Leanna Hue, Casey Poon
Professor Hu
3/13/2020
CS 396

Introduction and Motivation

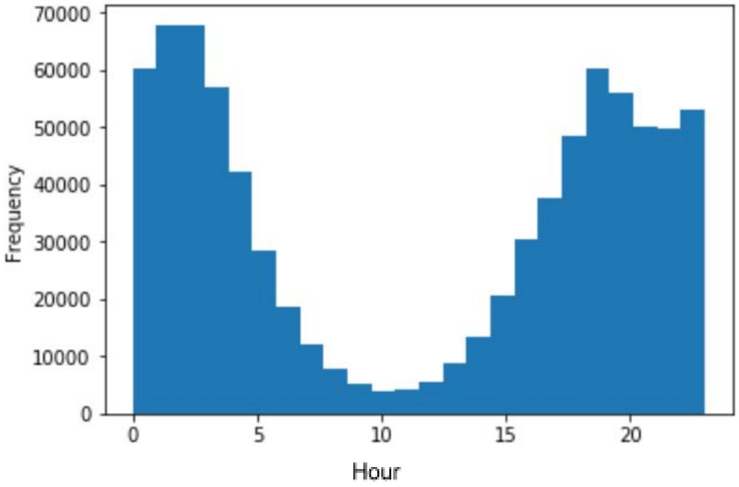
We were interested in studying the relationship between people's sentiment towards a cuisine type versus the local time of day as a heuristic to help restaurant owners gauge public sentiment on a cuisine type. Our motivation for this project stems from the fact that in the restaurant industry, owners want to optimize their hours of operations in order to maximize profits and decrease overhead. Everyone has different food preferences based on the time of day, and we want to gain a better understanding of people's sentiments and biases for different cuisines of restaurants based. We decided to yelp's dataset due to the fact that it contains information pertaining reviews with time stamps for businesses across North America.

Data Cleaning & Preprocessing

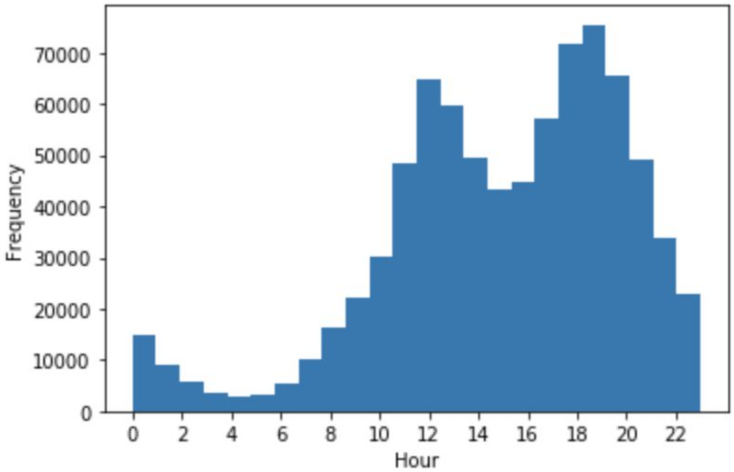
For data cleaning and preprocessing, we first extracted all the businesses from business.json to filter for restaurants. Initially, the tags were all embedded in a string, so we cleaned the dataset by parsing each category list and converting them into a list of categories by deliminating the string. We also deliminated the attributes field in the json in order to extract more features such as 'Classy', 'Hipster', 'Romantic' etc as features for machine learning. We did some preliminary analysis to search for the most common categories in the dataset and concluded the most common but still relevant category were 'restaurants'. The tag for 'food' contained entries such as grocery stores, and thus was omitted. From there we collected all the entries with those tags and gathered their 'business_id', and retrieved all the corresponding reviews from 'tip.json'. We also noticed from EDA testing that the times for each tip were in GMT, so we converted the GMT times into local times.

EDA

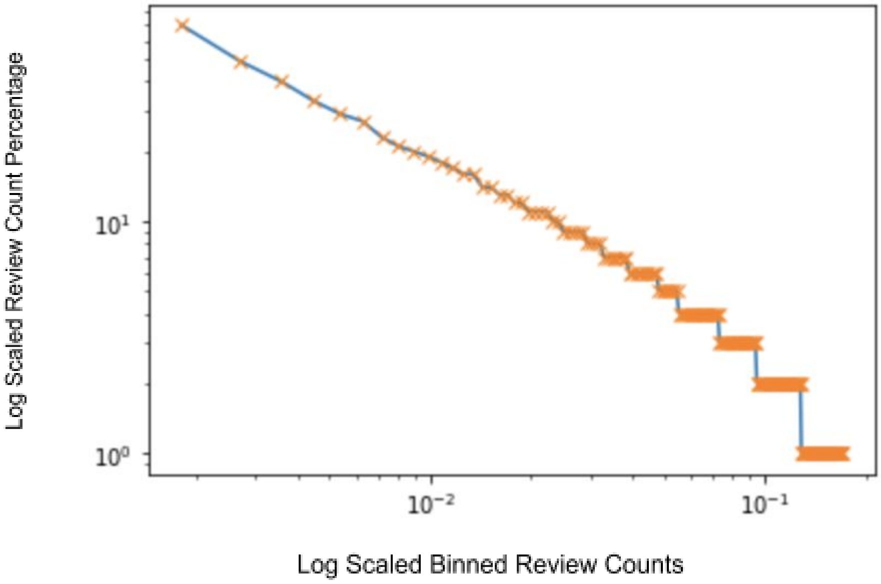
For our EDA, we plotted a histogram showing the distribution of review count against timestamps for reviews in tip based on businesses with the restaurant tag.



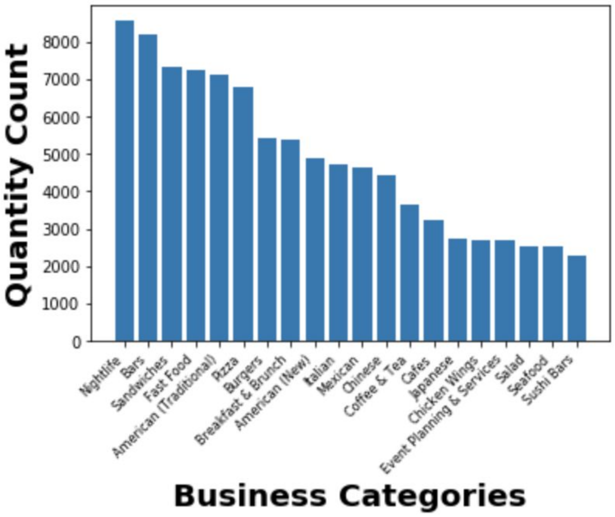
From this histogram and the unusual dip around hour 10, we learned that the tip times were in GMT, and after cleaning the data and converting the tip times into local times, the histogram was as expected with spikes in tips around meal times (12pm and 6pm).



The second EDA test we performed was a power law test on the distribution of tip counts per unique business ID. From this test, we concluded that the relationship is Very Strongly Negatively Correlated and that it follows a Power Law Curve. This helped us learn that our dataset is not uniformly distributed, and it helped us decide to add weighting to our machine learning classifiers to try to even out the distribution.



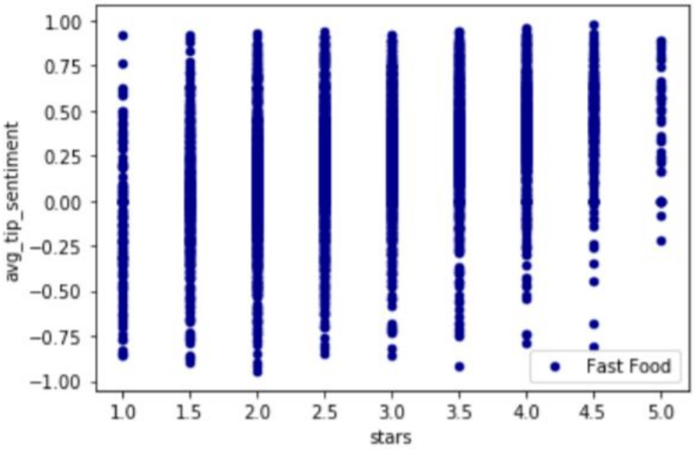
The third EDA test we performed was one to find the most popular business categories in the business.json.



Using this data, we could see that the food cuisines with the most data are Fast Food, American (Traditional), and American (New). We used this information to better focus our analysis and machine learning on only these cuisines with the most data.

Machine Learning

We applied machine learning to conduct sentiment analysis of reviews and tips as well as to potentially predict cuisine types. To gather the average sentiment of a restaurant we explore tips.json for each unique business id at different meal times. Our initial approach was to apply a bag of words technique by getting the count and map to positive words in each review in tips.json and then averaging them for each business id. However, this proved to be successful due to the fact the reviews with more words skewed higher in general sentiment. Our second approach to extract average sentiment was to use VADER sentiment analysis to extract a polarity score for positive sentiment (https://github.com/cjhutto/vaderSentiment). We chose this approach because the VADER's sentiment analysis model was specifically trained on social media text dataset, and Yelp's reviews and comments are very similar to how people post social media. Although sentiment analysis is considered an unsupervised learning problem, we were able to evaluate the VADER model's accuracy by examining average star rating for each business. Typically reviews with higher star ratings would have more positive sentiments in their reviews so we looked to see if there was a positive correlation/relationship between a restaurant's star rating and average positive sentiment. In order to evaluate this we decided to examine the Pearson and Spearman correlation coefficients between average sentiment and star rating. In the subset examining only Fast Food, we discovered that Spearman and Pearson correlation coefficients were .41 and .39 respectively. From this we were able to conclude that although there is some sort of positive relationship.



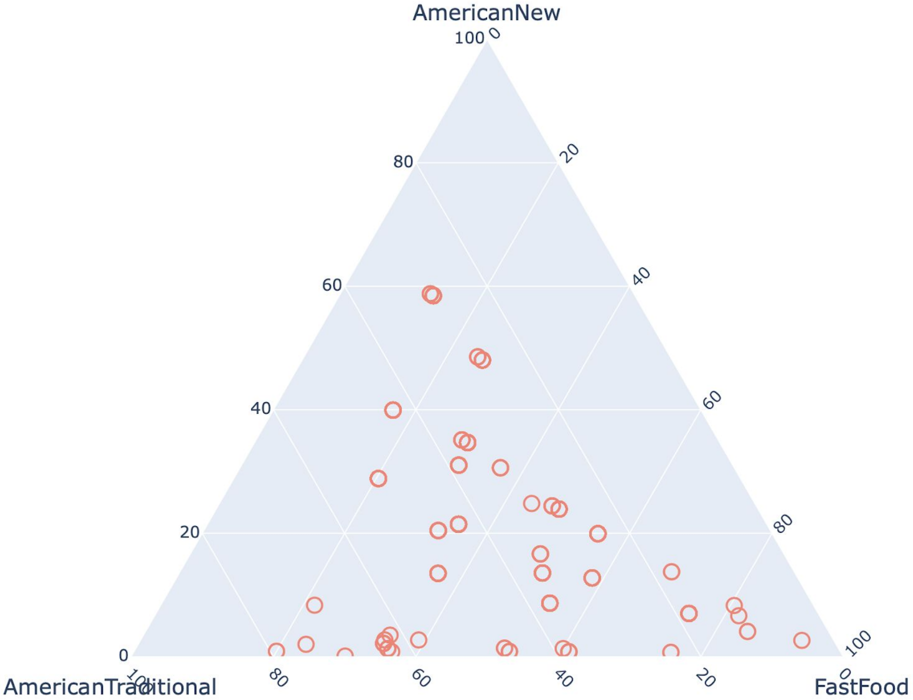
In order to predict cuisine types, we had to engineer and gather more features. From business.json we were able to pull out 4 more features in addition to average sentiment such as ‘Classy’, ‘Hipster’, ‘Romantic’, and ‘Casual’ without further compromising the size of the dataset. The average sentiment was a continuous feature ranging from -1.0 to 1.0 where -1.0 describes a negative sentiment and 1.0 describes a positive sentiment. The rest of features were discrete binary variables that indicated whether or not that feature is true or not for that instance. We tried six different approaches for prediction: Naive Bayes, Naive Bayes with weighting, Decision Trees, Decision Trees with weighting, K-Nearest-Neighbor, and K-Nearest-Neighbor with weighting. The reason we included weighted algorithms was because from our EDA, we noticed that our dataset follows a Power Law distribution and the dataset is heavily skewed. The results for our machine learning algorithms are the following:

	Naive Bayes	Naive Bayes w/ Weighting	Decision Tree	Decision Tree w/ Weighting	KNN	KNN w/ Weighting
Accuracy	0.5154	0.4638	0.8629	0.8769	0.8623	0.4146

Decision Trees with weighting performed the best with an accuracy score of .8729. This makes sense because Decision Trees are prone to overfitting for a dataset which would skew the results towards an abnormally high accuracy.

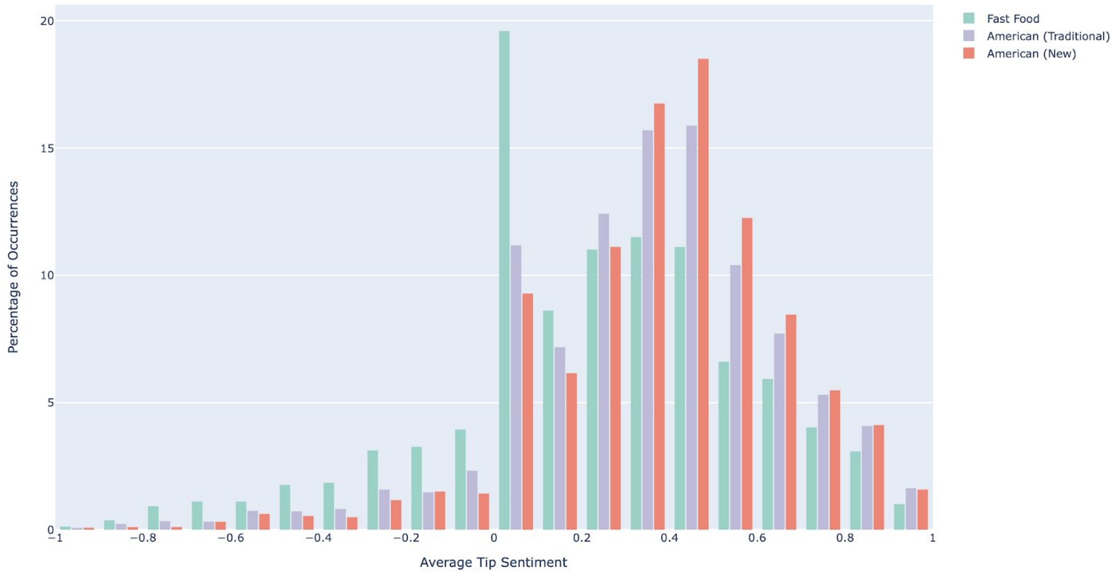
Summary of Findings

The results of our analysis were mostly inconclusive. After engineering and extracting features for each business and plotting our data on a ternary plot, we learned that there does not seem to be any clear distinction or clustering toward businesses of a certain cuisine, and there may be overlap or unclear semantic distinctions between the cuisine categories we chose to analyze: Fast Food, American (New), and American (Traditional).

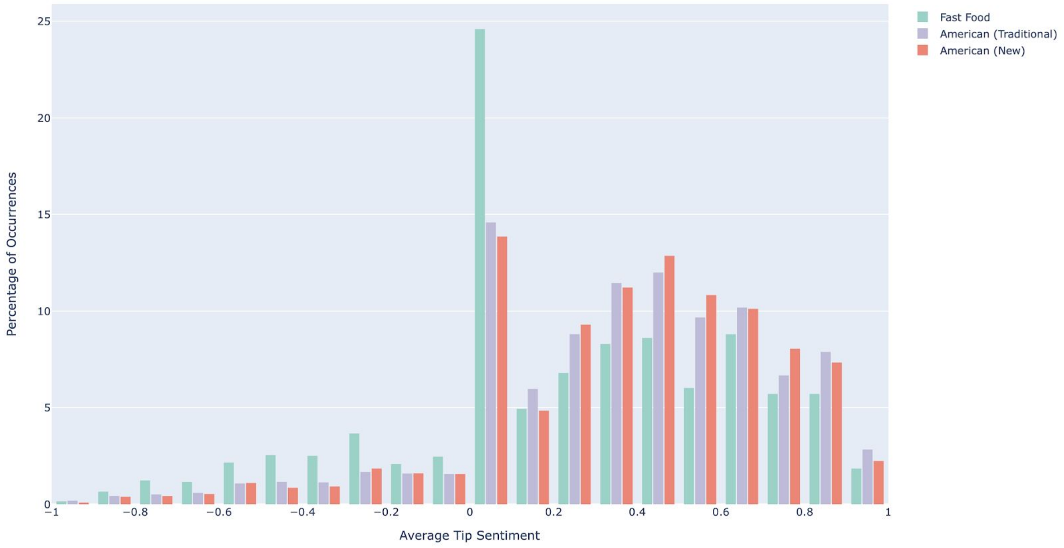


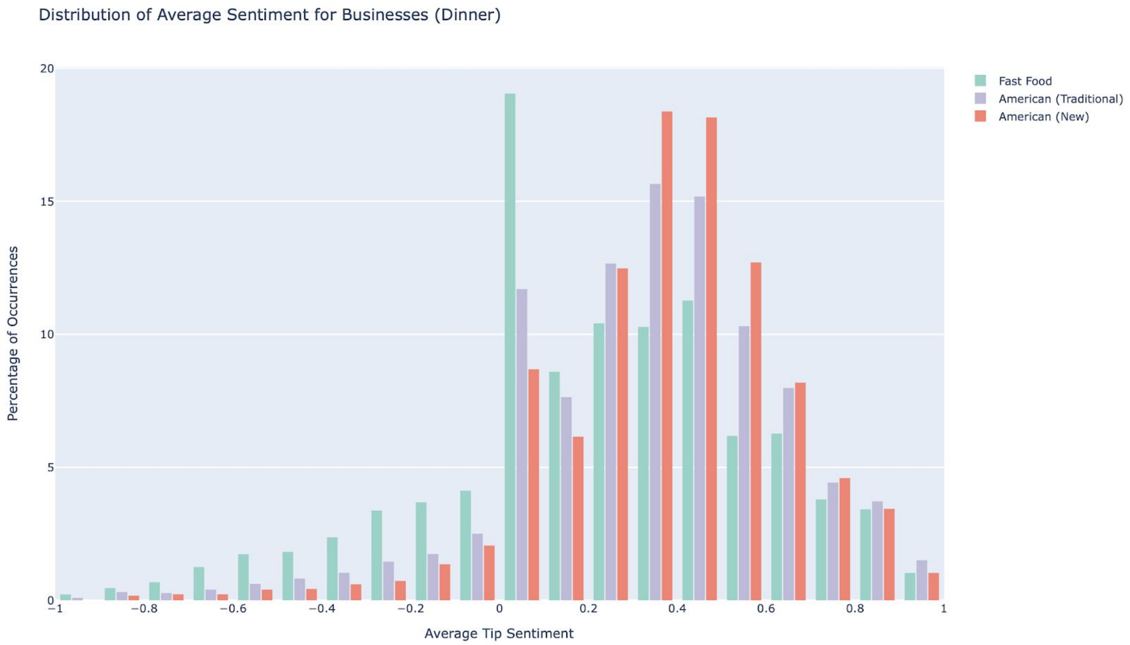
After some analysis on average tip sentiment per business, we learned that the modes for average sentiment in increasing order are consistently Fast Food, American (Traditional), and American (New) across all meal times. Our results did not show a large difference between meal times for average sentiment of tips.

Distribution of Average Sentiment for Businesses (Lunch)



Distribution of Average Sentiment for Businesses (Breakfast)





Potential Implications and Improvements

Part of the reason why we had inconclusive results was we ran into a few problems while using the Yelp dataset. The first was removing the incomplete business data greatly reduced the sample size of our dataset and did not leave us with a lot of data to work with. We also wanted to use review.json, but because the file is over 5gb, our computers were not able to process the file. Instead, we used tip.json, but the tips tended to be short and included less information than review.json.

Potential improvements for our project would be using data interpolation to fill in the missing gaps in our dataset,exploring more cuisines, and using review.json to analyze different features and how they vary across meal times. Additionally, our machine learning model was trained and evaluated on the feature ‘average sentiment’ that we extracted using the VADER Model. Although there was a positive correlation, it did not seem to be a very strong positive correlation with star count. More analysis can be conducted to verify that our VADER model is giving us accurate sentiments or an entirely different model can also be used.