data from **boston_clean** to create a new table, **boston_health**.

Remove Extra Punctuation Marks

before

latitude	longitude
“(42.278590000	-71.119440000)”

after

latitude	longitude
42.278590000	-71.119440000

During data profiling, we discovered that the field *result* had the following distinct entries, *HE_Fail*, *HE_Pass*, *HE_Filed*, *HE_FailExt*, *HE_Hearing*, *HE_NotReq*, *HE_TSOP*, *HE_OutBus*, *HE_Closure*, *HE_VolClos*, *Fail*, *HE_FAILNOR*, *HE_Misc*, *DTAERR*, *HE_Hold*, *PassViol*, *Closed.*

After consultation with the Boston Data Portal, it was shown that only *HE_Pass* was a successful health inspection, other entries were different forms of failed inspections. Many businesses had repeated inspections, therefore, it was indispensable to know the big picture. We created a new table **boston_health**, table 3. We used function *COUNT()* with *GROUP BY* clause from Hive to create new fields, *n_pass* and *n_fail*. The field *n_pass* and *n_fail* represented the number of *HE_Pass* and the count of the rest of entries. We also developed a health score called, *pass_rate,* to better measure a restaurant's health readiness. There are 1, 851, 492 rows in the table, **boston_clean**, after data cleaning, and there are 5, 518 rows in the table, **boston_health**, after data aggregation.

$$\text{health\_score} = \text{pass\_rate} = \frac{\text{HE\_Pass}}{\text{Total}} = \frac{\text{HE\_Pass}}{\text{HE\_Pass} + \text{Not\_HE\_Pass}}$$

Table 3: **boston_health**

Column Name	**Data Type**
name	string
address	string
city	string
latitude	string
longitude	string
n_pass	bigint
n_fail	bigint
pass_rate	double

3. Data Merging

Merging data was the créme de la créme of data analysis. After trial and error, a matching combination of multiple conditions were performed to produce the best outcome.

- first 2 characters in the name
- first 2 characters in the address
- 5 characters with 4 digits and 1 decimal point in latitude
- 6 characters with 4 digits, 1 decimal point and 1 minus sign in longitude

We created a table **boston_stats** from **boston_health** and **yelp_business** by matching their respective name, address, latitude and longitude with appropriate conditions described above.

In order to assess the criteria for matching. We developed metrics to measure our matching outcomes. We conducted other matching using different conditions and compared the relevant metrics to decide which conditions to use in the end.

The *business_id* from yelp dataset was unique. By counting the number of distinct *business_id*, we could find the *distinct_rate*, which was the ratio between distinct counts and total counts. By counting the number *business_id* that appeared only once, we could find the *unique_rate*, which was the ratio between unique counts and total counts. The reason that distinct counts and unique counts were different was because the distinct *business_id* could appear more than once. We tried to maximize the distinct matches using different number of characters

Visualizing Data Merging

yelp_business

Column Name	Data Type
business_id	string
name	string
address	string
city	string
stars	decimal(2,1)
review_count	int
is_open	int
latitude	string
longitude	string

boston_health

Column Name	Data Type
name	string
address	string
city	string
latitude	string
longitude	string
n_pass	bigint
n_fail	bigint
pass_rate	double

Criteria

metric 1

distinct_count	total	distinct_rate
2761	3520	0.784375

metric 2

unique_count	total	unique_rate
2260	3520	0.642045

in name, address, latitude and longitude. No other alternative combination had more distinct matches (2165) or *unique_count* (2260) than the current numbers we provided.

$$\text{distinct matches} = 2761 \times 0.784375 \approx 2165$$

During the data merging, we also used function "*LIKE %,%*" to filter out the comma signs in name and address to make it easier for further data analysis. From **boston_health** and **yelp_business**, we created a new table called **boston_stats**.

After data merging, we exported data from Hive to HDFS in a default file name *000000_0*. Afterwards, we used command line to rename the file as CSV file for our Apache Spark.

Table 4: **boston_stats**

Column Name	**Data Type**
business_id	string
name	string
address	string
city	string
stars	decimal(2,1)
review_count	int
is_open	int
n_pass	bigint
n_fail	bigint
pass_rate	double
latitude	string
longitude	string

4. Apache Spark Regressions

4.1 Linear Regression on health score and ratings

We conducted a linear regression with our created index of health score as the dependent variable and overall yelp ratings of individual food establishments as the independent variable. The results we obtained were as follows.

Regression 1: **health score and ratings**

Coefficients	**Intercept**	**r-squared**
0.02988437905866312	0.2543787098282755	0.0156945754049429

A consequent analysis of the goodness of fit gave the r-squared value recorded in the table above. This analysis showed that only a very small portion of restaurants' health scores are represented by their online ratings.

4.2 Linear Regression on health score and popularity

We performed a linear regression with our created index of health score as the dependent variable and number of reviews of individual food establishments as the independent variable. The results we obtained were as follows.

Regression 2: **health score and popularity**

Coefficients	**Intercept**	**r-squared**
- 0.007075239993046872	5.386810461496709	0.00346864406594638

Note that the negative coefficient might be an indication of the expectation disconfirmation bias discussed earlier. A consequent analysis of the goodness of fit gave the r-squared value recorded in the table above. This analysis showed that a even smaller portion of restaurants' health scores are represented by their online popularity.

4.3 Logistic Regression on restaurant openness and health score

We ran a logistic regression with the Boolean algebra of restaurant openness as the dependent variable and our designated health score as the independent variable. The results were the followings.

The interpretation of coefficients and intercepts should come after interpreting the AUC, area under the ROC curve (receiver operating characteristic curve). AUC provided an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC was as the probability that the model ranked a random positive example more highly than a random negative example. AUC ranged in value from 0 to 1. A model whose predictions were 100% wrong had an AUC of 0.0; one whose predictions were 100% correct had an AUC of 1.0. The AUC value of 0.5004 meant that the the health score, pass_rate could not be used as the basis for predicating the restaurant openness in the future. It was no good than predicative capability of a coin toss.

Regression 3: **restaurant openness and health score**

Coefficients	**Intercept**	**AUC**
0.05366519262928552	0.5653731412215562	0.5004848063008069

Conclusion

Exploring the representativeness of a restaurant's online reviews on a crowd-sourced platform, our regression analysis conducted on an original integration

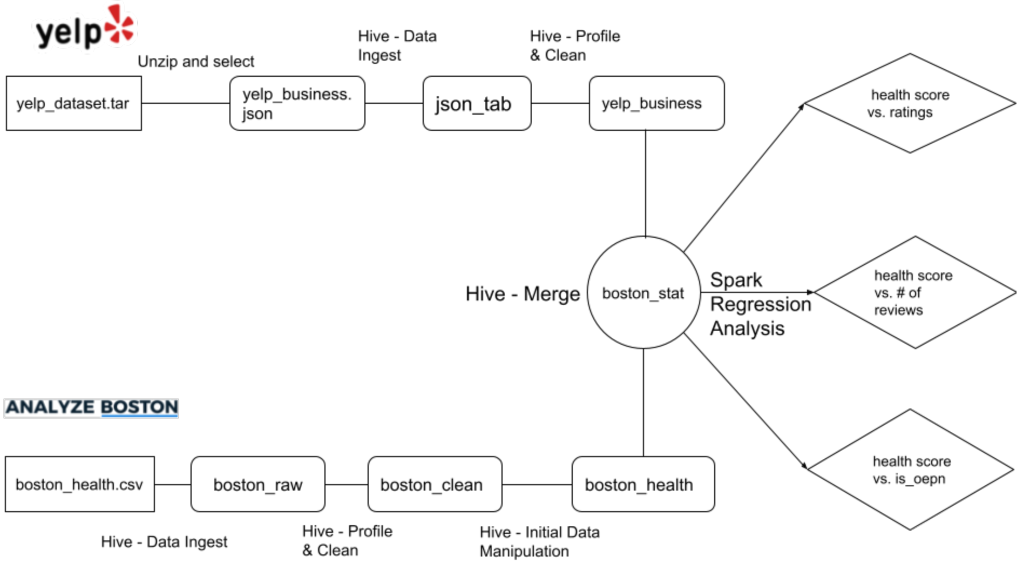

Figure 2: Data Analytics Summary

of data sources revealed a surprisingly weak correlation between a restaurant's health condition and its performance on the popular online food review platform. The contribution of this study is two-fold. In methodology, our proposed method of merging data sources with wildly different schemas was successfully implemented in our study and could be used to integrate more datasets related to restaurant quality. In the analytical sense, our finding of the inability of a restaurant's current online profile to represent health scores accurately provided insights for the end-users of the online review platform. This significant finding should propel a positive feedback loop by encouraging users to leave reviews on sanitary conditions and a lookout for these signals on the platform. Doing so would further improve the platforms' quality.

Directions for future research can be categorized into horizontal and vertical aspects. Horizontally, more data sources can be identified and incorporated with the methodology we have developed — forming a more comprehensive basis for analysis. Vertically, future work can explore information hidden in the linguistic content of individual posts related to food safety. A possible starting point could be to explore an inexpensive method of conducting linguistic identification with a vast database.

References

[1] Siering, M: Leveraging online review platforms to support public policy: Predicting restaurant health violations based on online reviews
Decision Support Systems, Volume 143,
2021, 113474, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2020.113474.
(https://www.sciencedirect.com/science/article/pii/S0167923620302293)

[2] Sadilek, A., Brennan, S., Kautz, H., & Silenzio, V.:
nEmesis: Which Restaurants Should You Avoid Today?.
Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 1(1), 2013.
Retrieved from https://ojs.aaai.org/index.php/HCOMP/article/view/13069

[3] Yelp Dataset
https://www.yelp.com/dataset/download

[4] Boston Food Establishment Inspections
https://data.boston.gov/dataset/food-establishment-inspections