Analysis of service quality in restaurants based on Yelp reviews

Manuel Walser

22nd November 2015

1 Abstract

This report analyzes the service quality of restaurants based on Yelp reviews and compares the differences between different regions and users groups. It is my result for the final Capstone Stone project of the Coursera Data Science specialization track. It uses the public data set of the sixth round of Yelp Dataset Challenge 2015. All English sentences about the service and staff have been rated with a sentiment dictionary regarding there positive or negative polarity. Since no residence was given for the users, a simple predictor has been constructed to predict the residence based on review counts and doctor visits. Finally, the obtained service ratings have been aggregated and compared by regions and customer groups. No significant differences in service quality could be found between the cities and customer groups. However the results show a strong correlation between the overall rating of a restaurant and the calculated service quality.

2 Introduction

The main goal of this analysis is to extract information about the service quality from Yelp reviews. This requires the application of text mining and sentiment analysis methods to obtain a quantitative rating for the service quality of each business. Afterwards the ratings can be aggregated and compared between different cities and visitor groups. The first goal is to investigate if there are any significant differences between the service quality in different cities and customer groups. One would assume that customers are very sensitive to the service quality of a business. This hypotheses should also be verified with the obtained data.

In general, the calculated service quality could be interesting for all Yelp users who would like to have a simple rating of the service quality of each business. Before reading through all reviews, this would give a quick summary to the user about what he can expect. The results are also interesting for owners of a business who would like to understand and maybe improve their service quality.

3 Methods and data

3.1 Data source

The data set is publicly available and can be downloaded from the homepage of the 6th round of the Yelp Dataset Challenge. Otherwise no other additional data has been used in the analysis. The ZIP file includes five JONS files containing data about businesses, reviews, users, tips and checkins. The total size of the unpacked files is about 2 GB. The structure of the original JSON files is documented in detail on the Yelp Dataset Challenge homepage. Beside normalizing columns with multiple values and converting datatypes, no special cleanup or filtering steps have been performed on the raw data.

The results shown in this report only use data from the business, user and review files. The other two files, as well as as the information about the social network, is not used. All calculation have been done with version 3.2.2 of R Revolution Enterprise. This distribution uses multiple cores for some mathematical operations. To improve the performance of some other functions, the `parallel` package has been used occasionally. However the performance of the scripts is still painful and many steps take hours to evaluate.

A quick analysis of the given data set shows that 63% of the reviews are written for businesses in the category Restaurants. The other two categories with many reviews are Nightlife (13%) and Food (12%). All other categories have 6% or less reviews. Since restaurants are the most interesting category to analyse service quality to me, the results in this report are restricted to restaurants only. This data subset contains about 990'000 reviews of 20'000 restaurants from 220'000 user in ten major cities.

3.2 Language detection

At the beginning all reviews have been included in the analysis. However it turned out that the results of the region Karlsruhe were significantly distorted by this approach. Therefore all non-English reviews have been excluded from the analyses. It would require additional effort to incorporate a German sentiment dictionary or any additional languages. Over 99% of the reviews are written in English. Only about 1% of the reviews are written in German or French. Further there are a few hundred reviews written in other languages like Chinese, Spanish or Dutch. Ignoring non-English comments is only a problem for Karlsruhe, because there most reviews are written in German. This is the reason why Karlsruhe does not appear in the city ranking shown later.

To identify the language of each review the `textcat` package has been used. It compares the n-gram profile of a text with an included profile db. This method works well for longer reviews, however shorter reviews are often misclassified. For example the statement 'terrible service' is very hard to classify correctly. After restricting the languages to the most common three languages the misclassification rate decreased significantly.

3.3 Sentiment analysis and service quality rating

First, all reviews have been split into sentences. For the splitting the sentence splitter from the `openNLP` package has been used. Next all sentences with the words 'service' and 'staff' have been filtered in order to separate them from other topics such as food, location, etc. However some longer sentences may still contain statements about other topics. Finally all sentences has been rated with the sentiment dictionary from Hu and Liu into a positive or negative polarity. To calculate the polarity the `qdap` package has been used. In principal the algorithm counts the number of positive and negative words and sums them up. Additional factors are included for words that amplify or negate the meaning. Each word could also be weighted with a different factor. However the included dictionary has only the weights ± 1. The final result of the sentiment analysis are 510'000 service quality ratings for 20'000 restaurants. A few hundred ratings have been check manually to check if the algorithms works as expected.

The obtained quality ratings have been aggregated by different groups, e.g. by business, by user or by region. For each group the mean polarity, the standard deviation and the standardized mean polarity has been calculated. The standardized mean polarity is given by the ratio of the mean polarity divided by the standard deviation of each group. For businesses with a large number of reviews this seems to be a more significant indicator. However for restaurants with a small number of reviews ($n \lesssim 20$) the average polarity is a more stable indicator for the polarity. Therefore this report only shows the mean polarity.

3.4 Grouping of cities into region

The data set contains businesses in 378 cities. The cities have been cluster with a k-means algorithm into ten regions around the eight major cities in the USA (Montreal, Phoenix, Charlotte, Las Vegas, Pittsburgh, Madison, Waterloo, Urbana), Edinburgh in Great Britain and Karlsruhe in Germany.

3.5 Residence predictor

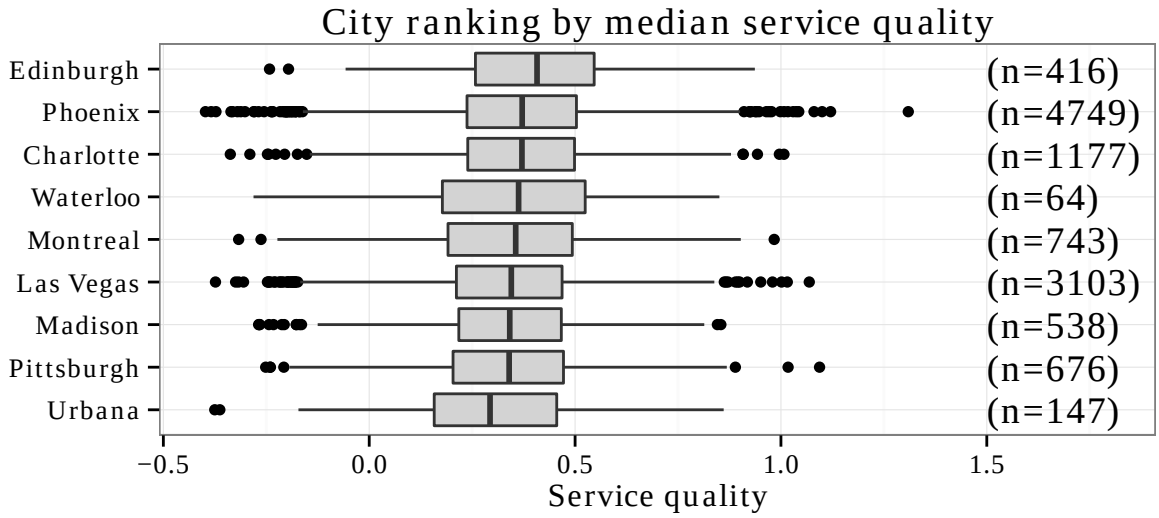
Unfortunately the given data set does not contain any information about the residence of the users. Since I was interested to know if one can find any difference in the rating of users from different cities, the residence

of the users had to be predicted. Since no labelled test data was available, supervised algorithms were no option. The implemented predictor is based on the following two assumptions. First, users write most reviews in the region were they live. Second, users usually visit a doctors in their home town. The implemented algorithm counts the reviews for each user in each region and assigns the region with most reviews as his residence. To increase the reliability of the predictor only users who have written at least 5 reviews or 1 doctor review are included. Both predictors, the one based all reviews and the one based doctor visits, agree for over 99% of the users. This is a very good indicator that the predictor does a good job. However it is not perfect, and does not consider when people relocate.

4 Results

4.1 City ranking by median service quality

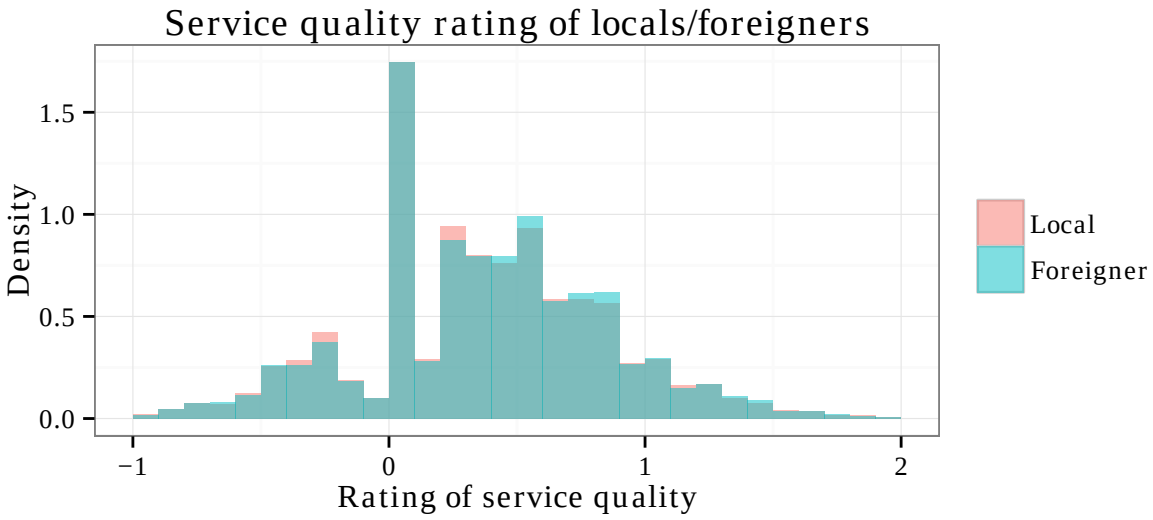
The first goal of the study is to calculate the average service quality in the ten major cities. First all service ratings have been aggregated by restaurants. Only restaurants with five or more comments have been included. Next, all business ratings are aggregate by region. The following box plot shows the distribution of the ratings of the restaurants in each region. The numbers in parenthesis indicate the number of rated restaurants in this region. The German city Karlsruhe does not appear in the plot because there where not enough English comments for this city. Since the distributions of the ratings are slightly skewed, they are ordered according to the median service quality.



From the box plot we can see that the average quality is best in Edinburgh (median=0.40), and worst in Urbana (0.30). However the differences between the service qualities are not significant. The standard deviation is to large and around 0.2 for all regions. If the restaurants with less than five comments are included in the aggregation, the main result does not change. The variance increases and there are more extreme ratings. However, the best and worst regions are still the same. But the ordering of the regions in the middle changes slightly.

4.2 Service quality rating of locals and foreigners

The second goal is to investigate the rating behavior of locals and foreign customers. The following plot shows the distribution of the ratings of locals and foreign customers.



The histograms of both groups are very similar. The median service quality rating of locals is 0.35 (n=323521), and 0.37 (n=9537) for foreigners. The difference is statistically not significant. Therefore we conclude that people apply the same criteria when they review local or foreign restaurants.

4.3 Correlation between restaurant rating and service quality

The last goal was to analyse the impact of the service quality to the overall star rating of a business. One would assume that the service quality is a very important factor for the users. This is confirmed by the following plot showing the correlation between the average stars of a restaurant and the rating of the service quality. Each comment about the service quality is represented by a point in the plot. Since the businesses are rated by one to five stars, the points cumulate around the integer and some typical fractional values.

A linear model fits very well to this distribution. The linear model is given by the equation *service quality* = -0.4 + 0.22 *stars*. The R-squared value of the correlation is 0.39. The correlation is strongly significant. However any further statements about the relevance of other factors is beyond the scope of this analysis.

5 Discussion

The study shows the application of text mining techniques to extract quantitative information about the service quality from text based reviews. It assigns a service quality rating to each business. This information could be used to provide additional filtering options to all users of the Yelp platform. The service quality is a very important factor for the overall rating of a business. Both factors are strongly correlated with each other. This finding is not surprising, however it is very important for business owners who care about their customer happiness.

No significant differences could be found between the average service quality in different regions. So it is not possible to make any statements about cultural differences between the cities and countries under investigation. The winner of the city ranking is Edinburgh. This city would be a good tip for a visit at the next opportunity. Since the residence of the users was not given in the data set, it was more difficult to investigate the rating behavior of different user groups than expected. A simple predictor based on review counts has been used to predict the residence. A comparison of the reviews of local and foreign customers does not show any significant difference. It is nice to get the confirmation that people use the same benchmarks for rating local and foreign businesses.

The analysis provides many options for improvements. The first would be to include other languages such as German. This seems to be a mandatory step to compare cultural difference in more detail. Further, many comments contain typos and slang words. It would be interesting to know if the rating quality could be improved by applying additional auto-correction or stemming algorithms.

Finally one short comment about my experiences with the performance of R. Many simple transformation, e.g importing the data, split sentences or calculating polarities, take hours, if not days, to run. It is possible to speed up the calculations by sampling the data, adding extra parameters or using a package to parallelize some steps. However it takes a lot of time to identify the bottlenecks and to optimize the performance. Sub-sampling the data helps at lot at the beginning, but many problems only appear if one works with the complete data set, e.g. special values that can break the code. Since text extraction algorithms are not very precise, one needs to work with bigger data set to get reasonable values and compare individual cities and user groups. Therefore it seems to me that R is not the best tool to make this kind of analysis. In the future I would like to invest some more time to use others tools that perform and scale better on big data sets.

6 References

- Coursera Data Science Specialization: https://www.coursera.org/course/dsscapstone
- Yelp Dataset Challenge: http://www.yelp.com/dataset_challenge