California State University, Long Beach

STAT 576 : Data Informatics

An Analysis of the Yelp Dataset

Detecting Suspicious Reviews

Author:
Yale Quan

February 5, 2020

Contents

List of Figures

List of Tables

1 Abstract

Using a subset of the Yelp Dataset which only contains restaurants from Arizona this analysis compares the ability of Random Forests classifier and Stochastic Gradient Descent classifiers to accurately classify a restaurant as open or closed.

After cleaning the data and performing sentiment analysis using VADER and TextBlob it was found that combining VADER with Random Forest classification results in a 97.48% accuracy when classifying open or closed restaurants.

2 Introduction

People use Yelp everyday to decide where to eat, where to shop, and where to have fun. Users rely on the Yelp star system which uses proprietary algorithms to apply a 'Star Rating' to each business. This rating (a number between 1 and 5) has great influence on a persons decision about where to spend their money.

With the rise of social media and instant communication it becomes easier and easier for business to offer rewards for users who leave 5 star reviews. This influx of 5 star reviews can boost the overall star rating but mislead potential customers. These 5 star ratings may look good but not offer any insight on how much value the business provides.

Using Machine Learning and Sentiment Analysis/Natural Language Processing can help address this issue. Sentiment Analysis can be used to classify a review as positive, negative, or neutral. Machine Learning can then be applied to use this information in determining if a review is suspicious, or if a restraint is open or closed.

This analysis will compare two popular Sentiment Analysis packages VADER and TextBlob and two popular binary classification algorithms Random Forests and Stochastic Gradient Descent. Once compared a classifier will be constructed to classify restaurants as being open for business or closed.

3 Dataset Description

The yelp dataset contains 6,685,900 user submitted reviews about 192,609 businesses. These business span 10 metropolitan areas. The dataset is separated into 6 JSON files:

1. business.json
   Contains business data including location data, attributes, and categories.

2. review.json
   Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

3. user.json
   User data including the user's friend mapping and all the metadata associated with the user.

4. checkin.json
   Checkins on a business.

5. tip.json
   Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.

6. photo.json
   Contains photo data including the caption and classification (one of "food", "drink", "menu", "inside" or "outside").

This analysis will use the business.json and review.json files. CSV files were prepared using Python 3 prior to the analysis.

4 Exploratory Data Analysis

The business dataset is comprised of 14 variables with 174,567 observations and the review dataset contains 10 variables with 5,261,668 observations. Both datasets contain the variable *business_id* which will be used to merge the two datasets later in the analysis. The exploratory data analysis begins with checking for missing variables that can effect the analysis.

4.1 Yelp Business Dataset

Inside the business dataset the largest variables with missing values are neighborhood with 106,552 missing values and postal_code with 623 missing values. The review dataset contains no missing values. The neighborhood and postal_code columns are dropped from the business dataset.

After plotting the latitude and longitude it is clear that the dataset contains worldwide business. To narrow the scope of the analyses all business not in the United States of America were dropped from the analysis. This reduced the business dataset from 174,567 observations to 128,302 observations.

Figure 1: Plot of Business Location

4.1.1 Top 10 Business Categories

To continue the exploratory data analysis the top 10 business types inside the business dataset are calculated and plotted.

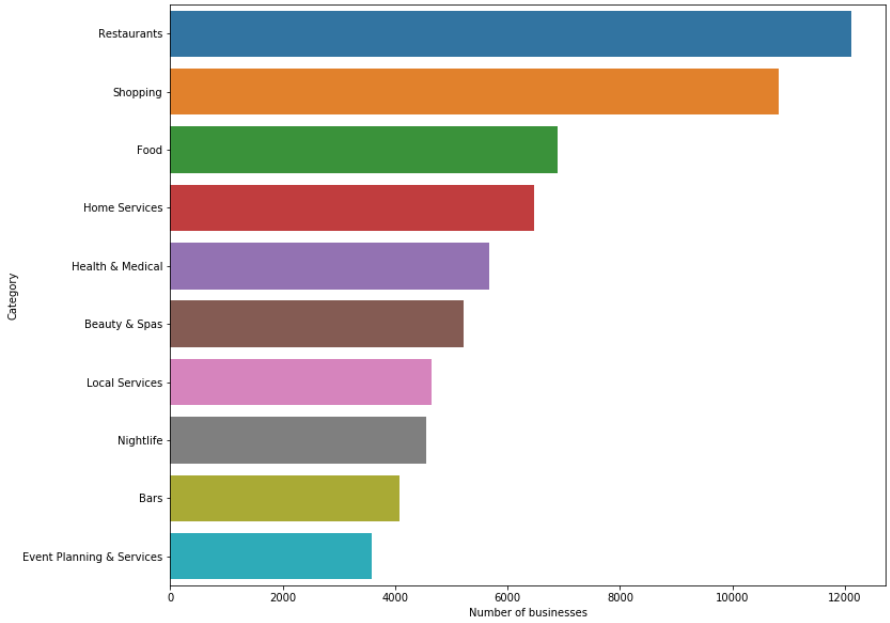


Figure 2: Top 10 business categories

Business Type	Number of Businesses
Restaurants	12,120
Shopping	10,822
Food	6,890
Home Services	6,467
Health & Medical	5,674
Beauty & Spas	5,218
Local Services	4,642
Nightlife	4,559
Bars	4,074
Event Planning & Services	3,583

Table 1: Top 10 Businesses

Looking at the above table there might be overlap between categories (i.e. Resturaunts and food) however, this analysis is not concerned with the overlap. Instead the dataset is subsetted into a restaurant only dataset. After performing the subset the USA Restaurant dataset contains 32,484 observations.

4.1.2 Star Ratings

Yelp uses a star rating to help users determine the 'best' business or restaurants. The exact way Yelp calculates a particular star rating is confidential. However, the star variable in the USA Restaurant is a natural variable to explore. Inside the USA Restaurant dataset the star rating assigned is the overall star rating. The overall star ratings are counted and then plotted in the below figure and table:

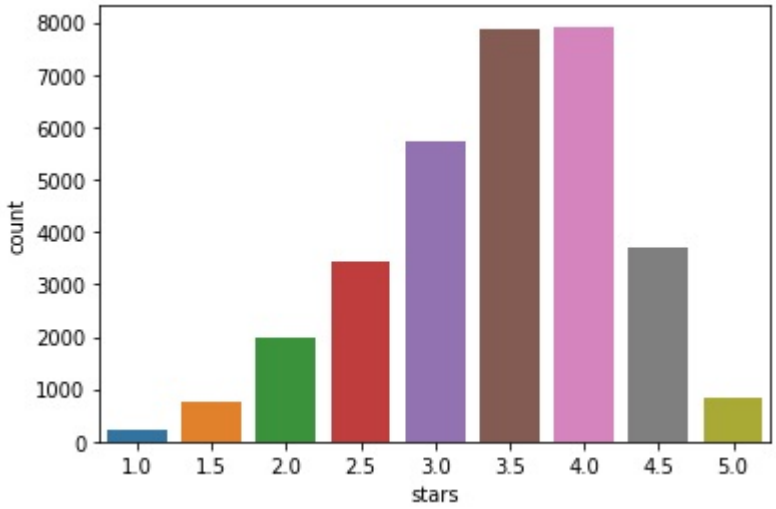


Figure 3: Distribution of Overall Star ratings

Overall Star Value	**Number of Occurrences**
1.0	3,788
1.5	4,303
2.0	9,320
2.5	16,148
3.0	23,142
3.5	32,038
4.0	33,492
4.5	24,796
5.0	27,540

Table 2: Overall Star Counts

From the above table there are 9 unique star ratings: 5, 4.5, 4, 3.5, 3, 2.5, 2, 1.5, and 1. Logically, users of Yelp might bin the stars before deciding on a place to go. For example: a rating of 1 - 2.5 might be considered low tier, 3 - 4 might be a middle tier restaurant, and 4.5 - 5 is a high tier restaurant. The figure and table below show the result of the binning.

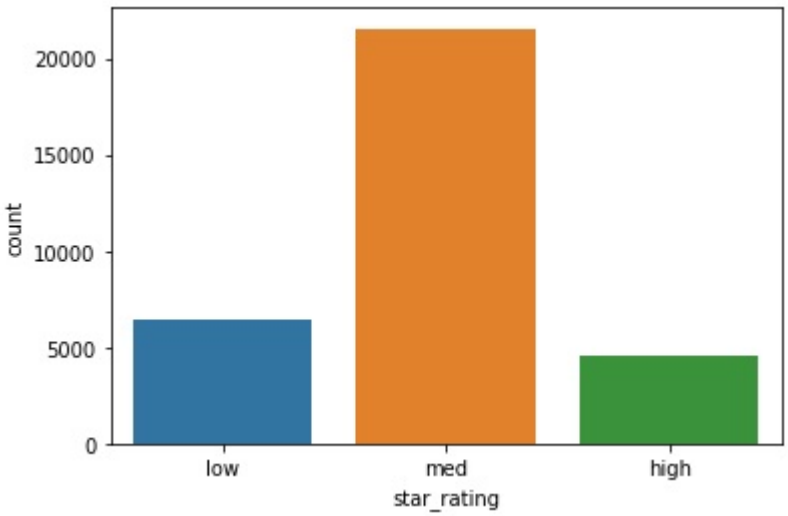


Figure 4: Distribution of Star Bins

Star Bin	Number of Occurrences	Percentage
High	52,336	29.98
Med	88,672	50.80
Low	33,559	19.22

Table 3: Overall Star Bin Counts

4.1.3 Yelp Star Rating Distribution

The majority of USA based restaurants receive a medium star score. By itself this information is not very helpful, however there might be a relationship between star value and open and closed status of a restaurant. After separating the USA dataset into open and closed restaurants the mean star value is calculated for each dataset.

Business Status	Star Average
Open	3.423
Closed	3.425

Table 4: Average Overall Star Values for Open and Closed Restaurants

This is not very informative so the percentage of star types is also calculated. The open restaurants has 23,393 observations while the closed dataset has 9,091 observations. Using the percentage of star values will help address the imbalance.

Overall Star Rating	Percentage in Open	Percentage in Closed
1.0	0.80	0.53
1.5	2.71	1.48
2.0	6.50	4.90
2.5	10.50	10.68
3.0	16.38	20.99
3.5	23.54	26.17
4.0	24.99	22.82
4.5	12.07	9.84
5.0	2.51	2.57

Table 5: Average Overall Star Percentages for Open and Closed Restaurants

4.1.4 Correlation Between Yelp Overall Stars and Open Status

Looking at the above table there is no clear distinction between the percentage of star values received and the open and closed status of the restaurants. This is further confirmed by analyzing the correlation between the open/closed status and the star value. The Pearson Correlation and the Point-Biserial Correlation Coefficient were calculated.

The Pearsons Correlation Coefficient (PCC) is the standard way of determining correlation between two variables. PCC is a measure of the liner strength between two variables and is reported as a value between -1 and +1. Generally,

PCC is used to calculate the correlation between two interval variables. In the USA Resturaunt dataset the *is_open* variable is binary; 0 = Closed and 1 = Open. Thus, the Point-Biserial Correlation Coefficient is also calculated. The Point-Biserial Correlation Coefficient is used when one of the two variables in the correlation calculation is binary. Mathematically the Point-Biserial Correlation Coefficient is equivalent to PCC.

	Stars	Open Status
Stars	1.0	0.001022
Open Status	0.001022	1.0

Table 6: Pearsons Correlation between Overall Star values and Open Status

Point-Biserial Correlation	P-Value
0.001022	0.8538

Table 7: Point-Biserial Correlation between Overall Star values and Open Status

From Table 6 and Table 7 it is evident that there is no strong correlation between overall star values and open status. The Point-Biserial Correlation returns a value of 0.001022 with a P-Value of 0.8538 which indicates that the correlation is not significant. Thus, it can be concluded that the overall star value is not a good predictor for open and closed status. The exploratory data analysis continues with the Yelp Review Dataset

4.2 Yelp Review Dataset

The Yelp review dataset contains 10 variables with 5,261,668 observations. These observations are written by individual Yelp Users and will contain multiple reviews per restaurant. The variables of interest in this dataset are review_text, length, and star value.

The analysis begins with merging the datasets and splitting the dataset into open and closed restaurants. Using an Inner Join will preserve the

multiple reviews and creates two variables for star value: star_x for the overall star rating per business, and star_y which is the star value the individual reviewer gave the establishment.

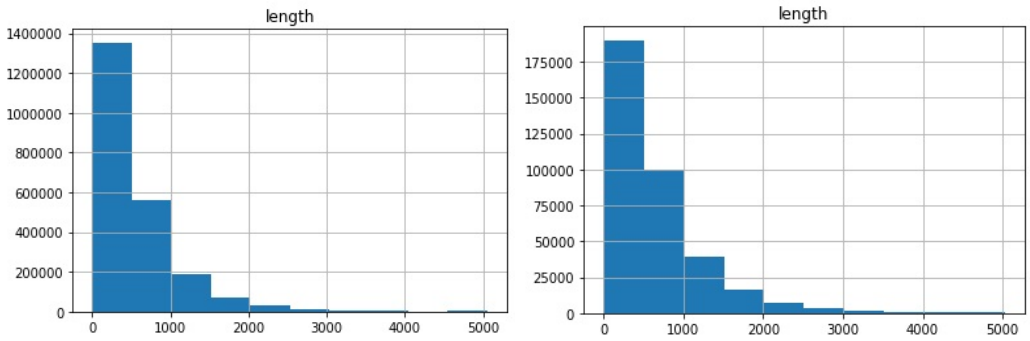
The analysis begins by looking at review length and its possible correlation to open and closed status of the restaurant.

4.2.1 Exploratory Analysis of Review Length

Business Status	Average Review Length (rounded to whole number)
Open	570
Closed	665

Table 8: Average Review Length

Of note is that users who reviewed a restaurant that is currently closed, on average, wrote 95 more words than users who reviewed a restaurant that is currently open. Graphing the length of reviews also provides useful information about the data:



(a) Open Restaurants

(b) Closed Restaurants

Figure 5: Comparison of Open and Closed Review Lengths

The open restaurants have considerably more reviews then the closed

restaurants. The distribution of reviews is skewed left which indicates that the majority of the reviews are short. Next is to look at the correlation between review length and open closed status. Again, the Pearson Correlation Coefficient and the Point-Biserial Correlation Coefficient are calculated.

	Length	Open Status
Length	1.0	-0.059374
Open Status	-0.059374	1.0

Table 9: Pearsons Correlation between Review Length and Open Status

Point-Biserial Correlation	P-Value
-0.059374	$\alpha < 0.05$

Table 10: Point-Biserial Correlation between Review Length and Open Status

From the above tables it is clear that the review length has a medium negative correlation with open status with a significant P-Value. This can be interpreted as longer reviews are correlated with a closed restaurant. This strongly suggests that review length can be used as a predictor variable for open and closed status.

4.2.2 Exploratory Analysis of User Stars

The analysis begins by plotting the overall user given star ratings.

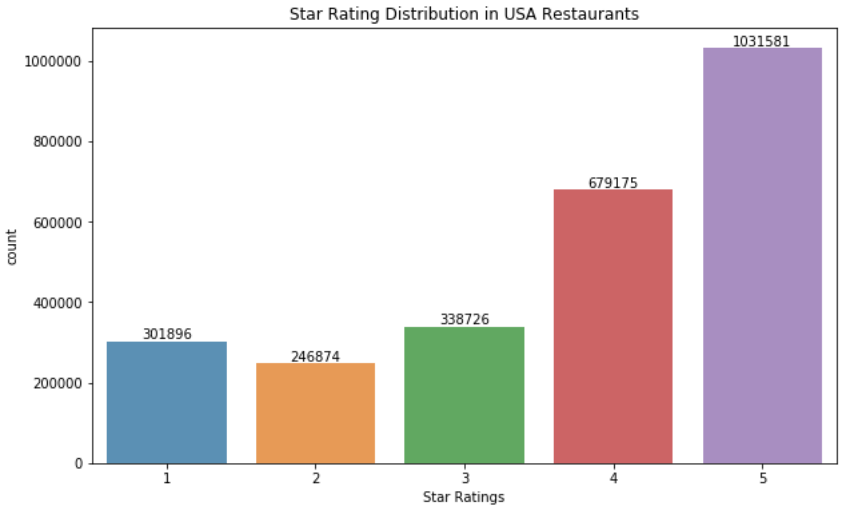


Figure 6: Distribution of User Star ratings

Overall Star Value	Number of Occurrences	Percentages
1.0	301,896	11.62
2.0	246,874	10
3.0	338,726	13.04
4.0	679,175	26.14
5.0	1,031,581	39.70

Table 11: User Star Counts

From the above graph and table the star ratings are heavily skewed left with 5 star ratings making up 39% of the data. This skew might affect analysis later and should be taken into consideration when interpreting results. Next is to analyze the correlation between user given stars and open closed status. Again, the Pearson Correlation Coefficient and the Point-Biserial Correlation Coefficient are calculated.

	Length	Open Status
Length	1.0	0.050481
Open Status	0.050481	1.0

Table 12: Pearsons Correlation between Individual Stars and Open Status

Point-Biserial Correlation	P-Value
0.050481	$\alpha < 0.05$

Table 13: Point-Biserial Correlation between Individual Stars and Open Status

From the correlation analysis there is a significant medium positive correlation between the value of individual stars given and the open/closed status of a restaurant. Therefore, higher individual star ratings correlate with a restaurant remaining open. This is expected because higher star ratings should indicate a better performing restaurant.

4.3 Subsetting by State

The last part of the exploratory data analysis is to reduce the dataset size due to computational limits. After counting the observations in each state the data will be subsetted so only the state with the largest observation will remain.

This subsetting was performed due to computing power. The below sentiment analyis and modeling was performed on a laptop with an Intel Core i7-8750H 2.20GHz(12 CPU) with 16gb of RAM.

Performing sentiment analysis on the original dataset took between 7-10hrs depending on the function being performed. After subsetting the analysis took approximently 3 hrs.

State	Number of Restaurants in Dataset
NV	1041833
AZ	920356
NC	201422
OH	169961
PA	158521
WI	75039
IL	24174
SC	6808
NY	79
IN	25
AK	21
CO	6
CA	4
VA	3

Table 14: Number of Restaurants in Each State

5 Sentiment Analysis of Yelp Restaurant Reviews

The next part of the analysis will focus on different types of sentiment analysis. The sentiment analysis will be focused on the review text partitioned

into open and closed restaurants. VADER (Valence Aware Dictionary and sEntiment Reasoner) is generally regarded to be one of the best methods for sentiment analysis. The goal of this section is to apply VADER to identify a baseline and compare VADER to other methods.

5.1 Classifying Negative and Positive Reviews using VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner)is comprised of 2 separate sections: A lexicon of terms and a rule-based sentiment analysis tool. VADER provides an positive and negative score and also provides output on how positive or negative a sentiment is. An example analysis is below:

sentiment_analyzer_scores("The phone is super cool.")
('neg': 0.0, 'neu': 0.326, 'pos': 0.674, 'compound': 0.7351)

Figure 7: Example of output from VADER

The positive, negative, and neutral scores represent the percentage of the sentence that fits those categories. In the above example the sentence was 67% positive, 33% neutral, and 0% negative. The compound measurement determines the strength of the highest percentage component with a value between -1 and +1. with 1 being completely positive, -1 being completly negative, and 0 being neutral. Thus, a value of 0.74 indicates the sentence is strongly positive.

Before continuing the analysis it is important to know that VADER also incorporates punctuation into the analysis. For example: using an exclamation mark **(!)** will increase the magnitude of the sentiment without changing the semantic orientation. Additionally, capitalization will also emphasize the sentiment without changing the semantic orientation.

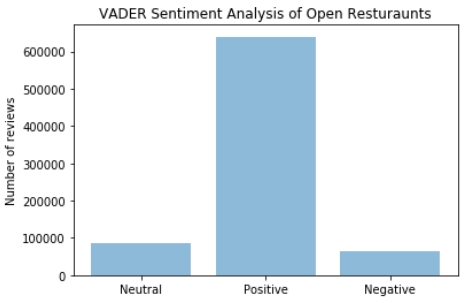
The first step in the analysis is to use VADER to count the positive, negative, and neutral reviews in the dataset. According to the VADER documentation we will use $x > 0.5$ as positive, $x < -0.5$ as negative and

$-0.5 \leq x \leq 0.5$ as neutral.

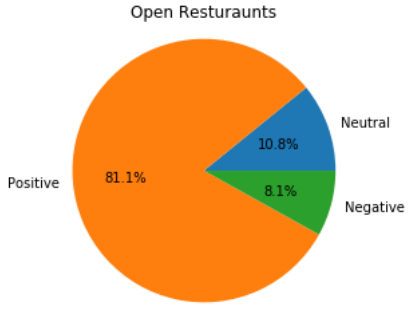
VADER was first used to count the number of positive, negative, and neutral reviews. This was calculated by looking at the compound value and incrementing a counter for each type of review. Neutral was incremented if the compound value was between -0.5 and 0.5, positive was incremented if the compound value was larger than 0.5, and negative was incremented if the compound value was below -0.5. The results of that analysis is below:

	Positive Reviews	Negative Reviews	Neutral Reviews
Open Restaurants	639,939	63,599	85,518
Closed Restaurants	105,376	11,291	14,633

Table 15: VADER counts with Punctuation and Capitalization

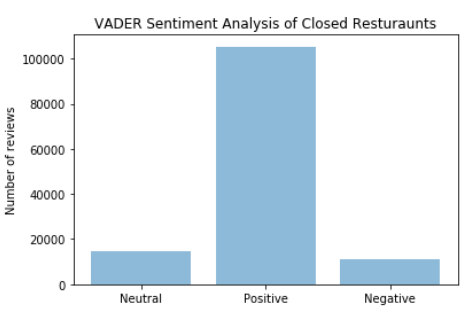


(a) Histogram of VADER counts

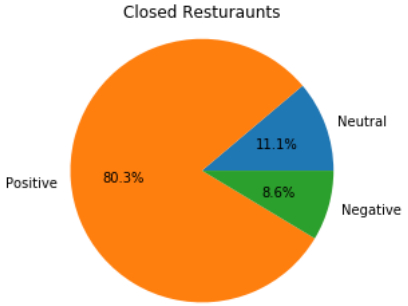


(b) Pie Chart of VADER counts

Figure 8: VADER counts on Open Restaurants



(a) Histogram of VADER counts



(b) Pie Chart of VADER counts

Figure 9: VADER counts on Closed Restaurants

After performing the VADER analysis there appears to not be a significant difference between the amount of positive and negative reviews an establishment receives and its open and closed status. This will still be used as a baseline to compare against another sentiment classifier.

5.2 Sentiment Analysis using TextBlob

TextBlob is another popular sentiment analysis library in part because it is very simple to use. TextBlob is a sentiment analysis library that uses the NLTK library and allows to easy parts-of-speech tagging, sentiment analysis, language translation, and text classification.

TextBlob requires that input text strings be converted into a proprietary string type called a 'TextBlob' Once converted the TextBlob can be passed into a sentiment analyzer which produces a sentiment score which ranges form -1 to +1. The more negative a string is the lower the score will be and the more positive a review is the larger the value is. A neutral review will have a score of zero. The subjectivity score will not be used in the analysis, but it is a measure of how subjective the review is.

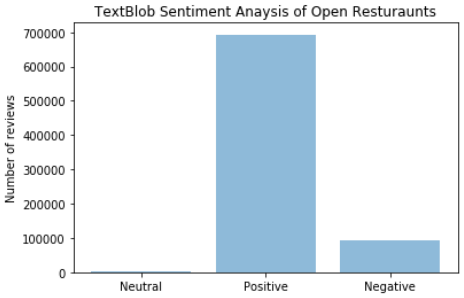
blob = TextBlob("The phone is super cool."); blob.sentiment
Sentiment(polarity=0.34167, subjectivity=0.65833)

Figure 10: Example of output from TextBlob

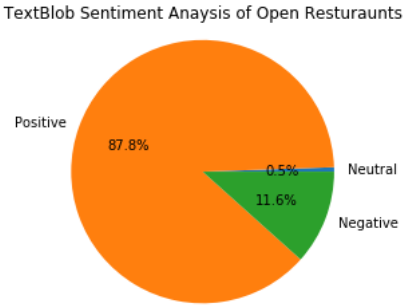
After passing the review column into a TextBlob converter and applying the sentiment analysis the below table and charts were produced:

	Positive Reviews	Negative Reviews	Neutral Reviews
Open Restaurants	693,148	91,683	4,225
Closed Restaurants	114,033	16,550	717

Table 16: TextBlob counts with Punctuation and Capitalization

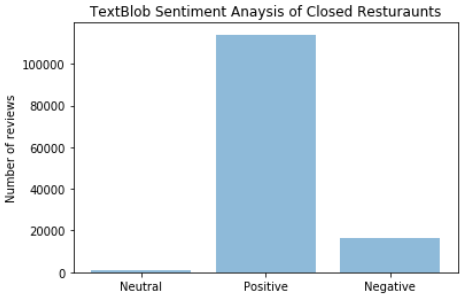


(a) Histogram of TextBlob counts

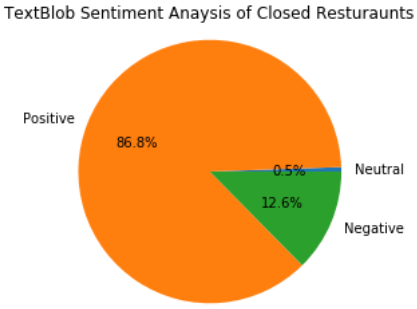


(b) Pie Chart of TextBlob counts

Figure 11: TextBlob counts on Open Restaurants



(a) Histogram of TextBlob counts



(b) Pie Chart of TextBlob counts

Figure 12: TextBlob counts on Closed Restaurants

5.3 Comparison Between VADER and TextBlob

From comparing the output of VADER and TextBlob there is not much difference between the sentiment classifiers. Therefore, the analysis will use the VADER sentiments for model building

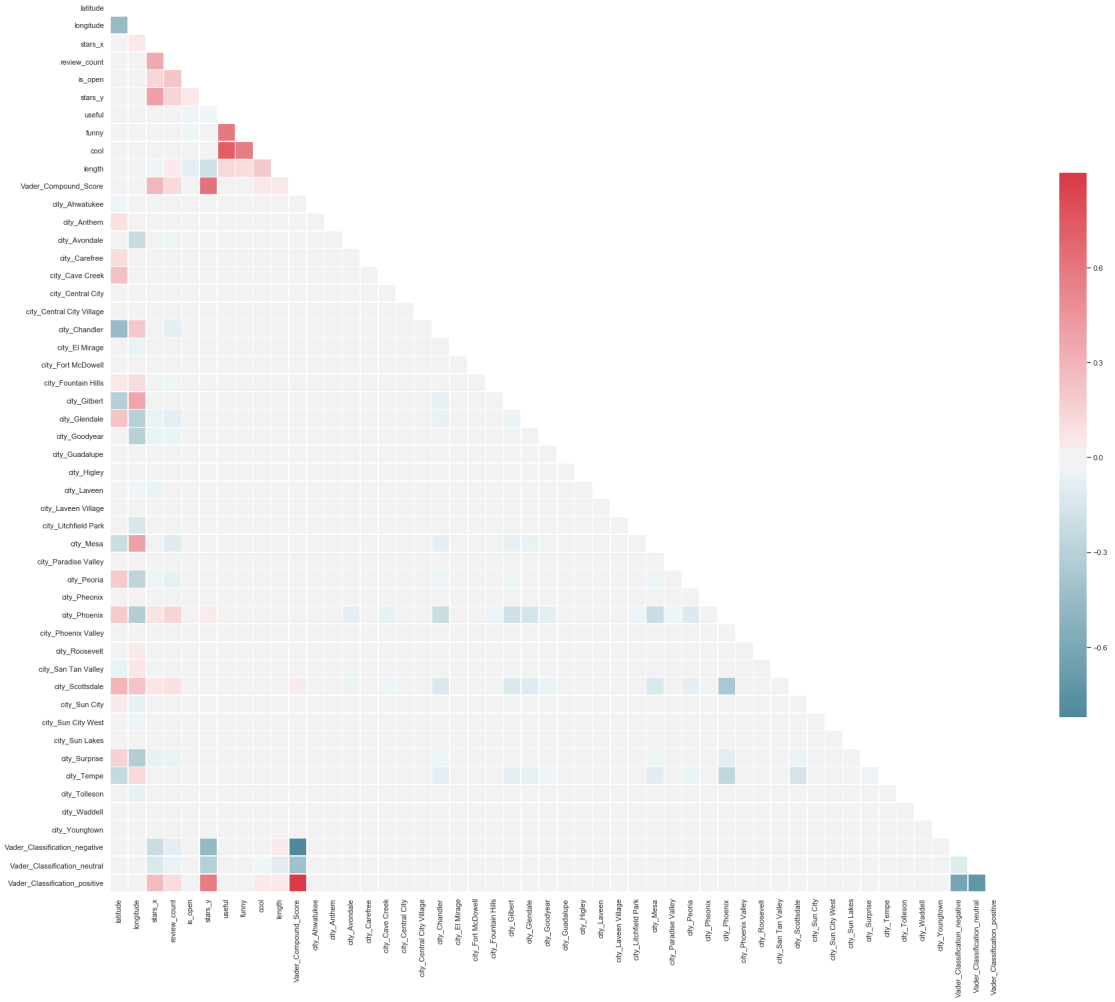


Figure 13: Correlation in Cleaned Data

Using the above graph and Python output there is no significant correlation in our features. Therefore we can proceed with building a model to classify open and closed status.

6 Modeling

The goal for the modeling is to find a classifier that is a good predictor for open and closed status. The following variables were used for the model:

6.0.1 SMOTE

The Yelp dataset is very unbalanced with over 20% more open restaurants than closed restaurants. To address this imbalance SMOTE is applied to the dataset. SMOTE stands for Synthetic Minority Over-sampling Technique. SMOTE will oversample the lesser data to create a balanced dataset which should be representative of the entire dataset.

6.1 Random Forest

The Random Forest algorithm is a supervised classification algorithm. The Random Forest classifier creates multiple decision trees and each tree only considers a random subset of input features and training points. For classification the Random Forest will take the majority vote of all the trees constructed.

The Random Forest constructed for this classification has 100 trees and uses the entropy option for decision making. After applying SMOTE and fitting the Random Forest the following confusion matrix was generated:

	Predicted	**closed**	**open**
Actual			
closed		161307	4168
open		2331	163472

Table 17: Random Forest Confusion Matrix

Predicted	**closed**	**open**
Actual		
closed	0.974812	0.025138
open	0.014087	0.985941

Table 18: Normalized Random Forest Confusion Matrix

Looking at the above confusion matrices the Random Forest the model has an sensitivity of 97.48% and a specificity of 98.59%. This can be interpreted as 97.48% of the time this model will accurately classify a restaurant closure and 98.59% of the time will accurately classify a restaurant as open.

It is important to note that these numbers are suspiciously high and should be investigated in further analysis. The assumption is that the model is overfitting and more variables should be introduced.

6.2 Stochastic Gradient Descent (SGD) Classifier

The Stochastic Gradient Descent classifier is an unsupervised machine learning process where samples are selected randomly from the testing and training sets. In SGD, the algorithm determines the gradient of the cost function of a single example at each iteration. This iterative process uses less processing power than Batch Gradient Descent (BGD).

The main advantage of SGD over BGD is that the Gradient Descent Classifier is sensitive to ordered data. If data is presented in an order to the process the algorithm will learn bias based on the order. With SGD the samples are randomly selected from the testing data to reduce the possibility of bias.

6.2.1 Default SGD

Using the default Stochastic Gradient Descent model in sklearn produces a model with 85% accuracy. This is acceptable accuracy for the model, however parameter tuning might increase the accuracy of the model.

	Precision	Recall
Actual		
closed	0.77	0
open	0.86	1

Table 19: Stochastic Gradient Classifier Classification Report

Accuracy	0.8584

Table 20: Stochastic Gradient Classifier Accuracy

6.2.2 Hyperparaemeter Tuning

Using the GridSearchCV in Python an algorithm is run to determine the optimal hyper parameters. GridSearchCV methodically iterates through possible fittings and return the best parameters for the model.

Parameter	Value
α	0.00001
loss	Hinge

Table 21: Stochastic Gradient Parameters

	Precision	Recall
Actual		
closed	0.97	0
open	0.97	0.19

Table 22: Stochastic Gradient Classifier Classification Report, $\alpha = 0.00001$

Accuracy	0.2966

Table 23: Stochastic Gradient Classifier Accuracy, $\alpha = 0.00001$

Tuning the parameters had a negative effect on the model The accuracy reduced from 86% to 27%. Therefore, only the default model will be considered for comparison. The model is then adjusted by hand with different α levels. $\alpha = 0.0001$ is the default and will be skipped.

Parameter	Value
α	0.001
loss	Hinge

Table 24: Stochastic Gradient Parameters

	Precision	Recall
Actual		
closed	0.84	0
open	0.86	1

Table 25: Stochastic Gradient Classifier Classification Report, $\alpha = 0.001$

Accuracy	0.85833

Table 26: Stochastic Gradient Classifier Accuracy, $\alpha = 0.001$

6.2.3 SGD Conclusion

The final step for SGD is to compare the performance of the various α levels:

Alpha	Accuracy
0.001	0.85833
0.0001	0.8584
0.00001	0.2966

Table 27: Stochastic Gradient Classifier Classification Comparison

Comparing the accuracy levels it is clear that using $\alpha = 0.0001$ is the ideal choice for SGD.

6.3 Model Comparison

Comparing the two classification techniques it is clear that the Random Forest is the superior classifier. The Random Forest classifier had an accuracy of 97.48% which is higher than the 85.84% that the default SGD provided.

7 Suspicious Reviews

For the purpose of this section and subsequent analysis a suspicious review will be defined as a review where the sentiment classification does not match the user star rating value. An example would be a user giving a restaurant a 1 star but the sentiment classification is positive with a high VADER compound score.

7.0.1 Designing the Filter

Using a boolean filter with the classification that if user stars were less than or equal 2 and the VADER sentiment is positive the review is flagged as suspicious. Likewise, a negative review is suspicious if the user gave a star rating greater than or equal to 4 and the VADER sentiment was negative. A FALSE flag will indicate the review is not suspicious while a TRUE flag indicates a review is suspicious.

Filter Example
if Stars = 4 and VADER Sentiment = Positive then Filter = FALSE
if Stars = 4 and VADER Sentiment = Negative then Filter = TRUE

7.0.2 Filter Output

	Suspicious	Not Suspicious	Percent Suspicious
Positive Reviews	76,824	843,532	8.34%
Negative Reviews	4,075	916,281	0.44%

Table 28: Suspicious Reviews

From the above table less than 1% of negative reviews were flagged as suspicious while 8.34% of positive reviews were flagged as suspicious. This could be due to restaurants offering incentive for high star reviews while not being concerned about the content of the review. This also exposes a flaw in the Yelp star ratings where reviews could be incentivized to leave high star reviews without positive content.

8 Conclusion and Limitations

From the analysis Random Forests are the best classifier to use for classifying open and closed restaurants in the Yelp Dataset. When compared to a Stochastic Gradient Descent classifier the Random Forest was 97.28% accurate compared to the 85.8% accuracy from the SGD classifier.

Comparing VADER to TextBlob the analysis revealed that both perform similarly well but VADER is stronger at handling neutral reviews. Using VADER for sentiment classification the algorithm classified $807,181$ reviews as positive, $108,233$ reviews as negative and $4,942$ reviews as neutral. Of those reviews $76,824$ of the positive reviews were suspicious and $4,075$ of the negative reviews were suspicious.

8.1 Limitations

The largest limitation to this analysis was processing power. The full Yelp Dataset has over 6 Million observations with over 30 different variables. This dataset could not be analyzed on the current laptop being used. When a subsection of 3 Million was originally used the analysis took over 7 hours to run. Using only 1 Million observations allowed the data to be analyzed in a meaningful timeframe.

8.2 Future Work

Future research should be performed using Logisitc Regression and other classifier types. Currently the scykit package produces an error when fitting Logisitc Regression that could not be fixed in the time allowed.

Additional analysis should be performed on the complete dataset as well. Sentiment analysis should be performed on the complete dataset to determine if sentiment is significant to a restaurants being closed or open.

The last limitation to this analysis is the amount of positive reviews. The dataset contains a majority of positive reviews which might affect further research with this dataset. Further analysis should include the full dataset and discuss the distribution of reviews.