DEPARTMENT OF COMPUTER SCIENCE
CS 4501: Special Topics in Computer Science - Machine Learning
University of Virginia Spring 2019
Tim Kim (tsk8va), Vish Panagari (vp5zg), Rahul Zalkikar (rz7vv)

A Machine Learning Analysis of Virginia Restaurant Reviews: Future Failure or Potential Phenomenon?

Abstract
This report summarizes the statistical modeling and analysis results associated with the study of the restaurant industry in Virginia and the associated Yelp reviews. The purpose is to establish a framework of the inherent correlations and relationships that exist between reviews, or better referred to as sentiment, and the restaurant industry. The underlying impact this study could have on local businesses of the future is enormous, as current and new players in the State can utilize essential insights of the study to structure their business plans in order to efficiently target the Virginia population. Using SVM and Naive Bayes classification, we are able to effectively predict if a restaurant is going to foreclose in the next two years.

Introduction
It is important that Virginia businesses focus on building the trust of their customers and stakeholders through heightened awareness of the social, environmental, and cultural impacts of their business processes on surrounding communities. Local businesses can better understand how communities perceive them through new, innovative machine learning technologies and sound statistical analyses that deliver data-driven insights regarding sentiments about their business practices.

The problem we are tackling is that of improving consumer-business relationships. With a greater understanding of the opinions and preferences of the local consumers, a business can modify its business plan to target the specific communities/regions that optimize their revenue. For example, if a business was able to see its ten most negative reviews, and vice versa, it could easily identify the biggest weaknesses or strengths in its day-to-day operations.

In addition, this approach to sentiment analysis will not only analyze numerical variables like rating, but also include insight from the review text itself. While there have been many studies investigating the relationship between ratings and restaurant success, many fail to take the actual text in the review into consideration. To be specific, this is a direct application of Natural-Language-Processing based analysis of online consumer reviews of Virginia restaurants. These restaurants represent the identities of Virginia communities, and are important for the wellbeing of the statewide economy, local farmers, food industries, and residents alike. Therefore, through a Support Vector Machine (SVM) algorithm based on NLP sentiment analysis of Yelp restaurant reviews, this study will provide essential insights into consumer opinion on local businesses for the betterment of the state of Virginia.

Method

Our method of analysis will consist of using a standard natural language toolkit dictionary of positive and negative word lexicons (specialized dictionary to the restaurant industry if possible). The process will include the standard cleaning, filtering and stemming of consumer review words prior to any analysis on unstructured data. SVM is conventionally used with sentiment analysis, as text analysis can provide a large number of possible features as used by the machine learning algorithm. We plan to use the area under the curve of the ROC curve, and conventional machine learning metrics such as accuracy, recall, and precision, to gauge the performance of our classification algorithm.

By selecting failed, or bankrupted restaurants and successful restaurants as our two binary classifiers in this project, we will use lagged customer reviews (those one year prior to bankruptcy or closing) in our classification model to add a prediction element to the study, predicting future restaurant failure from sentiment in annual customer reviews of the previous year. After a valid, unbiased model that satisfies relative metric requirements as defined by us is finalized, it can be used to forecast a Virginia restaurant's failure before it happens, based solely on consumer perceptions. This way, businesses can give way to positive societal externalities to improve their customers' perception of their business and avoid shutting down.

Experiments

After initial data cleaning, we were left with the columns business_id, stars, review_count, is_open, categories, and is_restaurant (Figure 1). Out of these, we will treat only stars (rating) and review length as learning factors. Review_count (number of reviews per business) will require additional exploration. Business_id is a foreign key unique to every restaurant and will not be useful in our predictive model. Is_open is the feature we are trying to predict, and is_restaurant will always be true. Restaurant category is not needed for this particular experiment as we are treating the entire sample of restaurants the same.

The distribution of star counts in Figure 2 indicates that most reviews are relatively positive (4 or 5) for both closed (failed) and open (successful) restaurants. Therefore, it seems as though star counts for reviews might not be a useful feature for predicting whether a restaurant will close in the future.

Figure 3, or the distribution of reviews per business indicates that some businesses have a lot more reviews than most others, and skew the average number of reviews per business. This could be good (more data about local businesses in general) or bad (biased data that caters to certain local businesses which have a lot of reviews). We will assume the first, and will not treat this as a feature in our model.

The distribution of open vs closed restaurants (Figure 4) indicates that appropriate class weighting will be needed in SVM, as there are many more open businesses than closed businesses in our dataset.

The charts in Figure 5 provide key insight into using the length of a review as a feature. They show that review counts follow a logarithmic distribution for each star rating, and that most reviews fall between 0-1000 in length for each star rating.

Judging by the boxplots of review length for each star rating (Figure 6), it seems as though the star ratings have approximately the same mean and quartile review lengths associated with them, despite many outliers across the board, and especially amongst 5 star reviews. Therefore, text length might not be a particularly useful feature to consider in our classification model, and we will proceed by using review text as our primary feature. For classification with SVM, we will subset the data into bad reviews (1 star), neutral reviews (3 stars), and good reviews (5 stars). This, in turn, also incorporates the dimension of review sentiment through star-rating.

The first step was to clean the data by removing stop-words, punctuation, and word stems as defined by conventional NLP practices and the standard NLTK package in Python. We used a bag-of-words approach with vectorized unigrams (single words) to train our SVM model.

Results

After vectorizing the reviews, we obtained a sparse matrix with a shape of (5000, 14432) and density of 0.368. We split our word vectors into a training set of 80% of the data and a testing set of 20% of the data. After fitting the SVM algorithm on our training data and grid searching for optimal hyperparameters (C = 10, gamma = 0.001), we obtained impressive benchmark results when predicting restaurant closure from testing data.

Accuracy	0.973
Precision	0.974
Recall	0.998
F1-Score	0.986
AUC-ROC	0.910

The following results are from k-fold cross-validation (5 folds) using the same SVM model:

Scores	[0.953, 0.956, 0.955, 0.954, 0.964]
Average Cross-Validation Score	0.956
AUC-ROC	0.906

For comparison with our SVM model, we decided to train classic Naive Bayes variants used in text classification (Multinomial, Complement, Bernoulli) and obtained similar results. However, our SVM model outperformed these variants in accuracy, recall, and f1-score.

	Multinomial	Complement	Bernoulli
Accuracy	0.964	0.964	0.934
Precision	0.991	0.989	0.963
Recall	0.973	0.975	0.969
F1-Score	0.982	0.982	0.966

Code Available Here: https://github.com/zalkikar/SVM-SentimentAnalysis

Conclusions

Our optimized SVM model performs remarkably well, with accuracy, precision, and recall metrics above 0.97 and an ROC of 0.91. Our hypothesis that sentiment and text features in restaurant reviews can forecast future restaurant closure was validated by the impressive performance of our predictive model during k-fold cross-validation. The resulting implication for local restaurant owners in Virginia is that they can apply our model on online reviews of their restaurant to gain insight into consumer perception. By running our model on historical reviews in real time, owners can record the probability of restaurant closure over time. This allows them to react preemptively to an increasing probability of closure and take necessary steps to improve consumer experience. Another conclusion that can be made from the results of the model are that online reviews matter. The performance of the model definitely demonstrates that the contents of an online review influence the restaurant's future, and shows that it is important to provide detailed and honest reviews to restaurants to provide feedback.

Future recommendations include adding relevant data to the training set, historical backtesting to validate the presence of an unbiased model, and introducing a lag time component where the model is trained with reviews from two years prior to when restaurants close or remain open. Although the bag-of-words approach with SVM or Naive Bayes is conventional for these types of problems, it is possible that other algorithms (custom variants of Naive Bayes, Neural Nets) perform well too. In addition, it might prove useful to train a model using bigrams (pairs of adjacent words) and trigrams (triplets of adjacent words) instead of unigrams (single words) to incorporate a different linguistic structure. An additional improvement could be to find a way to analyze the "strength" of a review. Due to the binary classification of "positive" or "negative" that our model employs, the model cannot distinguish between a positive review with a more neutral tone and a positive review that is gushing with support and praise.

Appendix

Figure 1: *Data head sample*

	business_id	stars	review_count	is_open	categories	is_restaurant
4	PfOCPjBrlQAnz__NXj9h_w	3.5	116	1	American (New);Nightlife;Bars;Sandwiches;Ameri...	True
5	o9eMRCWt5PkpLDE0gOPtcQ	4.0	5	1	Italian;Restaurants	True
10	XOSRcvtaKc_Q5H1SAzN20A	4.5	3	0	Breakfast & Brunch;Gluten-Free;Coffee & Tea;Fo...	True
14	fNMVV_ZX7CJSDWQGdOM8Nw	3.5	7	1	Restaurants;American (Traditional)	True
15	l09JfMeQ6ynYs5MCJtrcmQ	3.0	12	0	Italian;French;Restaurants	True

Figure 2: *Distribution of review star rating*

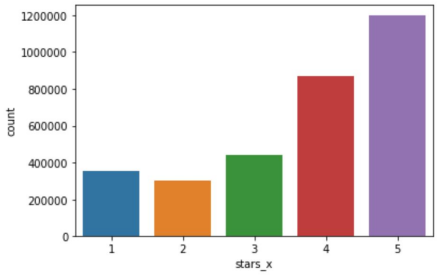


Figure 3: *Distribution of Reviews per Business*

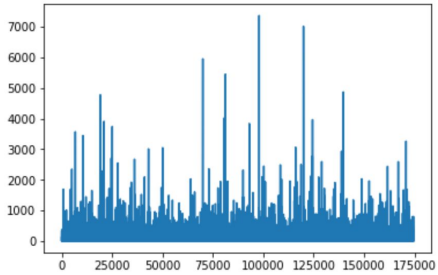


Figure 4: *Distribution of Open (1) vs Closed (0) Restaurants*

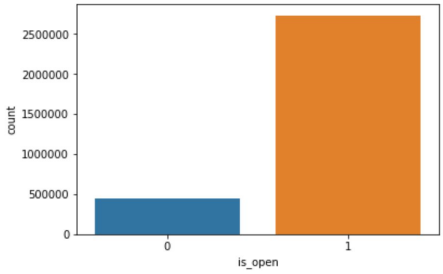


Figure 5: *Distributions between stars and the length of the review*

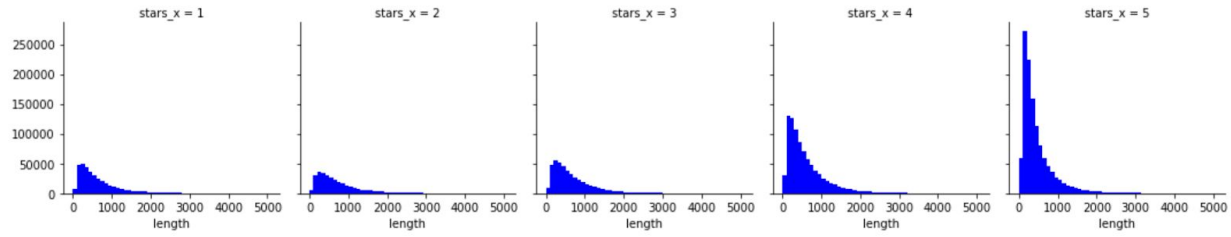
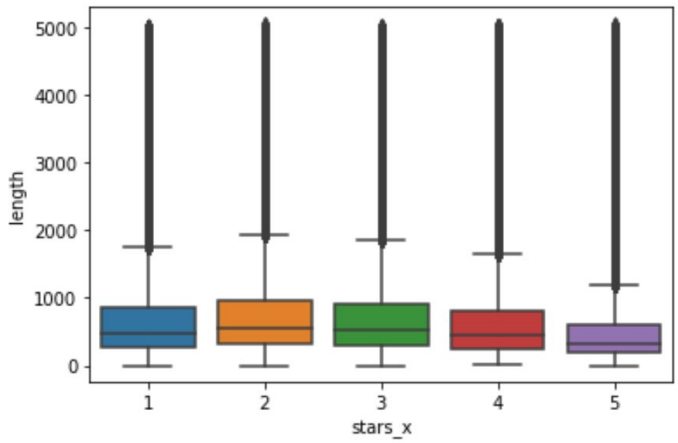


Figure 6: *Boxplot of review length for each star rating*



References

https://github.com/ageron/handson-ml
StackOverflow
Professor Rich Nguyen

Team Contribution

Code: Rahul Zalkikar
Video: Tim Kim, Vish Panagari
Checkpoint: Vish Panagari, Rahul Zalkikar, Tim Kim
Final Paper: Vish Panagari, Rahul Zalkikar, Tim Kim