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

ABSTRACT

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.

Initially we set out to answer four questions based on the yelp database. Do external factors affect average yelp reviews, such as economic or weather data. Our findings indicate that economic data, specifically higher unemployment correlates to lower reviews.

An attempt to characterize potential pitfalls and areas of improvement based on user review information indicated certain trends such as one star Chinese reviews tend to have the word ‘panda’ in them.

We also attempted to characterize and identify important aspects of businesses to a given regional population, in this case Phoenix Arizona. We discovered that one star reviews tend to have more negative words, and five star reviews had a very large emphasis on excellent customer service.

Yelp users can also rate individual yelp reviews. A review can be selected as funny, cool, and/or useful. We determined, using an algorithm based on naïve Bayes theorem that cool reviews are slightly more likely to result in a review being deemed useful by other users. Should a review be rated funny and cool, there is an extremely high probability of said review being voted useful as well.

**INTRODUCTION**

Yelp tracks many different attributes, and with such a large user base, the data set is immense. The data set available for this contest is nowhere near their complete dataset; containing complete data from only a select few cities. This is why we chose to examine Phoenix Arizona for many of our objectives that were specific to a particular region.

Our first objective was to pull an outside data set and look for correlations with the yelp data. As this yelp data challenge has been offered for several years, we thought it would be interesting to look to outside data, particularly economic or weather data to determine if a local economy has a perceptible impact on yelp reviews. Our hypothesis being that if the local economy is down, that reviewers affected by the economy would rate businesses lower than when the local economy was on an upward trend. This could provide valuable insight to businesses looking to open new locations in certain areas with high or low unemployment ratings, or signal to wait until the stock market transitions from a bear market to a bull market.

Our second objective was to characterize and identify potential pitfalls for businesses based on their reviews. If successful, companies could utilize this knowledge to identify areas of dissatisfaction at an early stage and fix whatever shortcoming they may have earlier rather than later. This information could prove critical to struggling businesses.

We also wanted to examine in a broad sense what users in a specific city or region valued based on the reviews they submitted, and the stars granted. For our city, we selected Phoenix AZ, as the dataset provided large amounts of data from Phoenix. We decided this question would be an important question to investigate as that could provide extremely valuable insight for companies as to which areas of business to focus on. For example, should restaurants focus more on atmosphere, ingredients, presentation or focus on some other aspect. Perhaps certain industries perform better on average than other industries in terms of yelp reviews and stars.

Our last objective was to examine the relationship between reviews, specifically the relationship between funny, cool, and useful reviews. Our hypothesis was that comedic reviews would have a higher correlation with reviews being voted as useful than a review voted as cool. This question is significant because depending on the results, reviewers could tailor their reviews to be more comedic or cool in nature, in an effort to reach as many other users as possible. The goal for many yelp reviewers is to be useful for other users, and insight into which tone is deemed more useful could prove valuable for many reviewers.

**RELATED WORK**

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.

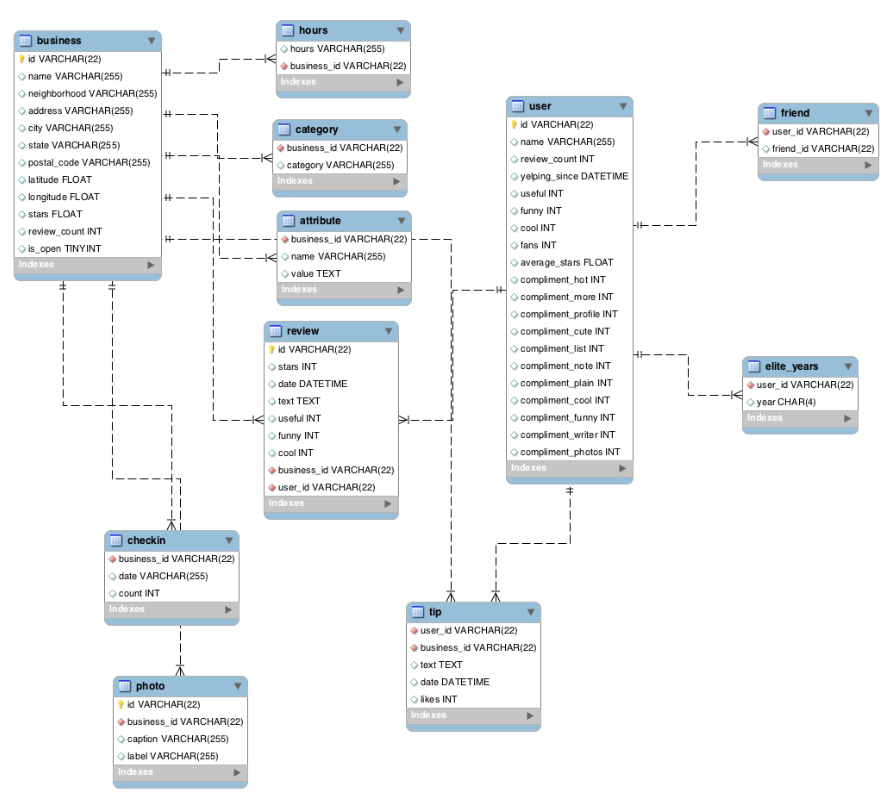
With the wealth of information within the dataset, it appears most researchers have searched for relationships wholly within the Yelp dataset, and few have drawn in additional information to correlate with information in the yelp dataset.

**DATA SET**

The Yelp dataset consisted of a subset of yelp reviews from many kinds of businesses from U.S. states as well as Canada provinces from between 2004 and 2018. States with significant representation in the dataset (>2000 reviews) included Nevada, Arizona, North Carolina, Ohio, Pennsylvania, Wisconsin, Illinois, and South Carolina. The two cities with most reviews were Las Vegas (>2 million reviews) and Phoenix (>700,000 reviews).

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. The tables attributes relating to information about the business, which contains geographic locations, business id, and review counts. There is a table consisting of business hours, a table for categorizing the type of business, such as restaurants or healthcare etc. There is a table consisting of variable attributes about the business, such as ‘tacos’ or ‘burgers’. There is a table tracking checkin information which is separate from the hours table. The review table is related to the business table through business id, where each tuple in the review table relates to one business id. The review table is reverenced by the user table, where each user has a user id, and each review they leave is referenced by the review table. A tip table exists, which tracks ‘tips’ which in this case are small snippets of advice, usually shorter than actual reviews. They reference the user table and the business table similarly to the review table. Users can have friends which is tracked in the friends table and elite years, that is years that a user has been an ‘elite member’ is tracked in a table labeled “elete\_years.”



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

Phoenix unemployment data was obtained from the Federal Reserve Bank of St. Louis[10]

**MAIN TECHNIQUES APPLIED**

Spark – Python distribution through Jupyter notebooks was used to load the json datasets and perform large querries with joins. Spark dataframes were then converted into pandas dataframes for smaller-scale analysis.

**OBJECTIVE 1 TECHNIQUES**

Linear regression and Pearson Correlation Coefficient were used to quantify the effect of economic health on reviews. Local unemployment was the key indicator of economic health used in this analysis. Other economic metrics were considered, such as stock inexes but local unemployment was a readily available metric which seemed an ideal indicator of economic status for a local community.

**OBJECTIVE 2 TECHNIQUES**

Common natural language processing techniques were used including removing punctuation, capitalization, removing stop words, and lemmatization. Phi coefficients, which are related to chi squared and commonly used for binary data, and raw word counts were used to find common sets of restaurant categories and words.

**OBJECTIVE 3 TECHNIQUES:**

The same common natural language processing techniques from objective 2 were used to process the review text to remove punctuation and process the raw text and tally the most common words in each star rating category. We were able to generate a list fot he 1000 most frequent words per star rating. That is we generated the 1000 most frequent words from 1 star reviews, then a separate list for 2 star reviews etc. We then looked for similarities among those lists by performing an intersect operation between the lists. We decided to then look for words from the frequency lists that were unique to that star rating. Next we manually classified the unique lists and tallied the classification entries and reported the top 10 from each star ranking. We decided to manually classify each word rather than using an algorithm to do so because classification of text is difficult, and a human will have a far greater understanding of context than an automated process. If we were to peruse this further, an algorithm for classification would be in order, but as it is, selecting unique words from the list of frequent words reduced the amount of data from millions of words to a couple hundred words, something manageable for a person to fill out. Surveying additional people to classify the word lists into categories would help reduce individual bias, and thus be a good next step as well as a classification algorithm.

Words unique to 2 star reviews: ['hertz', 'nissan', 'cab', 'walmart', 'apartments', 'camelback', 'mcdonalds', 'salesman', 'sons', 'ford', 'finance', 'neighbors', 'parker', 'express', 'station', 'keys', 'loan', 'managers', 'lines', 'solar', 'complaints', 'dealer', 'pull', 'upgrade', 'tenants', 'residents', 'dealing'

**Figure 2: List of words uniquely frequent in 2 star reviews.**

The idea being that if a word has a high frequency in all categories, that it is probably not significant. We feel that the method is statistically significant because if a word has a high frequency in one category, at least frequent enough to make it into the list of top 1000 most frequently used words for a rating category, but not make it into any of the other frequency lists, then it is uniquely predominant to that category.

We chose to examine the top 1000 frequent words per star rating arbitrarily. 100 was too few to obtain meaningful results, and 1000 seems to capture enough data to suit our purposes, though only investigating on this one sample size could skew the data somewhat due to our unique constrains. A solution to this would be to perform the same steps on different samples sizes and find an optimal number for frequent words to examine. As it stands, I believe that 1000 most frequent words can and does provide insight into our question of what do yelp reviewers in Phoenix Arizona value most.

**OBJECTIVE 4 TECHNIQUES:**

A correlation matrix was obtained through the pandas and numpy modules as a pilot investigation. A more in depth analysis was performed by a naïve Bayes theorem algorithm to predict whether a review would be voted useful if it was voted funny or cool or both funny and cool.

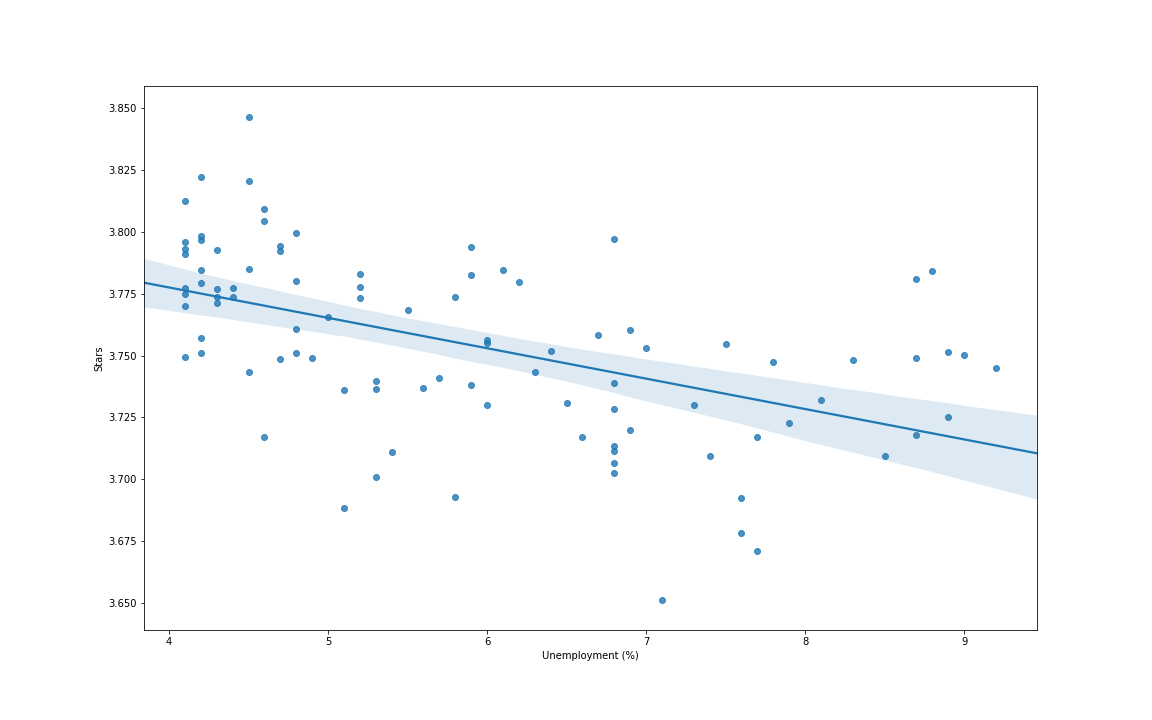
**KEY RESULTS**

Initial exploration resulted in total monthly average review stars between August 2010, which was the first month in the data set with over two thousand reviews and fall 2018 had and average review stars within the range of 3.65-3.85. The all time average stars in the largely represented cities was also in this range.

**OBJECTIVE 1 RESULTS:**

Our first objective was to examine a potential economic effect or correlation between economic success and review stars. We chose unemployment as a primary indicator of economic success [11]. We chose to focus on Phoenix, Arizona as a study city because it had the most data to work with other than Las Vegas, which we predicted would not be as affected by local unemployment due to its massive tourist presence.

Linear regression and correlation coefficient analysis were performed on monthly unemployment rate versus monthly average stars. A moderate negative relation was found between unemployment and stars, which confirmed our hypothesis that difficult economic situations would decrease overall satisfaction and ratings among yelp reviewers.



**Figure 2: visualization of stars versus unemployment percentage. A correlation coefficient of -0.51 was found.**

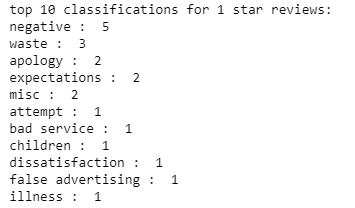
**OBJECTIVE 2 RESULTS:**

We also set out to determine whether we could identify important factors for different kinds of restaurants. We chose to inspect 1 star reviews because a review of 1 star can be particularly damaging for a business [12] and, in consumers’ eyes, can sometimes outweigh the effects of many 5 star reviews.

There were several findings when looking at word counts and correlations between words and restaurant categories. In Italian restaurants, common poor reviews involved pizza, slow delivery time, and poor service. Japanese restaurants with 1-star reviews often had sushi as the culprit of the bad review. After examining words correlated with reviews about sushi, it was evident that fish not being fresh causing sickness was very common in 1-star Japanese restaurant reviews. 1-star reviews at Chinese restaurants often mentioned chicken being soggy or dry.

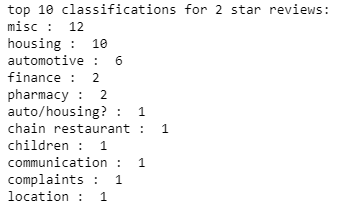
**OBJECTIVE 3 RESULTS:**

In order to attempt to determine which factors the population of Phoenix determines the most important, we identified the thousand most common words for each star rating. We then filtered that list down to only words unique to that particular star bin. As the lists of words were dramatically smaller we were then able to manually categorize which word belongs to which category, such as automotive, service, customer service, restaurants, housing, etc.



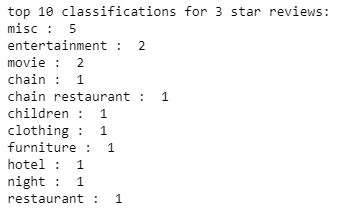
**Figure 3: top 10 classifications for 1-star reviews.**

The main aspects of 1-star review in Phoenix were generally negative words like ‘shame’, ‘nasty’, and ‘poorly’ and wastefulness. This does not necessary provide insight into how the community feels about specific industries or services, but it does make sense that low rating reviews would be very negative in nature.



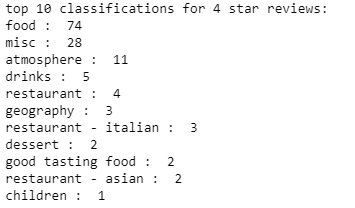
**Figure 4: top 10 classifications for 2-star reviews.**

Housing and automotive services seem to be uniquely predominant in the 2-star reviews bin for Phoenix. Pharmacies and finance also showed up multiple times though words such as ‘pharmacy’ and ‘prescription’ for the pharmacy classification, and ‘loan’ and ‘finance’ were categorized as finance.



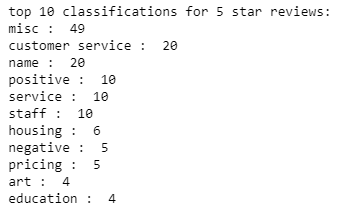
**Figure 5: top 10 classifications for 3-star reviews.**

The leading classifications for 3-star reviews were mostly related to entertainment such as sports, and many words were directly related to movies/cinema.



**Figure 6: top 10 classifications for 4-star reviews.**

4-star reviews mostly related to food. There were a large variety of ‘food’ words that were related to fresh ingredients. Italian and Asian foods were especially common in the word rankings with words like ‘pasta’ and ‘pho.’ Words relating to ingredients and atmosphere appear in this ranking much higher than any other star ratings bin.

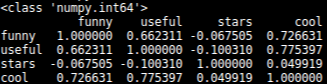


**Figure 7: top 10 classifications for 5-star reviews.**

Excellent service seems to be by far the most important thing to reviewers when granting a 5-star review. Many words were related to customer service. Individual names only showed up in this bin, and in no other bin. Words like ‘knowledgeable’ appeared frequently as well. Individual industries did not seem to be mentioned as much, though words relating to art and education showed up in this category while not as prevalent or obvious in other categories. Pricing was mentioned in this category as well, but with words like ‘bargain’ where the words mentioning pricing in the 1-star reviews were more associated with cheapness than receiving a good deal. These 5-star review words also had many positive words such as ‘happy’ and ‘love.’ Service related words such as ‘replace’ and ‘project’ were also fairly common to this category.

**OBJECTIVE 4 RESULTS**

A quick look at a correlation matrix between different attributes was our first results on this part of the project.



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

We immediately noticed that usefull and cool had a higher correlation than funny and useful, which was not encouraging for our hypothesis. We thought of also looking into whether funny reviews were more or less likely to have higher than average reviews or lower than average reviews, but the extremely low correlation value dissuaded us from investigating that aspect. We decide that a single correlation matrix, while informative would not suffice to prove to ourselves whether funny reviews were better, as in more helpful, than cool reviews.

Our next step toward completing this objective was to create a naïve Bayes algorithm accounting for the conditional probabilities that a review would be useful, funny, or cool.

Our Bayes classifier would assign both funny and cool reviews as being useful. And the accuracy of our model for funny reviews was 85.1%. Our accuracy of cool reviews was 86.6%. and the probability of a funny and cool review being useful was 96.8% accurate.

Our predictive model indicates that cool and funny reviews are a pretty good indicator that a review will be useful. Predicting based on cool reviews alone gives a slightly better model than funny alone but combined provide a very good indication that the review will be useful. This phenomenon is most likely explained by if a user is going to take the time to vote a review as funny and/or cool, they are very likely to vote the review as useful as well, as they have already invested time into voting/reviewing the review anyway.

To answer the question of which one is more useful, cool reviews are marginally better indicator that a review will be voted useful compared to funny reviews. If a review is voted both funny and cool there is a very strong chance that the review will be voted useful.

**APPLICATIONS**

Knowledge is power – Imam Ali

Some of our results themselves were not necessarily groundbreaking, though confirmation of our hypothesis can be useful to certain business owners. Knowing that reviews on average decrease during economic hardship does not necessarily help a business struggling in a tough economy, though that information may be valuable to a company looking into opening a new location in an area with high unemployment. Which could be especially important for emerging businesses as initial low reviews can be challenging for new businesses to overcome.

Knowing the potential pitfalls identified in 1-star reviews can also provide powerful insight to potential restaurant owners. If say someone is looking to start a Chinese restaurant, knowing that other Chinese restaurants have struggled with getting the chicken right can allow them focus on areas of common failure to ensure their business stands out and does not fall into the same pitfalls as other Chinese restaurants.

The information we discovered about the different yelp reviews by stars can be particularly useful to emerging businesses or even well-established businesses. It seems like common sense, but great customer service is rewarded with great reviews. No other star division had the same number of individual names mentioned. Hiring and training employees is one of the most important things a business can do according to this data set.

If you are a restaurant owner and you want good reviews, the results of our third objective indicate that you should strive to have fresh ingredients and authentic recipes. Atmosphere also appears important, but overall good quality food will earn higher 4-star reviews. If you are a restaurant owner and you want to avoid 1-star reviews, try not to make customers sick, as ‘illness’ and ‘sick’ only showed up as frequent words in the 1-star category.

Our hypothesis about funny and cool reviewers turned out wrong, and honestly, not very useful. Sometimes when searching for novel patterns in data, the results are uninteresting. Not only were they both very similar in correlation values and in accuracy from our Bayes classifier, but overall due to human nature, not unique values, or rather, correlated values, as many users will vote a review both cool and funny. What insight we can gain from this is that if you want your review to be voted as useful, attempt to engage readers by being cool and funny in order to encourage them to take the effort of up voting as such.

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

[10] <https://fred.stlouisfed.org/series/PHOE004UR>

[11] https://www.economicshelp.org/blog/10189/economics/key-measure-economic-performance/

**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.

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**

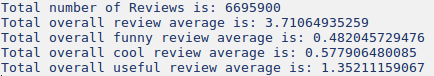
* November 1st We would like to have all of our data organized into a relational database for convenient SQL query.
* November 5th: We would like to have our data cleaned and preprocessed. This includes settling on a date format and formatting any additional data we wish to include.
* November 8th: We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

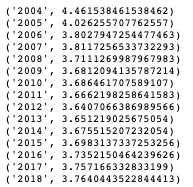
**SUMMARY OF PROGESS**

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.

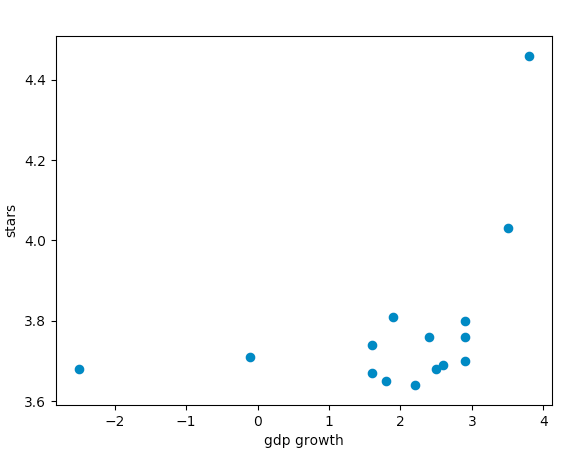
**RESULTS SO FAR**

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.

DOI:10.1145/1234567890

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?
* **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.

RRH: F. Surname et al.

Price:$15.00