COMP 347 – Final Project

Spring Semester 2022

Occidental College

David Zhang

According to the Atlantic Magazine, the first review submitted to Yelp was four stars for Truly Mediterranean. On October 12, 2004, a user wrote four words: "dirt cheap, good falafels." Since those four words were typed, Yelp's platform has become a colossal social media platform for business reviews. What do most people do when figuring out what and where to eat? They turn to Yelp. People search, eat, rate, and review restaurants and its food, making it an easy place to find local restaurants of all cuisines. To understand how influential Yelp is at guiding people to restaurants, let's take a look at some statistics surrounding Yelp. There were 92 million unique users on Yelp monthly and up to 184 million reviews worldwide (Marinova). Furthermore, 45% of customers were likely going to check a Yelp review before visiting a business and businesses see an up to 9% increase in revenue everytime they gain a star. Finally, on June 27, 2011, 32,245 reviews were added to Yelp in one day. It's easy to see how influential Yelp has been in helping people decide where to go and helping others make informed choices as well.

On the flip side however, with so many restaurant reviews it can also be difficult to decide where to go eat, especially if there were many restaurants of the same cuisine with similar ratings. A quick check of Indian restaurants in the Schaumburg, Illinois area, for example, found fourteen Indian restaurants with a three to four star rating within a three mile radius. Since all the restaurants were so similar, it would be challenging to know which restaurant to decide on. Therefore, it would be useful to evaluate the restaurants using different or modified metrics for the purpose of finding those restaurants that stand out from the pack. Since this project is looking to more accurately review a restaurant, this can be a way for restaurants to better understand if they need to improve their business or not, or to bring the spotlight to a restaurant that truly deserves the recognition. For customers, this new rating system would make it more convenient to find a quality restaurant because it would help filter out restaurants that may have an inflated rating and lead customers to a restaurant that is accurately rated or to one that is rated nicely. Convenience is really valued in today's society and by sorting Yelp reviews through a hierarchical system, the greater reviews were able to have the biggest impact instead of being "lost in the pack," while reviews with less weight were still valued and encouraged. In addition, this is an important topic because Yelp is a globally influential platform, so making improvements to it will come at the well-being of numerous businesses, customers, and Yelp.

One important aspect about the data being used is that there is a sufficient amount to train the model on and there's also enough to test the model on. This project only takes a look at Yelp reviews from ten different cities, but the data consists of 1.6 million reviews from 366,000 users for 61,000 businesses. Another important aspect about the data being used is that it's very detailed. The data contains several variables that can be used to determine whether or not a review is more or less impactful. For example, it's achievable to check on the profiles of those who left reviews and see how many reviews they've made, where they live, their ethnicity, their name, how long they've had a Yelp account for, and which places they've left reviews for. You can also check how many stars the restaurant had in the past, how many reviews they've had in the past, what categories the restaurant lies in such as "Night Life" or "Good for Kids." Furthermore, Yelp releases millions of reviews on their website for student/academic data challenge purposes. Thus, it's safe as well as ethical to use the data rather than illegally web-scraping data from websites. Yelp reviews have historically been used for academic research, so the data is more likely to be error-free. The data is clearcut and thus easy to utilize in order to perform a Machine Learning Project to try and improve Yelp's rating system.

The importance of the methods being used is that it creates a ranking system of reviews that prioritizes ones that were likely to be more accurate in evaluating the restaurant. This in turn can change the star rating of a restaurant and can help make the rating of a restaurant more accurate. Furthermore, it can help restaurants with the same star rating as the others around it be more noticeable because it shows customers that reviewers who may have more knowledge like this place. Another important aspect of the method being used is that it prioritizes those with seemingly more knowledge whether that be users with ethnically Indian names or users with multiple reviews under their name. As a result, future customers can better understand if the restaurant is worth going to.

As for the mathematical procedures of this project, the first step was to analyze the dataset to find the number of multiple reviewers there were for a cuisine. The point was, if there were very few, then adding weight to their opinions may ultimately have little impact on the overall rating. The project also wanted to see how much the number of multiple reviews varied from cuisine to cuisine. While the project was focused on Indian cuisine, they first wanted to make sure that usage patterns were similar enough across cuisines to be able to generalize. Taking a first look at Indian cuisine, an "is_indian" column was added to the table based on whether the word "Indian" appeared in "categories", using "grepl: rub$is_indian <- grepl("Indian", rub$categories) == TRUE", they then used a subset to create a data frame of just reviews for Indian restaurants, "indian <- subset(rub, is_indian == TRUE)." Once they had this, they used the select, group_by, and summarize commands from dplyr to create a table of the # of reviews of Indian restaurants each user had done, "num_reviews_Indian <- indian %>%

select(user_id, user_name, is_indian) %>% group_by(user_id) %>% summarize(tot_rev = sum(is_indian)).” After doing this, they used table, count, and mean to get review statistics: This yielded a result of 9,549 total reviewers, with 7,814 doing just one review, 1,048 doing 2 reviews, and the rest doing 3 or more. The highest number of reviews that one user wrote was 93. Overall, roughly 10% of the users had done multiple reviews of Indian cuisine. For this project, it seemed as if there were enough users in general to justify trying a multiple review rating (and not so many that the added weights canceled each other out). They then came to the real test: modifying the rating using these weights and seeing what impact they had. They then tried this on Indian restaurants. They had the # of reviews for each user in num_reviews_Indian. They then joined this back to the “indian” data frame containing all the individual ratings, which created new table which had the rating the user gave as well as the # of Indian restaurants the user reviewed. Afterwards, they stored this data in a variable called cuisine-indian (or cin for short). The weighted star ranking was created when the author multiplied the star rating with the number of reviews the user gave.Next, they used the group_by and summarized commands in dplyr to generate both the original average star rating (as a sanity check) and the new, weighted star rating. After doing this, they used “summary” to get a sense of the effect of the new rating, which showcased its minimum, first quartile, median, mean, 3rd quartile, and maximum. Looking at the new star rating, there was a disproportionately large upward effect on restaurants with 5 or less than 5 reviews compared to restaurants with more than 5. The author thus removed restaurants with 5 or less reviews and only looked at effects on restaurants with more than 5 reviews. The range of the min to max star increase decreased, making the new star rating be seemingly more accurate/reasonable in evaluating restaurant quality.

The second proposed method was to add an "Immigrant Rating." This section tries this on the Yelp database for Indian cuisine and analyzes the results. Since the Yelp database does not include ethnicity in its user database, the only way to try to guess ethnicity is through the user name. The first step was to generate a list of names that would qualify as being uniquely Indian. To do this, the user names in the Yelp "user" dataset were examined for potential candidates. When a name looked like a potential candidate, the question "Is ___ an Indian name" was googled to see if sites such as www.indiaparenting.com, www.modernindianbabynames.com, and www.indiachildnames.com would show up in the top of the search result. Anything that did so was included in the list. Many popular names in America were left out ("Daniel", for example, appears in the indiapwerenting.com database but is left out). This resulted in 608 names, ranging from "Aayush" and "Abhijeet" to "Yogesh" and "Yuvaraj". This list of names was read into R using the scan command: "inames <- scan("indian_names.txt", what = character())" Taking the "indian" dataset from Method 1 that contains the business name, reviewer name, and star rating for all Indian restaurants in the dataset, we add a "reviewer_indian_name" field by using "%in%: indian$reviewer_indian_name <- indian$user_name %in% inames." Also, to simplify calculation later, the author added an "istars" field which is simply the number of stars the reviewer gave the restaurant if that person has a uniquely Indian name, and a 0 otherwise. This is given by the math "indian$istars <- indian$stars * indian$reviewer_indian_name" After doing this, you can find out how many reviews fall under the "immigrant" category by using the command: ">table(indian$reviewer_indian_name)". Out of the 13,146 reviews analyzed, 1,274, or a little under 10%, of the reviews were eligible for the "Immigrant Rating." This apparently seemed like a substantial enough amount to go forward. Using "group_by" and "summarize" from dplyr, the author regenerated the average rating for a restaurant (as a sanity check) and also the "Immigrant

Rating," which would simply be the sum of the newly generated "istars" for each restaurant divided by the # of "immigrants" who reviewed that restaurant. After doing this and sorting the results by the difference between the new rating and the original rating, there were dramatic differences, but in many of those cases the differences was due to there being only 1 "immigrant" leaving a review. Thus, the author set an arbitrary value of needing at least 5 "immigrants" to be able to generate a useful "immigrant rating". The author used "subset" to screen for this: "ari5 <- subset (avg_rating_Indian, nin > 5)." When doing this and looking at the values with the greatest difference, the max difference decreased from -2.85 to -1.9. Using summary, you can see that the potential difference ranges from -1.9 to an increase of .31, with the mean value being -0.4, which is interesting because from the outside looking in it seems as if reviewers from India or reviewers of Indian heritage tend to be harder on Indian restaurants than reviewers from America, thus giving lower ratings. Using > summary(ari5$dif), you can see the data of the new rating, Min: -1.92300, 1st Qu: -0.72230, Median: -0.34500, Mean: -0.436, 3rd Qu: -0.09828, and Max: 0.30980. The other item to note is that, when imposing the requirement that there be at least 5 immigrant ratings, the number of restaurants with an "Immigrant Rating" decreases from 392 to 70. Looking at the results of both method 1 of weighted reviews and method 2 of immigrant reviews, both methods were meaningful enough to make some restaurants stand out from the rest even when those same 1-2 restaurants were rated similarly with the others initially. It seems like these two methods could be beneficial in the industry, but it may be difficult to apply outside of academia.

In the industry, Yelp faced economic struggles; and amid controversy, they've received complaints about their review system, which has hurt their reputation. In order to combat this, Yelp modified their algorithm. Somewhat recently, they began sorting reviews into two categories: “Recommended”, which appears in the immediate view of the business’ page, and “Not Recommended”, which is hidden from view and doesn’t impact the rating of the business, although you can click on a button to reveal the review. Another change they’ve made is showing whether or not a rating is trending up or down - a way to gauge whether or not a business is making changes based on the customer reviews.

Even with the changes, Yelp hasn’t been doing so well, thus, other companies have begun implementing new ideas. For example, Facebook, who wasn’t initially involved in the online review business, began experimenting with a feature that lets users search for local businesses, including nail salons, take-out restaurants and plumbers, and see results that include customer reviews and star rankings. The feature looks to add the credibility of reviews linked to someone’s social networking account as opposed to an anonymous profile. MunchAdo, a food-discovery startup, made advancements in data capture and storage. They help restaurants connect bad reviews with the dates, times, and circumstances so they can make changes when appropriate. Many review sites fail to give business owners key details to help them improve, but ding their reputations with poor reviews. MunchAdo’s CEO, Puneet Talwar, explained this idea, “They’re not able to connect the review with a transaction [...] That’s often the first thing we hear from restaurants.” This makes it easier for restaurants to connect the event with the bad review, helping restaurants make more informed decisions while trying to improve. Furthermore, IBM unveiled a new smartphone app — IBM Watson Trend — that puts an artificial intelligence spin on the consumer review. The app examines 10,000 data sources, including social media networks, blogs, shopping forums and review sites, to spot trends in consumer electronics, toys and health and fitness. The data is crunched into a 1-100 rating scale, which includes a sampling of reviews and an explanation of what’s driving the trend.

One of the few latest academic advancements made in the online consumer reviews was done this year. In the academic research, the study intended to utilize more data storage in order to advance its methods. Online hotel reviews were selected to be analyzed because travel reviews are important in helping make traveling decisions; thus, there are many reviews. A mix of both low and high budget hotels were selected. In total, 680 reviews, equally selected from 4 different budgeted hotels, were gathered, and 509 reusable reviews were selected for testing the model and the hypothesis. Significant but secondary factors such as reviewer name, age, gender, contribution badge, total number of reviews, and average amount of helpful upvotes was extracted. Primary factors such as the content of the review and latency of the review were extracted. For content, this meant images included in the review, the word/sentence count, and the rating given by the review; for latency, this meant the readability of the review, the grammar and spelling of the review (comprehensiveness), the sentiment of the review, the relevance of the review, and the clarity of the review. The dependent factors were the factors just mentioned, while the dependent variable was Review Helpfulness measured by counting the number of helpful votes which the review received. To extract the secondary and primary factors, content analysis techniques were used. However, for latency, content analysis was performed on the review text. Standard content analysis guidelines were used for coding (Hseih and Shannon 2005; Shriver, Nair and Hofstetter 2013). Two graduate students, studying at a prominent university (with English as their first language) were recruited to evaluate the online reviews on different dimensions. Comprehensive training was given to the coders. A TOBIT regression model was used to test the association between the independent variables and the dependent variable. The second study, in addition to the review factors and the reviewer factors, took into consideration the valence of the text. This included comprehensiveness, clarity, relevance, and valence. For comprehensiveness, a review is deemed comprehensive if it includes holistic and coherent opinions on information such as food, service, drinks, location, and atmosphere. Using Term Frequency, Inverse Document Term Frequency, and Latent Dirichlet allocation, text mining, and other methods, 3,000 words were extracted and then reviews were given a comprehensiveness score based on whether or not they contained these words. Review clarity was captured by looking at the grammar, spelling, and writing style of the review. The Hunspell command in R studio was used to capture any of these sorts of errors. Similar to the comprehensiveness method, capturing relevance used text mining techniques and reviews not containing any of the words that fell into relevant subtopics such as Food, Drink, Service, etc were deemed irrelevant. Finally for valence, AFINN-111 and NRC lexicons were used to extract review sentiment and to what degree. AFINN is a manually labeled list of words by Finn Årup Nielsen in 2009–2011, rated for valence with an integer value between negative five and positive five. The NRC Word–Emotion Association Lexicon contains more than 14,182 unigrams (words), with each word being subcategorized into the eight dimensions of Plutchik as well as positive or negative. It's conclusive that recent academic and industry methods have been advancing among data collection and methodologies. Going from analyzing the reviewer to analyzing the review was bound to happen, but taking into consideration the valence of the review was another step that not only helped in determining the helpfulness of a review, but also required more advanced models and more detailed data.

When implementing this project as my own, I got a more accurate star rating for the new weight star rating that I implemented. I focused on Indian restaurants in Tucson, Arizona. This data set contained 16 Indian restaurants from that city and I was able to create a more balanced and accurate star rating for each of those 16 restaurants. For my implementation, I imported all the datasets from kaggle instead of the yelp website because I could individually download the data sets that I needed instead of the entire zip file, which included irrelevant information to my topic. After downloading and importing the files to the python file, I merged each data set into one Data Frame. Next, number of times a duplicated user ID was counted and collected, then a new data frame was created that only included the index of the user id and the number of times the user id shows up which was placed under a column named “Num_reviews.” The number of times a user id showed up in the data frame was how many reviews that user had given and that number would be used to add weight to the rating they gave to each restaurant. Next, that new data frame was merged with the review data set in order to create a new data frame that included which business the user had reviewed, the star rating that the user gave that business, and the new weighted star rating that I calculated called ‘w_star.’ This new weighted star rating was calculated by multipying the number of reviews the user gave by the star rating they gave to the restaurant. The more reviews, the higher the weight, but each w_star value was divided by the mean value of the number of reviews in order for each w_star to be relative to each other and thus more balanced. This was the trial data set, however, so none of the data sets were cleaned up to have only Indian restaurants in the Tucson area.

Next, was the implementation of the data set that we wanted to test on, which was Indian restaurants in Tucson, Arizona. To clean the data set, we made sure to drop all values we did not need from each data set including columns such as latitude, longitude, how many useful upvotes, etc. After that, the data set was exported to Microsoft Excel in order to drop the cities we didn’t want and then the restaurants within Tucson that we didn’t want. The cleaned up data set ended up having 16 restaurants all in the Tucson area and that had ‘Indian’ in the category column indicating an Indian restaurant, the name of the restaurant, the star rating, and how many reviews it had. Then, this data set was merged with the review data set and then user data set in order to see how many stars each user gave the restaurant along with how many reviews that user had done in general. Similar to what we had done previously, we counted the number of times each user id showed up in the data set in order to count how many reviews each user did on Indian restaurants in Tucson. This is somewhat flawed however because we did not account for the Indian restaurant reviews that the user could've done outside of Tucson, this may not be fully accurate of the user's supposed impact, but Tucson has enough Indian restaurants to justify one's credibility on Indian restaurants. After finding out how many reviews each user did on Indian restaurants in Tucson, we multiplied that number by the star rating that they gave to the restaurant that they reviewed and then divided that product by the mean value of the number of reviews or num_reviews.

After this multiplication, we can see the new weighted star rating of the Indian restaurants in Tucson. The results include most Indian restaurants having a lower star rating. Only two of the 16 restaurants got a better weighted star rating. Spice Garden Indian cuisine went from a 4.5 rating to a 5.0 weighted rating, while Bombol went from 4.5 to 4.9. Every other restaurant dropped with some being significant as 1 star while others dropped by 0.2 stars. From this method, 2 restaurants stood out from the rest. In addition, the weighted method changed the original star rating ranging from -1 to +0.5 stars.

Bibliography

Luca, Michael. *Reviews, Reputation, and Revenue: The Case of Yelp - Hbs.edu*. Harvard Business School , 2011, https://www.hbs.edu/ris/Publication%2520Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf.

Marinova, Iva. “25+ Groundbreaking Yelp Statistics to Make 2022 Count.” *Review42*, 7 Mar. 2022, https://review42.com/resources/yelp-statistics/.

Saumya, Sunil, et al. “Ranking Online Consumer Reviews.” *Ranking Online Consumer Reviews*, 17 Jan. 2019, https://arxiv-export-lb.library.cornell.edu/abs/1901.06274.

Spencer Soper and Jing Cao, Bloomberg December 26th. “Yelp's Struggles and the Evolution of Online User Reviews.” *Skift*, 26 Dec. 2015, https://skift.com/2015/12/26/yelps-struggles-and-the-evolution-of-online-user-reviews/.

Srivastava, Vartika, and Arti D. Kalro . “Enhancing the Helpfulness of Online Consumer Reviews: The Role of Latent (Content) Factors.” *Sage Journals*, Journal of Interactive Marketing, 31 Jan. 2022, https://journals.sagepub.com/doi/full/10.1016/j.intmar.2018.12.003.