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](mailto:reidcc@colorado.edu)

Garrett Lubin

Applied Computer

Science University of Colorado Boulder

Boulder, Co U.S.A.

[galu7424@colorado.edu](mailto:galu7424@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**

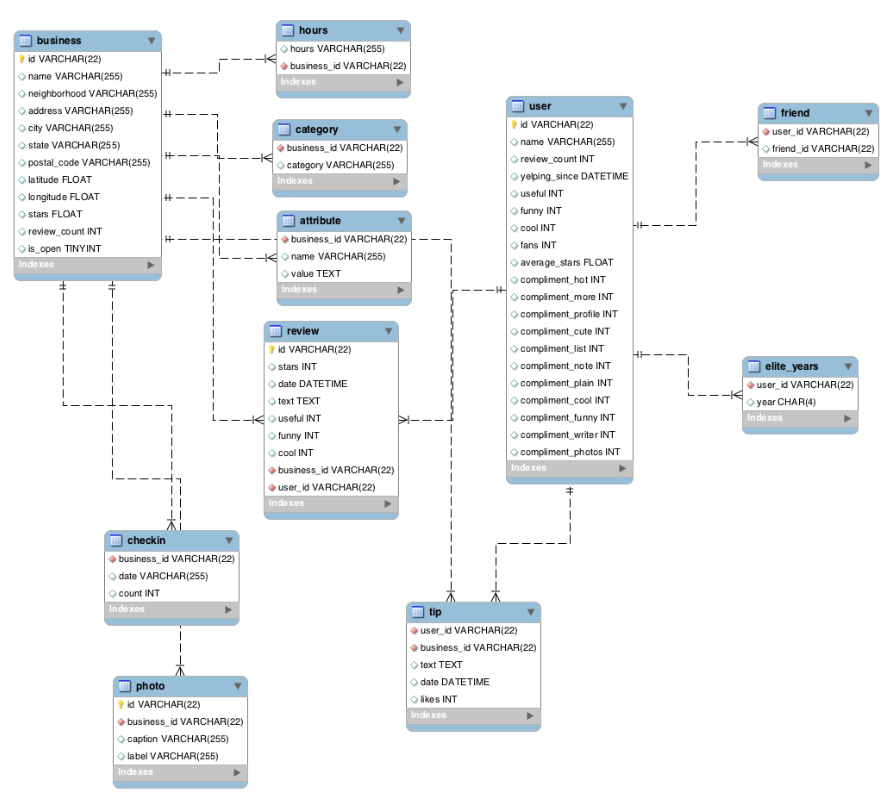
One of our initial tasks will be to design an SQL database to store the data and insert all our data. We may need to clean aspects of the data such as outliers and normalize the data using Standard Deviation method and resolve if there are any formatting discrepancies. Once the data is relatively cleaned, 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).

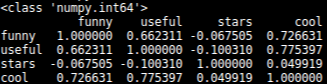
The data is packaged in the form of six json files, with multiple relations existing between the tables.



**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**

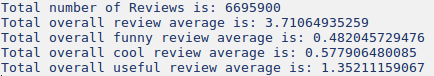
* November 1st We would like to have all of our data organized into a relational database for convenient SQL query.
* November 5th: We would like to have our data cleaned and preprocessed. This includes settling on a date format and formatting any additional data we wish to include.
* November 8th: We 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.

**MILESTONES COMPLETED (11/15/2020)**

Our initial goals of having the data in an SQL compatible format and having our data properly cleaned and examined have been met. Data integration of weather and economic data has not been met. Initial exploration of the dataset has been performed and tools to use have mostly been selected and implemented. We have completed several milestones behind schedule, which is unfortunate, however we believe that we are still on track to completing most of our goals and objectives.

**DATA ACCESS AND INITIAL STATISTICS**

Initial statistical points have been identified among specific attributes. We have computed the mean, median, and standard deviations of several attributes which we plan on using to use downstream in our data exploration.



**Figure 3: Example of precalculated metadata for the reviews table.**

**DATA FORMATTING AND DATABASE ORGANISATION**

Data has been organized and transformed into database files for SQL query. A problem arose in which the size of the database files prevented easy sharing over github and severe download delays for google drive inhibit file transfer, so we have elected to share database conversion scripts using the python json and sqlite modules, which avoid cloud hosting costs but could introduce version control issues if further refinement or changes need to be made. We are still investigating alternatives for database hosting. Google cloud services appears to be our best alternative at this point, though it comes with limited initial credits which a large database such as this could expend rather quickly.

**ADDITIONAL DATA SOURCES FOR ECONOMIC AND WEATHER DATA**

Additional stock data has been identified and determined to not require cleaning, however it has not been integrated into the relational database due to technical issues with the database formatting taking priority.

Weather data has not been pulled or analyzed. We are considering dropping this from the proposal at this point due to priority technical issues. We decided to stick with analyzing stock market data instead of weather data due to weather data being geographic as well as chronological requiring multidimensional correlations with our dataset as opposed to joining on only one attribute. Our main goal here being to correlate yelp data with outside data, and correlating stock market data to yelp review data will satisfy that goal, though may not be as interesting to the average user than weather data.

**MILESTONES TO DO**

Our highest priority currently is to iron out the database tools we will use. We have elected to stay with SQLite for database querying, however we are still investigating alternatives should other problems arise. We can currently query from several databases including the main three tables we will use: review, user, and business tables. These give us access to all raw attributes we will use in our project, including a review date to correlate with stock prices. Clustering and bagging still need to be performed on these datasets before additional classifications can take place.

**OBJECTIVE 1**

Our first objective is to attempt to correlate yelp reviews with an outside data source. We believe that due to the plethora of data within the yelp dataset few have attempted to integrate additional data into their yelp dataset investigations. We initially proposed using economic data from the US stock exchange and US weather data to attempt to correlate weather patterns and stock indexes with trends in reviews. As we have had some setbacks and delays in formatting our dataset to be completely compatible with the tools we intend to use, stock and weather data have not yet been integrated with our yelp dataset. Due to these delays, it is possible that correlating yelp data with US weather data may be dropped in favor of using economic data as the only outside data in our investigation. Our goal was to integrate additional data, and it makes sense to choose the data that will prove to be easiest to integrate. We still plan on integrating weather data if time permits, but due to delays previously experience, we may need to drop some of our previous goals. A more straightforward approach rather than using specific weather data, which may add to our already huge overhead, could be to study seasonal variation (i.e. month to month).

**OBJECTIVE 2**

In order to answer our second objective:

Can we characterize potential pitfalls/areas of improvement of a restaurant based on its reviews.

We will need to implement a parsing tool to extract frequent terms form restaurant reviews and then correlate those frequency values to review score. A parsing mechanism has been proposed however has not yet been implemented due to previous technical delays. We plan on using python dictionaries to store words as keys and frequencies as fields. We believe this to be the most efficient way of accomplishing this though we are wary of size restrictions. We may need to break up the data into several chunks and select the most significant terms for each chunk, therefore removing frequent, though not statistically interesting words such as articles or pronouns, those most likely being frequent words that most likely have little classification value. We expect this implementation to be challenging to implement and are therefore allowing time and resources accordingly. This parsing tool and mechanism will be applied later for objective 3, as discussed below.

**OBJECTIVE 3**

In addition to generating additional metadata for grouping clustering purposes, we also still need to extract restaurant attributes to determine regional user taste as per question 3:

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?).

We have access to restaurant identifiers in the business table, providing us insight as to what type of food, but more specific identifiers will be mentioned in the review table free text attribute. A more in-depth investigation of this question will require parsing the free text column and aggregating terms to look for frequency of terms to correlate with reviews in a specific geographic location. We plan on prioritizing the identifiers such as food type from the business table prior to parsing the free text review as that should provide a more interpretable and certain result. Once we are able to characterize the most important aspects of a restaurant based on food type and other information provided by the business in the business table, we will expand our investigation using the parsing tools previously implemented in objective 2. We have elected to implement the parsing mechanism for this objective last because we are anticipating the parsing and statistical analysis of free column text to be complicated and require additional testing and product tweaking before being ready to implement it into objective 3. We do not wish to ignore objective 3 entirely until all tools are ready, so we will start with the easily accessible data within the business table itself.

**OBJECTIVE 4**

We have calculated a correlation matrix for the review attributes indicating that cool reviews have a higher correlation with useful reviews than funny reviews. It appears that users frequently will upvote reviews for multiple attributes such as cool, funny, and useful, so more analysis will need to be performed to fully determine how a review being marked funny, cool, or useful will impact the overall review score. Current correlation values however do not indicate a very interesting surface pattern, so more in depth analysis will be required before we acknowledge that a novel and interesting pattern is not present. We may investigate clustering and comparing reviews to the average business review score to determine if a review is a good review or a bad review so that we can then apply Bernoulli and Bayesian statistical methods to determine the contextual class of the review, good or bad.

**DATA VISUALIZATION**

Tableau has not yet been attempted on the data as some of the raw csv files are too large to load directly into tableau, so some meta data generating and classification is needed prior to data visualization for the larger tables. Too much emphasis on visualization at this point could lead us astray from our proposed goals, so we have determined that more back-end work is required before data is visualized and eventually presented. Some data visualization should be attempted soon though, as pitfalls and unknown requirements are better discovered sooner rather than later, which we intend to have one of our group members attempt some form of visualization within the next two weeks. Licensing may also be an issue for some group members who no longer have access to a free student license or a free two-week trial.

Other visualization tools have also not been implemented or attempted. Our final goal is to not only answer our proposed questions and discoveries, but to also relay that in a visual manor to effectively communicate the findings from our data mining project. We initially proposed using the python modules plotly and seaborn. We have not attempted to use these tools with our dataset, however dataset size has been an persistent issue for many of our tools and this could become an issue once we attempt to visualize our results. While we still maintain that we wish to attempt to visualize our data using these tools, we are also aware that problems could occur, and we are actively looking for alternative methods should we need to peruse other data mining tools. As with the tableau software, an exploration of these tools within the context of our dataset should be completed within the next two weeks so that we will have insight as to how to apply these tools to our dataset or even if that is a possibility given the size of our data and potential size limitations of the python modules. If a compatibility issue is discovered, then time will be needed to find a work around method or locate tools more suited for visualizing large amounts of data points.

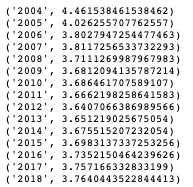
**SUMMARY OF PROGESS**

We are in fact behind schedule. Our milestones schedule was ambitious, and we have not managed to meet all milestones or were behind on our timetable. We still feel as if we will be able to complete all goals to a reasonable degree, though we may focus more on specific goals and allocate our time accordingly in order to have a more polished final product.

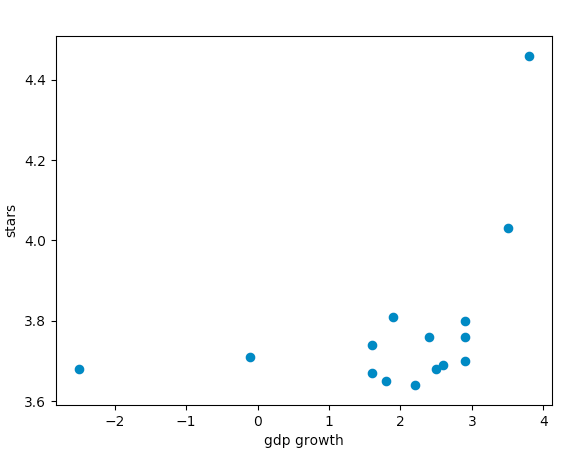
**RESULTS SO FAR**

From our previous work in calculating correlation coefficients on our reviews table (see figure 2), we can conclude that reviews marked as cool tend to be correlated with useful review more than funny reviews. This answers the initial question if funny reviews are more helpful. Yes, though it appears that if a user is going to take the time to upvote a review as useful they are likely to upvote other categories as well. The conditional probability of this hypothesis has yet to be determined, and that along with discovering other characteristics of funny reviewers has yet to be performed, however we intend to start with restaurant price, that is are funny reviews more frequent for less expensive establishments.

Initial exploratory analysis from short SQL queries has shown consistent overall average stars in reviews (approximately 3.7) with the month of December having lowest average reviews (3.68). Global average reviews are also fairly similar:



When compared against annual GDP growth, we get the following scatterplot:



With a correlation coefficient of 0.45.

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