What do good Yelp reviews look like?

Yelp Review Votes Prediction

Project Workflow

#1 Define question & Goal

Background & Goal

Yelp is using 3 community-powered metrics to track the review quality: Useful, Cool, and Funny.

The goal here is to understand what the high-quality Yelp reviews look like and make predictions on the good reviews in the future.

USEFUL Cool Funny

Why is this important?

- Always push the most recent and good-quality reviews in the "Review Highlight"
- Get insights for developing better content advertising

#2 Data Acquisition & Transformation

Review

Check-in

User

- Business name
- Business category
- Geo location
- Status: Open/Closed
- Business star: 1 to 5
- Review count

- Review date
- Review text
- Review Votes (3 Categories: Useful, Cool and Funny)

- Business id
- Check-in time
- # Check-ins

- User name
- User Average Stars
- # User reviews
- # Review Votes

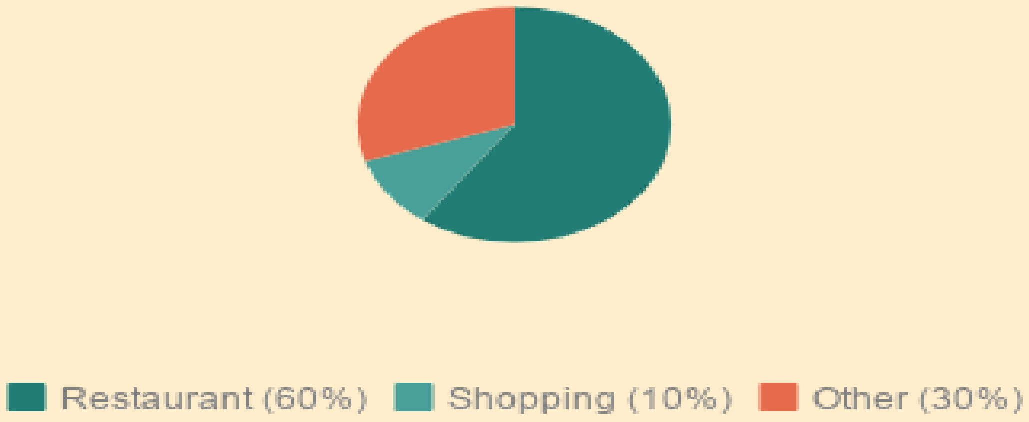
Data Tranformation for Modeling

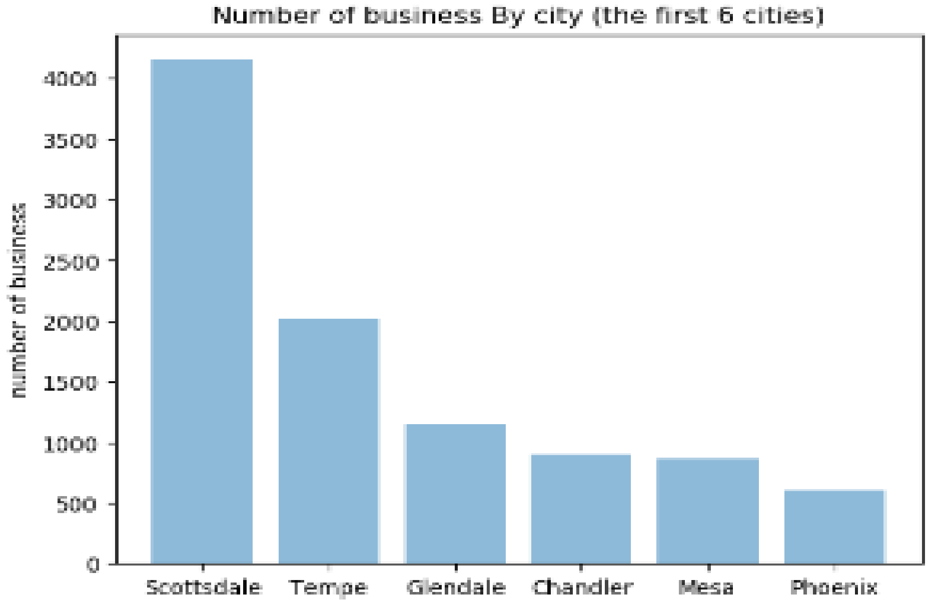
- Mutually exclusive review groupings: Useful vs Not useful, Cool vs Not cool, Funny vs Not funny
- Total Review Votes = Funny + Cool + Useful votes
- Groupings based on vote count: Below average votes Vs. Above average votes

#3 Exploratory Analysis

- 11,537 business
- 99% located in Arizona
- Most of the business (80%) located in Phoenix, Scottsdale, Tempe, Mesa and Chandler.
- 90% Open businesses
- 60% of the businesses are restaurants.
- 200,471 reviews and 43,873 Users
- 94 check-ins on average

Number of business by categories



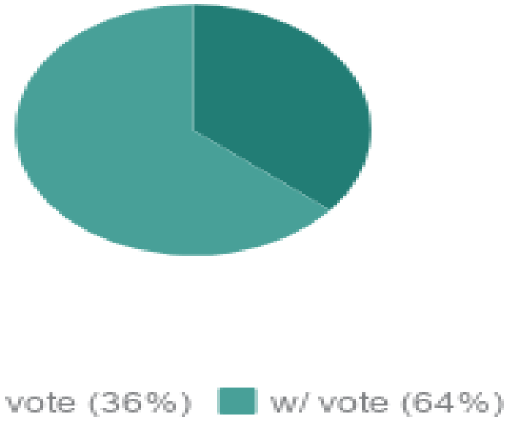


Screenshot of the Data Dictionary

Variable	Description	Type of Variable	Comment
business_id	unique identifier for the business.	text string	Overall 8881 business in the dataset
business categories	Categories of the business	categorical	Overall 1067 business categories. Example value:['Delis', 'Restaurants']
business city	The city where the business is located	categorical	65 unique cities. Example value: Youngtown
latitude	latitude of the business	continuous	
longitude	longitude of the business	continuous	
business name	Name of the business	categorical	Overall 5497 unique business names, business name to business id is one to many relationship
open	whether the business is open or closed	categorical	two unique values: True/False
business review_count	# of reviews a business has got so far	continuous	
business stars	# stars a business has got	categorical	1 - 5
review date	the date when the review was posted	date	
review_id	the id of the reviews	text string	Overall 200471 unique reviews
text	text of the reviews	text string	Overall 200297 unique review texts
user_id	User id	text string	Overall 41005 users
user average stars	the Average Star user has got	categorical	0 - 5
user review_count	# of reviews user has posted on Yelp	continuous	

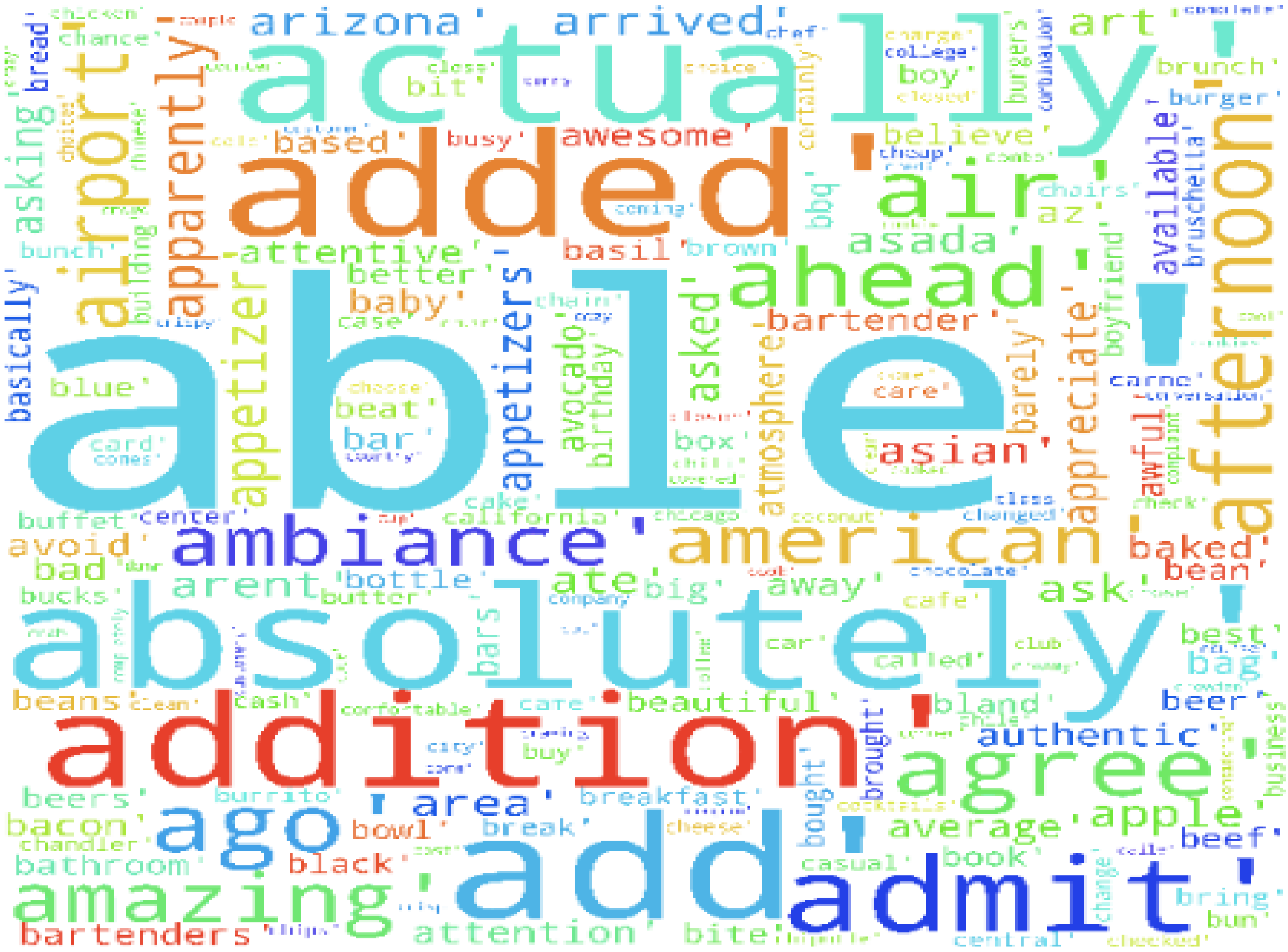
#3 Exploratory Analysis

36% of Reviews has no vote.



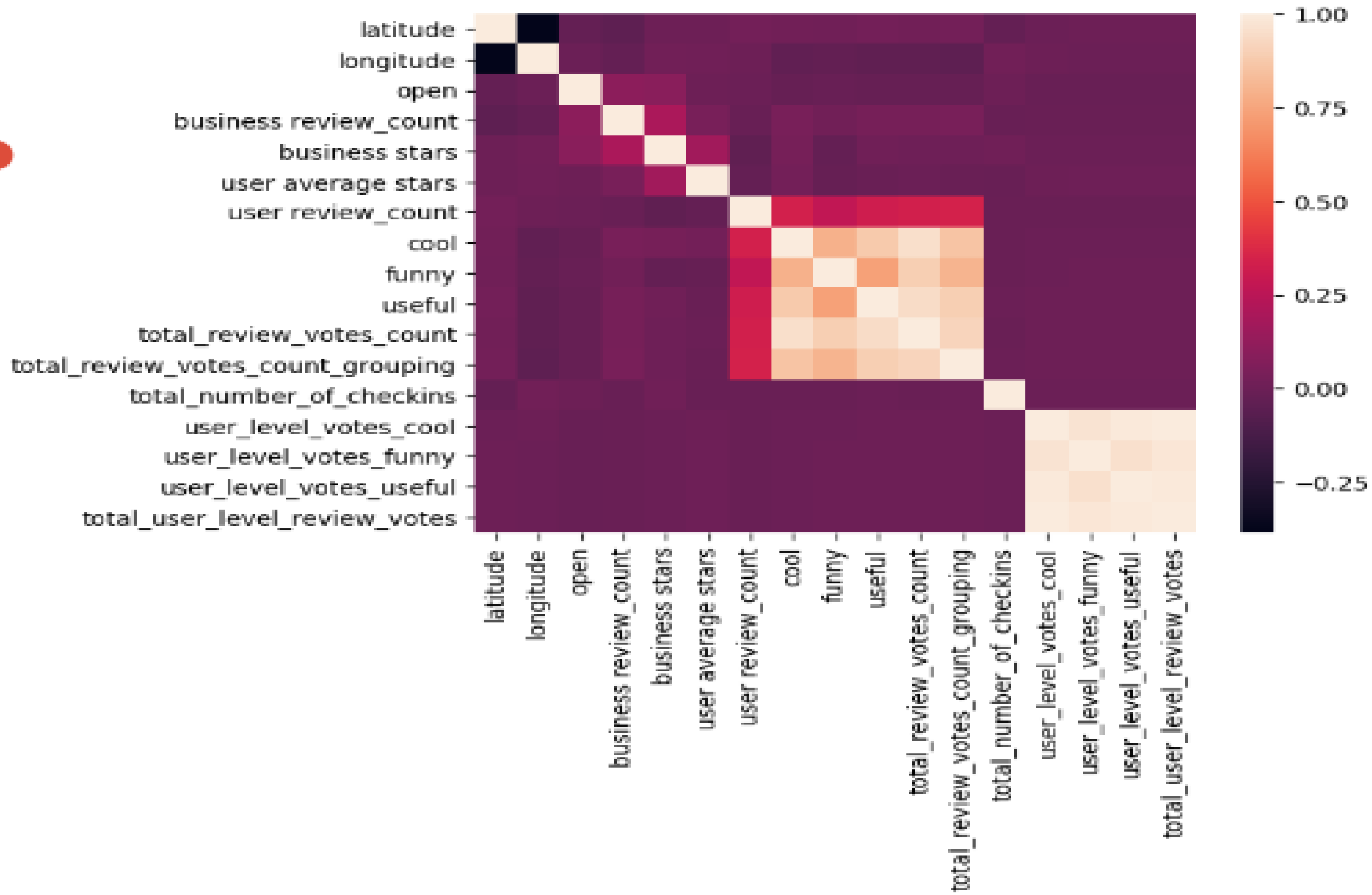
- 200,471 votes in total
- 30% useful vs 70% not useful
- 30% funny votes vs 70% not funny
- 37% cool votes vs 63% not cool
- Most reviews have up to 5 votes

Most Frequently-used words in reviews



#3 Exploratory Analysis

Any relationship in the data?



- Useful, funny and cool votes for the reviews are closely related.
- The count of the user reviews is related to the total number of review votes

#4 Modeling & Testing

Step 1: Transform all the review texts into 31,193 sets of keywords using count vectorizer and tfidf vectorizer

```
## Use the tfidf transformer to convert the review text to the vectors:
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_transformer = TfidfVectorizer(ngram_range =(1,3), stop_words='english', lowercase=True, min_df=50)
X_text_train_tfidf = tfidf_transformer.fit_transform(X_text_train)
X_text_train_tfidf.shape #(200471, 31193) # 200471 samples with 31193 features
```

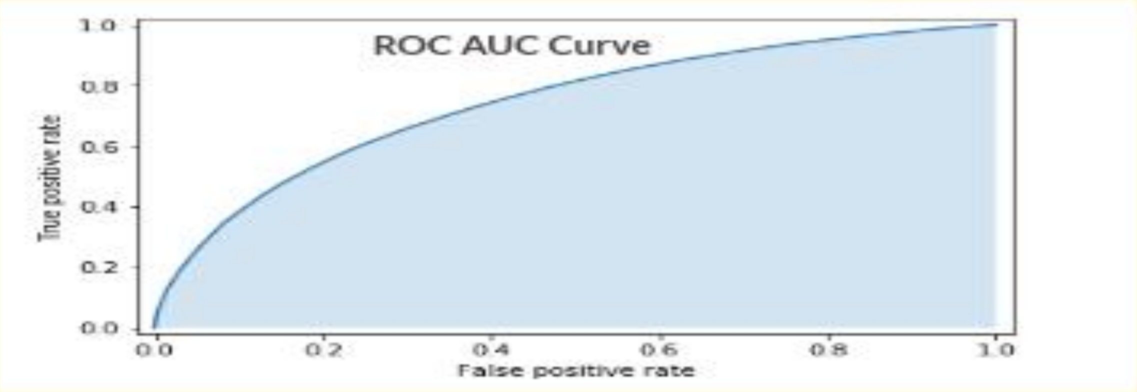
```
(200471, 31193)
```

Step 2: Classify reviews into categories using Naive Bayes, Random Forest and Logistic Regression

Useful vs Not Useful
Cool vs Not Cool
Funny vs Not Funny
Review Votes: Above Average vs Below Average

Step 3: Identify the right metric and evaluate the success of the model using cross validation

- Use ROC AUC score: Precision & Recall are both important in this case
- ROC AUC Score: Around 60 - 70% for each of the model



#4 Modeling & Testing

Step 4: Find out the keywords with positive and negative impact on the classfications

```
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("good"), 'review_good_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("time"), 'review_time_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("really"), 'review_really_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("just"), 'review_just_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("love"), 'review_love_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("like"), 'review_like_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("service"), 'review_service_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("place"), 'review_place_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("food"), 'review_food_dummy']=1
yelp_data_final_update.loc[yelp_data_final_update['text'].str.contains("great"), 'review_great_dummy']=1
```

Step 5: Utilize the high-impact keywords for building regression model to predict total number of review votes and leverage MSE, P value, R squared to evaluate model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     total_review_votes_count   R-squared:                       0.366
Model:                                  OLS   Adj. R-squared:                  0.366
Method:                       Least Squares   F-statistic:                     4829.
Date:                      Sun, 28 Jan 2018   Prob (F-statistic):               0.00
Time:                              15:04:12   Log-Likelihood:            -6.0957e+05
No. Observations:                    200471   AIC:                         1.219e+06
Df Residuals:                        200447   BIC:                         1.219e+06
Df Model:                                24
Covariance Type:                  nonrobust
=====================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
business_category_dummy_shopping   -0.1624      0.050     -3.249      0.001      -0.260      -0.064
user_review_count_sqrt              0.2286      0.002    133.723      0.000       0.225       0.232
text_string_length                  0.0025   2.63e-05     94.700      0.000       0.002       0.003
review_good_dummy                  -0.5031      0.024    -20.771      0.000      -0.551      -0.456
review_time_dummy                  -0.1449      0.025     -5.683      0.000      -0.195      -0.095
review_really_dummy                -0.1480      0.029     -5.157      0.000      -0.204      -0.092
review_just_dummy                   0.0134      0.027      0.495      0.620      -0.040       0.067
```

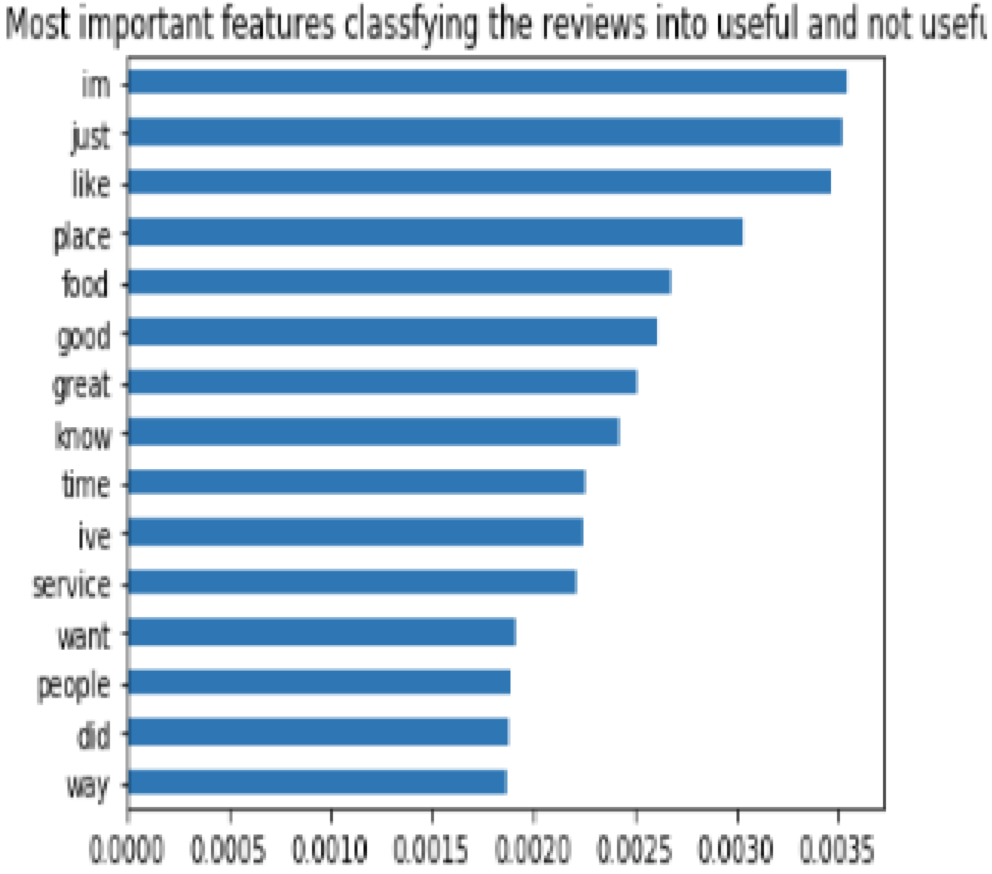
#5 Insights & Next Steps

Positive & Important keywords within the reviews for Classification

Positive/Negative keywords to group reviews into "Useful" vs "Not Useful"

(+) delicious, staff, new, come, friendly, salad, fresh, came, say, right, want, better, did, went, going, night, lunch, cheese, way, didnt, make, pizza, ordered, order, think, restaurant, try, menu, pretty, chicken, know, best, bar, got, people, nice, ive, little, love, service, im, dont, time, really, great, just, food, like, good, place

(-) food great ambiance, good seating, hour sushi, happy hour sushi, best indian food, wife chicken, great pho, good ribs, great bartender,toppings want, staff best, like spicy food, planning going, overall good place, tried, yummy service, place yummy, service staff friendly, time lot, phoenix week, places nearby, appetizer delicious, good wait staff, hampton, rice nice, chicken entree, location north, food old, just moved area, spinach pizza, crusted chicken, great wife, great food atmosphere, wasnt huge fan, beef little, pork burrito, fontina burger, expect lot, good wine selection, area lot, great going,service good atmosphere, right price



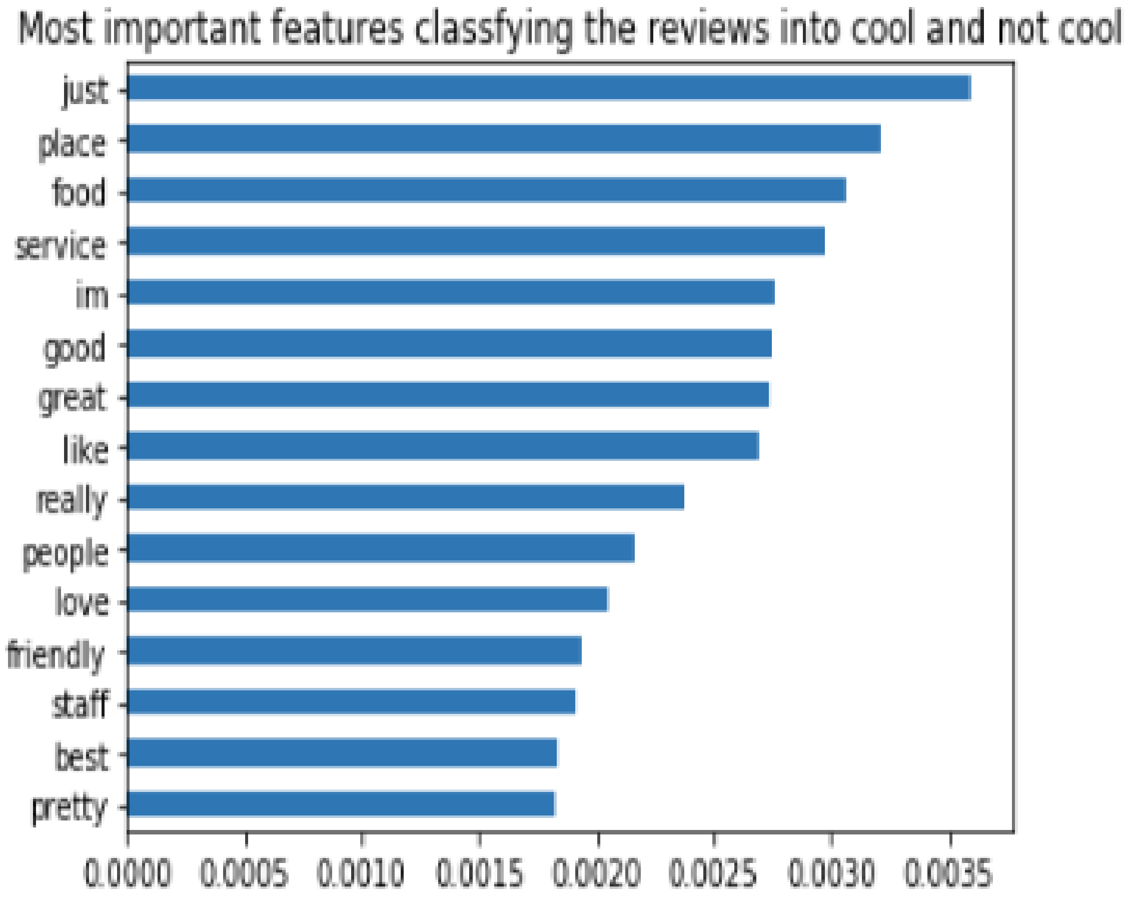
#5 Insights & Next Steps

Positive & Important keywords within the reviews for Classification

Positive/Negative keywords to group reviews into "Cool" vs "Not Cool"

(+) eat, new, staff, come, salad, did, say, delicious, right, want, fresh, better, going, went, friendly, didnt, night, ordered, order, cheese, way, make, restaurant, think, lunch, pizza, know, menu, try, chicken, pretty, people, got, bar, best, ive, nice, little, service, im, dont, love, time, really, just, great, food, like, good, place

(-) did apologize, brought attention, gratuity, spoke manager, meals great, rudest, half appetizers, restaurant closed, food old, inconveniencing, food needs, disappointing meal, disrespectful, server went, good seating, bad pizza, valley location, awful food, definitely fresh, great food atmosphere, hostess said, ordered entrees, items order, manager didnt, lost business, waitress finally, little smaller, location north, lousy service, waiting 30 minutes,



#5 Insights & Next Steps

Positive & Important keywords within the reviews for Classification

Positive/Negative keywords to group reviews into "Funny" vs "Not Funny"

(+) delicious, staff, new, come, friendly, salad, fresh, came, say, right, want, better, did, went, going, night, lunch, cheese, way, didnt, make, pizza, ordered, order, think, restaurant, try, menu, pretty, chicken, know, best, bar, got, people, nice, ive, little, love, service, time, really, great, just, food, like, good, place

(-)wife chicken, said manager, wife ate, did apologize, brought attention, recommend hotel, got orders, spoke manager, meals great, rudest, half appetizers, restaurant closed, food old, inconveniencing, disappointing meal, disrespectful, server went, prices great food, good seating, bad pizza, valley location, awful food, hostess said, helpful service, ordered entrees, items order, manager didnt, lost business, waitress finally, little smaller, location north, lousy service, food visit, spice flavor, received good, waiting 30 minutes, like canned, manager stopped, restaurant industry

#5 Insights & Next Steps

Positive & Important keywords within the reviews for Classification

Top positive/Negative keywords to group reviews into "Above Average Votes" vs "Below Average Votes"

(+) try, dont, little, chicken, ive, staff, friendly, pizza, nice, best, time, really, just, love, like, service, place, food, good, great

(-) beautiful carin, carin, uye, bec, certainly dont, really dont think, server comes, rand, cane sugar, giggling, robyn, inevitable, wasi, forgiven, goats, translate, bottle champagne, todayi, bastards, ridden

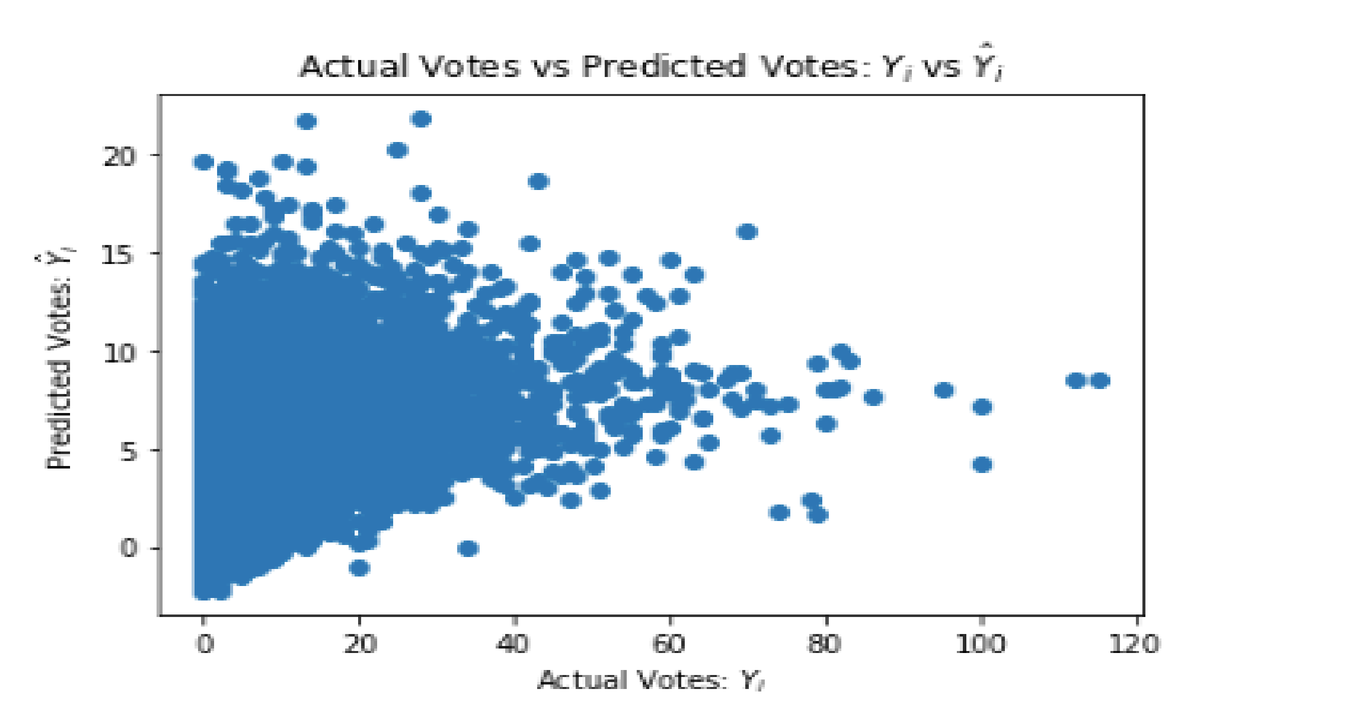
Next Step

- Clean up the text using more advanced techniques like stemming the word
- Acquire more reivew data from other state for further analysis.

#5 Insights & Next Steps

Relevant Variables identified within the regression model:

- Outcome Variable: Total Review Votes Count
- Predictor Variables:
  - Length of the reviews
  - Age of the reviews
  - User review count
  - Dummy variables created based on some important keywords identified from the classification model
- Result: R squared 36% with low p value. MAE: 2.6 votes



Next Step

- Develop more features (number of paragraphs in the review text, number of punctuation marks, number of business categories, number of insult words)
- Acquire more data (such as daily visits to the business page)

THANK YOU