Yelp Reviews

A Comprehensive Analysis of Review Trends

Dominic Deckys
University of Colorado
Boulder, Colorado
dominic.deckys@colorado.edu

Michael Gigiolio
University of Colorado
Boulder, Colorado
michael.gigiolio@colorado.edu

Dante Pasionek
University of Colorado
Boulder, Colorado
dante.pasionek@colorado.edu

1 PROBLEM STATEMENT & MOTIVATION

This paper and the research conducted seek to investigate factors that contribute to the performance and overall perception of a business. Through analysis of reviews on the popular restaurant critique website Yelp®, and the application of advanced data mining techniques, this work intends to reveal trends that define a restaurant in the public eye. This research draws motivation from businesses serving the public and their need to appeal to their customer base as effectively as possible. By determining what features are correlated with an establishment being viewed positively, this paper aims to provide a useful service to companies seeking to improve their operation.

2 LITERATURE SURVEY

While there currently has not been any work conducted that is similar to the research detailed here, many other studies have been performed using Yelp data. Notably, the company Yelp itself regularly hosts what is known as *Yelp Dataset Challenge* [1]. In this challenge, students are provided data from Yelp in a competition to help practice their data mining techniques. Other research includes:

(1) *What Yelp Fake Review Filter Might Be Doing?*[2]
An analysis on how Yelp determines fake reviews placed to artificially alter the standing of a business. An investigative report on the algorithmic analysis and determination of what qualities Yelp uses to determine the potential fraudulent nature of the review.

(2) *Integrating web-based data mining tools with business models for knowledge management*[3]
An analysis of how conducting data mining of a business' internal resources and information can allow companies to review and enhance the current business model and performance.

(3) *Yelp's Top 100 Places to Eat in Canada for 2018*[4]
Yelp conducted their own research to create a list of the 100 best rated restaurants in Canada based on both rating and volume. This study focused specifically on Canadian residents and was conducted only with reviews made by local people to better tailor the results to permanent Canadian residents.

(4) *What Was The Hottest New Restaurant in Your State in 2017?*[5]
Much like the study outlined in the previous article, Yelp's data team performed data analytics upon their American reviews in order to provide a list of the most popular restaurants started in 2017 by state.

(5) *Local Economic Outlook*[6]
Yelp conducted analysis of it's review data and produced a list of 50 cities, 50 neighborhoods, and 10 business categories that were found to be the most ideal places and business models for the success of small businesses. In addition, they determined the economic opportunity in these areas so as to aid current or prospective restaurant startup owners in making large scale choices that will most likely lead to their success. This analysis was performed using repetitive determinations of which small businesses were most likely to remain open in the subsequent three months over the course of two years.

This previous research will ideally aid in the analysis of the effects of location, hours of business and other factors on the overall impression of an establishment. Additionally, these studies will provide insight concerning specific methods of analysis on Yelp data in particular.

3 DATASET

The dataset used in this analysis was obtained through www.kaggle.com and is aptly named *Yelp Dataset*[7]. This dataset contains approximately 5,200,000 user reviews, with information on more than 174,000 businesses over a total of 11 metropolitan areas. This research will focus primarily on attributes such as `stars` for review ratings as well as `latitude`, `longitude` and `postal_code` to determine location of a business. Besides the attributes mentioned above, it's important to know what other attributes this research will be using.

In addition to the other tables, the data set contains a `business_hours` table, consisting of the hours of operation stored as a string of 24-hour time. Each day of the week is represented as an attribute, with each unique business id as a primary key. Establishments that are do not operate on certain days or are no longer in business have their hours represented as `None`.

The largest table in this dataset is `yelp_reviews` which contains information about the individual reviews. Specifically, the table

[1] https://www.yelp.com/dataset/challenge
[2] https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6006/6380
[3] https://www.sciencedirect.com/science/article/pii/S0167923602000982
[4] https://www.yelpblog.com/2018/02/109791
[5] https://www.yelpblog.com/2017/12/best-new-restaurant-every-state-2017
[6] https://www.yelpblog.com/2017/10/local-economic-outlook
[7] https://www.kaggle.com/yelp-dataset/yelp-dataset

contains the business reviewed, the number of stars given, the date the business was reviewed and comments made. This research will necessarily combine numerous attributes drawn from many of the tables provided in order to best identify trends.

4 PROPOSED WORK

Initially, the dataset will need to be cleaned and preprocessed. In order to accomplish this, numerous modifications are required to be made so that the dataset can be most effectively and efficiently worked with. These methods include:

- Several of the attributes, such as the name and address of the businesses, contain information represented as a string surrounded by quotation marks. In order to properly process and represent this data, the quotation marks will need to be removed so that only the relevant data is left.

- The `postal code` data column will need to be removed. As the data set includes reviews from several different countries, the postal codes often take different formats. For example, a postal code such as `L3P 1X3` represents a location in Canada, whereas `85286` is a postal code in the United States. This will require any algorithm processing the postal code to be able to distinguish between multiple different styles to properly determine location, and as the dataset already contains the latitude and longitude coordinates of the business, there is no need to expend so much effort to obtain less exact information.

- The `neighborhood` attribute will likely need to be removed. This results from the fact that this column has a large enough amount of missing values that render it functionally useless for processing. In addition, this information, even if it were to be completely present, does not provide the same utility as the other location services due to its inherent ambiguity.

- It would be useful to add a column to the `business_hours` table to represent the total number of hours open per week as well as average per day. This information, now easily accessible, would facilitate the comparison of other attributes to open hours.

- Creating a normalized score of a business by weighing the average star rating, the number of reviews left and number of check-ins. This is beneficial as a business with a high rating and very few or a single review may not have qualities allowing it to compare to other highly rated businesses. Conversely, a business with a low score and only a few reviews may be improperly represented as worse than its actual merit.

5 EVALUATION METHODS

This research is conducted primarily using Python and its peripherals. This includes, though is not currently limited to, Pandas, NumPy, and SciPy. Pandas will allow the CSV files to be converted into a dataframe, permitting easy operation of NumPy and SciPy upon the dataset. By using mathematical tools such as Ordinary Least-Squares Regression (OLS), confidence intervals and p-values, this analysis intends to find meaningful patterns regarding the interaction between business hours, ratings, and location. Specifically, the use of confidence intervals will aid in the determination of possible correlations between the aforementioned attributes. In turn, multiple regression techniques will allow for more accurate predictions of a business' performance in regards to certain criteria.

Methods similar to those executed in the prior research described in the literature survey will be implemented. For example, weight of the overall rating of a restaurant will be determined not only by the number of average stars, but also by the number of reviews, so as to not skew the data due to a restaurant having a insufficient sample of reviews. This results from the proposition that the comparison of average ratings for specific types of businesses, compounded with the various attributes that contribute to a higher or lower overall rating, will assist establishments in improving their consumer appeal.

6 TOOLS

A variety of tools will be used in this analysis, the programming tools to be used are detailed below.

- Python 3: A high level programming language

- Pandas: An open source Python library used for data manipulation and analysis. This library serves to format the data so that the other libraries used can operate upon it.

- NumPy: An open source Python library used for numerical computation on datasets and other forms of information. This will allow the performance of operations on the millions of values contained within the review data.

- SciPy: An open source Python library used for technical computing. This will allow the performance of complex computations such as regression and the calculation of confidence intervals.

In addition to these programming tools and peripherals, mathematical tools will also be required for this analysis to determine correlations and trends.

- Regression: This research anticipates to use a variety of regression techniques to aid in the modeling of trends found in the data, as well as to assist in the accurate prediction of future results.

- Confidence Intervals: Confidence intervals will aid in the determination of the statistical significance of any patterns that are located.

7 MILESTONES

The current main milestones of this project are detailed below, and will be updated with each project sprint.

(1) Remove the neighborhood attribute

(2) Distinguish between United States and Canadian postal codes.

(3) Clear out NaN, Null and Missing values.

(4) Parse open business hours to a single unit to help later computations.

(5) Create normalization scores for each review to use in overall analysis.

8 PEER REVIEW SESSION

During the peer review session, several ideas were provided that apply very strongly to this particular research. Principally, the point was made that, regardless of the topic at hand, there is always prior research done. This was specifically taken into account as evidenced by the numerous other works considered in the "Literature Survey" section. In addition to this, remarks were made about considering the business and real world applications of the research. This was also considered and, as a result, the research that will be conducted is tailored directly to the utility it can provide businesses.