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

Garrett Lubin
Applied Computer
Science University of Colorado Boulder
Boulder, Co U.S.A.
galu7424@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 was that if the local economy is down, 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 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 [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.

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."

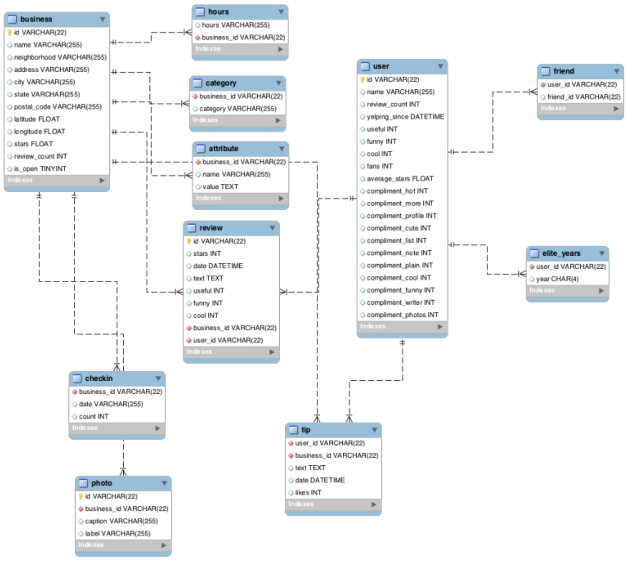

Figure 1: A relational database model of the yelp dataset constructed in MySQL workbench. [7]

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 queries 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 indexes 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 for the 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:  ['her
tz', 'nissan', 'cab', 'walmart', 'apar
tments', 'camelback', 'mcdonalds', 'sa
lesman', 'sons', 'ford', 'finance', 'n
eighbors', 'parker', 'express', 'stati
on', 'keys', 'loan', 'managers', 'line
s', 'solar', 'complaints', 'dealer', '
pull', 'upgrade', 'tenants', 'resident
s', 'dealing'
```

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

We presumed 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 do 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.

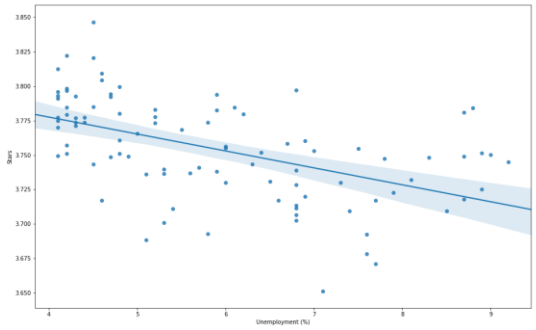

Figure 2: visualization of stars versus unemployment percentage. A correlation coefficient of -0.51 was found.

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.

```
top 10 classifications for 1 star reviews:
negative :  5
waste :  3
apology :  2
expectations :  2
misc :  2
attempt :  1
bad service :  1
children :  1
dissatisfaction :  1
false advertising :  1
illness :  1
```

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.

```
top 10 classifications for 2 star reviews:
misc :  12
housing :  10
automotive :  6
finance :  2
pharmacy :  2
auto/housing? :  1
chain restaurant :  1
children :  1
communication :  1
complaints :  1
location :  1
```

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.

```
top 10 classifications for 3 star reviews:
misc :  5
entertainment :  2
movie :  2
chain :  1
chain restaurant :  1
children :  1
clothing :  1
furniture :  1
hotel :  1
night :  1
restaurant :  1
```

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.

```
top 10 classifications for 4 star reviews:
food :  74
misc :  28
atmosphere :  11
drinks :  5
restaurant :  4
geography :  3
restaurant - italian :  3
dessert :  2
good tasting food :  2
restaurant - asian :  2
children :  1
```

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.

```
top 10 classifications for 5 star reviews:
misc :  49
customer service :  20
name :  20
positive :  10
service :  10
staff :  10
housing :  6
negative :  5
pricing :  5
art :  4
education :  4
```

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.

```
<class 'numpy.int64'>
          funny    useful     stars      cool
funny   1.000000  0.662311 -0.067505  0.726631
useful  0.662311  1.000000 -0.100310  0.775397
stars  -0.067505 -0.100310  1.000000  0.049919
cool    0.726631  0.775397  0.049919  1.000000
```

Figure 2: A correlation matrix of review attributes

We immediately noticed that useful and cool had a higher correlation than funny and useful, which did not support 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 decided 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

[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.

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

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