The Global Foodie: Predicting International Restaurant Success with Yelp

Angela Kong, Vivian Nguyen, Catherina Xu
Stanford University, CS 229

Abstract

Bon appétit! In this project, we aim to identify the key features people in different countries look for in their dining experience. 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? These findings will help businesses make informed decisions when expanding globally. To achieve this goal, we tackled the problem in three phases:

Obtain and **pre-process** restaurant data from around the world → For each country, perform **feature selection** to pick the attributes that correspond to high star ratings → Run various **classification** and **regression** models to evaluate the strength of selected feature sets

Dataset and Preprocessing

We selected **Yelp**, one of the largest and most popular platforms for crowd-sourced reviews about restaurants, to be the primary source of our data. The main dataset comes from the Yelp Dataset Challenge, and contains a total of 77K businesses, 2.2 million reviews by 552K users, and 566K business attributes. Restaurants come from one of four countries: the United Kingdom, Germany, Canada, and the United States.

We ran a Python script to convert the raw JSON data into CSV files. We converted categorical feature values, such as casual, dressy, or formal for restaurant attire, into numerical values, such as 0, 1, and 2, respectively. For feature values that are true and false, we converted the Boolean into an integer so that 1 represents true and 0 represents false. To account for missing data attributes, we used a method similar to sci-kit's **Imputer** - we replaced empty feature values with the average of existing values in that feature column.

Feature Selection

Step 1: Out of the 98 total business attributes per restaurant, we hand-selected 28 of the ones that we believed would have the greatest impact on business rating. Some of the features we selected include **attire, ambience, delivery, noise level, parking, outdoor seating, price range, review count, has take-out, and Wi-Fi.**

Step 2: Next, we conducted univariate feature selection on each country's restaurant data. We used **Chi-squared** and **Anova F-value** as the scoring functions for selecting the top 20 most important features from our previously hand-selected 28.

Bonus Step: We conducted **Binary Naïve Bayes** (unigram, tokenized, stemmed) using NLTK's built-in classifier to determine the word features in restaurant reviews that were most indicative of restaurant success.

Classification Methods

We considered two different methods of restaurant classification based on their star ratings: binomial and multinomial. In the binomial 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, Decision Tree, Random Forest, and Gaussian Discriminant Analysis.** In the multinomial case, restaurants are classified from 0 to 5 based on the nearest integer value to their star rating. The machine learning models used are **Multinomial Logistic Regression, Decision Tree, and Random Forest.**

Model	United States	United Kingdom	Canada	Germany
Naive Bayes	0.75373	0.58955	0.57813	0.64179
Binomial Logistic Regression	0.70149	0.56717	0.57813	0.56717
SVM	0.69403	0.49254	0.57813	0.44776
Binomial Random Forest	0.67164	0.47015	0.53125	0.49254
Binomial GDA	0.67164	0.50000	0.60938	0.54478
Binomial Decision Tree	0.58209	0.50746	0.56250	0.48508
Multi Logistic Regression	0.50252	0.50549	0.44616	0.49254
Multi Random Forest	0.43444	0.44231	0.36923	0.41791
Multi Decision Tree	0.40344	0.38187	0.23077	0.39552

Figure 1: Test accuracies for various classifications performed on 20 features

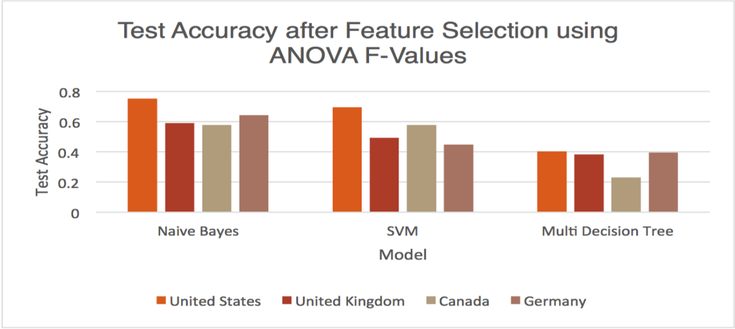

Figure 2: Best, mediocre, worst models respectively for all countries

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

Figure 3: Word features most indicative of restaurant review sentiment

Discussion

We found that in both the Chi-squared and Anova test results, there are 6 features that are highly weighted in all four countries: **parking street, takes reservations, review count, casual ambience, noise levels, and attire.** Other features that were generally correlated (which we quantified as appearing on three of the four countries' feature selected lists) include: **hipster ambience, garage parking lot, Wi-Fi, intimate ambience, good for kids, good for groups, allows smoking, and has TV.**

Next, we examined features, if any, which correlated to success for restaurants situated in a *specific* country. We found that **the divey ambience** was a feature that was solely important for restaurants in the United States, whereas a **touristy ambience** was highly valued only in the United Kingdom. In addition, in the North America region, customer satisfaction correlated with the existence of a **parking lot and parking valet services.**

After running a variety of binomial and multinomial classification models after selecting 20 features, we discovered that **Anova F-value is a more consistent scoring function.** The results of modeling using features selected by Anova F-value show that **Naive Bayes** is generically the best model for restaurants in all four countries. The **Multinomial Decision Tree** is the worst. This suggests that features have low dependency upon each other. Also, Naive Bayes could also be a robust model in this case because there are fewer data samples for most countries. Finally, it appears that test accuracy is generally higher for binomial models than for multinomial ones, which makes sense since multi-class classification usually yields lower accuracy.

From the most significant word features retrieved from sentiment analysis, we note that Americans reviewers tend to **be more negative** in their reviews, with rampant capitalization and emphasis on lack of personal attention (acknowledge, acknowledgement).

Future Work and Acknowledgements

One of the neat aspects that our project focuses on is identifying any differences in the success of a restaurant across multiple countries. One of the challenges that we faced in doing so was performing sentiment analysis on German restaurant reviews, since the majority of the reviews were written in German. While this proved difficult to process, we were still able to train our models on the features given in the dataset; in the future, however, we would like to improve our accuracy through leveraging **human-annotated multilingual sentiment datasets** online and explore **language-independent sentiment analysis.**

In addition, one of the challenges we faced earlier on was preprocessing the Yelp dataset, which was highly unbalanced and contained many unpopulated cells. Although there were a lot of possible features to consider, most of the restaurants did not contain data for these features. As a result, we decide to impute empty values in a column with the mean value - however, **more data should be collected** to improve and confirm our results.