Predicting Restaurant Review Sentiment from Yelp Reviews

Soorya Paturi (sopaturi@gmail.com)

Abstract— **Yelp reviews are a popular reference for gauging a restaurant's ability to provide a positive experience. A Yelp review consists of a text review as well as a star rating that is on a range of 1-5 stars. The shortcomings of a text review is that a review may contain sarcasm, ambiguity, or a play on words that can be misconstrued as positive or negative. A sentiment classifier for reviews will remove the need for sentiment interpretation and replace the star rating system with the percentage of people that left a positive review. The goal is to predict sentiment (positive or negative) and the likelihood of a positive review from a text review and certain restaurant attributes because this prediction removes the need for a star rating system. Star rating systems offer too many options and their interpretation is more subjective than the binary option of sentiment.**

I. INTRODUCTION

Leaving a review on Yelp is the most common way reviews for businesses are shared with potential visitors to a business. These reviews can greatly influence a potential visitor's decision to visit the restaurant. Sentiment prediction on a text review in Yelp provides clarity for the user on the meaning of the review and adds another metric in addition to the star based system.

Sentiment prediction is not deterministic because a review can take on many forms. However, the objective is to find a relationship between text review to sentiment, so that the likelihood of a review being positive can be predicted.

The client interested in the results of this report is any company that currently accepts text reviews for products or entities. The client should care about sentiment prediction for these reviews because ambiguous reviews can be identified as positive or negative. A viewer of the review is now clear on the intentions of the review when the sentiment is listed alongside the review. With a sentiment classifier, a user of a review site will not have to leave a star rating because a positive or negative sentiment will automatically be generated from the review. A site with reviews can offer a more conversational input style when describing a business that does not always require someone to state a star rating.

The outline of this report and the final report is summarized here with Section II introducing the data set. In Section III, the data wrangling of the data set is summarized by explaining how the data set was obtained, cleaned, and wrangled. In Section IV, exploratory data analysis of the data is summarized by a field by field analysis of restaurant attributes that affect sentiment as well as identification of words that are common in positive and negative reviews. In the final report, Section V will be added, and in it the prediction task is described. In addition, the last section, Section VI, conclusions and next steps are discussed in the final report.

II. DATA SET

The data set was downloaded from, `https://www.kaggle.com/yelp-dataset/yelp-dataset` [1]. The data set was collected by Yelp and contains data concerning business attributes, reviews, users, checkins, and tips left by the user. The data set contains 5.2 million user reviews across 174,000 businesses in 11 metropolitan areas. For this project, only data from the business attributes and user files were used.

The dataset is subsetted to only feature restaurants that are in Las Vegas in order to decrease the computation time in the prediction task. The business and review dataset were merged so that every review for every restaurant in Las Vegas was in one table. Both the review and business table have a column named stars and were renamed so that when the merge occurred, it's clear in the resulting dataset which column refers to an overall business rating or a individual review rating. In order to decrease exploration and prediction time, a representative sample of the population was chosen by selecting 50,000 reviews randomly without replacement from the dataset.

The prediction task is to use restaurant attributes and individual reviews to predict a review sentiment. In Figure 1, is a distribution of the star ratings of reviews for restaurants in Las Vegas. The distribution illustrates a common trend in reviews in that there are more one star, four star, and five star ratings than other ratings because users tend to either give an extreme rating or not give a rating at all. Users tend to want to post if the service was either very good or very poor. In the data wrangling section, the star ratings will be converted into either a positive or negative sentiment.

TABLE I
STAR RATING COUNTS

Star Rating	Review Count
1	6037
2	4401
3	6387
4	12145
5	21030

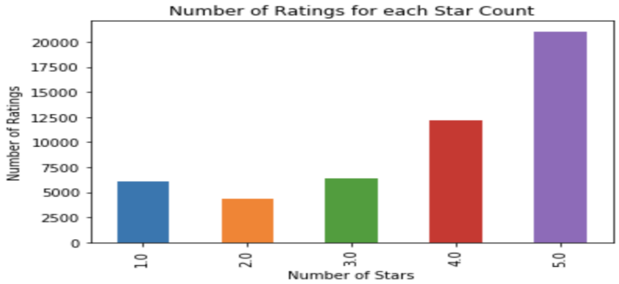

Fig. 1. Bar plot of star ratings for Las Vegas restaurant reviews

A selected number of fields, six to be exact, that may cause confusion, are described below so that interpretation of the data set above is easier. The full description of all fields is shown in the link below.

-ATTRIBUTES: dictionary of business attribute values such as price range and parking conditions

-CATEGORIES: type of business

-IS_OPEN: 0 or 1 if restaurant is closed or not

-REVIEW_COUNT: number of reviews for a business

BUSINESS_STARS-number of stars for the business(1-5 stars) and is average of individual reviews

-REVIEW_STARS: number of stars for an individual review (1-5 stars) text-text review left by user

Description of all the fields in the data set can be found at `https://www.yelp.com/dataset/documentation/main` [2].

III. DATA WRANGLING

Star ratings are converted into sentiment based on how a review's star rating is positioned on the scale of 1-5. If the rating is above three, the review is classified as positive. However if the rating is three or lower, meaning it is not positive, and either negative or neutral, the review is classified as negative. A user should go to a restaurant that is positively rated if they are choosing judiciously and want a good experience. A three star rating is not good enough to be positive because it is neutral. A metric that can be listed on the Yelp website next to each review is the percentage of users that left a positive review, meaning a 4 or 5. This dataset does not explicitly state whether all the reviews for a business are listed so calculating this value for now is not meaningful. Rate of positive reviews for a restaurant is important because a four star rated restaurant could have 100 3 star ratings, not positive reviews, in the average for restaurant rating but does not show that 50% of users found the restaurant not positive.

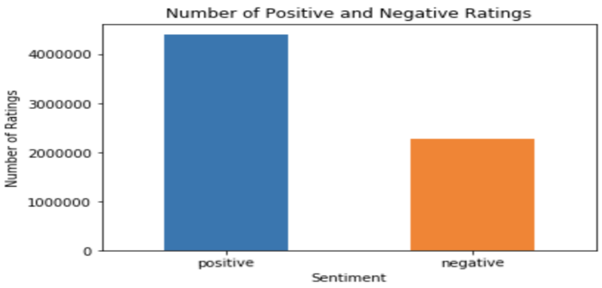

Fig. 2. Bar plot of review sentiment for Las Vegas restaurant reviews

After converting star rating to sentiment, review count by sentiment can be seen in Figure 2. There are approximately twice the number of positive ratings compared to the number of negative ratings. The positive and negative ratings are mapped to numerical format (1 and 0) respectively so that the machine learning models can process the target variable. This dataset is not heavily imbalanced and accuracy as a classification metric in sentiment prediction can be used because the sentiment classes are not for example, 95% positive and 5% negative.

The columns with missing values include ['address', 'attributes', 'categories', 'city', 'hours', 'is_open', 'latitude', 'longitude', 'name', 'postal_code', 'review_count', 'business_stars', 'state'] and rows with missing values in these columns were dropped from the dataset. The following columns, ['address', 'categories', 'city', 'hours', 'state', 'review_id', 'user_id'], are dropped because they do not aid in the prediction or data analysis sections. The categories column is dropped because even though it does give information about the restaurant type, there are too many categories that a business can take on at once and makes it hard to group restaurants into a small number of groupings. The text column with the reviews was already confirmed to have no missing values and doesn't have any empty strings as well. In addition, outliers were checked for in numerical columns and there appears to be no negative values and in the case of ratings no values above five.

A. Feature Engineering

The average star rating by itself does not give a full picture of how good a restaurant by itself. The number of reviews need to be taken into account as well since it is a measure of restaurant popularity. A Bayesian average formula smoothed the restaurant rating and review count by accounting for the prior rating as well as the total number of stars given to a restaurant [3]. The formula for this is better way to rank restaurants because it accounts for both review count and average star ratings and is listed below.

$$BayesianAverage = \frac{C * m + totalstars}{C + numberofreviews}$$

where,

C=prior number of observations
m=prior average of stars

The attributes column contains dictionaries with many attributes for each restaurant and not all of them apply to each restaurant. An interesting attribute that may affect restaurant sentiment is the price range that is given on a scale of 1-4. The price range was extracted from the attribute dictionary of each restaurant and added as a new feature.

IV. Exploratory Data Analysis

The distribution of values in a restaurant's attributes and how it affects restaurant sentiment is explored in this section. Also, the "best" restaurants will be ranked according the the Bayes average feature created in the data wrangling section. In addition, a Word2Vec model will be trained with the text reviews and similar words in the text for positive and negative words will be found.

The Bayes average of restaurant rating is a more representative measure of how good a restaurant is compared to just the number of reviews or star rating. The Bayes average was computed for all of the restaurants and the top five scores were plotted in Figure 3. These scores reflect the restaurants with the best combination of popularity and score.

The top restaurants by the number of reviews left were also found in Figure 4 and when comparing to the Bayes average rating, none of the restaurants in the top five for both rankings systems are the same restaurant. Taking into account both star rating and number of ratings intuitively seems more prudent when deciding on a restaurant to visit.

The distribution of values in a field that may affect review sentiment are now explored.

A. Restaurant Operation Status

Restaurants are either open or closed and are listed as a feature in the business dataset. The graph in Figure 5 shows that open restaurants have a higher rate of positive reviews than closed restaurants. This makes sense because closed restaurants may have closed due to negative reviews. The difference in rate of positive reviews between the

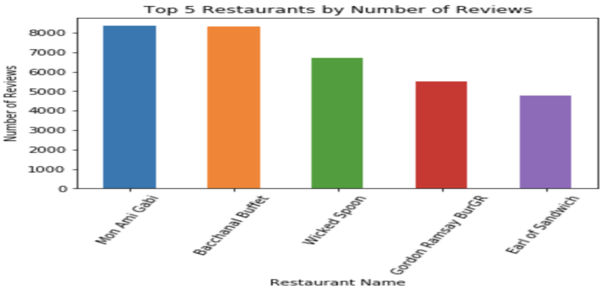

Fig. 3. Bar plot of the top 5 ranked restaurants by review count

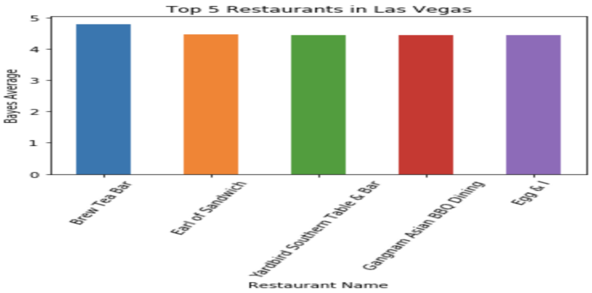

Fig. 4. Bar plot of the top 5 ranked restaurants by Bayes averaged ratings

values of the feature can be used as a predictor for review sentiment in a machine learning model.

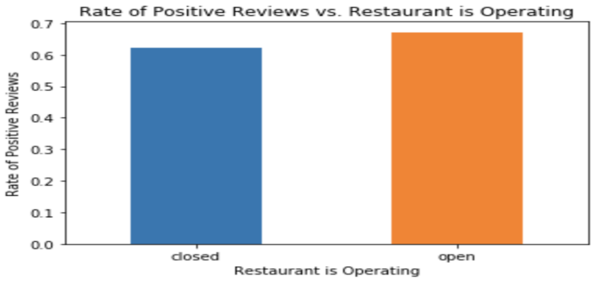

Fig. 5. Bar plot of average review sentiment for restaurants that are open vs closed

B. Restaurant Review Count

Restaurant review count is one measure of a restaurant's popularity and it is important to check if restaurant review count is positively correlated with rate of positive reviews for the restaurant. The graph in Figure 6 shows that the slope of the best fit line between the two variables is very small and is 2e-5. In addition, the Pearson correlation coefficient is small and does shows a very weak positive linear correlation. The positive linear correlation between restaurant review count and rate of positive reviews is low and shows that restaurant review count will be a weak predictor in a machine learning model.

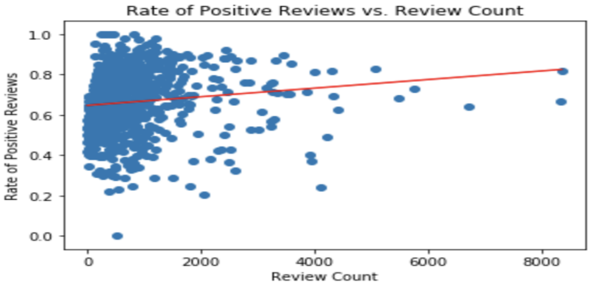

Fig. 6. Scatter plot of average sentiment vs the number of reviews for a restaurant

C. Price Range

Restaurants have a price range from a scale of 1-4 and the scale is listed as a feature in the business dataset. The graph in Figure 7 shows that restaurants with a price rating of four have a higher rate of positive reviews than price ranges between (1-3). This makes sense because more expensive restaurants tend to give better food and service. The difference in rate of positive reviews between the values of price range can be used as a predictor for review sentiment in a machine learning model.

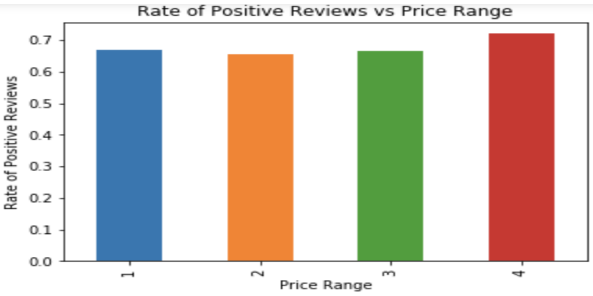

Fig. 7. Bar plot of review sentiment for restaurant price ranges

D. Review Impressions

Restaurants reviews have a review reaction of the possibilities that the review is useful, cool, or funny and the number of these reactions per review is listed as features in the review dataset. The graph in Figure 8 shows that there are more positive reviews with useful, cool, and funny ratings. Also, most reviews have a small amount of reactions, less than 20, and the distribution of reaction count for positive vs negative reviews is different. The difference in the distribution of cool, funny, and useful reviews for positive and negative reviews can be used as predictors for review sentiment in a machine learning model.

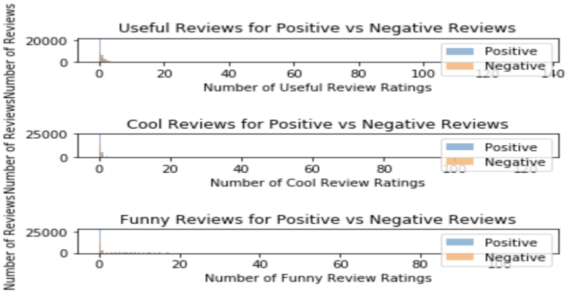

Fig. 8. Histograms of positive and negative reviews that had review reactions such as useful, cool, or funny

E. Geographic Representations

In Figure 9, there is a map of Las Vegas restaurants and their locations according to their latitude and longitude. There tends to be more reviewed restaurants in the center of the city as opposed to the edges. Figure 10 depicts a heat map of average restaurant sentiment across zip codes in the dataset. It is interesting to note that as the distance from the center of Las Vegas increases, restaurant sentiment seems to be going down. There are 63 zip codes in the dataset and it would not make sense to hot one encode all of these zip codes for prediction as too many new features would be created but it is still interesting to see that zip code affects restaurant sentiment.

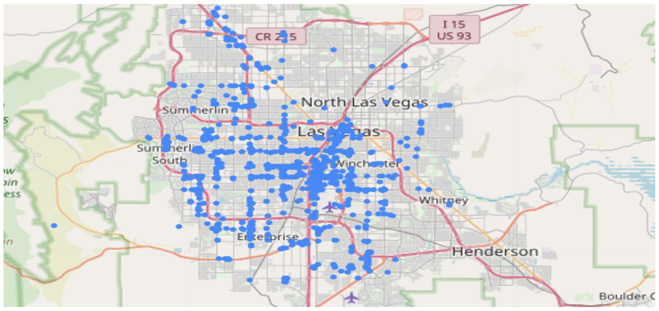

Fig. 9. Map of restaurants in the data set given their location coordinates

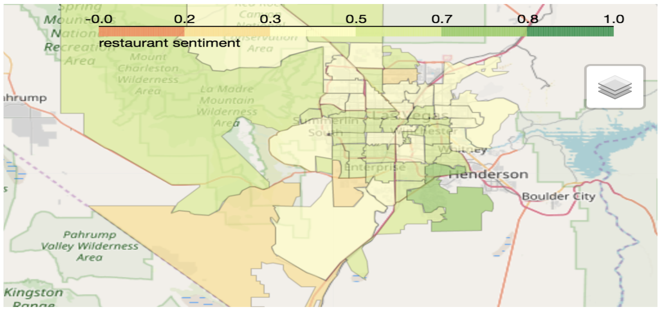

Fig. 10. Heat map of average restaurant sentiment by zip code

F. TextBlob

TextBlob is a NLP package that allows for sentiment analysis and was used to find the most common words in positive and negative reviews [4]. Before TextBlob could process the reviews, all of the reviews had non letter and space characters removed and the resulting strings were lower cased. Afterwards, all stopwords, common words in the english language, were removed to avoid giving significance to common words. TextBlob then returned sentiment polarity for each review based on the sentiment of the words inside of it. According to TextBlob, there were 43,879 reviews that were positive out of 50,000 reviews. A wordcloud was then constructed to show the most common words in positive and negative reviews.

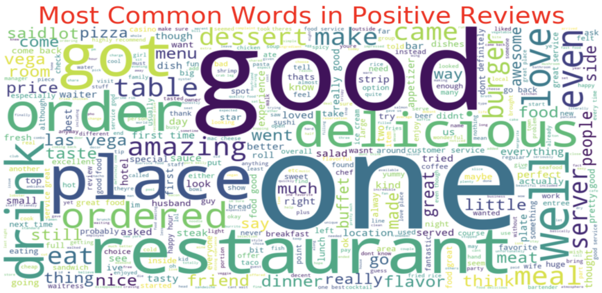

Fig. 11. WordCloud of common words in positive reviews

G. Word2Vec

Word2Vec is a neural network implementation that learns distributed representations for words and will eventually fed into a machine learning model to predict review sentiment [5]. A vector of a specified size, in this case 300, was used to represent each word in the model's vocabulary fed into by the sentences of the reviews. A Word2Vec model was run on the reviews after they were processed by first removing non letters. Stopwords were kept because these words give structure and context to a Word2Vec model that uses external text to create a vector for a word.

Parameters for Word2Vec model:

Downsampling of Frequent Words-Frequent words should not take on more importance just because they occur often because they are most likely stop words. Google documentation recommends values between 0.0001 and 0.001, the value chosen was 0.001.

Word vector dimension-The dimesnion of the word vector was chosen to be 300, the more features used results in longer run times but not always better models.

Minimum word count-Any word that does not occur at least this many times across all documents is ignored. The value 35 was chosen because each restaurant appears on average 20 times, and a value larger than that is used to limit the size of the vocabulary to meaningful words.

Window size-The number of words for context that the training algorithm should take into account, ten was chosen for this parameter.

The resulting WordVec model produced a matrix where each row is a word that appears in the review and the columns are the vector representation for that word. After the model was run, based on its vector representations of the words in the sentences of the reviews, the most similar words to great (positive) and awful (negative) were found. Word2Vec appeared to be working and trained on a large enough training set based on the synonyms produced.

V. Prediction Task

The dataset for this section was obtained from the exploratory data analysis section. The goal of the prediction task is to predict review sentiment, positive or negative, from just the text of the review. In the previous section, feature engineering was done to create the Bayes average rating feature and the price range feature. The Bayes average rating feature is a rating that takes into account the number of stars and the number of reviews for a restaurant. However, these restaurant attributes and others will not be used in classification and the features chosen will be described in the next section, Feature Selection. The prediction will determine the probability of the review sentiment being positive as well as whether the prediction for the review is positive or negative.

More wrangling steps need to be done to the data such as conversion of the text reviews into vectors that represent each word in the review as well as justification for dropping restaurants attributes from the dataset. Word2Vec will be used to create vector representations for each of the classification methods. In the deep neural network, the word2vec vector was used as an embedding for all of the words in each review while in the ensemble methods, the word vectors for each review were averaged by the number of words in a review. The methods used in this report include a Random Forest model, Gradient Boosted Classifier, Voting Classifier, and a Deep Neural Network.

A. Classification Metrics

The following metrics were used to compare the performance of the baseline classifier and other classifiers against one another.

Classification Accuracy: Fraction of reviews correctly classified

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

where

tp is the number of true positives,
tn is the number of true negatives,
fp is the number of false positives,
and fn is the number of false negatives.

False Positive Rate: Fraction of negative reviews labeled as positive.

$$FPR = \frac{fp}{fp + tn}$$

False Negative Rate: Fraction of positive reviews classified as negative.

$$FNR = \frac{fn}{fn + tp}$$

Balanced Error Rate: Average of the false negative rate and false positive rate.

$$BER = \frac{FPR + FNR}{2}$$

F1 Score: Another measure of accuracy that uses precision and recall.

- Precision is the ratio of true positives to reviews that were predicted to be positive.

- Recall is the ratio of true positives to reviews that were actually positive.

$$F1\,Score = \frac{2 * precision * recall}{precision + recall}$$

AUC: Area under the curve of the ROC-curve (True positive Rate vs. False Positive Rate). AUC represents the probability of a randomly drawn positive review having a higher probability of being positive compared to a randomly drawn negative review.

B. Feature Selection

Not all of the relevant features as well as the engineered features were used as inputs to the classifiers for several reasons. For one, the goal of this project is to predict sentiment from just the text review so that sentiment prediction can be used in any review website, not just Yelp. If features that describe restaurant attributes are used in this prediction, the models will not necessarily perform as well in another domain with different business attributes. Restaurant attributes other than reviews can change on a daily basis such as review count or number of people that rate the review as cool, funny, or useful. It is not practical to keep updating the dataframe of reviews and then reclassifying each review along with restaurant attributes as the restaurant attributes keep changing. The following business attributes were dropped: is_open, review_count, business_stars, cool, funny, useful, bayes_avg, and price_range.

A Word2Vec model was used on the text reviews to convert them into vectors. A vector with a dimension size of 300 was used to represent each word in the Yelp reviews. According to the shape of the Word2Vec model, not including stop words, 5715 words in the reviews had a corresponding vector representation. Every word in a review that had a word in the Word2Vec model was converted into a vector and then used as weights in an embedding layer in the deep neural network classifier. In other other models, the information contained in the sequential order of the words was not accounted because the average of the Wor2Vec vectors for each review was used.

A deep neural network with the same architecture used on the text review was used with multiple features including restaurant attributes. The metrics for this model such as accuracy, F1-score, and AUC were slightly lower when restaurant attributes were discarded. If restaurant attributes want to be added to the model in the future, input and outputs need to be created for categorical and numerical variables. Embedding would be used for categorical features and a single dense layer would be used for numerical variables. Aftwerwards, the overall architecture for all of the variables can be added similar to the one used later on just text reviews.

C. Models

A benchmark classifier is used to compare all the other classification methods against. ZeroR (Baseline): The ZeroR classifier used as a baseline simply classifies all reviews as the majority class which in this case is positive.

A.) Deep Neural Network: A deep neural network has multiple layers and finds the correct mathematical operations, whether they be linear or nonlinear to transform the input into the true prediction while minimizing error in the prediciton. A sequential model is advantageous for text reviews because it takes into account the order of words and positioning of the words when predicting sentiment. The word vectors in a review do not need to be averaged or concatenated in order to feed into a neural network because a list of words can be interpreted. Neural networks have many parameters that can be adjusted in order to improve prediction metrics such as the number of layers, number of outputs per layer, types of activation functions, and the number of epochs to run the model for.

B.) Random Forest: The random forest classifier is an ensemble method that uses many decision trees that have had bootstrap aggregating or bagging done on the samples. Bagging means that samples are chosen at random with replacement to reduce variance in any single tree. Random forests in addition take a subset of the predictors when constructing a tree so that each tree is more different from one another. The combination of the trees made from bagged samples and subsetted predictors is the random forest result. Compared to a decision tree, a random forest improves accuracy by determining class from the majority of votes for each sample in the many trees.

C.) Gradient Boosting Classifier: A gradient boosted classifier involves combining a large number of trees. Each subsequent tree that is constructed is done so based on the least pure splits created by the previous tree. The trees are added back to the previous tree so that the tree can learn slowly and improve classification accuracy. An advantage to gradient boosted trees is that compared to a decision tree that may overfit to one set of the data because gradient boosted trees learn by focusing on the worst split in a tree and improve over time. The disadvantages of gradient boosted trees is that they may overfit compared to random forests.

D.) Voting Classifier: A voting classifier involves combining multiple classifiers, in this case a random forest and gradient boosting classifier. A voting classifier using soft voting averages the probabilities for each sample being positive and assigns the prediction as positive if the average probability across all of the estimators is above 50% in the case of a binary class problem. An advantange of a voting classifier is that an estimator can be used to mask the errors in prediction of another estimator if the estimators are using different methods of prediction.

VI. Evaluation and Results

The original dataset of 50,000 Yelp restaurant reviews was split into a training and test set with 80

Before any of the machine learning models were used on the training data, stratified sampling was used so that ratio of restaurant sentiment was maintained in the train and test sets. The total number of samples in the training set was 40,000 and the number of sampeles in the minority class, a negative review, was the minority class rate multiplied with the size of the training set. Cross-validation was used with five folds so that the ML models would fit the training data in such a way thath the fit was representative of the whole dataset.

Two methods of converting the text review data to a numerical fields were used in the before ML methods were applied because these methods require that all fields are in numerical format. In the neural network and the ensemble methods, reviews were preprocessed by removing all non-letters, lowercasing the words, splitting the words in the review, and then removing common stop words in the english language. Then the word2vec representation was found for each word and concatenated into one array for each array. The vector for each review was padded to the max length of a reviw with empty word arrays so that the size of each of review was uniform. From there, in the tree-based ensemble methods, averaged the word vectors for each review while all of the word vectors for each review were fed into the neural network.

The average value of the word vectors were used in the ensemble methods because these methods cannot process the sequential flow of words and learn a specific order as well as a neural network. It was hypothesized that using the average vector would perform worse than the whole review in the neural network. The word vectors could have been padded in the ensemble methods as was done in the neural network but the difference in performance between average word vectors was explored instead.

The deep neural network was initialized with a sequential model. The first layer was an embedding layer than took in review vectors that were padded to the max review length and outputted vectors of size 300 for each word in the review because 300 was the size of the word2vec representation. The next layer was a contained a long short term memory (LSTM) network that could look backwards for up to 32 words in a bidirectional manner. The next layer was a globalmaxpooling layer that finds the maximum value of the word vector. Then there was a dense layer that outputted data 20 nodes with a relu activation function. Finally, five percent of the inputs were dropped before converging onto one node as the output. Only three epochs were used because the validation accuracy decreased from the second to third epoch.

The random forest ensemble method used 250 estimators and the function to measure quality of a split was gini impurity. The square root of the number of predictors were fitted to each tree. The minimum sample leaf was not tuned in order to decrease computation time.

The gradient boosted classifier also used 250 estimators. The minimum sample leaf as well as the learning rate were not used in order to decrease computation time.

Feature importance for this problem is not relevant as there is only one feature and the most that could be done is to find the most common words in positive and negative reviews. Once the words were encoded with their Word2Vec representation, the conversion back to English words and then finding the most common values would be unnecessary. A sentiment polarity was found in the exploratory data analysis section and the most common were visualized.

TABLE II

BENCHMARK CLASSIFICATION METRICS

Model	Accuracy	FPR	FNR	BER	F1-Score	AUC
ZeroR Baseline	0.6335	1.0	0.0	0.5	0.7977	0.5

TABLE III

MACHINE LEARNING CLASSIFICATION METRICS

Model	Accuracy	FPR	FNR	BER	F1-Score	AUC
Random Forest	0.8354	0.3733	0.0588	0.2160	0.8836	0.9118
Gradient Boosted Tree	0.8602	0.2547	0.0815	0.1681	0.8971	0.9261
Voting Classifier	0.8563	0.2871	0.0710	0.1790	0.8956	0.9236
Deep Neural Network	0.8774	0.1978	0.0836	0.1407	0.9078	0.9409

The best model for the prediction task was the deep neural network because it had the highest ROC-AUC of 94.09%. In addition this model had the best accuracy at 87.74%. The meaning of the ROC-AUC can be summarize with this statement: given a randomly selected positive review, it will have a higher probability for a positive review than a randomly selected negative review 94.09% of the time. In this prediction task, a false positive is worse to a viewer of a review because then that means they went to a predicted positive review but it ended up being a bad experience. The model with lowest false positive rate was the deep neural network. It is interesting to see that the performance of the averaged word vector in the ensemble methods was too much lower when compared to the neural network. The ROC-AUC for the gradient boosted model was less than two percent lower than the neural network.

VII. CONCLUSIONS AND FUTURE WORK

The best performing machine learning classifier was the neural network because it had the highest ROC-AUC of 0.9409. The neural network had the best performance metrics compared to other models for all of the metrics except the false negative rate. A false negative is not harmful as a false positive because it means a person saw a negative review, tried the restaurant, but found that their review aligned with a positive review.

The benefits of implementing sentiment prediction adjacent to a review is that the prediction will remove the need for sentiment interpretation that can be difficult in the case of ambiguous sentiments. Text reviews contain shortcomings in that it can be difficult to ascertain sentiment when the review contains sarcasm, ambiguity, or a play on words. The mean star rating for a restaurant can be replaced by the percentage of users that rated the restaurant as positive. The percentage of users that rated the restaurant positive includes information of the percentage 4-5 star ratings. An average star rating for all reviews for a restaurant does not include include information about the variance in the star ratings of the reviews for the restaurant.

In the future, in order to improve the accuracy for the neural network, more samples need to be used to train the classifier as this neural network only performed slightly better than a random forest model that was using the averaged word vector for each review. In addition, the number of positive reviews can be downsampled so that the review sentiment class ratio is 1. An even amount of positive and negative reviews would decrease the false positive rate because less positive reviews would lead to less false positives. Variations of the number of nodes per dense layer neural network can be tested in the future to see how more nodes will improve performance. Finally, more hyperparameters can be search over the tree based ensemble methods such as the learning rate and the minimum sample leaf but were not tried in order to decrease fitting time.

REFERENCES

[1] https://www.kaggle.com/yelp-dataset/yelp-dataset

[2] https://www.yelp.com/dataset/documentation/main

[3] https://fulmicoton.com/posts/bayesian_rating/

[4] https://textblob.readthedocs.io/en/dev/

[5] https://www.kaggle.com/c/word2vec-nlp-tutorialpart-3-more-fun-with-word-vectors