Predicting Ratings Based on Yelp Tips

by Angelique De Castro, Andrew Du, Aieswaryasayee Manicka

June 2, 2015

Abstract

Yelp tips are a highly valuable feature because they inform new consumers how to enhance their experience with the business. They could provide insider information about a business or restaurant's services and deals. For example, tips can provide information on wait times on specific days, difficulty of finding parking, secret menu items, and more.

In this report, we use the Yelp Challenge dataset to model the relationship between tip text data and a business' rating, and apply to model to predict a business' *overall* rating. Tips could reflect on the a business' fortes and deficiencies in a different manner than reviews. By nature, tips are more concise. Tips have a more specific purpose than reviews, so the conclusions drawn from tips would vary between reviews. Additionally, tips manifest summaries of a user's review thereby reducing the amount of data used to predict a business' rating.

Our goal is to discover what kinds of tip features correlate most with certain overall business ratings. Ultimately, we explore how a business can leverage data from its Yelp user tips section to enhance how it can better serve its customers. In addition, businesses may be better able to interpret the underlying meaning in their tips with insight from the trends of phrases and language we observe.

1 Introduction

The Yelp dataset is divided into five parts: reviews, users, businesses, tips, and check-in. For our predictive task, we decided to use 'tips' and 'businesses' because tips are short impactful statements offering users rich insight about the business or restaurant.

Dataset Characteristics

In the dataset, there are 495,107 tips total. There are 42,111 out of 61,184 businesses that contain at least one tip. Of those businesses that do contain tips, there are 11.76 tips on average for each business from April 2009 - January 2015.

Each tip consists of the fields:

```
'type': 'tip',
'text': (tip text),
'business_id': (encrypted business id),
'user_id': (encrypted user id),
'date': (formatted date: '2012-03-14'),
'likes': (count)
```

And each business consists of the fields:

```
business_id, name, neighborhoods, full_address,
city, state, latitude, longitude, stars (Star
ratings out of 5, rounded to half-stars),
review_count, categories, hours, open, and
attributes.
```

Because we are dependent on using business IDs for the predictor, and not all businesses have tips, we decided to omit any business that do not have any tips in our dataset.

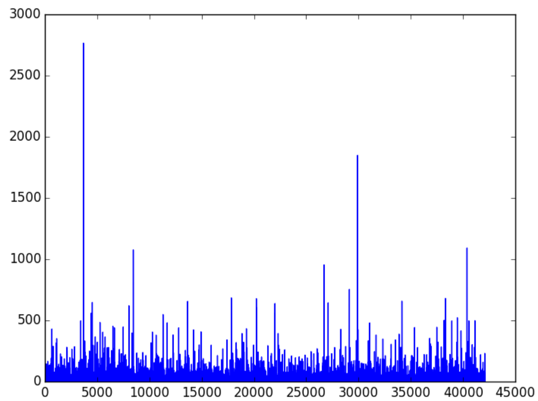

Fig.1: X-axis: Businesses, Y-Axis: Number of Tips

The average rating of all businesses is 3.67 which implies that there are more positively rated businesses. This most likely means that most tips will indeed be something positive about the business or restaurant. To get a better understanding of this, we visualized the most common words (excluding stop words) for all tip text.

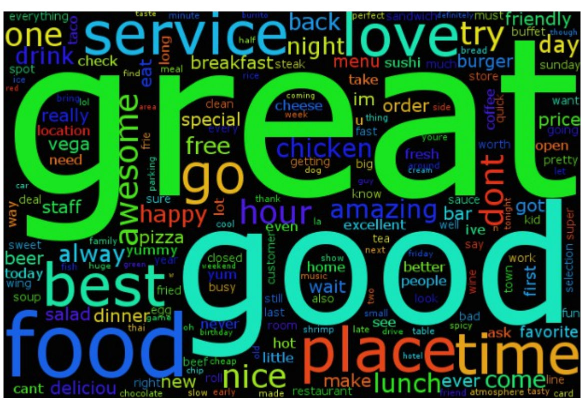

Fig2: Exploratory Analysis of Subset of Yelp Dataset Data

# of tips: 495107
# of businesses: 61184
 # business there are tips for: 42111
average number of tips for restaurants (only including restaurants with tips): 11.757189333
time period covered by data: '2009-04-15' -> '2015-01-22'

```
Interesting findings:
Min tip count: 1
Max tip count: 2765
```

2 Predictive Task

The goal of this task is to predict business' overall rating based on the tips users have left on its Yelp page. We selected tips as our main data set because they are more concise than reviews, but could still reflect strongly on the business in a unique manner. As a result, we wanted to explore how potentially meaningful and impactful tips could be in predicting overall business ratings.

The relevant features that we considered are the most popular words in tips based on our training set of tip data. We started by training our model based on 75% of the data and testing it on 25% of the remaining data, and we assessed the validity of our model's predictions against the test set using Mean Squared Error.

Baseline

In addition a relevant baseline for comparing our predictor is simply averaging the ratings of all businesses. The baseline computed average was ~3.66 stars and when checked on the test set had an mse of 0.41.

Approach 1: Linear Regression - Bag of Words

We are using a bag-of-words approach in which we gather the top 1000 common words in tips and fit each feature in the vector to a rating using linear regression. To avoid getting meaningless features, we discarded punctuations, and stop words.

Model: rating = θ_0 + ? θ_i × [top i popular word]

Approach 2: Linear Regression using Unigrams and Bigrams from All Tips for a Business

In our second attempt, we decided to take a slightly more sophisticated approach by accounting for more of the syntax and pairs of words that may be used together often in tips. We iterated through the tips to capture the top 1000 most popularly used nltk library's unigram and bigram functions.

We decided to allow stop words to be included because there could be significant bigrams where a stop word changes the context or meaning. With the top 1000 popular phrases, we created 1001 index feature vectors for each business. The feature vector includes the number of occurrences of each popular word in the tips for that business, followed by an offset term. We then performed linear regression on the training data to observe how each of these popular phrases affects the business' overall rating.

The result was as expected except for a few outliers which as presented in our in our visualizations. For example, “las vegas” is one of the phrases which has a relatively strong positive correlation with a business' rating. This may be skewed by the sample of data we were given. An outlier which surprised us on the other spectrum, was the phrase “free wifi”, which was surprisingly in the top negatively correlated phrases with a business' overall rating. This could be a result of reviewers' disappointment in there being *no* free wifi.

Fig. 2 and Fig. 3 demonstrate which phrases are most positively or most negatively correlated with the business' overall rating.

Model	MSE
Baseline	0.4125759581590231
Approach 1	0.3825190003757996
Approach 1 (Rounded to nearest half int)	0.4226673776226601
Approach 2	0.67327680133994816

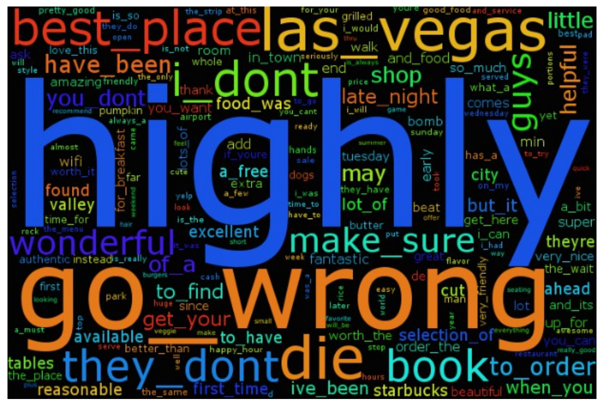

Fig. 3: Most positive impactful unigrams and bigrams

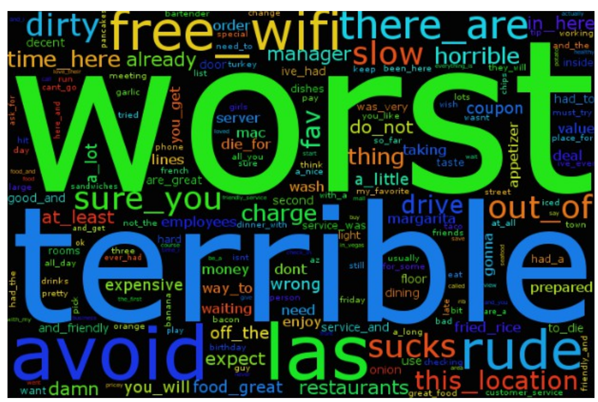

Fig. 4: Most negative impactful unigrams and bigrams

3 Literature

We explored literature on various ways to train predictors for restaurant ratings. Jeff Han, Justin Kuang, Derek Lim investigated how to predict user ratings based solely on business and user characteristics [1]. For each user investigated, they found a group of N similar users who visited the same restaurants and used their ratings to predict what the user in questions will rate those restaurants. They also ran linear regression on reviews using different combinations and permutations of features. They found that user average stars, business average stars, and business review count had the lowest average MSE and highest impact on prediction. Using cosine similarity for grouping N=all similar users resulted in their lowest MSEs regardless of a weighted or unweighted prediction.

In contrast, Mingming Fan and Maryam Khademi predicted business reviews based only on user review text [2]. Their best results came from a linear regression model that used the top frequent words from raw data compared to top frequent words after Part-of-Speech. They concluded that there is a linear correlation between raw user review text and the business rating.

Jason Jong implemented a naive bayes classifier on reviews and then built a bag-of-words model for regression [3]. Though this used reviews instead of tips, we incorporated a similar model to correlate popular words in tips to the rating of businesses.

Unfortunately, our implementation inspired by a naive bayes counting algorithm fared worse than the baseline. This is possibly because most Yelp tips are generally positive. This indicates that tips act as suggestions and are meant to help readers. Another possibility is the model may have actually been good, but the predictor wasn't implemented properly: the threshold could be too low/high, or we shouldn't have counted words in the feature vector. If we were to go forward with improving our model, this is how we could enhance it.

4 Conclusion

In summary, we chose the Yelp dataset and tackled the predictive task of applying text mining concepts from class towards analyzing Yelp tips to predict overall business star ratings. We attempted incorporating similar sentiment analysis models on review text that we read from other research papers and incorporated them into tip text; but some results didn't fair as well as others did. For example, classifying tips as positive or negative and then regressing on the most popular words in each set actually turned out to be worse than the baseline (which simply predicted the global average for each business). We then took a look at using the most popular words in tips overall and then built a naive regressor which fared better than the baseline.

With one of our approaches we observed an MSE as low as ~0.383 which suggests that tip sentiment could be a strong predictor of Yelp business overall ratings, but more sophisticated features and more tailored predictions to reduce error.

1 http://cs229.stanford.edu/proj2014/Jeff%20Han,%20Justin%20Kuang,%20Derek%20Lim,%20Predicting%20Yelp%20Ratings%20From%20Business%20and%20User%20Characteristics.pdf

2 http://arxiv.org/pdf/1401.0864.pdf

3 http://cs229.stanford.edu/proj2011/Jong-%20PredictingRatingwithSentimentAnalysis.pdf