Sentiment Analysis of Yelp user reviews

Introduction

We are a group of 3 people with very different personalities but the one thing that connects us is our love for food! Therefore, for our final group project, we decided to analyze the publicly available Yelp datasets to draw insights about the correlation between user review sentiments and user ratings by employing sentiment analysis techniques. We believe this kind of an analysis will be able to provide foodies like us with a more realistic view about the reviews on Yelp. At the end of the project we expect to be able to answer questions like: How reliable are these reviews? Do the reviews and the star ratings always match up?

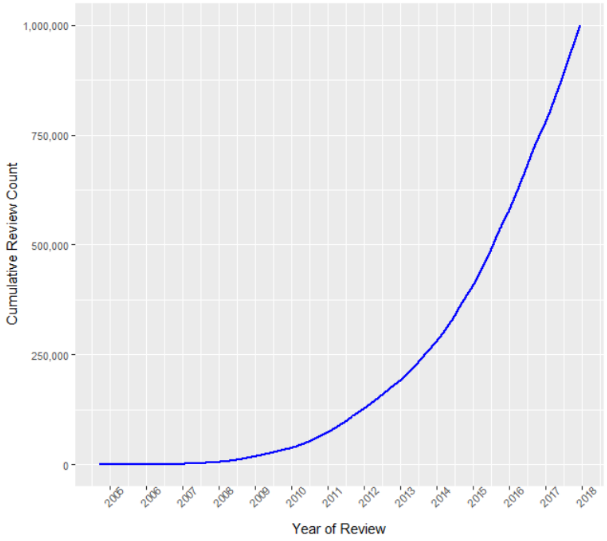
Yelp has over 135 million restaurant and business reviews worldwide. Whether one is looking for a new pizzeria, a great coffee shop nearby, a new salon, or the best handyman in town, Yelp is the city guide to finding the perfect places to eat, shop, drink, relax, visit and play. However, Yelp is mostly used by us for quickly browsing through popular restaurant reviews and finding our favorite cuisines in town. No wonder this topic is of great relevance to us and fellow gastronomic aficionados.

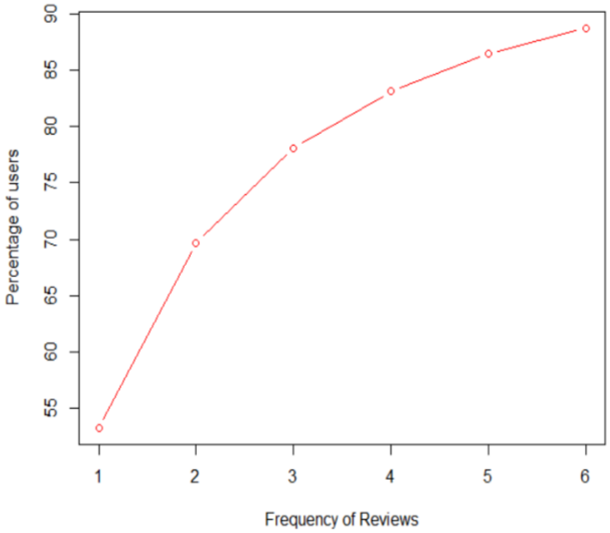
Data Acquisition

We used the publicly available Yelp “business” dataset (in JSON format) https://www.yelp.com/dataset/challenge along with Yelp “review” dataset from Kaggle (csv format) https://www.kaggle.com/naveenkhasa/yelp-dataset/data and integrated both for our analysis. For joining the datasets, we used the matching criterion, Business ID which happens to be the common variable in both the datasets.

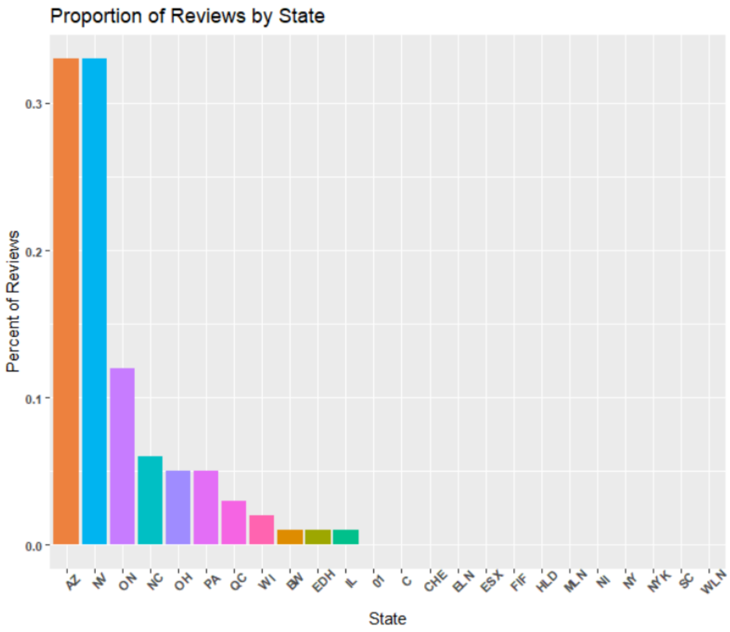
Data Cleaning & Exploration

We all have worked equally in the whole process and we started with data exploration. The imported “review” dataset was huge with over 5 million records therefore, we decided to take 1 million records for our analysis. Since, “business” dataset was in json format therefore we converted the dataset into csv using R. The reason behind taking “business” dataset from Yelp.com was that the “business” dataset on Kaggle was trimmed which was limiting our analysis. Next, we wanted to check the distribution and growth of the reviews over time and we prepared the following graph showing the cumulative review count for the time-span. As expected, the reviews seem to have grown exponentially in the past few years reflecting the growing popularity of Yelp.



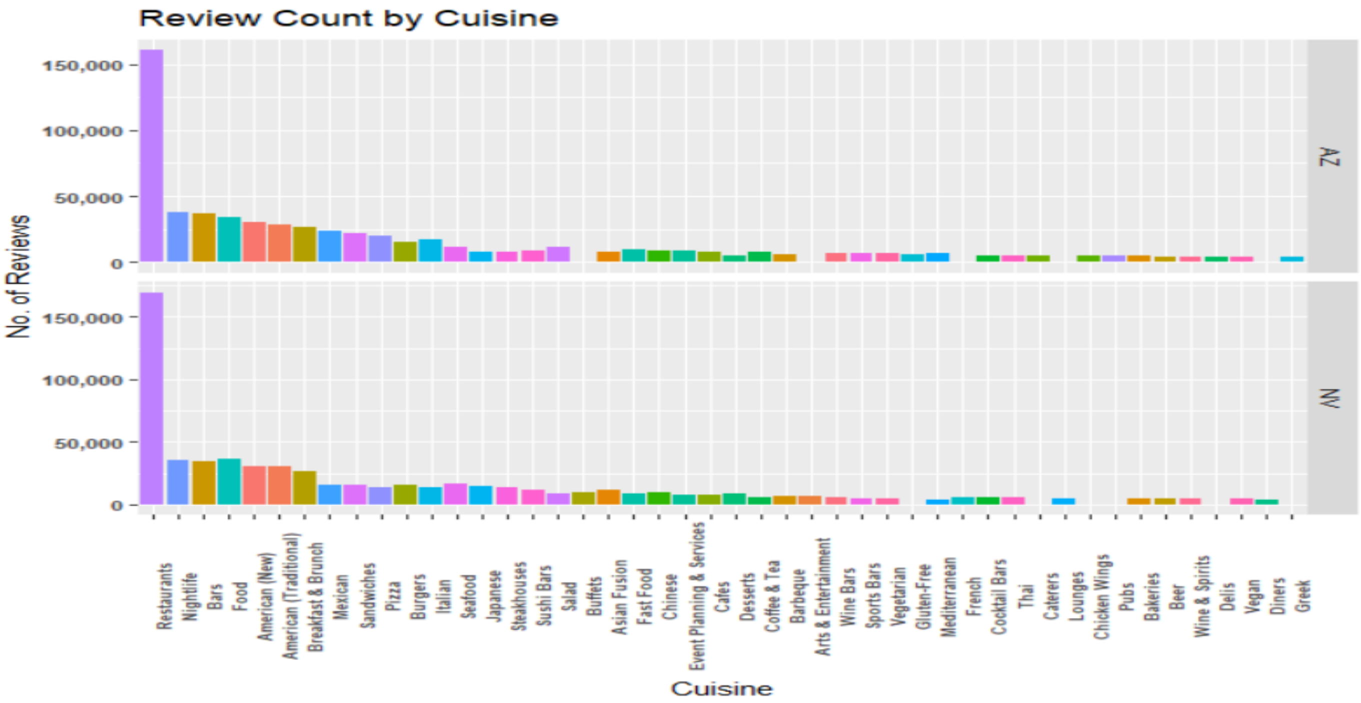


The next step was to scrub the data. For improving the reliability of our dataset, we decided to get rid of the inactive users from the dataset. We assumed that user accounts with very low no. of reviews are either inactive or are fake accounts. Our initial plan was to remove users with 2 or less reviews but as can be seen from the plot above, that would have resulted in large amount of data loss because 70% of the users have 2 or less reviews. Therefore, after a lot of deliberation we decided to drop only the users with 1 review i.e. around 55% of the users. By doing this we lost almost 14% of the reviews and were able to retain 86% of our review data.



The next step was merging the Business and Review data to produce a table of all reviews with matching business information. We then checked the different kind of cuisines and the percentage reviews pertaining to those cuisines. Once we had the information, we decided to choose Pizza as our cuisine of interest (being graduate students we all essentially survive on Pizza!) and also the amount of data available was sizeable enough (about 15% of the total number of reviews) for our analysis.

Next, we wanted to see the state-wise distribution of user reviews. While looking at the state wise count we wanted to be sure that we are looking at the reviews for restaurants, therefore we looked at all the categories of reviews present and came to a conclusion that American restaurants have a possibility to be marked as a Bar or a Lounge. To include them in our plot, we added these categories while looking at the state wise count. Also, we filtered out 90th percentile for our categories from the whole to make the state wise plot. A view by state showed Nevada and Arizona contributing an overwhelming 66% of total user reviews as shown in the graph. For these 2 states, we again checked the no. of reviews given for Pizza as shown in the plot below. We found out that around 40k-50k reviews were for Pizza and this was a considerable amount for our analysis.



Building the Corpus

Based on our analysis we decide to work on the reviews that were for Pizza from states Nevada and Arizona and this was the corpus for our sentiment analysis.

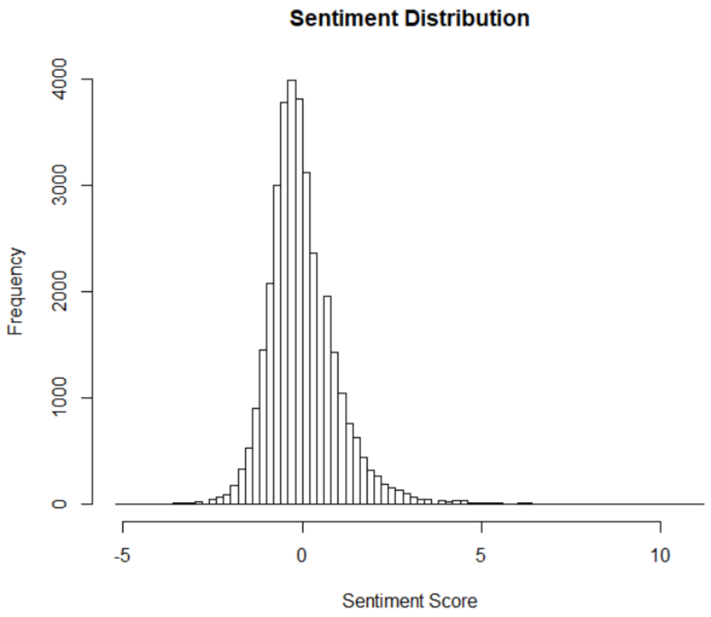
Sentiment Analysis

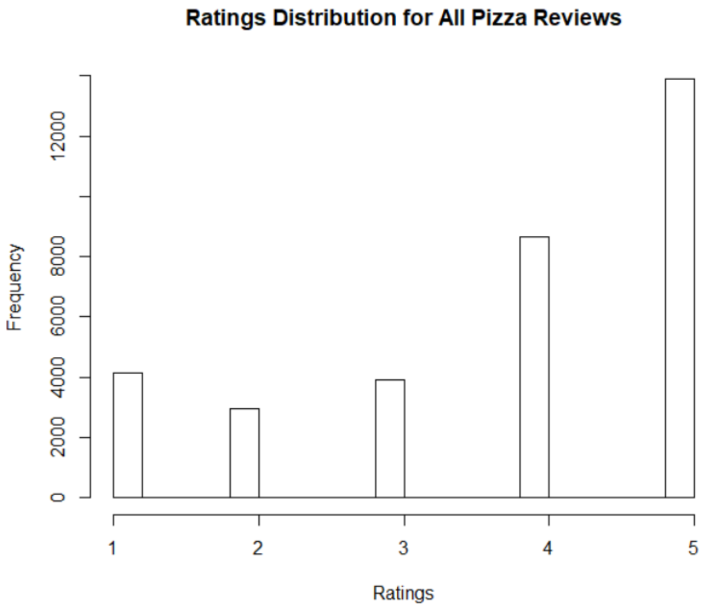
Once our corpus is defined, we are now ready to carry out Sentiment Analysis of the user reviews. For this we employ the bag of words technique. The steps of Sentiment Analysis are listed below:

1. We first removed the stop words like punctuation, articles etc.
2. We then split the remaining text into words and arrange them in the form of vectors.
3. For categorizing our words to positive and negative words, we used the Sentiment Lexicon by Prof. Bing Liu from UIC *(cs.uic.edu/~liub/FBS/sentiment-analysis.html)*.
4. For calculating the overall sentiment score we used a very basic approach which is essentially the difference between the summation of positive and negative scores

$$Sentiment\ score = \Sigma(positive\ matches) - \Sigma(negative\ matches)$$

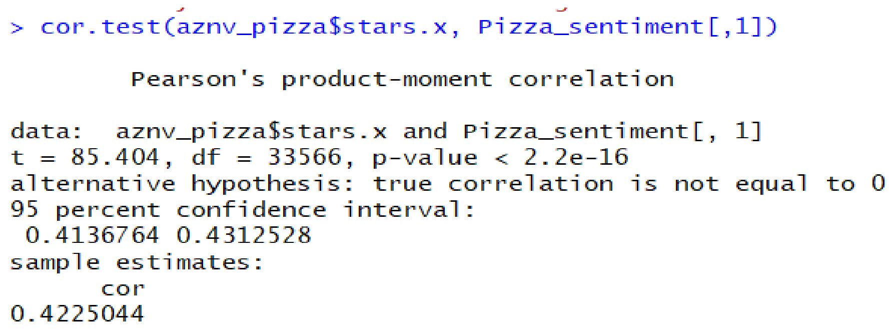
5. The score for each review is then scaled and plotted as a histogram.





Findings

The histogram of the sentiment scores is normally distributed which implies that most of the reviews are neutral. On plotting the distribution of the corresponding user ratings, we see contradictory results i.e. results are biased toward 4 and 5 stars. The first impression is that, review sentiment is probably not the best predictor of user ratings. To confirm this, we carried out the Pearson correlation test between the two, the result shows that there is a 42% correlation which is not strong enough.

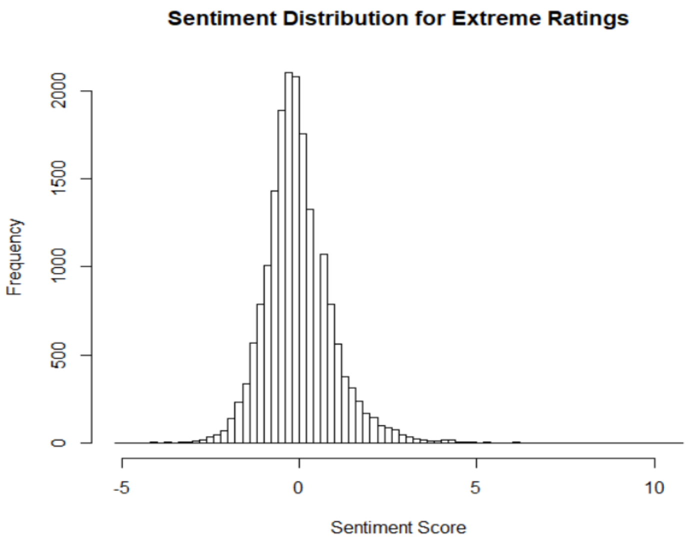


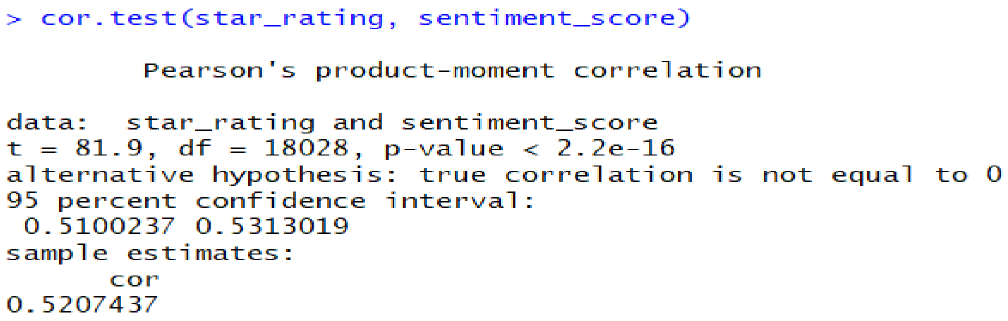
```
> cor.test(aznv_pizza$stars.x, Pizza_sentiment[,1])

        Pearson's product-moment correlation

data:  aznv_pizza$stars.x and Pizza_sentiment[, 1]
t = 85.404, df = 33566, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4136764 0.4312528
sample estimates:
      cor 
0.4225044 
```

We then decided to concentrate only on the extreme ratings (1 or 5 star ratings). The histogram of the sentiment scores for these ratings too is normally distributed. This time the outcome of the Pearson correlation test between the two, improved slightly to 53% correlation which is still not strong enough.





```
> cor.test(star_rating, sentiment_score)

        Pearson's product-moment correlation

data:  star_rating and sentiment_score
t = 81.9, df = 18028, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5100237 0.5313019
sample estimates:
      cor 
0.5207437 
```

From the above, we are now quite certain that user review sentiments are not strongly correlated to user ratings, hence the reviews can often be misleading. This also got us thinking about the probable limitations of our study, which if addressed in future might result in better results.

Limitations

1. Assigning equal weightage to all positive words (or all negative words) may not produce accurate sentiment scores. For example, ‘extraordinary’ and ‘nice’ are both positive words but they should not ideally carry the same weightage.
2. Analyzing single words without context does not give us a true picture. For example, ‘Very nice’ and ‘nice’ are not same. Another example is in the sentence ‘the food was not worth’, the word ‘worth’ is a positive word but the word ‘not’ makes it negative.
3. Reviews might comprise of reviews of food, ambience, service etc., so calculating sentiment score for food alone is challenging.

References

- https://medium.com/tensorist/classifying-yelp-reviews-using-nltk-and-scikit-learn-c58e71e962d9
- https://www.kaggle.com/lohitharcot/sentiment-analysis-nlp/code
- http://varianceexplained.org/r/yelp-sentiment/
- https://github.com/snehabangar/Sentiment-Analysis-NLP/blob/master/Project_Report.docx