Project 2 – Exploration of Restaurant Reviews

Team 2: W. Ching, D. Goon, J. Ling, M. Speckmann, O. Shapira, L. Vidiella, X. Wang

I. Introduction

Whenever one seeks a new place to eat, online reviews play a big factor in where they end up choosing to dine. How big? A 2017 report published by websitebuilder.org claimed that approximately 61% of customers have read online restaurant reviews and about 34% of diners choose restaurants based solely on online customer review websites[1]. Furthermore, a 2016 Harvard Business School study found that a one star improvement on Yelp leads to an average increase in revenue of anywhere from 5-9%[2].

But while accessing an overall star rating for a restaurant is easy enough, it is not easy to figure out factors that contributed to the star rating like food quality, cost, ambience and service. Restaurant reviews may also vary between different websites, such as Yelp, TripAdvisor or Google. Overall, making a well-informed, yet quick decision on where to eat may be a harder task than initially realized.

Through the use of web scraping, social media mining, and text mining analysis, we hope to better understand customer sentiment associated with different star ratings. We will stratify this sentiment analysis by different websites, locations, cuisines, and specific aspects of a restaurant. Lastly, we plan to compare how the association between sentiment and star rating differs between websites like Yelp and Google reviews.

II. Gathering Data & Pre-Processing

Social Media Mining – Yelp Fusion API

Yelp is the most popular social media website for restaurant reviews and the company offers an open-source "Yelp Fusion" API [3] library developed for R. By directly interfacing with the Yelp dataset, we were hoping to extract real-time reviews rather than using a static dataset that may contain outdated information. To use the API, we created a new application, which generated a unique client and token ID. Those tokens are used anytime a "call" was made to the Yelp website. From there, we can run queries on different search criterion.

---

[1] https://www.modernrestaurantmanagement.com/the-impact-of-reviews-on-the-restaurant-market-infographic/

[2] https://www.hbs.edu/faculty/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf

[3] https://www.yelp.com/fusion

Figure 1: This is a sample Yelp API search result for Mexican restaurants in Phoenix, AZ and a few reviews for a sample restaurant.

business_id	alias	name	rating	review_count	price
frCxZS7IPhEnQRJ3UY6m7A	la-santisima-phoenix	La Santisima	4	1978	$$
9MVKjEMN5T59uzG1xoD2BQ	cocina-madrigal-phoenix	Cocina Madrigal	5	198	$$
DaVTuhzi6EgWStb2eAjNjA	presidio-cocina-mexicana-phoenix	Presidio Cocina Mexicana	4.5	559	$$
ELxoOY2F-1jSb8mqyN7Zbg	baja-fresh-mexican-grill-phoenix-2	Baja Fresh Mexican Grill	4	37	$
mss-LiOfL1vtoNo3WoEoJw	kiss-pollos-estilo-sinaloa-phoenix	Kiss Pollos Estilo Sinaloa	5	175	$
Cd2ERofxOeCbCi-0yDs0aw	marias-frybread-and-mexican-food-phoenix-2	Maria's Frybread & Mexican Food	4.5	417	$
odIM0YsqEuZ_3kW44A4h_w	casa-corazon-restaurant-phoenix	Casa Corazon Restaurant	4.5	149	$$
pDewiJY6KCcZgLxxgxg13Q	comedor-guadalajara-phoenix	Comedor Guadalajara	4	498	$$
1JF9TbJ2d5hH8xsQvvkIHg	irmas-kitchen-phoenix-2	Irma's Kitchen	4.5	250	$
iXV2kGGyXJIKM8JKek-HNg	centrico-phoenix	Centrico	4.5	113	$$
rbywwnfMjFjJwLQkgpdIEw	ritos-mexican-food-phoenix-phoenix-3	Rito's Mexican Food- Phoenix	4.5	351	$
89uU51kOiQXbJHVA3C6XMQ	the-original-carolinas-mexican-food-phoenix	The Original Carolina's Mexican Food	4	646	$
eS29S_06IvsDW04wVrIVxg	barrio-café-phoenix-2	Barrio Café	4	1418	$$
T-KniGykrZ46ZC9pIOTspw	phx-burrito-house-phoenix	PHX Burrito House	4.5	563	$
_iEI9sCLsvXEFHUWPvgsAg	chico-malo-phoenix-3	Chico Malo	4	378	$$

reviews.id	reviews.url	reviews.text	reviews.rating
RDFb3tZwuArGXkICEiCGQg	https://www.yelp.com/biz/la-santisima-phoenix?hrid=R...	When we are in the mood for mole tacos, this is where we end up... on top of that ...	5
RDFb3tZwuArGXkICEiCGQg	https://www.yelp.com/biz/la-santisima-phoenix?hrid=R...	When we are in the mood for mole tacos, this is where we end up... on top of that ...	5

There are a few key limitations with the Yelp API that led to the determination to not use this source for subsequent text or sentiment analysis. Firstly, when running a query to find businesses, only a maximum of 50 rows can be returned. For example, if searching for Chinese restaurants in Boston, the API will only return 50 distinct restaurants which omits many other restaurants. Secondly, the review text field has a character limit of 160 characters for extracted reviews, therefore often truncating the full review. This is unreliable for any subsequent text or sentiment analysis. Lastly, only up to 3 reviews can be extracted for a specific restaurant, which would make it difficult to make a substantial conclusion on specific businesses with such a small sample size of reviews.

Social Media Mining - Twitter API

Another social media source that we attempted to use is the Twitter API tool to get customer reviews from their tweets. The first step to extract tweets from Twitter is to set up a new application and token. The “rtweet” package was used to access tweets.

When using the chain restaurant “Legal Seafood” as a keyword search, the Search_tweets function returns 88 variables which included user id, created time, sources, etc. The most important aspect is the text column which are the actual tweets mentioning “Legal Seafood”. Once again, there are limitations with the Twitter API that make it an unreliable method for capturing data for text analysis. Firstly, only tweets from the last 6-9 days can be accessed, which would introduce a recency bias. Secondly, only 18,000 tweets can be requested in one call. While 18,000 tweets would likely have sufficed for this analysis, the 6-9 day limitation was one of the main reasons that we decided to not go through with using the Twitter API. Secondly, there isn’t a way to clearly indicate which “Legal Seafood” restaurant the tweets

apply to. Another possible avenue would have have been to identify the twitter handle (username) of the restaurant--if it existed--but we did not find a review site that had this information readily available for us to use.

Figure 2: This is an example result by using "Legal Seafood" as a keyword search

	user_id	status_id	created_at	screen_name	text
1	2912695064	1069310568211902464	2018-12-02 19:21:20	Marcoangel2000	#Dolphin cottage located near downtown #Goodland. ...
2	6736882	1069024450589483009	2018-12-02 00:24:25	monovalent	Very Legal Seafood, Very Cool Seafood
3	949436952784310272	1068868531951939584	2018-12-01 14:04:51	LindseyKook	@petridishes Seafood is very legal?
4	3002215918	1068735492559122432	2018-12-01 05:16:12	maddsnacks	IT'S DEC 1! Wrapping up the year with bday, green car...
5	2361046392	1068711923410264064	2018-12-01 03:42:32	mikedolanindc	@JahHills Unless you're Legal Seafood in 1973.
6	15339975	1068685183266824193	2018-12-01 01:56:17	RedsArmy_John	Pro Tip: If you want $50 to Legal Seafood, bring a bab...
7	348174495	1068684166034845697	2018-12-01 01:52:15	DavidFutrelle	legal seafood and cool seafood
8	17000080	1068660778679582720	2018-12-01 00:19:19	Loroma	Yes I can. Because you're my favorite male skater. Leg...
9	298221517	1068632045469995009	2018-11-30 22:25:08	RaqWinchester	@petridishes I appreciate Very Legal Seafood.
10	1056243391355248641	1068629588475969536	2018-11-30 22:15:22	zneeley25	@petridishes Something very legal "Seafood" You are ...
11	22036565	1068621110223220737	2018-11-30 21:41:41	zpleat	@petridishes lol legal seafood
12	350704176	1068621092162392064	2018-11-30 21:41:37	real_KFab	@petridishes how did seafood make it onto Very Legal
13	105348643	1068546110371872769	2018-11-30 16:43:39	msjenNjuicee	@Xhelsea_ @Modern_Pharaoh That's an OUTING to yo...
14	61667933	1068541660047527937	2018-11-30 16:25:58	h4nd	@MrPope sketchy ass competitor to legal seafood
15	2349807652	1067878357642223624	2018-11-28 20:30:15	kieshakiesha391	@blvckngld Or fish....I swear white people always sme...
16	3267589964	1067835517599772672	2018-11-28 17:40:01	adaughma	I will sell my firstborn child for fried clam strips. Legal...

Web Scraping

The purpose of using web scraping was to aggregate a corpus of user reviews for specific restaurants across different websites (e.g. Yelp, Google, TripAdvisor). Web scraping was done in both Python and R.

I. Web Scraping – Python

In Python, the "beautifulsoup" web mining package was used for the scraping of text reviews from TripAdvisor. First, the HTML code of the webpage was retrieved using the "urllib2" package. Next, the page is then parsed into "beautifulsoup" format so that the package can be called to work on it. The coding was done on a Jupyter notebook hosted on the Anaconda Platform for the ease of reading which makes it easier for collaboration on Github and to avoid using pip to manage the packages.

While we were able to scrape the restaurant reviews from the TripAdvisor website, we encountered limitations with web scraping as well. The first issue is that our code was only able to scrape the summary of each review which was the text displayed on the HTML of the website when "urllib2" retrieved it. The second issue is that we were unable to loop through all the pages of the review portion of the website to retrieve all reviews. The way that the website is structured means that all the different pages still share the same URL. This means that the urllib2 library is not able to retrieve the HTML code of the other pages.

II. Web Scraping – R

For R, we used the tidyverse package to scrape the reviews from the XML files of the website of interest. We used a web crawler tool called WebSPHINX which is written in Java to retrieve the XML file of the website. The WebSPHINX looped through and imported a defined list of web URLs in XML or CSV format and converted the extracted content into a data frame. Next the files were put into a directory whose file path is then passed on to our R code for scraping to occur. The result was a text output in .txt format with the information required for the analysis.

The ultimate goal was to see how text and sentiment analysis differed between websites for the same handful of chosen restaurants. Although we attempted the scraping in both Python and R, we encountered the common key limitations in that we are only able to scrap what is displayed on the html of the website.

For these reasons above, we once again had to disregard web scraping as a reliable method for subsequent analysis. We also couldn't compare reviews for specific restaurants across different websites.

Yelp Dataset

The primary data source we ended up using for text analysis was a Yelp dataset that is publicly available for free [4]. In this .tar collection of JSON files, we used two datasets:

1. A dataframe of ~200,000 distinct businesses and associated metadata (e.g. business name, category, location, etc.)
2. A dataframe of ~6 million rows containing full text of user reviews and star ratings.

Below are screenshots of the "businesses" and "reviews" datasets, which were eventually merged by joining on the primary key "business_id", and then reformatted and queried as needed.

	business_id	name	neighborhood	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open
1	Apn5Q_b6Nz61Tq4XzPdf9A	Minhas Micro Brewery		1314 44 Avenue NE	Calgary	AB	T2E 6L6	51.0918130155	-114.031674872	4	24	1
2	AjEbIBw6ZFfln7ePHha9PA	CK'S BBQ & Catering			Henderson	NV	89002	35.9607337	-114.939821	4.5	3	0
3	O8S5hYJ1SMc8fA4QBtVujA	La Bastringue	Rosemont-La Petite-Patrie	1335 rue Beaubien E	MontrÃ©al	QC	H2G 1K7	45.5405031	-73.5993003	4	5	0
4	bFzdJJ3wp3PZssNEsyU23g	Geico Insurance		211 W Monroe St	Phoenix	AZ	85003	33.4499993	-112.0769793	1.5	8	1
5	8USyCYqpScwiNEb58Bt6CA	Action Engine		2005 Alyth Place SE	Calgary	AB	T2H 0N5	51.0355914	-114.0273656	2	4	1
6	45bWSZtniwPRiqlivpS8Og	The Coffee Bean & Tea Leaf		20235 N Cave Creek Rd, Ste 1115	Phoenix	AZ	85024	33.6713751	-112.0300171	4	63	1

	review_id	user_id	business_id	stars	date	text	useful	funny	cool
1	x7mDIiDB3jEiPGPHOmDzyw	msQe1u7Z_XuqjGoqhB0J5g	iCQpiavjjPzJ5_3gPD5Ebg	2	2011-02-25	The pizza was okay. Not the best I've had. I prefer Biaggi...	0	0	0
2	dDl8zu1vWPdKGihJrwQbpw	msQe1u7Z_XuqjGoqhB0J5g	pomGBqfbxcqPv14c3XH-ZQ	5	2012-11-13	I love this place! My fiance And I go here atleast once a ...	0	0	0
3	LZp4UX5zK3e-c5ZGSeo3kA	msQe1u7Z_XuqjGoqhB0J5g	jtQARsP6P-LbkyjbO1qNGg	1	2014-10-23	Terrible. Dry corn bread. Rib tips were all fat and mushy ...	3	1	1
4	Er4NBWCmCD4nM8_p1GRdow	msQe1u7Z_XuqjGoqhB0J5g	elqbBhBfElMN5rjFqW3now	2	2011-02-25	Back in 2005-2007 this place was my FAVORITE thai place...	2	0	0
5	jsDu6QEJHbwP2Blom1PLCA	msQe1u7Z_XuqjGoqhB0J5g	Ums3gaP2qM3W1XcA5r6SsQ	5	2014-09-05	Delicious healthy food. The steak is amazing. Fish and p...	0	0	0
6	pfavA0hr3nyqO61oupj-IA	msQe1u7Z_XuqjGoqhB0J5g	vgfdVK81oD4r50NMjU2Ag	1	2011-02-25	This place sucks. The customer service is horrible. They d...	2	0	0

There were some technical obstacles and limitations with this dataset, which included:

[4] https://www.yelp.com/dataset

1. The dataset was so large, that our computers didn't have enough memory to read all the rows in the JSON reviews file without crashing. We chose to only read the first 2 million rows as a compromise.
2. Because this is a static dataset, our analysis on specific restaurants may be slightly outdated.

III. Text Mining Analysis

Analysis #1 - Sentiment Analysis

Data Preparation – For our initial sentiment analysis, we looked at all the text reviews in rows extracted from the Yelp dataset. For efficiency purposes, we combined the character vector into a single JSON string and the flattened it. We referenced an AFINN compendium that gives a list of English words an integer score of 1-5 based on how positive the sentiment is. We were curious to see how text sentiment compared to the star ratings given.

Figure 3 – Words vs Av Stars & # of Review

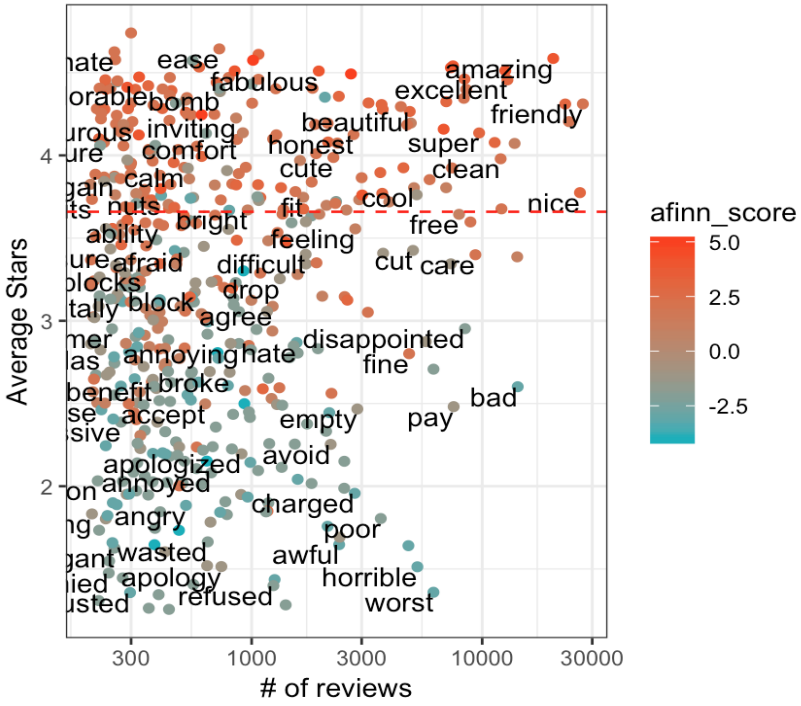
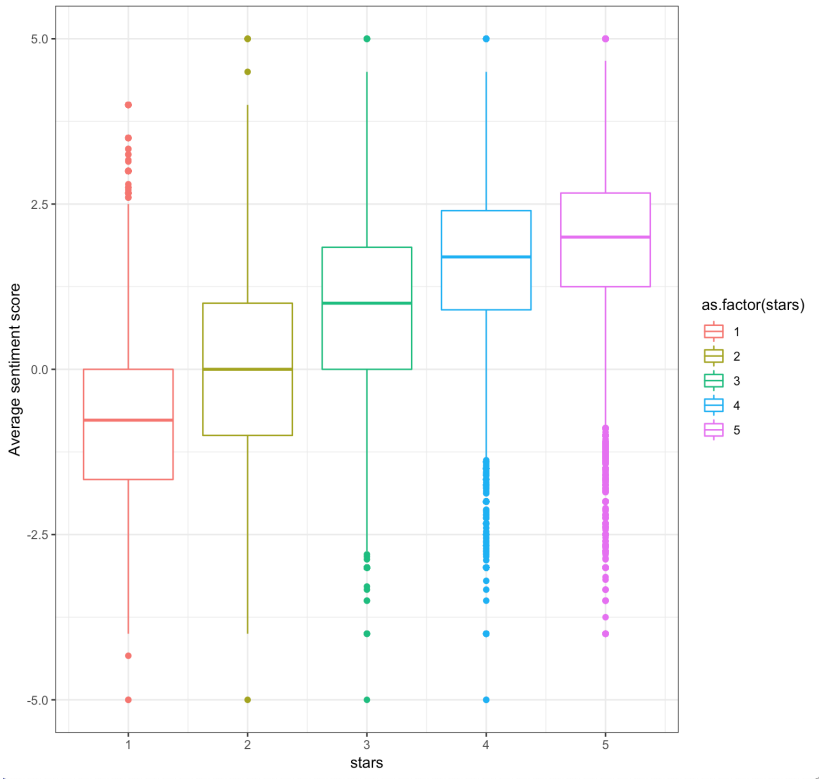


Figure 4 – Average AFINN score vs Stars



Data Visualization - The **Figure 3** above plots the rating of AFINN words compared to the average stars rating change when the number of reviews increase, as well as a line showing where the average score and rating is. **Figure 4** shows how reviews are correlated with the AFINN words using box plot.

Observations - In **Figure 4**, we can clearly see a correlation between how the AFINN score and customer's star rating. As one would expect, restaurants with more stars have a higher median sentiment score. However, there are a lot of outlier data points.

Since it can be tough to digest visualizations with so many data points, we decided to narrow our focus on the two cities with the highest number of restaurant reviews in the dataset rows we extracted: Las Vegas and Phoenix. We conducted a few types of analyses for these cities.

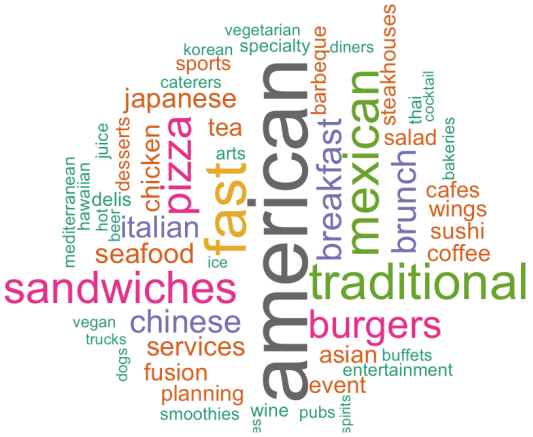
Analysis #2 – When comparing Phoenix, AZ and Las Vegas, NV, what type of categories are most frequent? Are there any similarities between the two cities?

Data Preparation – We extracted all businesses from the business dataset that has the city of Las Vegas or Phoenix listed as well as ensured the word "Restaurant" is included in the yelp category. Then we extracted all categories from the new dataset and removed the following words as they do not have value to highlighting the cuisine type: Restaurants, Food, Nightlife, Bars, New.

Data Visualization – A wordcloud was created to show the most frequent categories across the two cities. In order for the category to be considered, it must have shown up a minimum of 100 times in the dataset. The most frequent categories appear in the center in large text and as we move outward, the word frequency decreases with the size of the word and its position.

Figure 5- Most Frequent Categories for Phoenix, AZ

Figure 6- Most Frequent Categories for Las Vegas

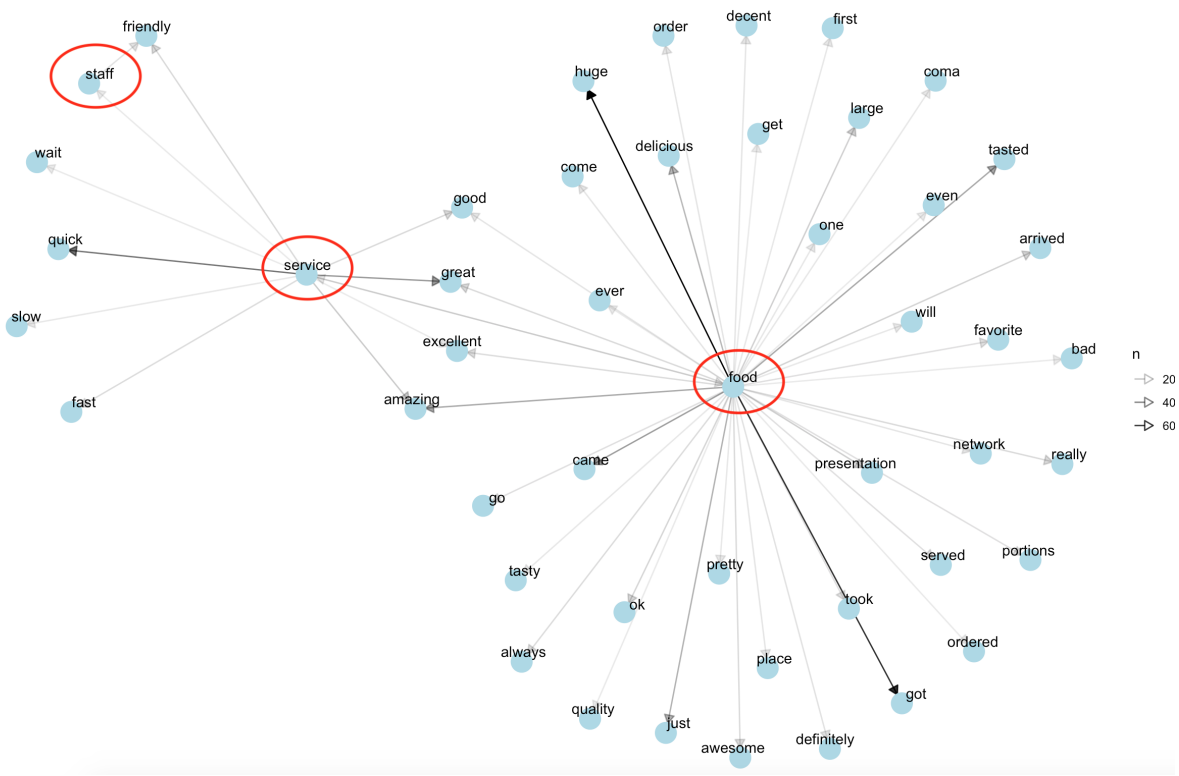


Observations – Right away, it is obvious that both Phoenix and Vegas have equal representation of American themed restaurants. It is also apparent that while Mexican food is more popular in Phoenix, Las Vegas has better representation of all types of categories of food. Surprisingly, brunch seems to be more popular in Phoenix than in Las Vegas, despite all of the popular and well-known buffets on the Las Vegas strip.

Analysis #3 - Sentiment Analysis for the two top reviewed restaurants in Vegas: Which restaurant is rated higher for food, staff, service and atmosphere?

Data Preparation – By using the R filter and table functions, we identified that the top 2 restaurants with the most reviews were Hash House A Go Go with 1,253 reviews and Mon Ami Gabi with 1,110 reviews. A dataset was then created for each restaurant, and review texted was tweaked so that all characters were set to lowercase, and so all numbers and stop words were removed. Next, we analyzed the words that appear around the words food, service, ambience and staff and made a network graph. In order to reduce noise of insignificant relationships for these bigrams, the descriptive words had to appear at least 5 times to be considered. Below is a sample of the network graph in **Figure 7**:

Figure 7 - Network Graph of Hash House A Go Go in Las Vegas, NV



The thickness of the arrow correlated to the number of times the word appears around the category. Service is linked to both quick and great with thick lines. There are large number of reviews centered around food quality. Service and food also have some common words used in the reviews. Atmosphere is not showing up in this network graph which means the volume of words around that category is less than 20.

Now that the bigrams have been collected, an analysis of these four areas of a restaurant can be performed. Using the review stars, AFINN lexicon that assigns words a score of -5 to 5 for sentiment, and the bigrams, the top 5 positive and negative words for an attribute can be shown across review star ratings.

Figure 8 - Sentiment Analysis – Services: Hash House A Go Go in Las Vegas, NV

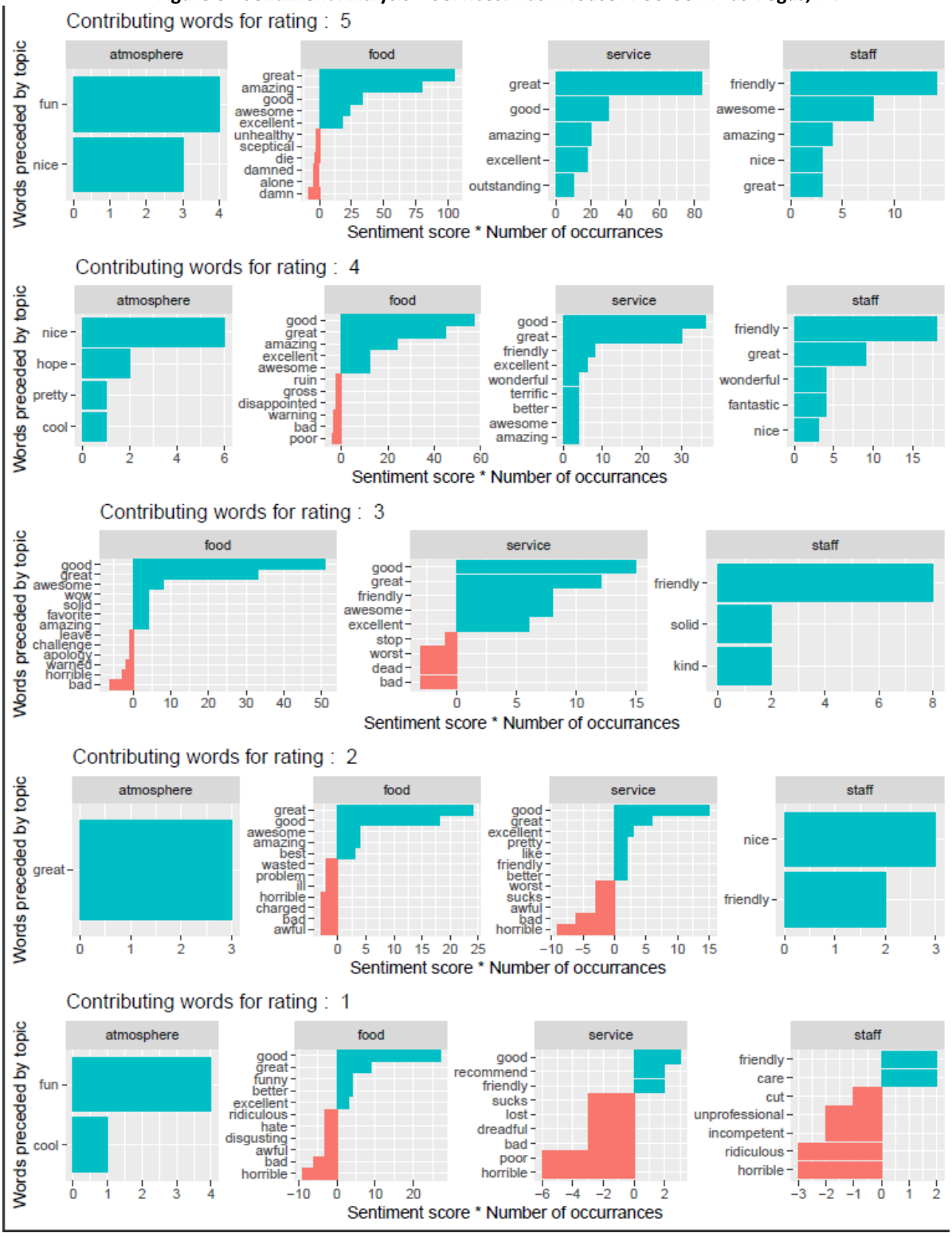


Figure 9 - Sentiment Analysis – Services: Mon Ami Gabi in Las Vegas, NV

Observations – When comparing sentiment for food across reviews for both restaurants, we see that the food is considered “great” and “good” in high volume even though the overall

review may be rated poorly (1 or 2 stars). Atmosphere received only positive words but was nonexistent in Mon Ami Gabi's 1 and 2 star ratings. This could mean that the word atmosphere isn't a good indicator for the decor and vibe a restaurant as it may only be associated with happy customers. Both received overall good score in service and both had some negative words associated to service as the star rating dropped. Mon Ami Gabi performed a bit better in staff sentiment as some of the negative words may not be considered negative like "pay" and "shy" depending on context. Hash House A Go Go also had 5 negative words in the star rating of 1 that confidently implied that the staff did not perform well in the eyes of the reviewer.

IV. Conclusion

The research goal was to investigate sentiment of restaurant reviews across different website platforms as well as investigate various factors (e.g. food quality, cost, atmosphere, service, etc.). that cause sentiment to be positive or negative. Throughout this process, our team encountered many limitations and some of the conclusions we found were unexpected.

Our limitations with accessing data from various websites created our inability to assess sentiment across different website review platforms. Twitter and Yelp API both were limited with recency bias and did not give us access to a large amount of review data for restaurants. Web scraping also proved to be ineffective as again we were unable to gather the full set of reviews--and even a single review with all its text--for any given restaurant. This then left us with a free static Yelp dataset that we were also limited in using as it was too large to digest all of the data with only the memory and storage of our laptops.

However, we were able to use the Yelp dataset in a productive way and draw some important conclusions that might be useful for restaurant owners on how to improve their online reviews and what factors might influence a customer giving a higher star rating. We found that positive words that related to the restaurant atmosphere aligned with a higher star rating.

We also learned that sentiment within user reviews seemed to align with star rating for the most part (though our scope of analysis was limited). However, sentiment of words seemed to be a bit more scattered when trying to associate it with different aspects of a given restaurant.

With more time and resources, next steps would include accessing more current and relevant data. We were able to build some tools for restaurant analysis with text mining, accessing APIs and web scraping. However, data has proven to not be easily accessible and sites want to charge for access. Thus, the next steps would be to pay for a subscription for products to access the full and live datasets of the review websites. Another option would be to speak with company representatives and see if we could be given access for research and educational purposes only. Lastly, in an effort to continue our analysis of reviews, we would next look at emotions across star ratings to see if we could deepen our understanding of sentiment.

V. Sources

1. Arevalo, Megan. (2017, April). The Impact of Reviews on the Restaurant Market (infographic). Retrieved from www.modernrestaurantmanagement.com

2. 2. Luca, Michael. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.com. Retrived from www.hbs.edu

3. Vallikunnel, A. (2018, March 25). Sign In. Retrieved November 21, 2018, from https://rpubs.com/AnithaVallikunnel/RestaurantAnalysis

4. Robinson, David. (2016, July 21). Does sentiment analysis work? A tidy analysis of Yelp reviews. Retrieved from: http://varianceexplained.org/r/yelp-sentiment/

5. Borole, Kaustubh. (2018, October 8). Web Scraping Yelp, Text Mining and Sentiment Analysis for Restaurant Reviews. Retrieved from: https://medium.com/@kborole7/web-scraping-yelp-text-mining-and-sentiment-analysis-for-restaurant-reviews-ea500e1ef84d

6. Coursera. (2015, October 18). Analysis of Mexican Restaurant Reviews with Yelp Data Challenge Data set. Retrieved from: http://rstudio-pubs-static.s3.amazonaws.com/121639_3364a2eb69b54ed9b85faf1ecf21cd7f.html

7. Kaggle. (N/A). A Very Extensive Data Analysis of Yelp. Retrieved from: https://www.kaggle.com/ambarish/a-very-extensive-data-analysis-of-yelp