2021

Generating Insights from Business Review Data

SCRAPING AND DATA ANALYSIS

PRABAL CHHATKULI

FACULTY SPONSOR: PROFESSOR SCOTT FREES

FACULTY READER: PROFESSOR BENJAMIN FINE

Overview:

The widespread use of internet and massive mount of accumulated data present a great opportunity for businesses to analyze these data to improve their services and better cater to the need of the customers. Most of these data, including review data for businesses, from some of most visited web services play a vital role in the longevity and success of the business.

This project emphasizes that such data may not be readily available for analysis and extraction methods may be used by businesses to extract these data from their sources. Scraping enough data for analytics is vital to generate concrete evidence for generated insights. Similarly, an automated pipeline for data-scraping can greatly reduce manual efforts as well.

A variety of insights can be generated about a business from its customer reviews. Some of the insights that has been included in this project are analyzing customer sentiments and getting satisfaction of customers on a product or services. Examples of Analysis like these can be used by businesses to give better service to their customers. Similarly, advanced implementation of review data may include using review data for robust recommendation systems. Furthermore, proper analysis of user reviews can give a business an edge over their competing counterparts. Hence, we can see clear evidence that proper analysis of review data clearly ties to the success of a business in the future.

This project summarizes a simple structure of data collection, filtering, analysis, and machine learning to give simple but effective insights on business and their services.

Tech Stack used:

Language/Platform used:

- Python, Jupyter Notebook

Webpages for scraping/Datasets:

- Google Reviews, Businesses' websites, Yelp Reviews

APIs:

- Google Places API
  - Extract business data based on type of businesses/zip code.

Frameworks and Libraries:

- Scrapy
  - Spider Crawler for business' website data.
- Selenium
  - Scraping Algorithm for review data from Google Maps.
- Scikit-learn:
  - Machine Learning for sentiment Analysis.
- Pandas/NumPy/Matplotlib:
  - Dataset management/ Exploratory data analysis

Methods:

Complete Project GitHub: https://github.com/prabalchhatkuli/Review-Analysis

1) Data Scraping
    - The Data scraping section works on devising an algorithm for data extraction.
    - Initially, Google Places API was used to get a list of certain type of businesses based on the zip/area code and/or the type of business. This returns useful information related to the rating, contact information, URL of businesses' website, hours etc.
        - The Google Maps URL retrieved from this process is passed to the function built using Selenium. This function extracts individual reviews and rating information as well as the customer's names from the respective listing in Google Maps/ Google Reviews.
        - After this process, the businesses' website received from the Places API is crawled to extract meta tags, description, and keywords as well as email contacts for the business.

    Limitations:

    - This data flow takes a lot of time and compute power to extract enough data for multiple businesses. So, it was decided that for the later processes, we will move forward with sample review data provided by Yelp reviews.

2) Data Analysis
    - The initial phase of the data Analysis features an Exploratory Data Analysis (EDA), which is done to get a complete look of the data before analysis.
    - Since the Dataset is tens of gigabytes, it takes enormous amount of RAM capacity and CPU processing to analyze the data. One solution to this problem is filtering to exclude missing data, outliers, and only focus on a certain category of data. For, the Yelp Reviews Dataset, most of the data was around Austin Texas. This are had high number of businesses as well as high reviews per business.
    - The data analysis process includes reducing the dataset to include only the data from Austin, Texas. This was the city with largest number of business with corresponding review.
    - Upon greatly reducing the dataset, it was used to construct a model was constructed to see how a certain business was doing in different year, month, and word clouds were constructed based on the positive and negative reviews.
    - Similarly, the businesses webpage was crawled using the crawler described in (1). The extracted meta-tags were observed, and a word cloud was constructed to find different properties of the businesses.
    - Then, a process to see customer satisfaction on a production was made to understand the sentiments of customers on a certain subject over a period.

- Finally, a sentiment analysis model was constructed using the review text and rating. A Random forest analysis was done to explore try if there was a way to connect the ratings to the sentiment of the reviews.
    - The Random Forest method in this project uses a vectorizer based on the feature of a text. The train data (review text) is vectorized and turned into an array and trained on a Random Forest Classifier. Random Forests operate by constructing a multitude of decision trees that predict a class prediction and the mean of these become the actual prediction. In this project 50 trees are being used.

Understanding the Results:

1) Two Datasets were loaded for Analysis. The column structure of the dataset are as follows:
   Business Dataset:

   business_id, name, address, city, state, postal_code, latitude, longitude, stars, review_count, is_open, attributes, categories, hours

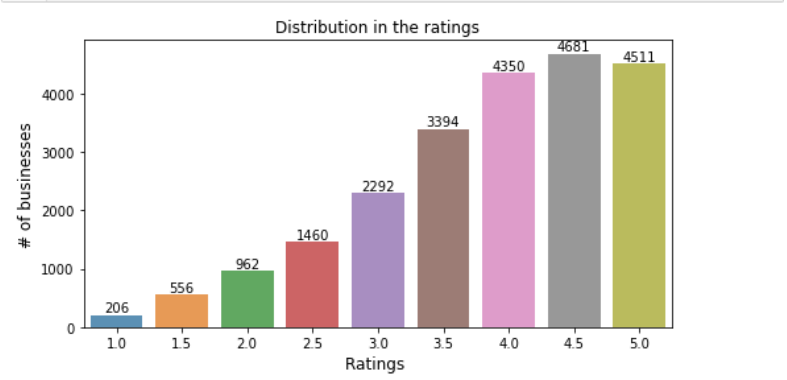
   Reviews Dataset:

   review_id, user_id, business_id, stars, useful, funny, cool, text, date

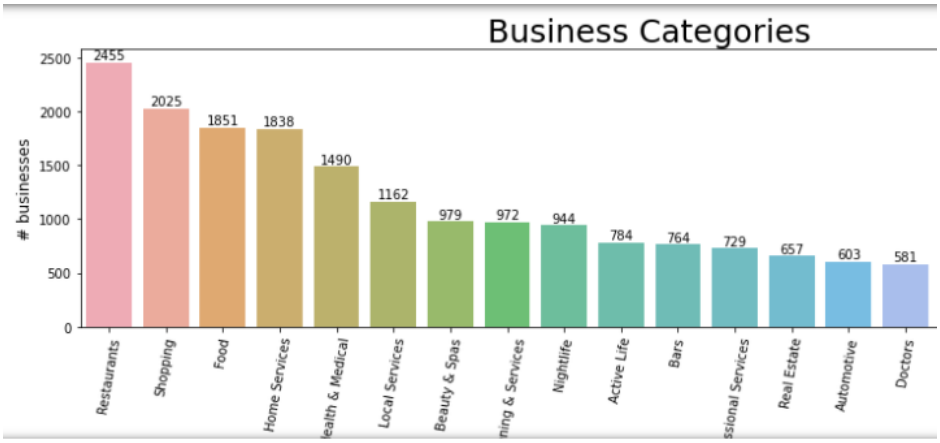
2) In the EDA, it the following information was found about the number of businesses in the dataset:

	city	count
8	Austin	22416
2	Portland	18203
5	Vancouver	13330
4	Atlanta	12612
6	Orlando	10637

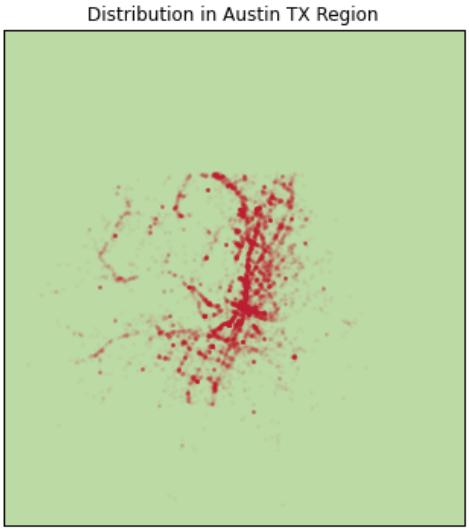
- Other cities like NYC had only a couple businesses, hence, we decided to go with Austin.



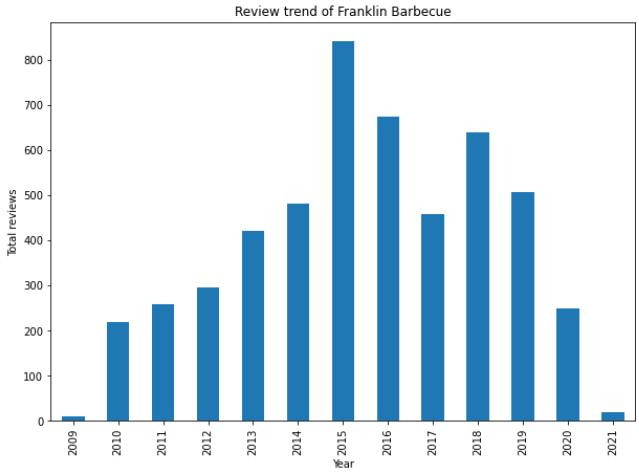
- We can see from above how most businesses have high average reviews.



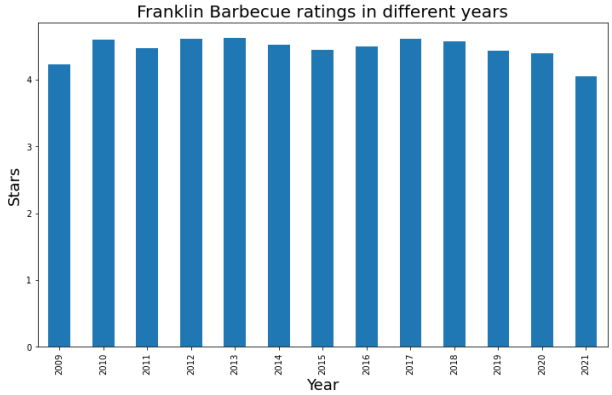
- Similarly, we see the distribution of categories among businesses.



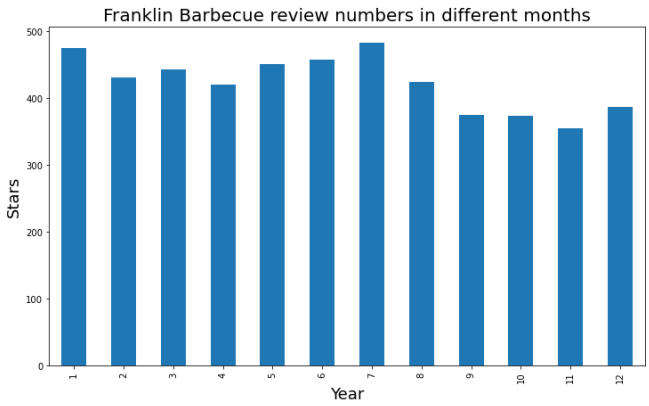
3) After a simple EDA and filtering procedure the following results were obtained from the analysis:
   a) Yearly review trend for one of the top reviewed restaurants



b) Average Rating differences in different years

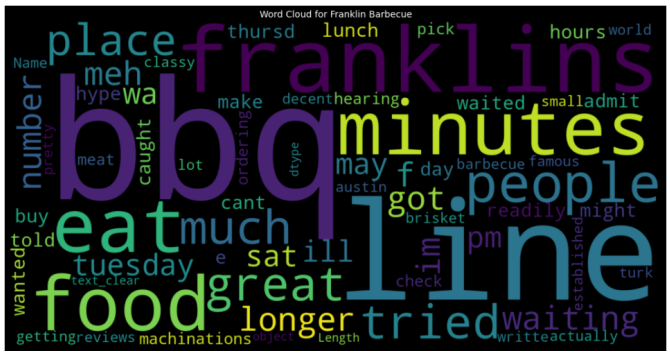


c) Differences in Review based based on months.



d) Positive word Clouds

e) Negative word Clouds

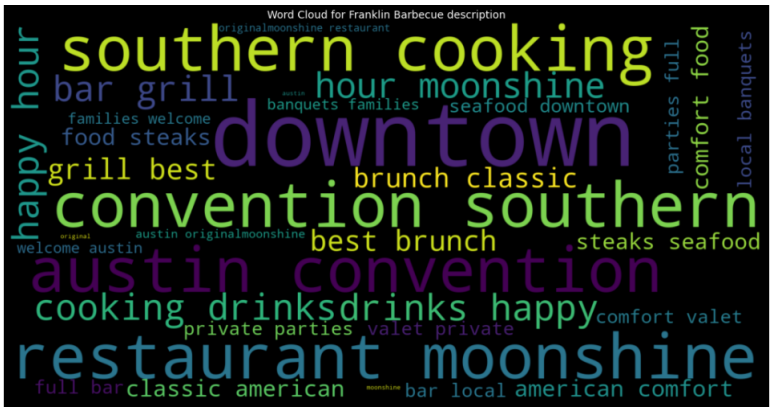


4) Text Analysis

a) First a set of description was taken by the crawler, it returned data in the following format:

```
{'description': "Moons
Waterloo Compound.Moon
c Waterloo Compound.Mo
ric Waterloo Compound.
toric Waterloo Compoun
istoric Waterloo Compo
historic Waterloo Comp
s' historic Waterloo C
 'keywords': 'moonshin
oonshine bar & grill,
te parties, full bar,
ustin, convention, sou
steaks & seafood, down
iginalmoonshine, resta
bar & grill, best brun
s, full bar, local, ba
nvention, southern coo
seafood, downtown, Sou
```

b) Running a word Cloud on the keyword and description for Moonshine Patio Bar & Grill gave the following word Cloud:



c) The results were observed and for this sample restaurant keywords were created and separated into four categories: bar, service, food, place.

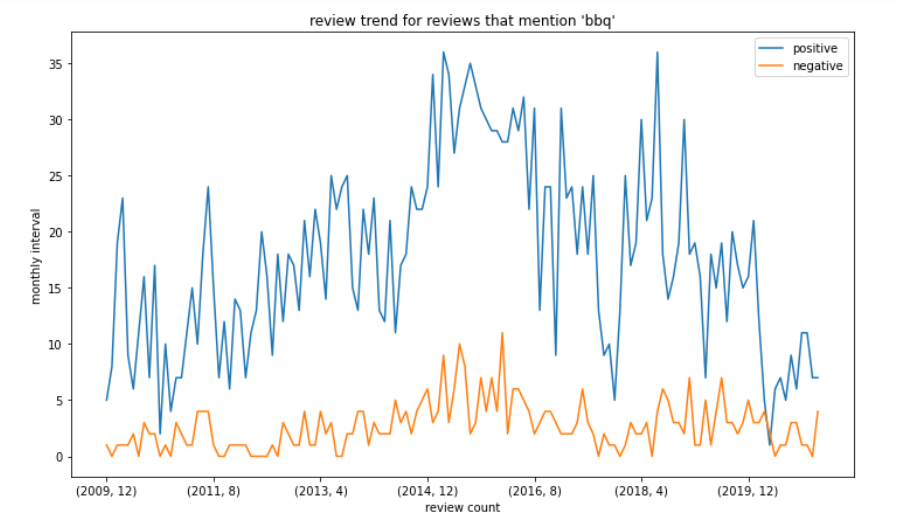
d) Each review for Moonshine Patio Bar & Grill was categorized into certain category based on the corpus of the keywords. The following result was received for this business:

```
False    3222
True     1849
Name: bar, dtype: int64

True     3905
False    1166
Name: service, dtype: int64

True     4659
False     412
Name: food, dtype: int64

True     3404
False    1667
Name: place, dtype: int64
```

e) Then food category was chosen, and the rating of each reviews was calculated for each item inside this category, and the following result was obtained:

```
{'brisket': {'count': 3490, 'positive': 3201, 'negative': 290},
 'delicious': {'count': 718, 'positive': 692, 'negative': 27},
 'bbq': {'count': 2728, 'positive': 2371, 'negative': 358},
 'food': {'count': 1747, 'positive': 1512, 'negative': 236},
 'chef': {'count': 43, 'positive': 38, 'negative': 6},
 'lunch': {'count': 245, 'positive': 216, 'negative': 30},
 'chicken': {'count': 30, 'positive': 27, 'negative': 4},
 'southern': {'count': 24, 'positive': 22, 'negative': 3},
 'sandwiches': {'count': 90, 'positive': 84, 'negative': 7},
```

f) Then a general trend analysis was done to see how the positive or negative reviews are increasing or decreasing for ‘bbq’:



5) Sentiment Analysis Model

- Finally, a sentiment analysis model was created to see if we could predict whether a rating was positive or negative based on the Random Forest Approach. The following output were received:

```
[[ 8630  3504]
 [ 1044 38656]]
              precision    recall  f1-score   support

          -1       0.89      0.71      0.79     12134
           1       0.92      0.97      0.94     39700

    accuracy                           0.91     51834
   macro avg       0.90      0.84      0.87     51834
weighted avg       0.91      0.91      0.91     51834

0.9122583632364857
```

1)

- Limitations for sentiment model:
  - We can see how even for positive word clouds there are some words that generally carry a negative connotation. This is a huge limitation for our sentiment analysis model that uses rating to determine the emotion of the text.
- Improvements for sentiment model:
  - For each review our dataset has funny, useful, and cool ratings which can be incorporated into the model to improve the model based on our method of construction.

Discussion:

The analysis presented plenty of results that could be used to generate insights. From the result in 3 (a), we can see how the number of reviews has been decreasing in the recent years, especially in 2020 and 2021. The business was doing much better and getting huge numbers of

customer reviews. If there exists a correlation between number of reviews and the number of customers, we can assume that certain changes after 2015 might have reduced business activities.

In 3 (b), a mapping of average rating per year is done for this business. There is a slope of decrease in the average rating starting from 2016. This means that the business has been receiving some less rating as well. This caused the average rating to fall. Even the slightest of the slope could mean huge changes in the customer satisfaction level.

In 3 (c), we can see how the business gets less reviews in the fall when compared to the spring/summertime. There might be several factors affecting this which can be generalized by keeping in mind the type of business, location, and service updates. Similarly, in 3 (d) and 3 (e) we have the word clouds that are basically the words that appear more frequently in reviews with the more positive and more negative, respectively. The word clouds are excellent for inspection about what a general customer might say about the business whether positively or negatively. One example is finding similarities between negative reviews. This can be used by businesses to improve their services as many people are talking negatively about the same product or services. Another example would be to search for a certain product or service what a customer does not like about it. We will discuss more about this further on.

Starting from 4 (a) we focus on the textual data in the reviews dataset. Initially, the webpage of the business is crawled using the crawling script described in the Methods section. When we get more information about the business a word cloud is constructed. This gives more idea about the business as a collection of keywords. From the data from the crawler and the word cloud it is easier to predict what the overall purpose of the business is. One more important concept about word clouds implementation on business is that it is sometimes easier to discover new Search Engine Optimization (SEO) terms and keywords that the business can use on their websites to gain more online customers to their webpage.

In 4 (c) we created different categories that might be related to the business. Here, we chose four categories: bar, service, food, place. These were chosen as flags for each review. With categorical flags it can be easier to track what subject each review explains. This can be related from finding out what a customer likes about a certain category of a business to getting information about what can be improved on a certain item. Furthermore, more flags can be set for items inside these categories and positive and negative feedback can be taken for each item. For example, in 4 (f), we get a graph explaining the number of positive and negative feedback on the ‘bbq’ products of the restaurant. And from the graph, it looks like customers are quite dissatisfied in recent times as the number of negative reviews is increasing while the number of positive reviews related to ‘bbq’ is decreasing. This method is a very simple method to get insights about some overlooked aspect of a business as well. For example, a business might be proud of the customer service, however, one of the employees may have been misbehaving with the customers. This can be easily found if the business investigates the service flag set for the restaurant. Similarly, they can also identify services that could technically increase the number of customers as well.

Finally, in 5, a Sentiment Analysis was done. Sentiment Analysis in this case was done to examine if the sentiment(negative/positive) could be predicted using just the text of the review. The result was very promising as there was an accuracy of 91%. Sentiment analysis models can be used by businesses to validate ratings and evaluate reviews to better understand how customers feel about their brands and products. In addition, it can be used to categorize certain text about the business to determine. Similarly, it can be used to derive information about the business from different textual sources like news articles, provide result in real-time based on user feedback, Influencer content, etc. Finally, sentiment analysis can also be used to evaluate competitor's business as well.

Conclusion:

Data Analysis is a pipeline of methods from data production, collection, preparation, and evaluation. There are several methods that must be carefully processed to obtain the simplest datasets. Every Dataset must be understood properly to proceed with building a robust model using its information. Even after this step, there are infinite number of insights that can be generated from such data. This project finds insights from business reviews that can be used by the business to create policies or change its current practices so that they become more attractive to customers.