Mexican Restaurant Yelp Reviews:

Sentiment Analysis and Topic Modeling

Catherine Magsino

Yelp provides a wealth of information that can be used by both consumers and business owners to understand the business, as well as customer experiences. One important aspect of a Yelp page is the customer reviews. As a business becomes more popular, the more reviews it receives. However, as the number of reviews grows, it becomes difficult to read through it all and to get a full picture of the customer experience. By performing natural language processing on Yelp reviews, and providing sentiment analysis and topic modeling, this project has two goals:

1. To make it easier for customers to make restaurant decisions.
2. To allow restaurants to action off of customer reviews.

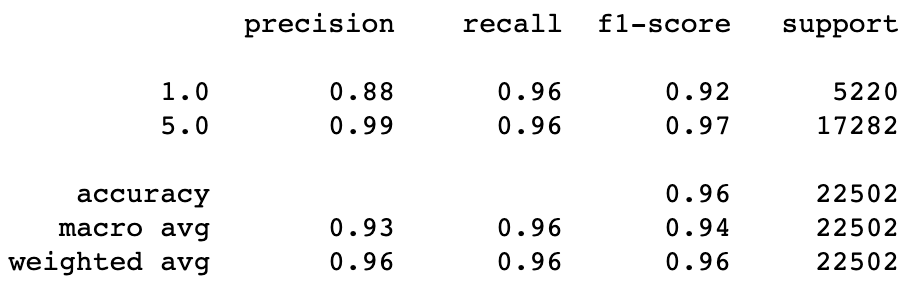
Data

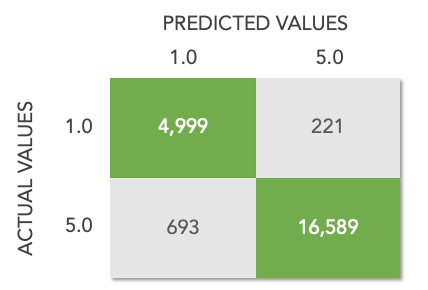
The data for this project was retrieved from the [Yelp Dataset Challenge](https://www.yelp.com/dataset/challenge), using the business.json and review.json files. Since it came in JSON format, I first converted the files into CSVs and combined the two files into one dataframe. With the original dataset consisting of 6 million reviews, I filtered the data down to only including Las Vegas Mexican restaurants. This resulted in a dataset that included 1,019 restaurants and 129K+ reviews from 2013-2018.

Sentiment Analysis

Performing sentiment analysis required using both unsupervised and supervised learning. First, I pre-processed the text data by removing punctuation, stop words, numbers, and converted everything to lowercase. I then converted the text documents into a matrix of token counts using CountVectorizer.

Once my data was ready for supervised learning, I trained the data only on the 5 and 1 star ratings, which I corresponded to be positive and negative reviews. Since the data consisted of 76% 5 stars and 24% 1 stars, I had to balance it using random oversampling. After validating across several different machine learning models, Multinomial Naïve Bayes gave me the best results, with F1 scores of 0.92 for 1 star ratings and 0.97 for 5 star ratings. I then applied this final model to my entire dataset to predict positive and negative sentiment.





Topic Modeling

My next step was to perform topic modeling based on sentiment, therefore splitting the positive and negative text data that was predicted by the model. I pre-processed, tokenized, and vectorized the same way I did with the sentiment analysis, and then performed dimensionality reduction using LSA, LDA, and NMF. NMF (non-negative matrix factorization) gave the best results. The 6 positive topics included: ordering, meat, service, sides, atmosphere, and high praise. The 5 negative topics included: service, quality, sides, wait, and meat. A few of the same topics appeared for both the positive and negative sets.

Positive Topics:

|  |  |
| --- | --- |
| **Topic** | **Top 10 Words** |
| **Ordering** | like, one, ordered, chicken, time, salsa, would, also, menu, restaurant |
| **Meat** | tacos, taco, asada, carne, pastor, meat, al, get, el, best |
| **Service** | food, service, restaurant, best, amazing, always, delicious, fast, friendly, staff |
| **Sides** | good, really, pretty, salsa, service, chips, nice, got, also, little |
| **Atmosphere** | great, service, back, amazing, friendly, awesome, definitely, atmosphere, staff, drinks |
| **High Praise** | place, love, go, try, like, get, always, amazing, burrito, best |

Negative Topics:

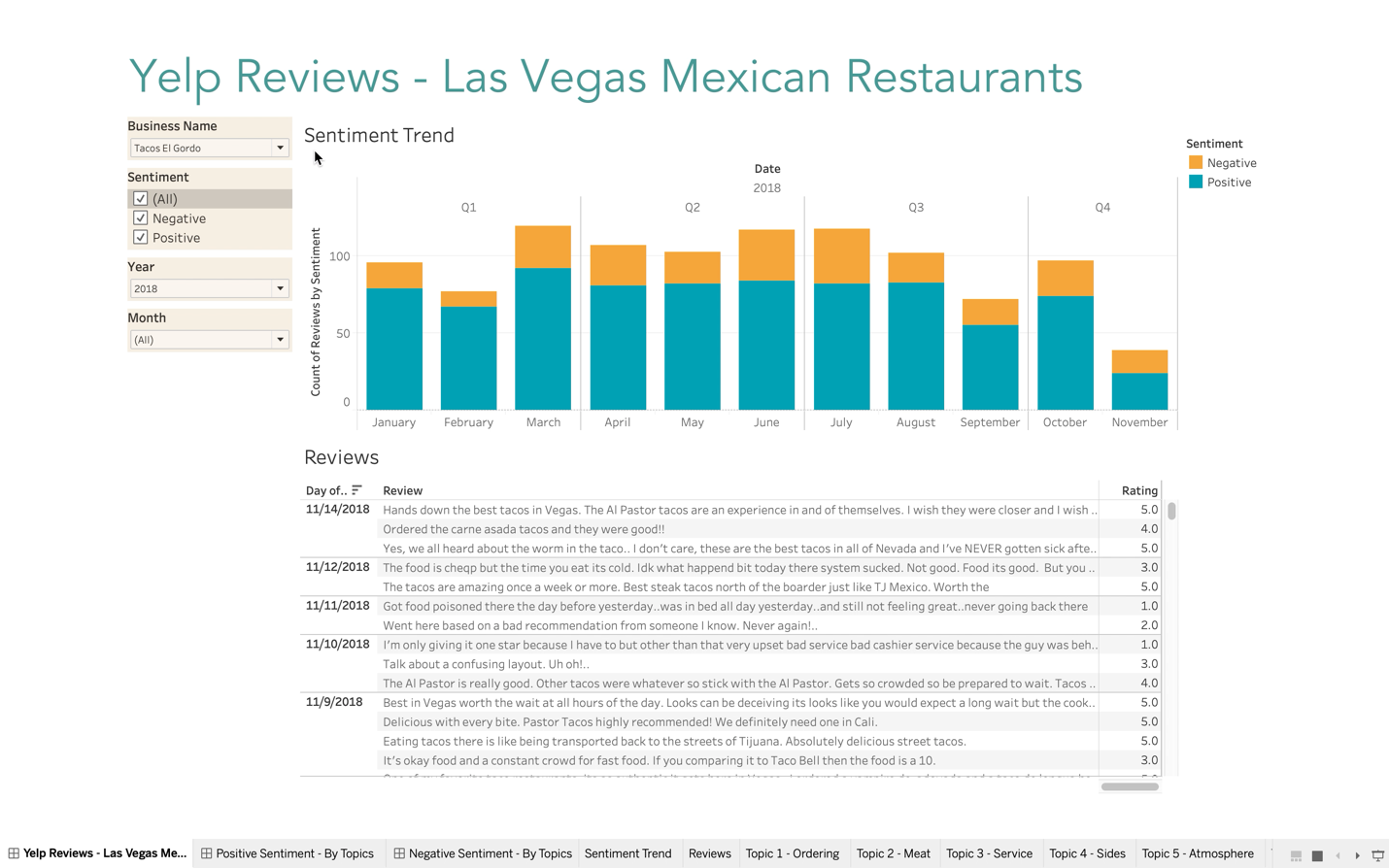
|  |  |
| --- | --- |
| **Topic** | **Top 10 Words** |
| **Service** | us, came, table, asked, said, service, would, server, one, minutes |
| **Quality** | food, service, place, good, great, restaurant, bad, better, eat, quality |
| **Sides** | good, like, place, really, chicken, salsa, ordered, burrito, chips, one |
| **Wait** | order, time, get, back, go, one, location, said, got, never |
| **Meat** | tacos, taco, asada, meat, carne, ordered, fish, line, bell, one |

Recommendation

Based on the results, I would recommend that:

* Customers/restaurant owners review sentiment over time and during specific periods.
* Restaurant owners create action plans based off of customer comments by topic.

In addition, since it is helpful to be able to see the verbatim comments for each of the sentiment and topic groupings, I created a Tableau dashboard that would easily allow a user to do this. Double click on the video below to play.



Future Work

Future improvements of this work would include:

* Additional cleaning of data (e.g. adding more stop words).
* Summarizing results with representative sentences, since reading verbatims still requires extensive reading.
* Creating a recommendation system for restaurants with similar positive comments.