Predicting International Restaurant Success with Yelp

Angela Kong[1], Vivian Nguyen[2], and Catherina Xu[3]

***Abstract*— In this project, we aim to identify the key features people in different countries look for in their dining experience. Using the Yelp Dataset [1], we performed feature selection to identify the business attributes that correspond to high star ratings for each country. Then, we classified the data using models such as Naive Bayes, support vector machines (SVM), decision trees, logistic regression, and Gaussian Discriminant Analysis (GDA) to evaluate the strength of the feature sets we selected. We used univariate feature selection with a chi-square scoring function to choose the most important features. GDA was the best-performing model, with test accuracies of around 60% for binary classification. Lastly, we used natural language processing methods to identify the most informative features from restaurant review texts. Using these features, an accuracy of around 69% was achieved for review classification.**

I. INTRODUCTION

Yelp is one of the largest and most popular platforms for crowd-sourced reviews about local businesses, with over 145 million monthly unique visitors and 102 million reviews to date [2]. A restaurant's Yelp page has become its first impression to customers, and strongly influences an individual's dining decisions. Consequently, success on the platform, in the form of positive reviews and high star ratings, is coveted by businesses worldwide.

The ability to identify business features that are most indicative of success on Yelp can help restaurants devise sensible strategies to improve their own ratings. Specifically, we will explore how features gain and lose importance as we vary geographical location by country. For example, are Americans more inclined toward a late-night snack than their German counterparts? Do Canadians value a take-out option more than those who live in the United Kingdom? Through our work, we aim to bring to attention the various, and sometimes non-obvious, cultural nuances that impact the dining experience. In the future, these methods can also be utilized to determine how varying other attributes (restaurant size, price range, etc instead of location) affect the critical features chosen and overall classification performance.

II. RELATED WORK

As a modern, information-rich dataset, the Yelp Dataset has been a valuable resource for predicting a restaurant's star ratings and success. For example, Gingerich and Bochkov have previously conducted similarity analysis on restaurants based off of text analysis and word vectors [3]. In another seminal paper, Yun, Wu, and Wang have explored using Part-of-Speech (POS) analysis to predict a business' rating based off of user-generated text alone. While they claim that this representation cancels out subjectivity [4], users' reviews tend to be reactive instead of constructive. Therefore, while the accuracy of these sentiment based predictions are high, they may not necessarily provide restaurant owners with specific improvements to increase their chances of success. In our project, we focus on not only modeling a restaurant's success through textual analysis of user reviews, but analyzing which features are better predictors of success among restaurants in different countries to provide data-driven predictions of international trends.

III. DATA

We obtained our dataset from the Yelp Dataset Challenge webpage, which contains a total of 77,000 businesses, 2.2 million reviews by 552,000 users, and 566,000 business attributes. Because our project focuses only on predicting restaurant success, we filtered out all non-restaurant businesses. This left us with 25,071 restaurants from four different countries: the United States, the United Kingdom, Canada, and Germany. The datasets for the businesses and reviews are stored in two separate JSON files, with one object per line.

A. Format of Business Data

```
{
  'type': 'business',
          ...
  'name': (business name),
  'neighborhoods': [(hood names)],
  'city': (city),
  'state': (state),
  'latitude': latitude,
  'longitude': longitude,
  'stars': (star rating, rounded to half-stars),
  'categories': [(localized category names)]
  'open': True / False (corresponds to closed, not
  business hours),
  'hours': {
          ...
  },
  'attributes': {
          ...
}
```

B. Format of Reviews

```
{
  'type': 'review',
```

[1]Angela Kong, Computer Science, Stanford University akong2@stanford.edu
[2]Vivian Nguyen, Computer Science, Stanford University vnguyen2@stanford.edu
[3]Catherina Xu, Computer Science, Stanford University yuex@stanford.edu

...
'stars': (star rating, rounded to half-stars),
'text': (review text),
'date': (date, formatted like '2012-03-14'),
'votes': {(vote type): (count)},
}

IV. PREPROCESSING & FEATURE SELECTION

We ran a Python script to convert the JSON data into a CSV file. To prune the Yelp dataset, which contains 98 feature categories, we performed the following procedure: first, we removed features that are not relevant to predicting restaurant success, such as `business_id` and `business_name`, and features that pertain only to non-restaurant businesses, such as type of hair specialization. Additionally, we removed all features for which approximately fewer than 30% of the restaurants did not have a value. For example, while the majority of restaurants had information about restaurant ambience, very few had information about whether or not it accommodates halal dietary restrictions. We narrowed down the feature categories to 28, some of which include: attire, good for kids, noise level, outdoor seating, price range, review count, has take-out, and has Wi-Fi.

A. Preprocessing

We converted all categorical features, such as restaurant attire, to numerical values. For instance, we represented casual, dressy, and formal as 0, 1, and 2, respectively. For feature values that are true and false, we converted the boolean values into their respective integer values. In addition to the aforementioned 30% threshold we placed on feature selection, we considered several different approaches, including case deletion, regression imputation, and mean imputation, to account for missing data [5]. We decided against case deletion because it discards any incomplete data example, which significantly reduces the size of our dataset. We also decided against regression imputation since it requires the implementation of a model to predict missing values and assumes a correlation among features, which we cannot assume in the first step of our analysis.

Therefore, we chose mean imputation, a widely accepted method in the statistical community, to fill in our missing data values [5]. For each missing feature value, we averaged the existing values for that feature and replaced the missing value with the average. We use a variety of machine learning models, including both generative and deterministic; therefore, to ensure consistency, we conducted mean imputation for all of our classifiers. In addition, since we chose only features for which at least 70% of the restaurants had values, we avoided averaging over a small number of values. Instead, we averaged over the majority of the values for each feature and filled in missing blanks with the overall trend for that feature.

B. Feature Selection

We used univariate feature selection to identify which twenty features were most important in predicting success, for all the countries combined and for each country separately. We considered two different scoring functions that return univariate p-values in order to select the most important features: chi-square and ANOVA. Because both our features and classes (star-ratings) are represented as discrete values, we decided to use a chi-square test. ANOVA tests are mainly used when the feature variables are discrete and the classification variable is continuous.

V. MODELS

We considered two different modes of restaurant classification based on star ratings: binary and multiclass. In the binary case, restaurants with a star rating below 4.0 are classified as 0, and restaurants with a star rating of 4.0 and above are classified as 1. The machine learning models we used to train and predict the data are Naive Bayes, logistic regression, support vector machine (SVM), decision tree, random forest, and Gaussian discriminant analysis (GDA). In the multi-class case, restaurants are classified from 0 to 5 based on the integer value their star rating (rounded). The models used are multinomial logistic regression, decision tree, and random forest.

For every prediction model, we trained and tested on the data from all countries combined, and then on each countrys data separately. In total, there are 25,071 restaurants, of which we used 14,000 to train on and 11,070 to test on. We only considered six cities in the U.S. for which Yelp had the most data, instead of all of the cities. For each country, we trained on approximately 70% of the restaurants and tested on the remaining 30%: 8000 training and 3775 testing for U.S., 850 training and 364 testing for the U.K., 150 training and 65 testing for Canada, and 320 training and 134 testing for Germany. Furthermore, we conducted multiple iterations of training and testing for each model with randomized testing and training data each time. Therefore, our results are averaged using Monte Carlo cross-validation[6].

A. Naive Bayes

We implemented the Multinomial Naive Bayes classifier with a Laplace smoothing value of α = 1.0. Using all of the countries data, the test accuracy was 0.55174. The results for each separate country are included in the results section of our paper. The test accuracy ranges from 0.47253 for the U.K. to 0.58462 for Canada. The equations used are included below:

$$\phi_{k|y=1} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} n_i + |V|}$$

$$\phi_{k|y=0} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} n_i + |V|}$$

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

B. Logistic Regression

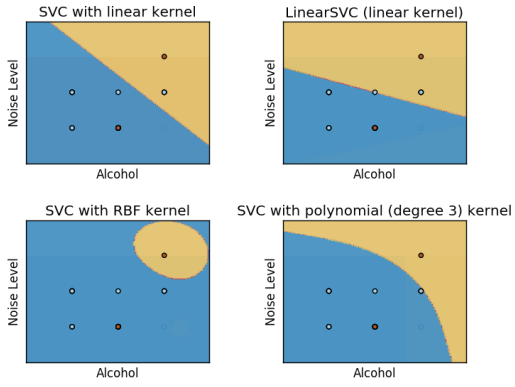
For logistic regression, our test accuracies varied depending on the strength of regularization. We implemented logistic regression using scikit, where the inverse of regularization strength is represented by parameter C. Increasing or decreasing C by factors of 10 does not yield significant differences in test accuracy, but we noticed that a higher C-value (less regularization), results in higher test accuracy. Conversely, in the multi-class model, a higher C typically results in lower test accuracy. Overall, our multi-class model performed worse, as shown in the table below:

C	Binary Test Accuracy	Multi-Class Test Accuracy
0.01	0.53984	0.52213
0.1	0.54851	0.51491
1.0	0.55032	0.51454
10.0	0.55059	0.51454
100.0	0.55068	0.51463

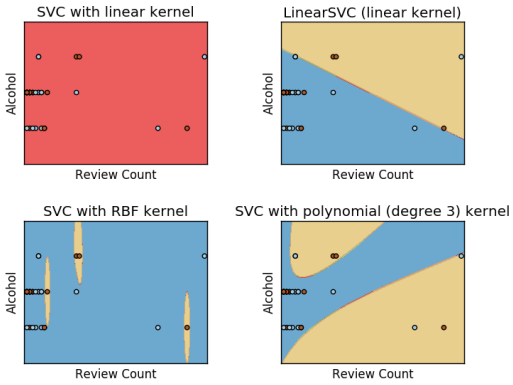
C. Support Vector Machine

To compare the results of different linear SVM classifiers graphically, we display the following 2D projection of the Yelp data set. Since we can only graph two features of the data set, we select two features that are among the most informative for a restaurant's success across all four countries: noise level and alcohol.

Since noise level and alcohol are both discretized variables, with alcohol having three possible values (none, beer and wine, and full bar) and noise level being quiet, average, loud, or very loud, there are only 12 possible combinations for a training example to possess, as seen below:



To consider how our SVM model would perform on variables that are not categorical, we chose to consider the review count feature, which has a mean value of 31 and a standard deviation of 100.



We see that both linear models have linear decision boundaries, while the nonlinear kernel models have irregular decision boundaries that are representative of the corresponding kernels and parameter values. However, there are slight differences in the decision boundaries produced by SVC with linear kernel and LinearSVC, which may be due to the fact that SVC minimizes the regular hinge loss, while LinearSVC minimizes the squared hinge loss.

SVM Classifiers	Accuracy
SVC	0.507
Linear SVC	0.515
RBF Kernel	0.529
Polynomial (degree 3) Kernel	0.512

We see that the RBF kernel produces the best results. Then, we experimented with tweaking the C and gamma parameter values to see if we can achieve even better accuracy.

C	gamma	Accuracy
0.01	0.1	0.50409
0.1	0.1	0.50231
1.0	1.0	0.50428
1.0	100.0	0.51026
100.0	0.1	0.50205
100.0	1.0	0.51852

We conclude that the best SVM classifier uses a rbf kernel with parameters as gamma = 100.0 and C = 1.0.

D. Decision Tree and Random Forest

We implemented decision trees for both binary and multi-class classification. The decision tree is more complex and is a generally more accurate model than decision stumps, which are the simplest version of a decision tree [7]. Benefits of decision trees include logarithmic growth in the cost of predicting data as the number of data points increase, and statistical validation of its reliability. However, decision trees are very susceptible to minor changes in the data and tend to overfit [8]. Therefore, we also implemented a random forest that averages ten trees in order to better model and predict the data. For binary classification and for all of the countries combined, the decision tree test accuracy is 0.55140, and the

random forest test accuracy is 0.55547. For the multi-class classification for all the countries combined, the decision tree test accuracy is 0.39855, and the random forest test accuracy is 0.44697.

E. Gaussian Discriminant Analysis

For binary classification, we implemented GDA using scikit. The test accuracy for the full dataset with all of the countries is 0.55257. The test accuracies for each of the countries is enumerated in the results section. The test accuracies range from 0.55497 for the U.S. to 0.6 for Canada. Thus, GDA is generally the best-performing model (explained in more detail in the sections below). The formulas used for this generative model are the following:

$$\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)}=1\} \\
\mu_0 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}} \\
\mu_1 &= \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
\end{aligned}$$

VI. ANALYSIS OF TEXT-BASED REVIEWS

Apart from our core analysis of business features, we also investigated the top review text features that were indicators of restaurant success. The purpose of this section is to identify important features that were not part of Yelps official set of business features. For example, slow is not in Yelps feature set, but may be mentioned many times in negative reviews - we can then infer that customers place high value on speed of service.

To preprocess the review data, we gathered all of the reviews available for each of the four countries. A review was labeled positive if the reviewer gave the restaurant 4 or more stars, negative otherwise. The data was tokenized using the the Natural Language Toolkits (NLTK) word tokenizer, which split off white-space and punctuation other than periods. We chose not to convert all letters to lower-case because upper-case often indicates strong emotion.

Naive Bayes is a generative supervised learning model that is commonly used for text processing and sentiment analysis tasks, and assumes independence between pairs of features. We used the NLTK Naive Bayes classifier to train and test our data, and to identify the most informative features. This classifier uses the same model as the multinomial Naive Bayes classification discussed above.

We performed 10-fold cross validation to acquire the test accuracies below:

Country	# Train	# Test	Accuracy
Canada	7355	819	0.68009768
UK	40505	4502	0.70146602
US	182833	20316	0.67203190
Germany	250	29	0.4827586

This approach seemed to work best with medium to large datasets. If there is too little data, the classification was made based on 'random' words that happened to appear more in one type of review. Thus, we decided to drop Germany from our analysis of informative features because its accuracy was worse than random guessing ($\tilde{5}0\%$).

The informativeness of a word is:

$$\frac{\max_{labels}\ P(word|label)}{\min_{labels}\ P(word|label)}$$

The top 20 most informative features of US, UK, and Canadian reviews are outlined in the table below.

Canada: Words	Canada: Ratios	UK: Words	UK: Ratios	US: Words	US: Ratios (all neg : pos)
worst	24.2 (neg) : 1	downhill	24.0 (neg) : 1	Terrible	44.8 :1
confused	16.4 (neg) : 1	microwaved	24.0 (neg) : 1	Horrible	41.7 : 1
horrible	12.3 (neg) : 1	Highly	21.3 (pos) : 1	downhill	38.1 : 1
minute	12.0 (neg) : 1	tasteless	20.5 (neg) : 1	TERRIBLE	33.7 : 1
disgusting	11.4 (neg) : 1	McDonald	19.7 (neg) : 1	rotten	33.5 : 1
stale	11.1 (neg) : 1	A-OK	18.8 (neg) : 1	Worst	32.7 : 1
terrible	10.5 (neg) : 1	luke	18.2 (neg) : 1	Awful	28.7 : 1
undercooked	10.2 (neg) : 1	Wo	17.9 (neg) : 1	pathetic	28.5 : 1
rude	9.6 (neg) : 1	delightful	17.9 (pos) : 1	Gross	28.1 : 1
45	9.3 (neg) : 1	Soba	16.7 (neg) : 1	RUDE	26.3 : 1
Itamae	9.3 (neg) : 1	Gie	16.7 (neg) : 1	WORST	25.9 : 1
health	9.3 (neg) : 1	redeeming	16.7 (neg) : 1	Wo	24.8 : 1
eg	9.3 (neg) : 1	Ong	16.7 (neg) : 1	filthy	24.2 : 1
delicious	9.3 (pos) : 1	unpleasant	16.2 (neg) : 1	acknowledgement	24.2 : 1
wonderful	9.1 (pos) : 1	lukewarm	15.9 (neg) : 1	MINUTES	24.2 : 1
Love	8.9 (pos) : 1	bouncer	15.3 (neg) : 1	HORRIBLE	23.5 : 1
East	8.4 (neg) : 1	dreadful	15.3 (neg) : 1	unacceptable	23.2 : 1
boxes	8.4 (neg) : 1	inedible	15.3 (neg) : 1	acknowledge	22.1 : 1
proceeded	8.4 (neg) : 1	unhappy	15.3 (neg) : 1	indedble	21.6 : 1
Which	8.4 (neg) : 1	avoid	15.3 (neg) : 1	unapologetic	21.1 : 1

Sentiment-heavy words such as 'disgusting', 'unhappy,' 'worst,' dominated all three lists. UK customers did not like food that tasted like it was microwaved or lukewarm, though reviewers often added a 'redeeming' quality to a negative review, softening the overall harshness of the review:

> "Their coffee is kind of **redeeming**, if not overpriced…"
>
> "On a **redeeming** note..."
>
> "The single **redeeming** feature…"

The strongest words for US customers were overwhelmingly more negative, with use of capitalization to emphasize intense negative emotion. US customers seemed to crave personal attention and focused strongly on a servers attitude:

> "I don't think she even **acknowledged** my presence…"
>
> "No one **acknowledged** that I walked in..."
>
> "Still no **acknowledgement** from our waiter or anyone else…"
>
> "He was **unapologetic** and...patronizing…."

We also tested three other models: a unigram model with stemming, stop-word removal, and a bigram model. However, stemming lowered the 10-fold classification accuracy by around 2-3% for each country. Removing stop-words (MySQL stop-word list) caused the accuracy to remain about the same for each country - for example, Canada's accuracy was at 0.67888 instead of 0.68009. This is likely because the stop word probabilities for each category are very similar and did not have high impact on the class decision. Bigrams also had similar accuracies, though the

majority of the most informative bigrams were extensions of the words on our unigram list, and less indicative of specific business features - some examples include ('tasteless,' 'and'), ('horrible,' 'service').

VII. DISCUSSION AND RESULTS

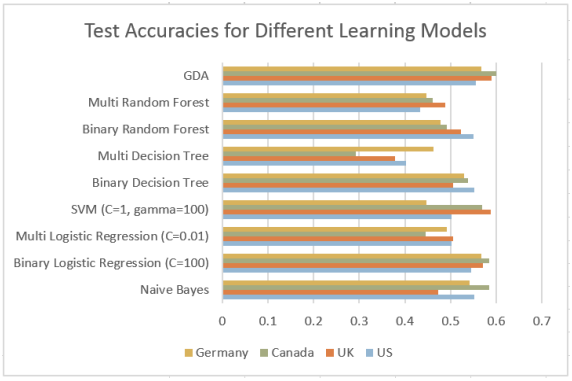
After running univariate feature selection with the chi-square scoring function, we selected the 20 most important features for the US, UK, Canada, and Germany. Ultimately, our goal was to identify features, if any, that are considered globally important indicators to a restaurants success, as well as possible features specific to a particular countrys idiosyncratic cultural values.

We found that there are 6 features that are highly weighted in all four countries: availability of street parking, ability to make reservations, review count, casual ambience, noise level, and attire. Other features corresponding to high star rating include: outdoor seating, classy ambience, touristy ambience, waiter service, hipster ambience, garage parking, trendy ambience, Wi-Fi, intimate ambience, good for kids, good for groups, allows smoking, and has TV.

Next, we examined features that are indicative of success for restaurants situated in a specific country. We found that the divey ambience was solely important for restaurants in the United States, thereby demonstrating an American appreciation for lower-end dining establishments. In addition, in the North America region, customer satisfaction is positively influenced by the existence of parking lots and parking valet services. We speculated that parking is more important in the U.S. and Canada due to a higher percentage of drivers, whereas in Europe, public transportation is more popular. In Europe, restaurant success is also positively correlated with availability of alcohol. This is perhaps due to the lower drinking age in Europe than in the US. Furthermore, by analyzing the review text, we were able to infer (from the most informative features) the tendency of Americans to be more vocal, negative, and service-oriented in their reviews.

Using the 20 most important features, we applied various models to the data. The resulting test accuracies are summarized in the table below. For most countries, the GDA model outperforms the others, and the multi-class decision tree performs the worst. This is because GDA assumes p(x—y) is distributed according to a multivariate normal distribution. If this assumption is correct, GDA is asymptotically efficient, which means that with large training sets, we dont expect many models to be strictly better than GDA [9]. In general, the binary models do better than the multi-class models, as it is more difficult to accurately classify with multiple possible outcomes.

		Country				
		US	UK	Canada	Germany	All
Model	Naive Bayes	0.55232	0.47253	0.58462	0.54179	0.55174
	Binary Logistic Regression (C=100)	0.54517	0.57143	0.58462	0.56716	0.55068
	Multi Logistic Regression (C=0.01)	0.50252	0.50549	0.44616	0.49254	0.52213
	SVM (C=1, gamma=100)	0.50225	0.58791	0.56923	0.44776	0.51852
	Binary Decision Tree	0.55258	0.50549	0.53846	0.52985	0.55140
	Multi Decision Tree	0.40212	0.37912	0.29231	0.46269	0.39855
	Binary Random Forest	0.55020	0.52198	0.49231	0.47761	0.55547
	Multi Random Forest	0.43364	0.48901	0.46154	0.44776	0.44697
	GDA	0.55497	0.59066	0.60000	0.56716	0.55257



VIII. FUTURE WORK

In the future, we plan to implement multiple imputation. In general, this method better identifies data variability than single imputation, but due to the time constraint, we reserve these models, such as mixture of Gaussians, for the future. Another way to make up for the missing values is to gather more data, perhaps from a future Yelp Dataset Challenge.

Furthermore, we would like to improve our textual analysis accuracy through leveraging human-annotated multilingual sentiment datasets online and exploring language-independent textual analysis, especially for datasets where the primary language is not English.

IX. REFERENCES

[1] "Yelp Dataset Challenge." Yelp. N.p., n.d. Web. 01 June 2016.¡https://www.yelp.com/dataset_challenge¿.

[2] "An Introduction to Yelp Metrics as of March 31, 2016." Yelp. N.p., n.d. Web. 26 May 2016. ¡http://www.yelp.com/factsheet¿.

[3] Gingerich, Travis, and Yevhen Bochkov. "Predicting Business Ratings on Yelp." *Stanford University*. 2015.

[4] Xu, Yun, Xinhui Wu, and Qinxia Wang. Sentiment Analysis of Yelps Ratings Based on Text Reviews." *Stanford University*. 2015.

[5] "Missing Problems in Machine Learning." Marlin, M. Benjamin. *Graduate Department of Computer Science University of Toronto*. 2008.

[6] Xu, Qing-Song and Yi-Zeng Liang. "Monte Carlo Cross Validation." Chemometrics and Intelligent Laboratory. 6 July 2000.

[7] Burges, Chris. "From Stumps to Trees to Forests." Cortana Intelligence and ML Blog Team, 10 Sept. 2014. Web. 06 June 2016.

[8] "1.10. Decision Trees." 1.10. Decision Trees Scikit-learn 0.17.1 Documentation. N.p., n.d. Web. 06 June 2016.

[9] Ng, Andrew. "CS 229: Generative Learning Algorithms." *Stanford University*.