Exploring the Yelp Dataset

Correlating Yelp reviews with Economic and weather trends

Carter Reid  
 Applied Computer Science University of Colorado Boulder  
 Boulder, CO U.S.A.  
 reidcc@colorado.edu

Maxwell Reynolds  
 Applied Computer Science  
 University of Colorado Boulder  
Boulder, CO U.S.A  
mare3521@colorado.edu

Surya Jatavallabhula  
 Applied Computer Science  
 University of Colorado Boulder  
Boulder, CO U.S.A  
 suja3865@ colorado.edu

PROBLEM STATEMENT

Yelp is a business directory service and crowd-sourced review forum. That is, the business revolves around the connections made between the consumers who read and write reviews and the local businesses that they describe [1]. Since the company’s founding in 2004, it has grown to include 4.6 million active claimed business locations and 192 million cumulative reviews for those 4.6 million business locations [1].

Despite the simplicity of the service offered, there are many attributes tracked and related to each other in the dataset allowing a vast opportunity for data mining.

Yelp.com has an extensive dataset gathered from their online review services, and Yelp has made this dataset available to students in the form of a contest aimed at encouraging students to explore their data and discover novel trends and relations among their reviewers and businesses. This contest has a cash incentive and is on its thirteenth iteration, ending in December 2019.

Some potential questions we would like to answer are:

* Can external factors such as local weather or global economy influence review ratings or sentiment?
* Can characterize potential pitfalls/areas of improvement of a restaurant based on its reviews.
* Can we characterize the most important aspect of a restaurant to a given regional population (i.e. what do reviewers in Austin, TX seem to value the most?).
* Are “funny” reviews considered more or less helpful? What are some characteristics of “funny” reviewers?

**LITERATURE SURVEY**

As this particular contest has had many previous iterations, there is abundant work performed on similar Yelp datasets. These previous works explore many aspects of yelp from determining user’s influence [3], finding local experts [9] examining an apparent warm-start bias for reviews of new business establishments [4], detecting deceptive and or fake yelp reviews [5], predicting whether a restaurant would succeed or close [2], and associating healthcare reviews with cervices offered [6].

The most recent contest winners have a public github linked from the contest landing page [here](https://github.com/Yelp/dataset-examples) [8]. The showcased winners created a positivity estimator based on review text and key words and created an automatic review generator that generates a review from an initial small text such as “They have the best…” using a Markov chain technique.

With the wealth of information within the dataset, it appears most researchers have searched for relationships wholly within the Yelp dataset, and few have drawn in additional information to correlate with information in the yelp dataset.

**PROPOSED WORK**

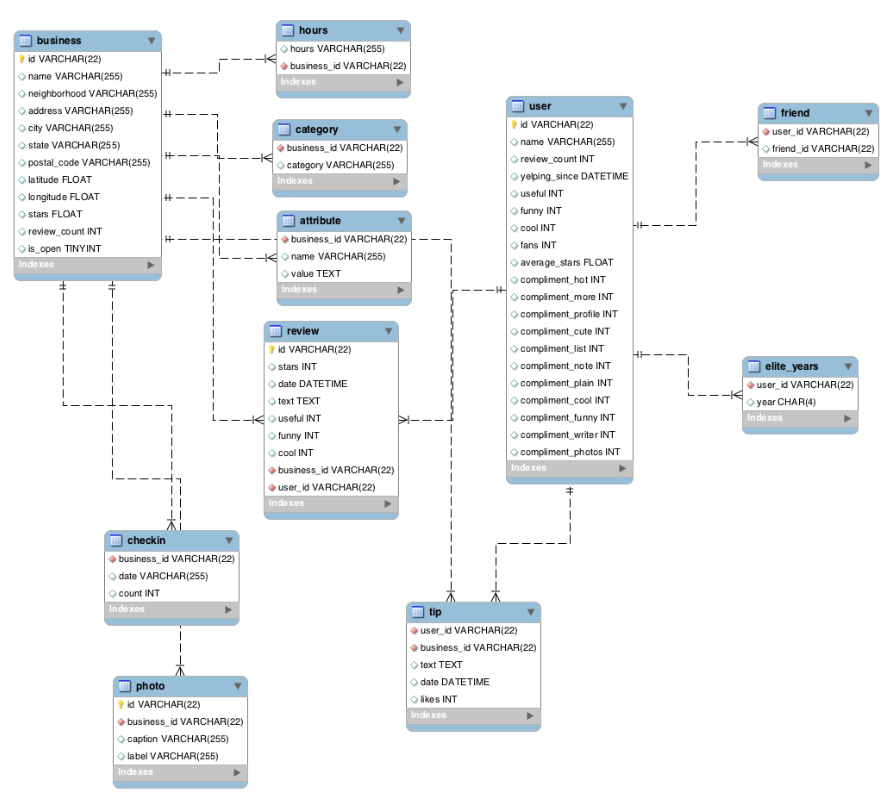
One of our initial tasks will be to design an SQL database to store the data and insert all our data. We may need to clean aspects of the data such as outliers and normalize the data using Standard Deviation method and resolve if there are any formatting discrepancies. Once the data is relatively cleaned, we should perform an exploratory data analysis to identify trends and potentially form hypotheses regarding what we would like to investigate. After this step, we will be able to start mining the data, building models, and testing our hypotheses.

The dataset itself contains other information as well which could reveal other interesting relationships and correlations. One aspect of the dataset which seems less explored than relating review scores with text, is comparing reviews with external data sources, such as economic data, or specific calendar dates like religious holidays. The yelp dataset provides dates which we plan to normalize and relate to public and religious holidays, economic data, and weather data. We also would like to explore relationships between user attributes and the attributes associated with the reviews they write (useful, funny, cool) to predict their review score of an establishment. We believe that predicting a review score based on other attributes could help direct establishments optimize their hours of operation and focus their marketing on specific demographics to increase average reviews and maximize marketing effectiveness. We might also explore what aspects of restaurants that reviewers care the most about across different regions.

**Data Set**

The available dataset is large. It is 8.69 gigabytes of business, user, and review data with another 7.67 gigabytes of business and customer photos. It is available form yelp directly with a valid school email address. Link [here](https://www.yelp.com/dataset/challenge).

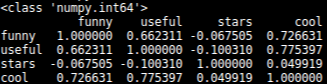
The data is packaged in the form of six json files, with multiple relations existing between the tables.



**Figure 1: A relational database model of the yelp dataset constructed in MySQL workbench. [7]**

**Evaluation Methods**

Statistical analysis. Looking at correlation coefficients and other statistical metrics for the dataset in the beginning will help guide our progress though the dataset. Certain calculations will take time and rely on statistical constants such as mean, median, standard deviation, variance, correlation coefficients etc. Generating those constants now may save time and computation in the future.



**Figure 2: A correlation matrix of review attributes**

Bayesian analysis of certain attributes associated with user and reviews will be our method of predicting review score based on user attributes and review attributes such as user history and review likes. A Bayesian prediction algorithm seems appropriate for a large complex dataset allowing for unanticipated patterns to emerge as well as specific target patterns.

Additional data sources will be related to the yelp dataset through SQL, using a join on economic dataset date and yelp review date. Date format may need to be normalized between the two datasets in order to create a join, and weekends may need to be treated specially as the stock market is not open over the weekend and after-hours trading has a different volatility than standard trading.

**Tools**

The current proposed tools for this exploration are Python, Github, SQL (sqlite3), and Tableau. Python will be used initially to load portions of the database into pandas dataframes due to ease of use. The dataset itself is too large to store the entire database and therefore a more scalable solution will need to be implemented, as solutions to the size constraint exist, they still leave a lack of optimization and are slow.

An SQL database is slated to be the scalable solution to the dataframe size and runtime constraint. The service for our relational database has yet to be determined but sqlite appears to be a reasonable choice unless we host an interactive web-visualization, in which case Google Cloud may be used. This will also allow other data sources to be incorporated into the analysis using a join on date. This does require the date attribute to be normalized to the format used in the yelp dataset (YYYY-MM-DD).

Tableau will potentially be used for visualization of the data and trends discovered.

**Milestones**

* November 1st We would like to have all of our data organized into a relational database for convenient SQL query.
* November 5th: We would like to have our data cleaned and preprocessed. This includes settling on a date format and formatting any additional data we wish to include.
* November 8th: We would like to have added any additional economic and weather data to our relational database
* November 15th: We would like to have our initial exploration phase completed and have a clear understanding of which tools we are using and how to use them.
* November 22nd: We would like to be finalizing our methods and reviewing our results.
* November 29th: We would like to have completed work and resolved any lingering issues so we can focus on the final writeup.
* December 13th: Project completed.

ACKNOWLEDGMENTS

Yelp.com for providing the data and incentive

REFERENCES

[1] Yelp.com. 2019. Investor Relations. Retrieved from <https://www.yelp-ir.com/overview/default.aspx>

[2] Xiaopeng L. Jiaming Q. Yongxing j. Yanbing Z. *Should I invest it?: Predicting Guture Success of Yelp Restaurants.* PEARC 2018 (July 22) DOI: [10.1145/3219104.3229287](https://doi.org/10.1145/3219104.3229287)

[3] Andres B. Agrima J. Bharat B. *Measuring user's influence in the Yelp recommender system.* PSU Research Review, Vol. 1 No. 2, pp. 91-104. <https://doi.org/10.1108/PRR-02-2017-0016>

[4] Michalis Potamias. *The warm-start bias of yelp.* Cornell University. arXiv:1202.5713 [cs.SI]

[5] Mahmudur R. Bogdan C. Jaime B. Duen H. 2015 *To catch a fake: curbing deceptive yelp ratings and venues.* The ASA Data Science Journal. Vol. 8 Issue 3. <https://doi.org/10.1002/sam.11264>

[6] Benjamin R. Rachel W. et. al. 2006. *Yelp reviews of hospital care can supplement and inform traditional surveys of the patient experience of care.* Health affairs Vol. 35, No. 4. <https://doi.org/10.1377/hlthaff.2015.1030>

[7] <https://github.com/2gotgrossman/yelp-dataset-challenge>

[8] https//github.com/Yelp/dataset-examples

[9] Jindal Tanvi. 2015. *Finding local Experts from Yelp database.* Masters Thesis. University of Illinois at Urbana-Champaign, Urbana-Champaign. U.S.A.Conference Name:ACM Woodstock conferenceConference Short Name:WOODSTOCK’18Conference Location:El Paso, Texas USAISBN:978-1-4503-0000-0/18/06Year:2018Date:JuneCopyright Year:2018Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00