Yelp Reviews Topic Modeling to Predict Business Latent Features

Abby Tisdale & Nick Gonzalez

Yelp Dataset

- 80,000+ Businesses with Latent Features such as “Good For Kids”, “Takes Credit Cards”, and “Happy Hour”
- 2 Million+ Reviews with specified businesses
- Each review and business stored as individual JSON objects in large JSON files

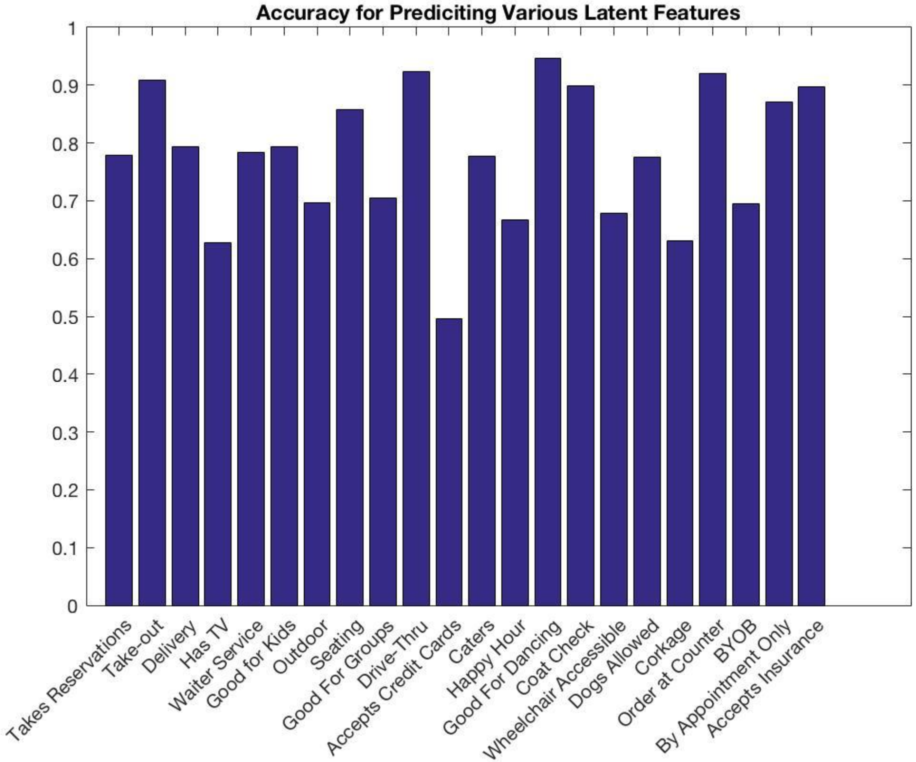
Topic Modeling on Reviews

- Aggregate business specific reviews
- Determine word frequencies
- Create sparse matrix with businesses as rows and words as columns
- Filter out overused and underused words
- Non-negative Matrix Factorization to create matrix of businesses as rows and 20 topics as columns

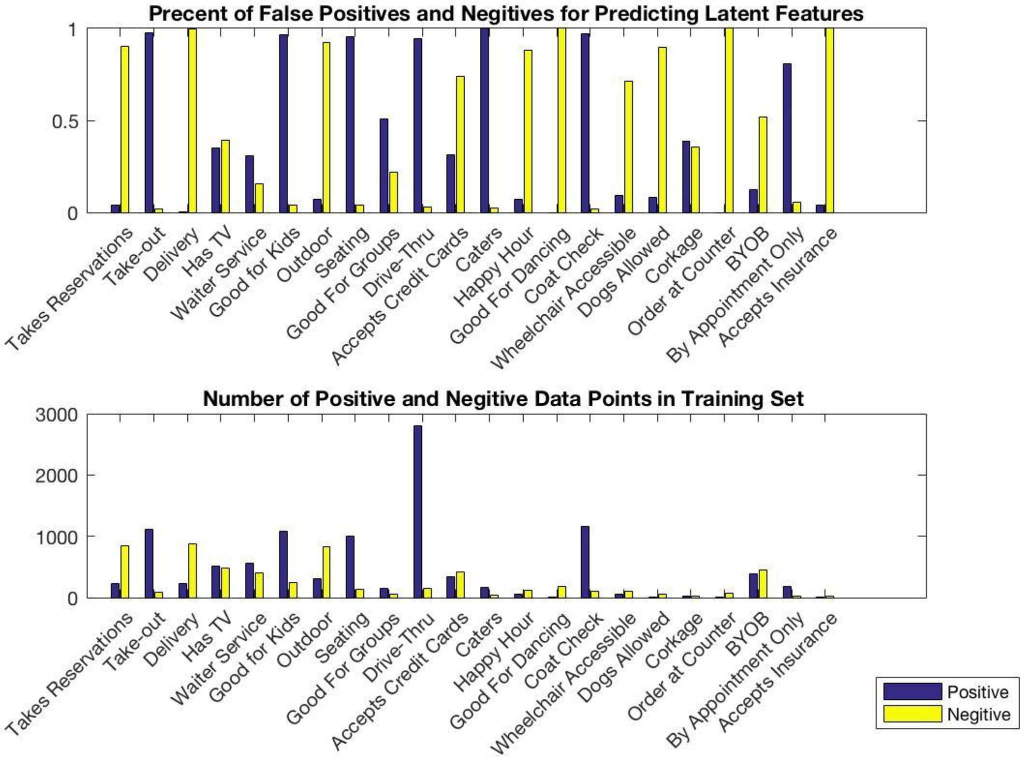
Construct Training and Testing Sets

- For each business in the topic modeling matrix, build a row in another matrix
- For each row in the new matrix determine latent features for the corresponding business so that each column corresponds to 1 latent feature
- Topic modeling matrix is X
- Latent feature column vectors are y
- Split data into Training set (66%) and Testing set (33%)

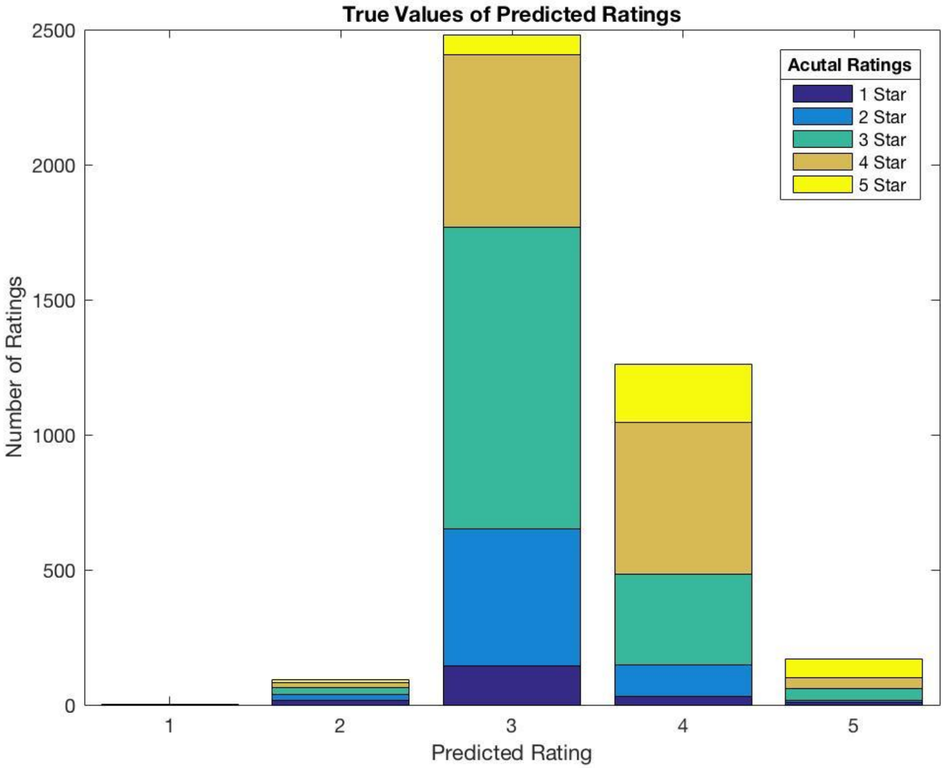
Logistic Regression on Binary Features



Logistic Regression on Binary Features



Multinomial Logistic Regression over Ratings



Results

- Able to predict Stars with 44% accuracy
- Most predictable latent features: Has TV, Waiter Service, Good for Groups, and Corkage

Future Work

- Make a larger data set
- Set lower word thresholds on NMF
- Look into improving classifier of lopsided data

Questions