Yelp Review Sentiment Analysis

Abstract

Yelp is an online social network service in which consumers provide reviews on their local business restaurants. Because our data is in the form of human text, we can utilize natural language processing (NLP) and machine learning to vectorize the text reviews and apply sentiment analysis through logistic regression. Through our research, we present a case study in which we examine relationships between sentiments and ratings as well as the aspects consumers talk most positively and negatively about from Yelp restaurant review data.

1. Introduction

In a world where data is becoming increasingly more important, people have a prerequisite demand to have more information before spending their money on a product or service. This is especially evident in review-based applications like Yelp. Yelp, founded in 2004, is an online social network service in which consumers provide reviews on their local restaurants by describing their individual experience with a review and rating on a 1 to 5 star-rating scale. Consumers can see what a new restaurant will entail in terms of type of food, amenities, etc. before commuting. Restaurant owners benefit by receiving valuable, direct customer feedback on service, quality, location, amenities, etc. Reviews may contain key factors consumers are looking for in their restaurants that the restaurants may be currently missing or are performing poorly on. Analyzing these reviews are an integral part of continuous improvement because they could result in future, higher reviews and positive feedback. Higher reviews will also enable the restaurant to become more recommended and popular in search engine entries.

Sentiment analysis is a vital technique in which we will be classifying the review text as either positive or negative by incorporating aspects of Natural Language Processing (NLP) and machine learning. Specifically, we will conduct logistic regression and determine the sentiment based upon the evaluation of a restaurant review. From the Yelp review data, consumers discuss their experiences and opinions through their reviews, thus natural language processing and sentiment analysis can be utilized. Ultimately, a review is what the opinion of the user is about the business in his or her own words and not a rating or mathematical predictive task, thus, it is essential to be able to predict what the user feels about a business from the review text (Channapragada & Shivaswamy, 2015, p. 1) [1]. In this project, we conduct data preprocessing, topic modeling, exploratory descriptive analytics, modeling, results and conclusions. Overall, sentiment analysis can help understand consumers as a whole; detecting business' day-to-day successes and faults, and remedying them over time.

2. Data

The dataset that we had used for our research was the Yelp Dataset that included JSON files of business, checkin, reviews, tip, and users. With our analysis, we had limited our dataset to that of the business and review JSON files. The business file includes business_id, stars, attributes, and categories, with around 8 million records. The review file includes the business_id and reviews at about 200,000 records. The two datasets will be combined to produce one dataset to work from. Our focus on Yelp is to understand the relationship between text reviews and ratings provided by the consumers' experiences at their local restaurants. We incorporate sentiment analysis to our research to create sentiments based on text reviews provided by the consumer.

3. Preprocessing

3.1. Overview

One of the bigger concerns with the dataset was the size and volume of the data. With the two utilized datasets, the task of joining both the business data and the review data proved to be very expensive. As each of the 209,393 unique businesses in the dataset had hundreds to thousands of reviews, it necessitated the requirement to narrow down our analysis to study a certain niche of Yelp restaurants. This would make both topic modeling and the application of the business problem more viable. Ultimately the study was conducted on Japanese restaurants' reviews in the city of Las Vegas, Nevada. In this city was the most heavily concentrated number of Yelp user reviews for establishments.

3.2. Basic Cleaning

Regarding the business dataset, all geographic and location-based features were dropped such as address and coordinates. Next, filtering was utilized to reduce the dataset down to Japanese restaurants in Las Vegas. After filtering, the dataset comprised 437 individual restaurants. Further dataset cleaning was done such as

querying for NA values. 80 missing values were found in the column "Attributes", and these records were removed. Finally, once the business dataset was cleaned and filtered, this made joining with the review dataset less demanding. The business dataset and the review dataset were joined on the column business_id. Finally, the joined dataset resulted in 148,224 records. This represented all the Japanese-style restaurants in Las Vegas and all their respective Yelp reviews.

3.3. NLP

For machine learning and data science tasks, the cleaning of text is a very important step. To clean the textual data of the reviews, a pipeline for normalization was implemented to conduct regex preprocessing and building the corpus. In this pipeline, special characters, punctuation, and white spaces were removed and lower case sensitivity was ensured. Also, stop words were filtered and removed while tokens were extracted into the created corpus. An important and final facet of this pipeline is the use of Porter Stemming. PorterStemmer employs the uses of suffix stripping in order to produce stems of words, providing a powerful, efficient, and useful way to avoid features from being overpopulated by words with similar meaning. From sci-kit learn, CountVectorizer was used to vectorize the tokenized text and create a Bag of Words (BOW). This is a necessary step because machine learning modeling requires that the text is represented as an encoded vector that represents the entire vocabulary and an integer count for the frequency of word appearance in the document.

Additionally, terms that appeared in less than 20% of reviews were ignored while terms that appeared in more than 80% of reviews were ignored. This is specified in the parameters for min_df and max_df.

TF-IDF is employed to compute an assigned weight to each word and denoting its importance of the word in the document and the corpus. While TF (term frequency) measures the frequency of a word in a document, DF (document frequency) measures the significance of the document in the corpus or the count of occurrences of term *t* in the document set *N*. Simply put, DF measures the frequency of documents where the word appears, and IDF is defined as the inverse of the document frequency, assessing the significance of the term *t*. In this step, the Bag of Words was transformed to TFIDF. Though BOW is sufficient to use text data for the model and may be easier to interpret, there are some evident drawbacks. For one, BOW in itself does not consider the order of words. Also, because BOW is based on absolute frequencies, more frequent terms could possibly overwhelm other terms. TF-IDF, on the other hand, encompasses information on both more significant features and less important ones. Similar studies conducted on movie reviews sentiment analysis have shown that TFIDF outperforms the accuracy given from the use of basic bag of word's representation (Bijoyan and Chakraborty, 2018, p. 6) [2].

Term Frequency:

$$tf(w, D) = fwd\ (frequency\ \ for\ word\ w\ in\ document\ D)$$

Inverse Document Frequency

$$idf\ (w, D)\ =\ 1 + log\frac{N}{1 + df(w)}$$

TF-IDF

$$tfidf = tf \times idf$$

Figure 1. TF-IDF Formulas

Finally, in order to conduct supervised learning, the Yelp reviews must be labeled. To do this, all the reviews that gave the restaurant experience a 3 and below were labeled as a negative review and the rest of 4 and above were labeled as a positive review. Again, 0 denotes negative and 1 signifies positive sentiment. Labeling the reviews in this manner provided over 50,000 more positive reviews than negative.

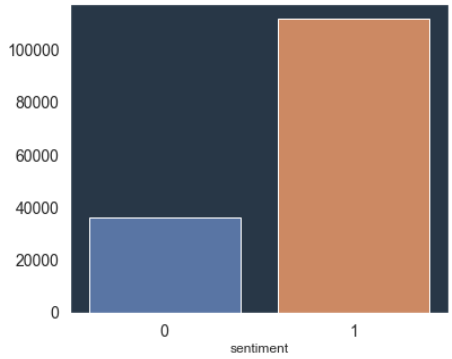

Figure 2. Sentiment (Neg/Pos) Value Counts

4. Topic Modeling

4.1. Overview

Through topic modeling we can understand the relationships between that of documents and terms, by creating a set of concepts for documents and terms of Yelp reviews. We implement methods of Latent Semantic Analysis (LSA) and Singular Value

Decomposition (SVD) with our normalized corpus. Our objective is to create concepts from the normalized corpus of Yelp text reviews. As Ashgar illustrates, It finds 'topics' in reviews, which are words having similar meanings or words occurring in a similar context, thus applying this method, we hope to find similar collections of two terms when running bi-grams belonging to a specific number of concepts (Ashgar, 2016, p. 4) [3]. With the implementation of SVD, we will be decomposing the normalized review corpus into three constituent matrices. The SVD function outputs three matrices: the word topic matrix U of size m x m, the rectangular diagonal matrix S of size m x t containing t singular values, and the transpose of the topic-review matrix V of size t x t (Ashgar, 2016, p.4) [3]. Our feature matrix will be represented as V and our S will represent our singular value matrix which will measure the importance of that topic in a descending order. Such method is important in comprehending the terms and their values of importance to the text, seeing if there are any discrepancies, or if there is a relationship between customer experiences that they have cited within their individual Yelp reviews.

Singular Value Decomposition (SVD):

$$X = U * \sum * V^T (singular\ value\ decomposition)$$

4.2. Latent Semantic Analysis (LSA)

We use TF-IDF and bi-grams to get the frequency of two terms with our max features at 12,500. Next, we incorporate SVD by decomposing the matrix into three constituent matrices setting the number of components to 5, iterating through it 100 times. Through the below figure, we can visualize the singular values and their relative importance of each component and the amount of variation captured from each concept.

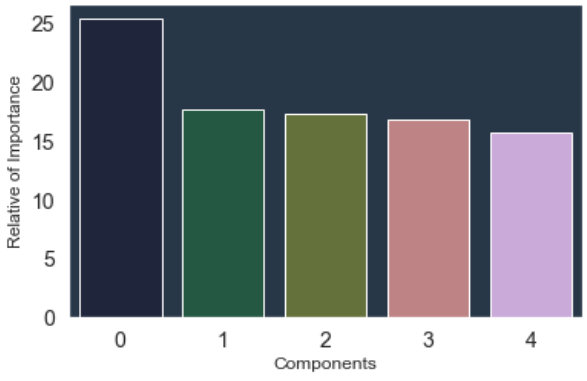

Figure 3. LSA Component Weights

From each of the 5 components, we will be combining the terms with their concept importance and sorting based off of the top 10 words within the concept. This provides insights on the importance of the terms within each concept. So when looking at the terms provided by Yelp reviews of customers, we can see the importance of each of the terms of these restaurants as a whole. Overall, LSA is giving an average conceptual idea of what Japanese restaurants in Las Vegas entails based on reviews and experiences expressed by customers on Yelp.

5. Exploratory Descriptive Analytics

Distribution of star ratings was viewed by a bar plot. Through the below figure, we can see comparisons among discrete categories of star ratings. Each category in the bar plot is a respective star rating. It is clear that the majority of reviews for Japanese-style restaurants in Las Vegas range from 3 to 4.5 star ratings. Based on this, we can assume that most of the reviews will be relatively positive. This may be due to performance or because consumers who seek out Japanese-style food in a city like Las Vegas may be more familiar with the style of food and know that they already enjoy it.

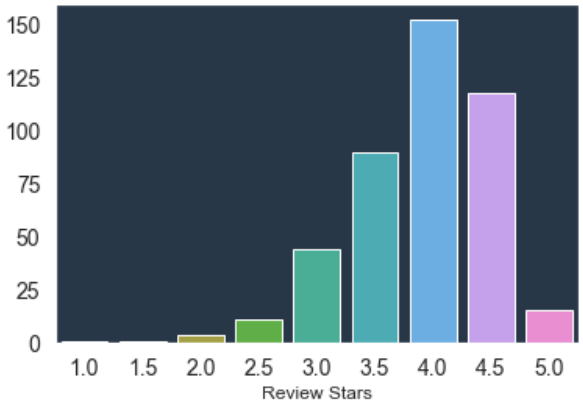

Figure 4. Review Stars Barplot

Continuing on the previous assumption, we can see in the Appendix Figure 1 the top 10 restaurants based on review count. The majority of this view are restaurants that cater towards the most well-known types of Japanese-style foods like BBQ, sushi, and ramen.

Distribution of the count of reviews for each restaurant was viewed by a histogram. Through the below figure, we can see that a majority of Japanese-style restaurants have less than 500 reviews. Lower review counts may be correlated to the generally lower, overall popularity of traditional Japanese food in a city like Las Vegas,

where Japanese-style food is tended to be fused with familiar styles of food in the area.

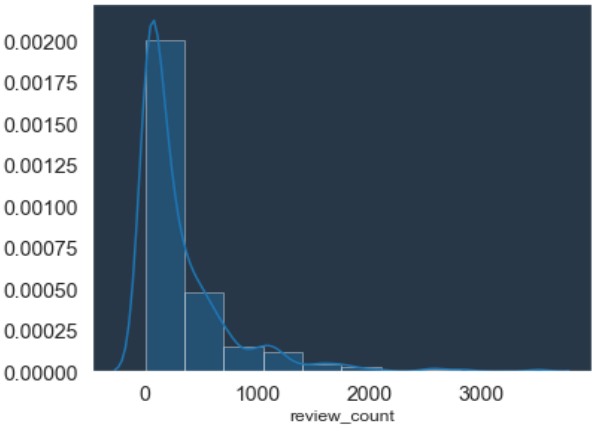

Figure 5. Review Count Histogram

Word clouds were produced in descriptive analytics to quickly visualize the most frequent words, in respect to top positive words and top negative words. These allow Japanese-style restaurant owners to swiftly view concepts or items consumers use most in their reviews. These can provide nimble insights on what people talk about most in their positive or negative reviews. Quick insights can be made due to the nature of the Word clouds. Word clouds visualize the most frequent words in a set of text easily by manipulating size. The larger the size of the text words in the word cloud, the higher the frequency of the words in the set of text. Also, stop words were filtered out of the reviews in the NLP cleaning step, so unmeaningful words were filtered out.

In the below "Top Positive Words" word cloud, a Japanese-style restaurant owner could quickly see the types of items or concepts consumers talked most positively about from Japanese-style restaurants. It should be made clear that these items or concepts are not in relation to a specific restaurant but to all of them included in the filtered dataset. For example, in the "Top Positive Words" word cloud, "ice cream", "sushi place", and "happy hour" are some of the most frequent words in positive reviews. These are just the concepts that consumers talked most about in positive reviews overall. A Japanese-style restaurant owner could integrate having quality ice cream on the menu or having competitive happy hours to possibly increase his or her number of positive reviews.

Figure 6. Positive Word Cloud

In the below "Top Negative Words" word cloud, a Japanese-style restaurant owner could swiftly see the most frequent words consumers talk most negatively about. Many of the most frequent, negative words are food types like "miso soup", "sushi roll", and "fried rice" as well as words like "customer service". Japanese-style restaurant owners can be informed that food quality and customer service are some of the attributes that consumers talk most negatively about in reviews. Through this, restaurant owners can make the appropriate adaptations towards these aspects.

Figure 7. Negative Word Cloud

6. Predictive Task

The predictive task for our study aims at predicting the sentiments of a user's experience at a Yelp restaurant solely given the textual data of their respective review. Our goal is to build a model that classifies the sentiment

of the review being positive or negative, without the supplemental categorization of a Yelp star or review rating. Additionally, we are interested in identifying the textual features that serve as the most significant for this method of classification. For our model, we utilize the discriminative model Logistic Regression. Logistic Regression encompasses the use of the sigmoid function, outputting a probability between 0 and 1. Furthermore, a binomial logistic regression predicts the probability that a specific observation is classified in one of the two categories, positive or negative.

7. Model

7.1 Splitting Data

The train-test split ratio was set at 80/20, respectively. The training data consisted of 118,579 reviews while the testing set consisted of 29,645.

7.2 Feature Selection

An important parameter for constructing our model is the number of features selected for the vectorized tokens in the corpus. We evaluated the accuracies based on the varying numbers of features allowed using another pipeline where the Logistic Regression model was fit to the training data and ran with 500 iterations using trigrams from the CountVectorizer. The evaluation of the selecting of n-grams will be discussed later in the study.

Table 1. Feature Selection Accuracy

# of Features	Accuracy
5000	91.99%
7500	92.18%
10000	92.22%
12500	92.31%
15000	92.30%

With the results above, the higher number of features starting from 5,000 results in higher accuracy. However, once the number of features exceeded 12,500, the accuracy dropped off in value. We decided to choose 12,500 features for further implementation.

7.3 N-Grams

When utilizing predictive modeling to predict sentiment labels of reviews, there can be potential drawbacks in using only single words as features or unigrams. Being able to understand certain negations such as “not good” and “not bad” will not be accounted for, possibly leading to poor classification. To take this into proper consideration, n-gram usage as features can lead to increased accuracy of the sentiment prediction model.

Table 2. N-Gram Selection

# of N-grams	Accuracy
1	91.41%
2	89.66%
3	82.76%

The results provided above were produced from a pipeline similar to the one used for evaluating feature selection. However, for this pipeline, the parameter for n gram range in the TFIDF vectorizer was run for n-grams 1, 2, and 3 or single feature, bigrams, and trigrams. From the results, we see that single features provide the most accurate model, with the use of trigrams having the least. Attaching a word such as a negation to another word that precedes or follows it can be a beneficial procedure that enables the improvement of a classification problem’s accuracy since negation can factor into opinion or sentiment expression. Many studies evaluate the performance of sentiment classification models based on n-grams. In a similar case that classifies polarity on Tweet sentiments, it was found that the best performance was achieved when using bigrams. Conclusively, bigrams maintain a solid mediary between both the comprehensive coverage that unigrams provide and also the ability to capture the sentiment expression patterns enabled through trigrams (Patodkar and I.R, 2016, p. 320-322) [4]. For our study, though unigrams had the higher accuracy, choosing bigrams for the model will prove much more effective. It will provide more conclusiveness in identifying phrases of words that are associated with both the positive and negative sentiment class.

7.4 Model Implementation

Following the evaluation and selection of our parameters including max_features and ngram_range, the final fitting of our data to the Logistic Regression

was implemented. For the model, max_iter, or the maximum number of iterations taken for the solvers to converge, was set at 100. Max_features was set at 12,500 and ngram_range was set for bigrams in (2,2).

8. Results

8.1 Feature Importance

We conduct feature importance for the purpose of gaining insights into the data and as well as the model. The implementation of feature importance assigns scores to input features based on how impactful they are at predicting a target variable. This is done by splitting the coefficients from the classifier into positive and negative features and plotting them to see their impacts on positive or negative sentiments. From the top positive and negative words, the most important features are labeled with scores on the y-axis.

From Appendix Figure 2, two of the most important features are "definite back" and "highly recommend". These inform Japanese-style restaurant owners that when positive experiences are had, consumers are more likely to return or spread their positive experiences with their communities which can be a form of free advertising by word-of-mouth.

From Appendix Figure 3, the opposite can be informed. Words like "won't back" and "never come" inform Japanese-style restaurant owners that consumers with negative experiences will not return and that they may spread those negative experiences with their communities which results in negative advertising by word-of-mouth. It can also be seen that "food poisoning" and "service horrible" are aspects that are highly correlated to negative sentiments. Japanese-style restaurant owners can adjust their business to steer away from these aspects.

8.2 Model Evaluation

Using the test data, the performance results from the predicted and actual values are visualized in the confusion matrix in Figure 8.

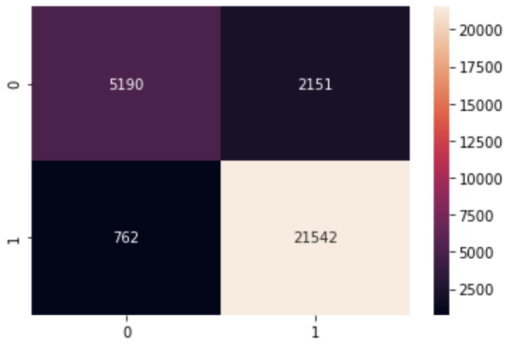

Figure 8. Logistic Regression Confusion Matrix

A confusion matrix is helpful for retrieving performance measures such as accuracy, recall, precision, and F1-score. Accuracy answers the question: "Out of all the classes, how much did we predict correctly?". Recall measures how much the model was able to predict correctly out of all the positive classes; a high value is desirable. Precision measures how much actual positive predictions were made from all the correct positive class predictions. The F1-measure allows the ability to measure both precision and recall simultaneously. These measures along with the classification report are provided in Table 3 and Table 4 respectively.

Table 3. Model Performance Measures

Measures	Score
Accuracy	90.17%
Recall	96.58%
F-1	93.67%
Precision	90.92%

Table 4. Model Classification Report

	Precision	Recall	F1-Score	Support
0	87%	71%	78%	7341
1	91%	97%	94%	22304
accuracy			90%	29645
macro avg	89%	84%	86%	29645
weighted avg	90%	90%	90%	29645

Another important performance indicator is known as the AUC-ROC curves. ROC is defined as a probability curve that plots the True Positive Rate (TPR) on the x-axis against the False Positive Rate (FPR) on the y-axis. The AUC conveys the model's capability of classifying the distinction between classes. A higher AUC that's close to 1 represents a more effective model in predicting the sentiment between negative and positive.

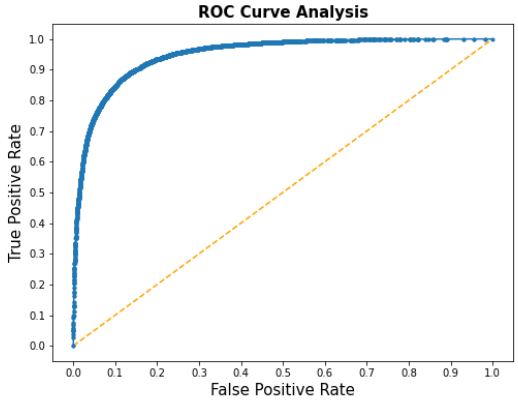

Figure 9. ROC-AUC

8.3 Performance Evaluation

The model produced very good results with the overall accuracy of the model being 90%. The recall score is also very high at 96%, denoting that our model is able correctly identify positive classes. Finally, the F-1 score at 93% conveys that the model is able to correctly predict positive classes very well. Looking at the classification report, we see that there are 22,304 occurrences of the positive class and only 7,341 of the negative class, even despite assigning the neutral star rating reviews to negative sentiment. We expected this imbalance of distribution of classes from exploratory analytics, as most of the reviews for these Las Vegas restaurants were positive. Ultimately, this is a concern as unbalanced support in the data may prove a weakness and could possibly necessitate resampling or further feature extraction.

By acknowledging our AUC score of 0.949 and its closeness to 1, we interpret this as a very good measure of separability, meaning there is a 95% chance the model will be able to distinguish between the two classes.

9. Conclusion

9.1 Findings

Throughout our analysis, we found that our sentiment words were more positive than negative of Yelp users' experiences at the designated restaurants. In relation to the positive sentiments of users, we can display a positive correlation between that of positive reviews with high ratings, and negative reviews with lower ratings. From segmenting the area to Las Vegas and categorizing off of Japanese restaurants, we gain insight on how they operate as a whole. Each of the individual users provide their opinions throughout their reviews; as the positive outweighs the negative, Japanese restaurants are providing great dining services and food to their customers, which increases the positivity disclosed in their review, as well as an input of a higher rating. Although, through this project, Japanese restaurant owners can also view aspects that drive more negative sentiments which they can take initiative on and remedy over time.

9.2 Future Steps

In our study, we have worked to fine tune the hyperparameters to develop an efficient logistic regression classification model that predicts binary sentiment very well. Though our model enabled us to answer our business to an extent, there are many potential areas of opportunities to expand our research and further deep dive into classifying sentiment. From preprocessing to modeling, we propose different methods that can result in many improvements. Firstly, it would be interesting to instead of performing a binary sentiment classification for negative and positive, to treat the neutral category of star rating 3.0 as its own distinct class. With this implemented, this would pivot the model results, identifying new and interesting features associated with customer sentiments that are rather on the fence about their particular experience at an establishment.

Another implementation that past text mining and natural language processing research have found success in is the utilization of optimized techniques for handling text representation. For example, delta TFIDF is employed to efficiently weight word scores before classification. Being easy to compute, implement, and understand, it has proven to be more accurate in Support Vector Machine models for sentiment analysis compared to the use of BOW and TF-IDF (Martineau and Finin, 2009, p. 1) [5]. Support Vector Machines and other algorithms such as Naive Bayes and Random Forest would also be areas of interest where we can explore better fit and functionality for this business problem. In a study conducted on the impact of exploiting both BOW and TF-IDF for using different machine learning classifiers, Support Vector Machines actually outperformed the other classifiers (Pimpalkar and Raj, 2020, p. 49-68) [6]. Ultimately, it would be much more beneficial to conduct ensemble learning, or using multiple machine learning methods to obtain better predictive performance results as opposed to a single model on its own.

Finally, another route of this analysis worth mentioning for future study would be to categorize sentiment analysis based on the topics we found using LSA. It would be insightful to look into segments such as certain foods, customer service, restaurant atmosphere, and social aspects. After doing this, measuring feature importance for binary sentiment classification in relation to each of these topics rather than as a whole would result in more intuitive insights for businesses to understand their customers.

10. References

[1] S. Channapragada and R. Shivaswamy, "Prediction of Rating Based on Review Text of Yelp Reviews", 2015, pp. 1.

[2] D. Bijoyan and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation", *ResearchGate*, June 2018, pp. 1-6.

[3] N. Asghar, "Yelp Dataset Challenge: Review Rating Prediction", 2016, pp. 4.

[4] N.V. Patodkar and I.R. Sheikh, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", *IJARCCE,* 2016, pp. 320-322.

[5] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis", *Proceedings of the International AAAI Conference on Web and Social Media*, March 2009, pp. 1.

[6] A.P. Pimpalkar and R.J. Retna Raj, "Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features", *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 2020, pp. 49-68.

Appendix

Figure 10 - Top 10 Restaurants on Review Count

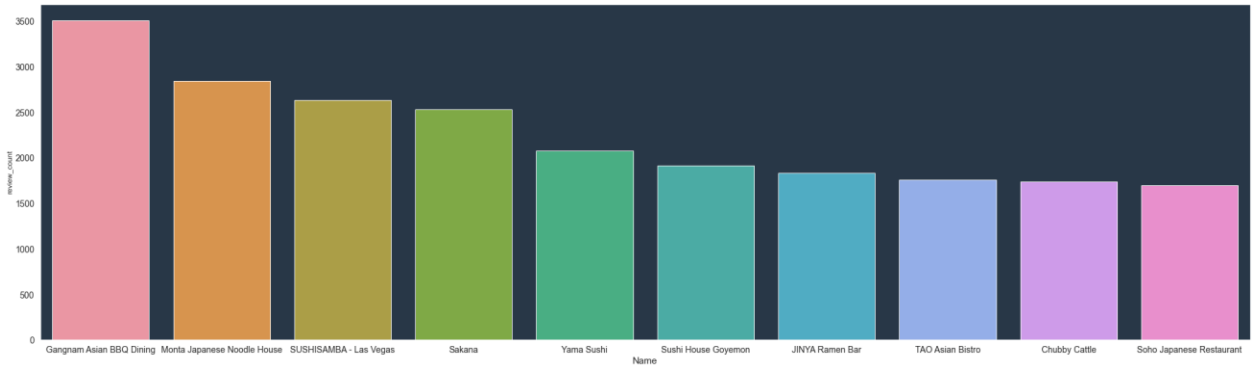

Figure 11 - Top 10 Positive Bigrams

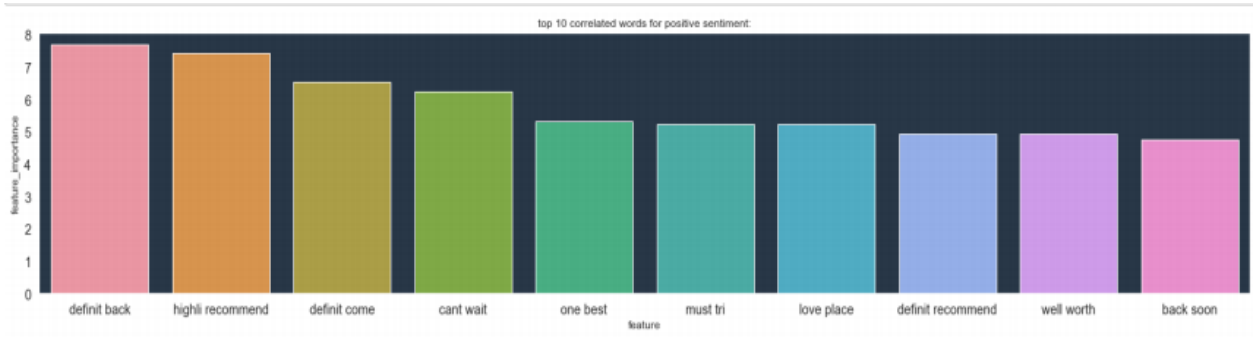

Figure 12 - Top 10 Negative Bigrams