Classifying Sentiment and Topic Differences Between Yelp Tips and Reviews

Marcus Scott, Abbas Tahirzadeh, Alex Ngo, Rhea Arora

***Abstract*— In this paper, we examine a generally underrepresented Yelp feature called "tips" through a variety of techniques to better understand their value proposition. Using the Yelp open dataset, we examined and compared the text content of nearly 1.6 million reviews and 500,000 tips through Latent Dirichlet Allocation (LDA) topic modeling and sentiment analysis to try and predict which feature affects businesses more. We present a breakdown of the topics and the sentiments of tips and reviews, revealing a subtle but noticeable difference in content. We conclude with an outline of possible future work, and suggestions for businesses and Yelp on how to better utilize tips with the results of our investigation.**

I. Introduction

Yelp's most ubiquitous feature is undoubtedly its reviews; within their open dataset alone, the provided 1.6 million reviews make up nearly half of the three gigabytes of data. However, another significant part of this open dataset are the five-hundred thousand tips that Yelp also makes available.

Yelp defines a tip as "a way to pass along some key information about a business -- such as the best time to go or your favorite dish -- without writing a full review about your experiences" [1]. A tip may focus on things unrelated to a customer's overall experience and highlight facts not quickly discernable from a review; for example, a user may leave a tip to inform users that a restaurant is open late and they prefer cash. However, these tips do not actually appear on Yelp's website; instead, users can only view these tips on their mobile device, and are located at the very bottom of the mobile applications or the mobile version of their website. While the reviews outnumber tips in the dataset 3.2:1, this ratio is actually quite low; none of the authors were aware that tips even existed, and for the dataset to provide five-hundred thousand tips was something worth exploring further.

Our motivation for focusing on tips came from the lack of clarity or obvious benefit that tips provided in comparison to reviews; this was especially of interest when we noted that tips were not featured prominently in any particular way. How do tips affect businesses, if at all? Are there any discernable difference between the content of tips and the content of reviews? If there are seemingly no differences, should Yelp discontinue support for tips, or should they ultimately provide a different way to look at these tips? We aim to answer these questions throughout our paper by extracting a variety of topics and general sentiments found in reviews and texts, in order to ultimately compare the differences found.

II. Related Work

In general, there are very few other related works that focus on tips in particular. In a previous submission for Yelp's data challenge, students from the University of California, San Diego also attempted build a regression around tip sentiment to predict the star rating of the particular business [2]. They used a variety of approaches when building their regressions, such as using common positive and negative unigrams and bigrams, which all had different levels of success in predicting the star rating through the tip sentiment.

Other entrants also took advantage of topic modeling with LDA; a group of students from the University of California Berkeley used reviews in the dataset to predict the demand of customers [3]. They extracted latent subtopics by running an online LDA algorithm in order to determine what customers care about the moss. They broke down hidden topics they determined through using LDA and predicted how many stars each hidden topic would receive. They concluded that many of these hidden topics – service, value, décor – had a noticeable effect on the overall star rating of the business.

Another paper by a group of students from Stanford University were interested in uncovering latent factors in review texts in order to predict ratings through these hidden factors/topics. [4]. While they did not use the Yelp dataset as their sole dataset, they were able to uncover many hidden factors in the product ratings and hidden topics in reviews that were aligned closely. This gave them a more accurate prediction than other current existing models when trying to predict the rating of a product based off of topics in reviews.

III. Data and Methodology

A. Dataset

The dataset used in our research was provided by Yelp for its Dataset Challenge [5], in which they provide roughly 2-3 gigabytes of JSON data for various facets of a business. Data was gathered by Yelp in ten cities in the US, Canada and Europe. In the JSON files, there are objects for 61,000 businesses, 1.6 million reviews, 366,000 users, check-ins, and 500,000 tips. The data begins for some business as early as 2004, and as recently as 2015. In our analysis, the three main datasets we used were reviews, business, and tips. A business consists of a variety of features, but our research focused

mostly on the star rating and review count for a given business. From each tip and review, we gathered the "text" feature to run our sentiment and topic analysis, and joined on the "business_id" feature in order to correlate specific businesses with the tips or reviews left.

B. Tools

The data was first put into a RethinkDB instance for easy querying and sorting of data. Python scripts were then used to create CSV files, create topics, and analyze sentiment. Packages used in the Python scripts were the Natural Language Toolkit (NLTK) for removal of stop words, textmining for creation of term-document matrices to be used with the LDA package, and TextBlob for sentiment analysis. R was used for the visualizations, statistics, and running regressions on our data.

C. Topic Modeling with LDA

The topics created were modeled in two different ways: random sampling of tips/reviews and topics built from businesses with the top number of tips and reviews. Stop words were removed when creating topics to better encapsulate user meanings in the topics. For each business or sample, ten topics were created, with eight words making up each topic. Due to some user error, however, some words are broken up or incomprehensible. For example, the following topic below was generated through our script:

```
Topic: s d c est l m e
```

This topic was generated through our Python script and is generally not comprehensible – this could be because one city in the dataset was Montreal, where French reviews are more common. In scenarios where topics are completely incomprehensible, but in cases where the word (e.g. "don" and "t" are in the same topic) is obvious, we have chosen to clean the topics up to ensure clarity. A raw topic may look like this:

```
Topic:
t time get would back one i didn got
```

While a cleaned topic would look like so:

```
Topic:
time get would back one i didn't got
```

Although these topics are not as clear as other forms of topic modeling, this may be due to the brevity of many tips (and reviews); our goal is not to necessarily have fully comprehensible single topics, but to understand the general ideas conveyed in multiple topics.

IV. Results

A. Preliminary Data Analysis

Before fully diving into topic modeling and sentiment analysis, we first investigated some of the basic features of the data. We first determined the top ten and top fifty words found in tips and reviews after removing stop words. Although the top ten words (Figure 1) contain a good handful of positive words – like, love, best, good, great – we cannot say that tips are, as a whole, positive. This is because factors like word combinations (e.g. "not great", "did not love") do not show up in any figure. Without the word combinations that imply negative connotation, we cannot completely come to the conclusion that tips are used positively. However, we can examine a broader number of words to gain a better sense of tips and their connotations. Looking at the top fifty words gives a better idea of the general sentiment of tips: the word "don't" appears at number thirteen, while "favorite," "amazing," and the ":)" emoticon also appear, implying that tips might usually be used positively rather than negatively.

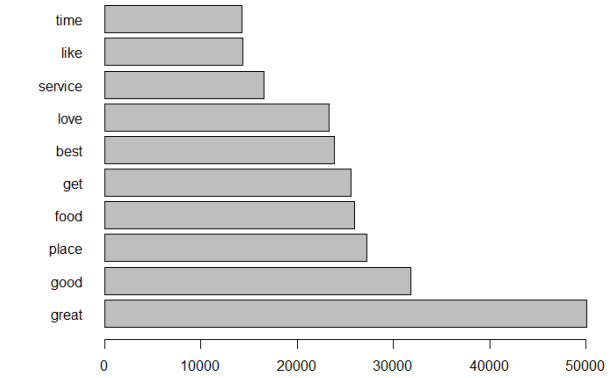

Figure 1 – Top ten words found in tips, stop words removed

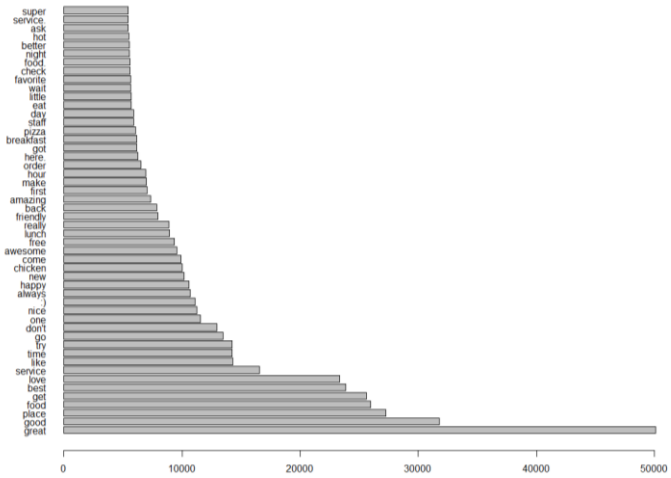

Figure 2 – Top fifty words found in tips, stop words removed

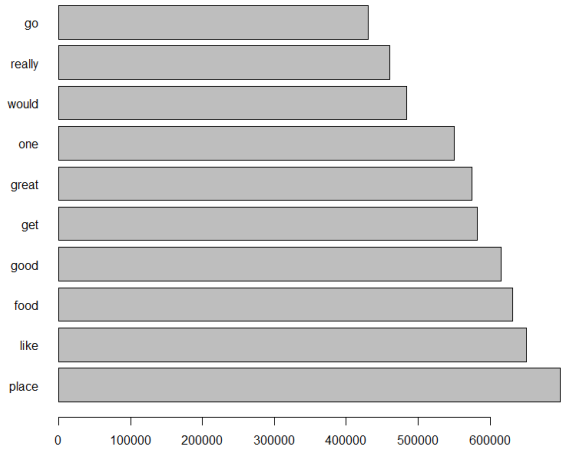

Figure 3 – Top ten words found in reviews, stop words removed

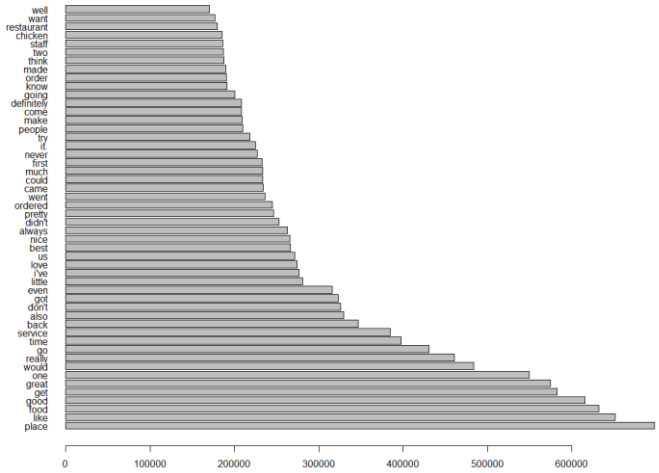

Figure 4 – Top fifty words found in reviews, stop words removed

Comparing the top ten words found in tips to the top ten words in reviews, a few noticable differences can be seen. Although many words are shared – "great", "good", "food", "like", "place" – some are unique to the top ten of others. The word "love," for example, is not found in the top ten words of reviews (it ends up being number 20), while "service" and "time" are not in the top ten for reviews (although they are 11th and 12th). From this first pass of text analysis, we can see that there are some small differences between tips and reviews that could be important.

Before diving deeper into the text content, it is worthwhile to note what businesses have the most reviews and tips when we do our topic modeling comparison to see what kinds of businesses have reviews and tips left frequently.

Name	Review Count	Star Rating
Mon Ami Gabi	4578	4
Earl of Sandwich	3984	4.5
Wicked Spoon	3828	3.5
Bacchanal Buffet	3046	4
Serendipity 3	3007	3

Name	Tip Count	Star Rating
McCarran Intl. Airport	2765	3.5
Phoenix Sky Harbor Intl. Airport	1849	3.5
The Cosmopolitan of Las Vegas	1092	4
Earl of Sandwich	1076	4.5
Wicked Spoon	954	3.5

Figure 5 – Top five businesses, ordered by review count and tip count

All five businesses with the most reviews are restaurants in the Las Vegas area, whereas the top three businesses with the most tips are two airports and a hotel. This also ends up being a key distinction for our comparison; because services have tips left more often, determining how the topics differ for services (like airports) when using reviews versus tips can shed light on how tips and reviews actually differ.

B. Topic Modeling

1) Comparing topics from top businesses

First, we wanted to compare topic differences between reviews and tips for the business with the top number of reviews (Mon Ami Gabi), the business with the top number of tips (McCarran International), and a business that they had in common within the top five (Earl of Sandwich).

```
Topic: good like food french place i
Topic: table time would seated ami wait
Topic: great food vegas bellagio place
service patio
Topic: breakfast eggs brunch good toast
benedict great
Topic: steak good bread butter sauce
delicious cheese
```

Figure 6 – Five topics for Mon Ami Gabi from review text corpus

```
Topic: benedict eggs steak try get
breakfast cheese
Topic: make dinner reservation first time
reservations come
Topic: patio view bellagio outside sit
great seating
Topic: french steak soup frites onion bar
escargot
Topic: food great service good best place
```

Figure 7 – Five topics for Mon Ami Gabi from tip text corpus

Topics from both tips and reviews naturally cover the food available at Mon Ami Gabi – French food, steak, breakfast eggs, and so forth. However, the tips have three topics not strictly related to the food itself; the second topic mentions the need to make a reservation, the third topic mentions the view from the patio, and the final topic seemingly praises the service. The reviews, on the other hand, also touch on those topics in the second and third topic, but do not seem to generally suggest as clearly that the patio is better seating nor to make a reservation.

We can also see this general trend when comparing the topics at the Earl of Sandwich restaurant in Figures 8 and 9.

```
Topic: sandwich good place got like
sandwiches
Topic: sandwich earl vegas one
sandwiches tuna
Topic: sandwich bread sauce beef turkey c
heese earl
Topic: sandwich food sandwiches place
strip planet good
```

```
Topic: sandwiches place sandwich vegas
good great get
```

Figure 8 – Five topics for Earl of Sandwich from review text corpus

```
Topic: good open lunch sandwiches hours l
ate food
Topic: tuna melt earl sandwich best
club
Topic: sandwich soup sandwiches tomato go
od best delicious
Topic: chicken full sandwich montagu
jerk chipotle turkey
Topic: line must vegas long fast first
```

Figure 9 – Five topics for Earl of Sandwich from tip text corpus

Again, we see the topics from the tips as more directly notifying other customers about the business itself – the final topic from the tips mentions "long" and "line," and the first topic mentions the hours and that the business is "open late." The topics generated from the reviews, on the other hand, solely focus on the sandwiches; every topic has the word "sandwich" and does not seem to directly relate to the business itself, but the food.

Finally, we look at the topics created for the McCarron International Airport in the figures below.

```
Topic: airport terminal security get food
time
Topic: airport vegas slot you machines
get
Topic: line security get gate flight
people
Topic: flight a burger put
Topic: baggage shuttle car claim get
rental terminal
```

Figure 10 – Five topics for McCarron Intl. from review text corpus

```
Topic: home back going time let's bound
Topic: flight airport slot machines get
vegas
Topic: terminal free airport new
wifi place
Topic: vegas bye las see baby hello time
Topic: security check get long early
airport line
```

Figure 11 – Five topics for McCarron Intl. from tip text corpus

Another direct suggestion can be found in the final topic from the tips – long lines at the security check. The third topic also mentions Wi-Fi, which many people may be looking for at an airport. Interestingly, the rental car services at the airport seem to be a topic covered in reviews, but not in the tips; this may be due to the fact that users are reviewing the rental car service desks, rather than the airport itself. In this case, users might simply not have suggestions about the rental car services and may have had a bad experience, leading them to review the airport instead. While there are more topics that overlap between reviews and tips in this instance, the distinction of users "suggesting" more often in tips remains evident.

2) Random Sampling

Although 1.6 million reviews and five-hundred thousand tips were provided, creating topics for the entire corpus of text for each feature was computationally exhaustive given the authors' current technology. As a workaround, we randomly chose either ten-thousand or fifty-thousand reviews or tips to then create ten topics from; we ran the ten-thousand sample multiple times while only running the fifty-thousand sample once. We determined that the ten-thousand sample topics and the fifty-thousand sample topics were similar enough to each other that we could run the ten-thousand samples multiple times.

```
Topic: place bar great good night
drinks really
Topic: room hotel vegas nice stay the
strip
Topic: I've like places don't
Topic: food good sushi chicken rice
restaurant fish
Topic: store s coffee selection shop
love items
Topic: good cheese ordered pizza burger s
auce food
Topic: great time i car work service
get
Topic: back would time one get didn't
Topic: show see vegas great kids time
Topic: food great good place service
love always
```

Figure 11 – Ten topics from ten thousand randomly sampled reviews

```
Topic: chicken salad try good cheese soup
fried
Topic: best vegas place ever i favorite
love
Topic: happy hour night day open pm lunch
Topic: good great food beer menu lunch
pizza
Topic: good try tea hot love chocolate
ice
Topic: get make re free you sure check
Topic: s nice place parking area like
location
Topic: time here first s home back
place
```

```
Topic: great service food good place
friendly nice
Topic: don't can get like it place
```

Figure 12 – Ten topics from ten thousand randomly sampled tips

We can see similar topics in both reviews and tips at the ten-thousand sample, mostly covering food and the users' likes and dislikes. But again, we can see some general suggestions that are more obvious in the corpus from tips; suggestions like "make sure you check" and to "try" specific menu items. References to the hours of business operation are also another topic that has been seen before. Similar trends can be noted in the fifty-thousand samples.

```
Topic: great i time place back staff
Topic: good like place burger
Topic: like i you place get
Topic: restaurant good salad great
steak meal dinner
Topic: store shop selection find
items
Topic: would get time service i back
Topic: food good place chicken i like
Topic: food service time back didn't
would
Topic: great food place good pizza
service always
Topic: room hotel vegas nice stay strip
```

Figure 13 – Ten topics from fifty thousand randomly sampled reviews

```
Topic: time get wait long first make
come
Topic: chicken good try salad cheese
get fries
Topic: great good food place menu lunch
pizza
Topic: great food service good place
friendly staff
Topic: don't can like place service
back
Topic: try coffee tea chocolate ice yum
love
Topic: free get check one card buy car
Topic: nice place great s parking new
like
Topic: best place ever vegas I've s
Topic: happy hour night pm day great s
```

Figure 14 – Ten topics from fifty thousand randomly sampled tips

Ultimately, we begin to see a general trend in how tips are seemingly used. For the most part, it is very akin to how Yelp defines what tips are: a way to pass a long key information. This key information usually comes in the form of a type of suggestion; whether this is for a menu item, when to arrive at the business, or what one should or should not do, all groups of topics had some form of suggestion among them. Reviews, on the other hand, were mostly based on experiences and menu items, reflecting the importance of overall experience at a business instead.

C. Sentiment Analysis and Regression

1) Sentiments – Polarity and Subjectivity

Using the Python package TextBlog, we then ran a sentiment analysis on all of the tips. Roughly forty-thousand businesses had tips, whereas all of the businesses had reviews of some sort. TextBlob determines two factors when analyzing sentiment: polarity (or, positiveness or negativeness of words) and subjectivity (how objective or subjective the text is). Polarity is defined on a -1.0 (negative) to 1.0 (positive) scale, while subjectivity is on a 0.0 (objective) to 1.0 (subjective) scale. After averaging the polarity and subjectivity for each business, the means for the polarity and subjectivity for reviews and tips are represented in the table below.

	Mean	SD
Avg. polarity of tips	0.24921	0.26316
Avg. polarity of reviews	0.21750	0.12197
Avg. subjectivity of tips	0.45405	0.22328
Avg. subjectivity of reviews	0.53948	0.07724

Figure 15 – Descriptive statistics of polarity and subjectivity of reviews and tips

On average, tips and reviews are generally seem to be slightly more positive and both are close to being in the center between objective and subjective. Tips, as a whole, tend to have a higher variance as well; plotting the polarity of tips against star rating of the business and comparing it to the polarity of reviews makes this particularly obvious (Figures 16 and 17 below).

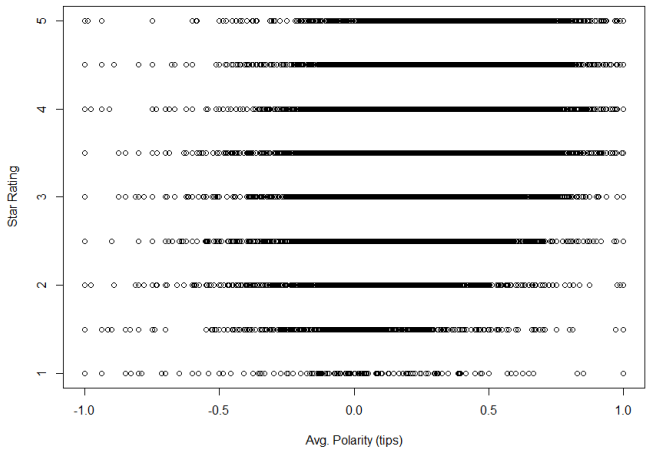

Figure 16 – The average polarity of a tip versus the star rating of a business

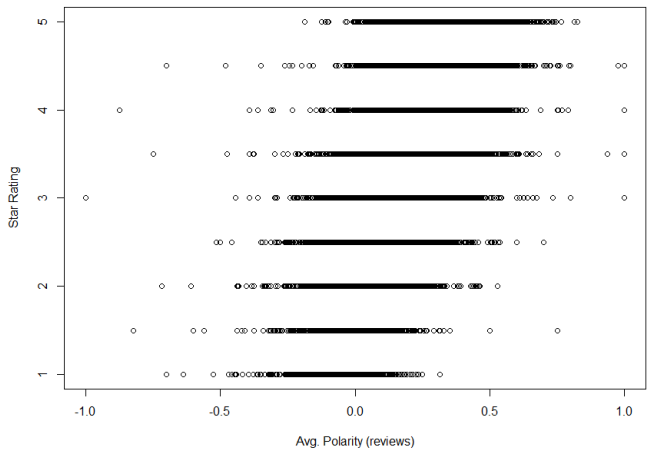

Figure 17 – The average polarity of a review versus the star rating of a business

Both reviews and tip both seem to follow the general trend that as a tip or review is more positive (higher polarity), the higher the star rating is. However, the reviews tend to be clustered more around a polarity of 0.0, implying that reviews tend to be more neutral. Indeed, when comparing the subjectivity of reviews and tips, tips have a high variance of completely objective tips and subjective tips, while reviews are usually closer to a subjectivity of 0.5, or directly in-between completely objective and completely subjective.

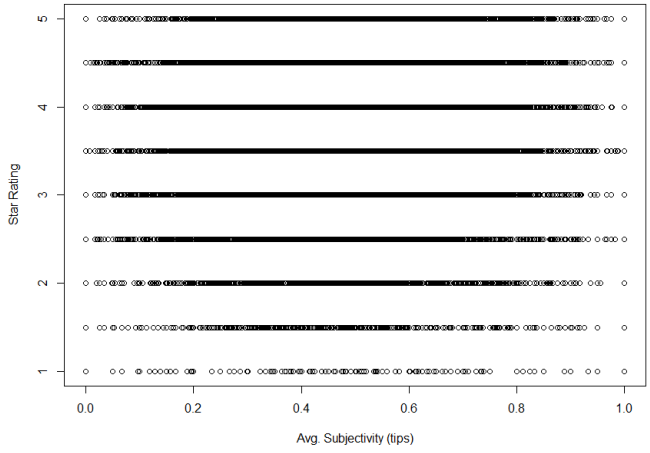

Figure 18 – The average subjectivity of a tip versus the star rating of a business

Again, this seems to fit the vision of tips that Yelp has defined; users are leaving an entire range of tips (objective tips like hours of operation to subjective tips regarding what they liked or disliked) while reviews are generally more neutral in their polarity and use a mix of objective and subjective words. This should not come as much of a surprise, since reviews are generally meant to be a user's personal evaluation of a business, which would likely have a mix of negatives and positives.

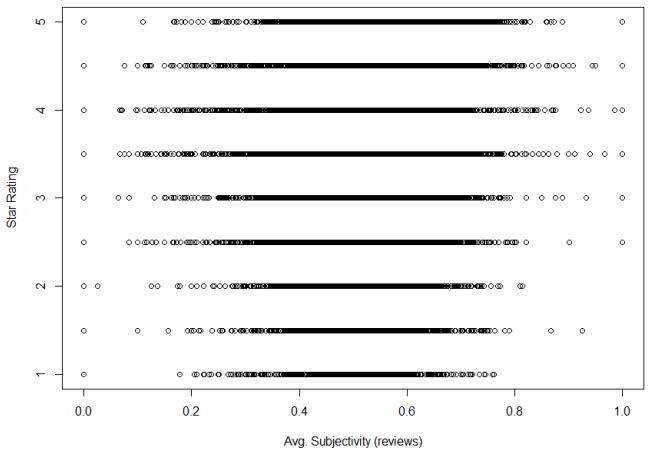

Figure 19 – The average subjectivity of a review versus the star rating of a business

2) Regression

We can then build a variety of regressions in order to better understand polarity its effect on the star ratings of a given business (or, whether star ratings have an effect on the polarity).

First, we built a regression to attempt to predict the star rating based on the average polarity of a tip or review. Using the polr() function in R to create an ordinal logistic regression with star rating as the dependent variable, we get the following intercepts and coefficients:

```
Coefficients:
tips_polarity 1.972

Intercepts:
1|1.5   -5.0322
1.5|2   -3.5621
2|2.5   -2.4646
2.5|3   -1.4312
3|3.5   -0.4636
3.5|4    0.5845
4|4.5    1.7501
4.5|5    3.0858
```

Figure 20 – Ordinal Logistic Regression results for tip polarity predicting star rating

```
Coefficients:
reviews_polarity 14.32

Intercepts:
1|1.5  -3.0816
1.5|2  -1.8210
2|2.5  -0.6127
2.5|3   0.7004
3|3.5   1.9136
3.5|4   3.2709
4|4.5   4.6044
4.5|5   5.9302
```

Figure 20 – Ordinal Logistic Regression results for review polarity predicting star rating

With these regressions, we can then check the goodness of fit of these regressions by running a chi-squared test on both regressions with the following hypotheses with a p-value of 0.05:

H_o : The current model is a good enough fit
H_a : The current model is not a good fit

In both cases, the p-value ends up being 0, meaning that we can reject the null hypothesis. Thus, the model for both tips and review polarity predicting star rating is likely not a good fit. Even when creating the same type of regression and including subjectivity as an additional dimension, the p-value also ends up being 0, meaning that the average polarity and average subjectivity are not likely to be good predictors of star ratings.

However, we also ran a categorical regression to see if the rating of a business can predict the average polarity of a business. After factorizing the star ratings, we can run the regression and view the plotted results below.

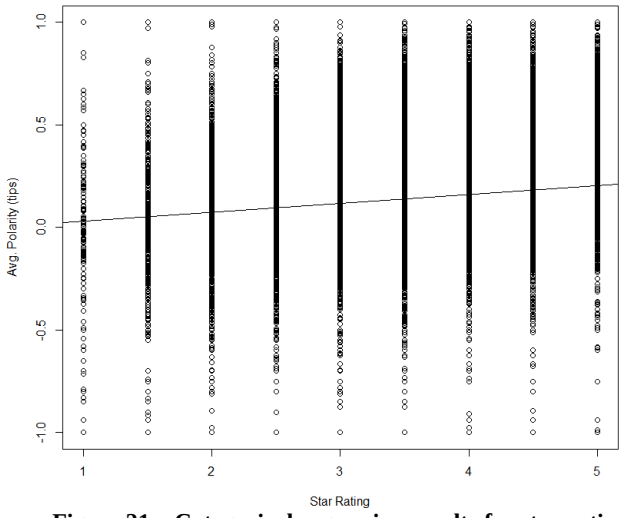

Figure 21 – Categorical regression results for star rating predicting tip polarity

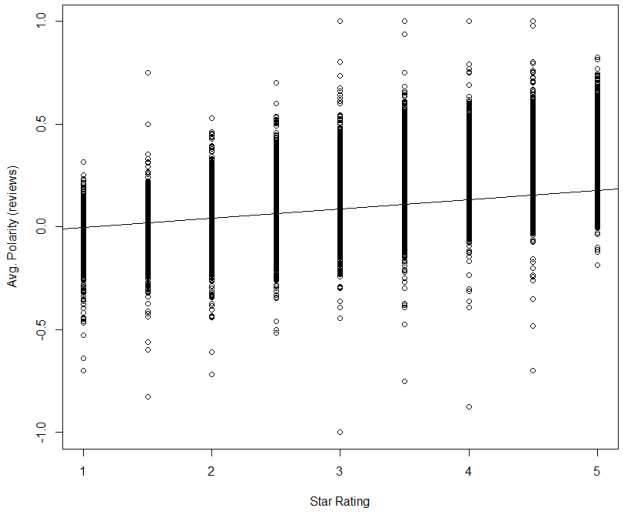

Figure 22 – Categorical regression results for star rating predicting review polarity

Although both lines look as if they fit the data well, examining the R^2 value for each reveals that star ratings are also not good at predicting tip polarity:

$$R^2_{tips} = 0.074$$
$$R^2_{reviews} = 0.453$$

While the fit for the regression for reviews is not particularly strong, it is much stronger than tips. This could potentially be because when a user leaves a review, they also leave star ratings; the higher a star rating, the higher polarity a review could possibly have. The generally low R^2 could be attributed to the fact that the variance of each category of data is particularly large, and that the "polarity" of a tip is solely determined by the TextBlob package; other packages could report polarity differently.

V. Discussion And Future Work

A. *Discussion*

After running our topic and sentiment analysis, we feel as if we better understand the true value proposition of tips. Although Yelp's definition of a tip is clear, through our analysis, we were able to begin to better understand how tips are used. Indeed, tips are generally used as explicit suggestions users leave for other users, whereas reviews are often less suggestion-oriented and more about menu items or experiences. Tips, too, tend to range from extremely objective or extremely subjective, but both reviews and tips are also generally more positive in their text. Important to note, however, is that tips generally do not seem to predict star ratings particularly well (or vice versa); this is likely due to the variety of tips that could be left (subjective/positive, objective/negative, etc.), whereas reviews generally are a more clear indicator of star ratings.

Although there are these similarities, we believe that the value in tips lies in the quick suggestion use-case and that Yelp should continue to support them. However, we also believe that to better take advantage of the information that tips have, Yelp should create more of a focus on tips on their mobile apps and website. Instead of having users scroll near the bottom of the screen to see tips, we believe that tips would benefit from moving them up above reviews (similar to where "review highlights" are located), in order to bring attention to the tips. For users that are particularly busy – e.g. business travelers who need to know how early to get to an airport they have never been to – giving them easier, more convenient access to tips would likely benefit them, possibly promoting more active use of the tip feature.

B. *Future work*

Two areas in which our research could be expanded upon are more in-depth topic modeling and machine learning k-fold cross validation when building regression to predict star ratings.

Our topic modeling was a good basis for investigation, but could have also used a more rigorous investigation. Removing

stop words, using n-grams similar to the UCSD students, and actually creating topics for the entire text corpus would likely give us a better idea of other hidden topics that we were not fully aware of. Hidden topic discovery, too, would like help clarify the differences in topics between tips and reviews.

Although we were able to build a regression model for both tips and reviews, this was only over the entire dataset. Being able to use more data and attempt to test the accuracy of the regressions build through k-fold cross-validation would also greatly benefit whether polarity of a tip affects star ratings in any way. Adding features to the regression would also likely help create a better predictive model for star ratings – users, time, and location are all additional dimensions that could significantly affect the star ratings of a business in ways that the text content of a review or tip could.

Outside of adding features to the regression, doing deeper analysis on the other factors is possible. Are the specific types of users that leave tips instead of reviews? Do tips influence reviews in any way over time, or vice versa? Does the time of day or day of week affect polarity of a review? Utilizing more of the dataset to supplement our current research would likely help refine the difference between how users' use tips and reviews.

VI. Conclusion

Through some basic topic modeling and sentiment analysis, we were able to determine that, though underutilized, tips do topically differ from reviews slightly. Tips ultimately attempt to provide suggestions regarding that business, rather than attempting to describe their experience at that business. Because of this, we believe that tips should be emphasized at a greater level within Yelp's mobile apps in order to better utilize their defining characteristics. More investigation needs to be done, however, to fully determine how different tips are and if they influence the star rating of a business in any particular way.

References

[1] http://www.yelp-support.com/article/What-are-tips?l=en_US

[2] A. De Castro, A. Du, and A. Manicka. "Predicting Ratings Based On Yelp Tips," June 2015.

[3] J. Huang, S. Rogers, E. Joo. "Improving Restaurants by Extrating Subtopics from Yelp Reviews," 2014.

[4] J. McAuley and J. Leskovec. "Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text," 2014.

[5] http://www.yelp.com/dataset_challenge