From Reviews to Networks: A Study of User Behavior in Yelp

Kristine Andersen (KA), Frederik Bingen (FB)

Technical University of Denmark

Author Contributions:

- Testing subsets (KA)
- Analysis of network (FB)(KA)
- Sentiment Analysis (KA)
- Community structures (FB)
- Writing the report (KA)(FB)

Computer Science | Network Analysis

Yelp | *Sentiment* | *Community Structure*

Significance Statement

Understanding the relationship between user sentiments, star ratings, and social graph structures provides valuable insights into online behavior and review dynamics. This study focuses on Yelp data, where users are connected if they review the same business, to analyze the interplay between sentiment analysis, ratings, and graph metrics like centrality. The findings, such as the strong correlation between sentiment scores contribute to the understanding of user-generated content in review platforms. These insights can inform improvements in recommendation systems, content moderation, and user experience design by highlighting how users' behaviors and connections reflect their sentiment and preferences. This research adds a novel network perspective to sentiment studies.

Abstract

This study explores the interplay between sentiment analysis, star ratings, and social graph structures in Yelp review data. By constructing a network where users connect via shared business reviews, we examine key metrics such as modularity, community structures, and sentiment scores. A subset of users with exactly 20 reviews was analyzed, compared with other subsets to validate representativeness. Our findings reveal that while sentiment moderately correlates with star ratings, network structure and community formations are strongly influenced by geographic business locations, offering insights into user behavior and network dynamics.

Introduction

Understanding user behavior in online review platforms like Yelp provides valuable insights into how sentiments, ratings, and social connections relate. With millions of reviews and users, Yelp offers a rich dataset for studying the dynamics of user-generated content. In this study, we adopt a network analysis approach to examine the interactions between users who reviewed the same businesses. By connecting users as nodes based on shared business reviews, we aim to understand how sentiments and star ratings correlate with the structural properties of the social graph.

To manage the scale of the dataset, we focus on a subset of users who have reviews exactly 20 business. To ensure representativeness we compare the subsets characteristics against users with slightly fewer or more reviews. Metrics, such as average sentiment scores, star ratings, and modularity of communities, are analyzed to identify patterns in user behavior.

The study also dives into community structures within the graph using the Louvain method, exploring how factors such as geographic locations influence user clustering. By combining sentiment analysis with graph theory, this research sheds light on the interplay between user

opinions and network connectivity, contributing to a deeper understanding of online behavior and its implications for recommendation systems and content moderation.

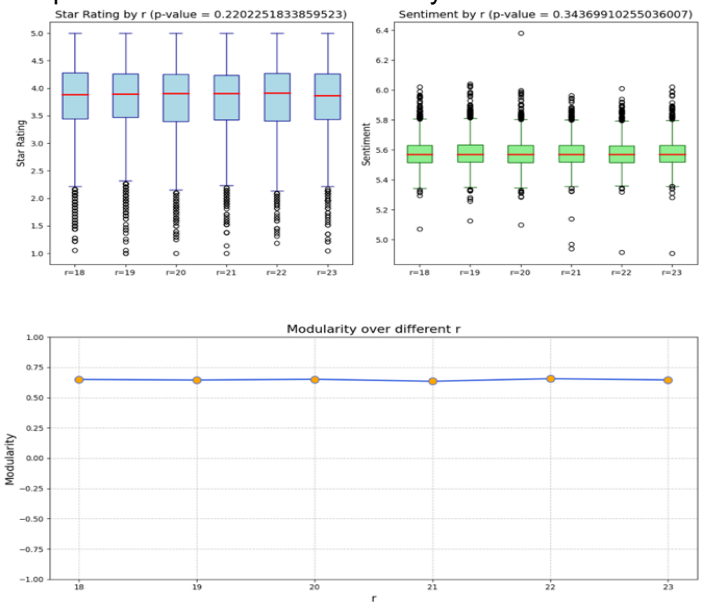

Figure 1: This figure shows 3 different plots. In the upper left corner is a box plot of the stars given per user is for different r, and a p-value that is indicating that the null-hypothesis that the different group are the same cannot be rejected. The same goes for the p-value for the box plot in the upper right corner, which shows how the sentiment score per user is when r changes. Here the null-hypothesis cannot be rejected either. The bottom plot shows how the modularity changes when r change.

Results

The dataset full dataset contains 6,990,280 reviews from 1,987,897 users that has reviewed 150,326 businesses. To make the data a workable size, a filter is applied to make a subset, where we only consider users who have made exactly 20 reviews (r=20). To make sure we can work with this subset, we first test out whether r=20 is representative for the whole dataset. We do this by looking at some of the important factors, such as average stars given per user, and sentiment of the reviews (all the reviews a user have done is together). These are compared with the subset where r=18, r=19, r=21, r=22, r=23, and then tested with a one-way ANOVA test, to check if we statistically can call them the same. After this, the modularity for Louvain communities is calculated for the different networks and compared.

From the ANOVA test, we derived the p-value for both variables, with the null-hypothesis that the groups are the same. And from the results (p-value = 0.22 and p-value = 0.34), we cannot reject the null-hypothesis, meaning when looking at these two variables, we will assume the groups are the same. This is also visible in the figure 1. Below the boxplots are the modularity plotted for the different r. It is plotted in the range modularity goes from -1 to 1, as that is the range the modularity can take[2], and we see it is quite constant around 0.65. This means we will also assume the modularity is the same for the different groups.

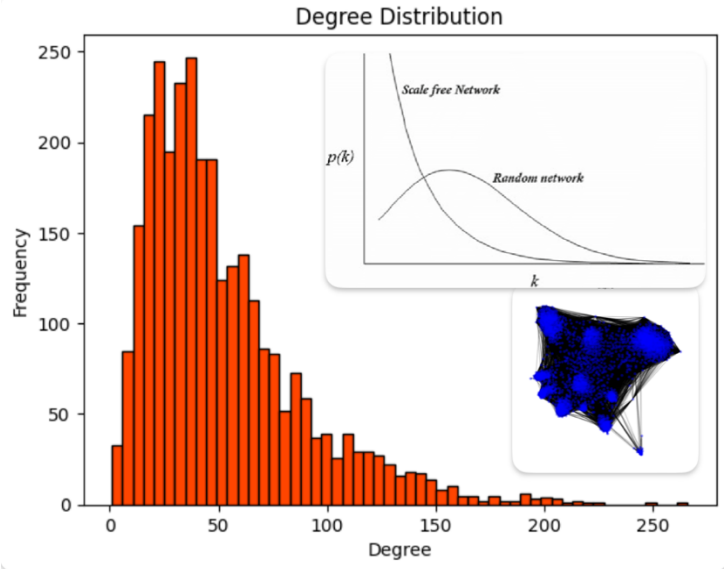

Figure 2: This figure shows the degree distribution of the graph G. In the upper right corner is the general distribution for a random network and a scale free network[3], this is to compare with the distribution from G. Right under this, the network is show, where hubs are visible.

Analysis of the network

Firstly, in the analysis, we will look at the network created from the subset

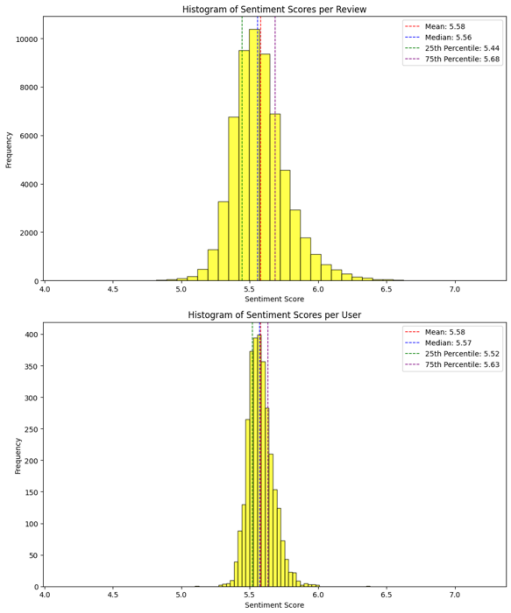

Figure 3: The sentiment score plotted. The top plot shows the sentiment plotted per review, while the bottom shows the sentiment score per user. Both have the bell shape from a normal distribution, but in the top plot, the range is a bit wider. For the top plot we have min=3.98 and max=7.38. For the bottom plot we have min=5.10, max=6.38.

r=20. The network will be denoted G, and is the greatest component in the subset. This network have 3,024 nodes (*N)* and 78,349 edges (*L)*. The lowest degree of a node in the network is 1, and the highest is 266, while the average degree (*k)* is 51.82. The degree distribution is plotted in the figure 2. The degree distribution plot reveals that most nodes have a low degree, with a few highly connected nodes. This right-skewed distribution suggests that the network is not uniformly connected but instead includes a small number of hubs, which also can be seen in the network that is shown in figure 2, while the majority have fewer connections. The degree distribution shows similarities to scale-free network but with notable differences. In a true scale-free network, the degree distribution follows a power-law, where the tail remains heavy, and a few hubs dominate the connections. Here, while the distribution is right-skewed, the tail tapers off more sharply, suggesting that the network does not fully conform to the scale-free model.

Sentiment Analysis

The sentiment score gives a scores based on how happy a text is. This is a score from 1 to 10, where 1 is the furthest from happy, and 10 is the happiest you can be. Now, looking at the sentiment score for the reviews, we look at that text written by the users. First, we look at the sentiment scores plotted. In figure 3 we see the sentiment scores per review and the sentiment scores when you concatenate all the reviews a user have made, i.e. sentiment score per user. Common for both plots, are they have the bell shape that is a key feature for a normal distribution. On the other hand, when you are only looking at the reviews individually (sentiment per review), then the sentiment score spread in a wider range (min=3.98, max=7.38) than the sentiment scores per user (min=5.10, max=6.38). This mean when you aggregate the sentiment scores for the reviews a user has made, that user becomes more neutral in the language. Thus, we don't have any users that are just mad all the time, neither do we have any users that are happy with everything.

Sentiment vs Stars Given

As described before, the sentiment score gives a scores based on how happy a text is. One would expect that the number of stars given in a review, is a similar measure, although it is quite subjective. This means, we will look at the sentiment score together with

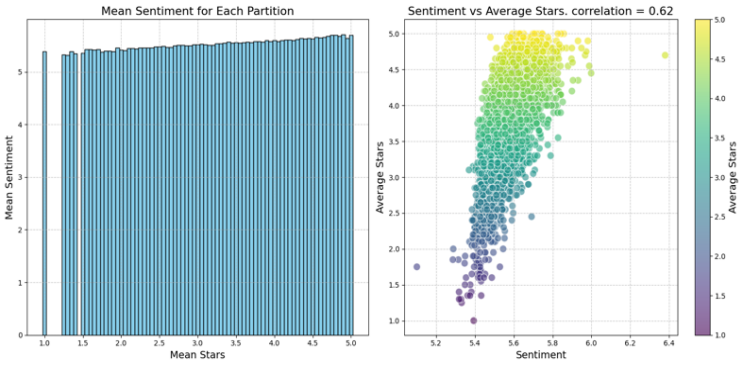

Figure 4: The left plot shows a bar plot where the average stars given is plotted with the average sentiment per user. We see a very slight increase in sentiment when stars given increases. The right plot is a correlation plot between stars given and the sentiment (both per user). Here we also see a small correlation which is verified with the correlation coefficient, that is 0.62.

the stars given. The stars given will be the average stars given for a user. The expectation for this would be that when the average stars given increases for a user, then the sentiment score per user also increases, meaning they are correlated. For this analysis, a bar plot was created in figure 4 was created. Here we see a slight increase in the sentiment score when the mean stars given increased. It is however a really small increase, which makes it hard to conclude anything, thus we created a correlation plot that is shown next to the bar plot. From the plot, we see tendency to correlation. Together with the correlation plot, the correlation coefficient is calculated, which is 0.62. This indicates a moderate correlation between the sentiment score and the stars given for each user[4].

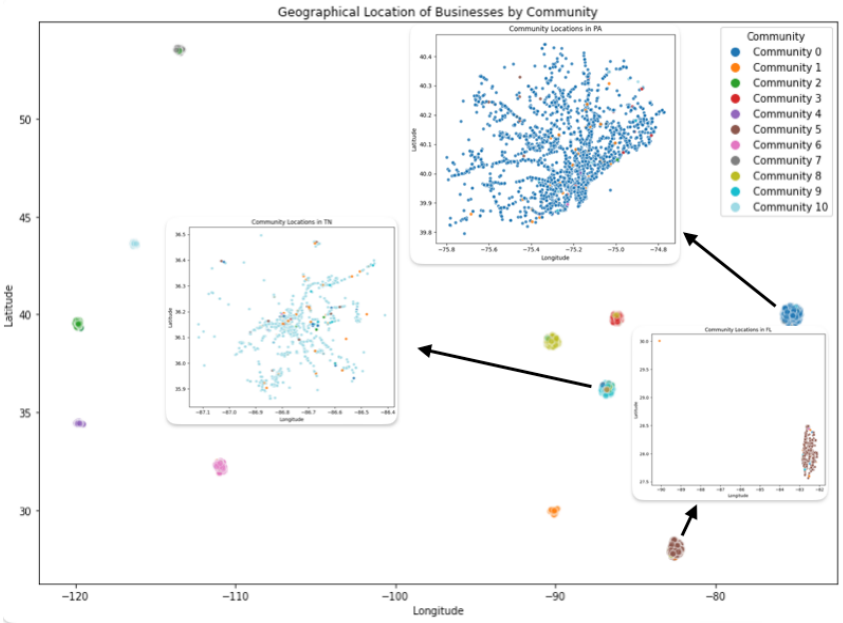

Figure 5: Louvian communities plotted on a longitude/latitude graph where 3 communities is zoomed in. Here we clearly see how the location of the businesses influence the communities created.

Structure of Network

As noticed earlier, the modularity was quite high, and are calculated to be 0.65 for this subset, thus we know the network have strong community structures. To analyze this structure and partition, we look further into the Louvain partition. For the sentiment score, each community in the partition has almost the same mean sentiment score, namely around 5.58 ± 0.02, which indicates the communities is based on something else. The location of the businesses that was reviewed. When plotting this as seen in figure 5, it is noticeable how the locations have a big influence on the communities. In the plot, 3 communities have been zoomed in.

Discussion

At the beginning, we assumed that the subset (users with exactly 20 reviews) represents the entire dataset, this is risky, due to selection bias. Users with exactly 20 reviews may show unique behavior, such as more consistent engagement or specific reviewing patterns, which might not be general to users which are less or more active. We tried getting rid of that bias by looking at specific variables and comparing subsets, and although the ANOVA test suggests no significant difference for some variables, these results might overlook nuanced behavioral or structural differences. We chose to compare variables such as sentiment score and average stars as these are a major part of the analysis. Modularity was also compared as this is a measure of the structure of the network and are scaled which makes the numbers comparable.

When looking at the sentiment scores, this measure do not account for the context or intensity of sayings (likely typos or linguistic nuances in reviews). For example, sarcastic language can skew sentiment analysis, misrepresenting the true emotion expressed. This skewness could be improved by incorporating a more advanced natural language processing method.

The wider range of sentiment scores per review compared to aggregated sentiment scores per user is quite logical. Individual reviews reflect specific experiences, which can vary, while

aggregated scores smooth out extremes. This aligns with the observation that users tend to appear neutral when their sentiments are averaged, indicating balanced overall behavior.
The moderate correlation (r = 0.62) between sentiment scores and stars suggests a relationship but not a strong causation. Variations in individual reviewing standards, cultural differences, and subjective interpretations of rating scales may influence this correlation.
The analysis of the network structure highlights the distinction of community formations driven by geographic locations rather than sentiment scores. Despite strong modularity (0.65), the Louvain partitions reveal consistent mean sentiment scores across communities. This indicates that local business clusters, rather than emotional tone, influence the observed community structure. However, the potential influence of other attributes, such as user demographics or business types, remains unexplored. Incorporating these factors into future analyses could provide a more holistic understanding of the forces shaping community formation. Additionally, refining sentiment analysis techniques to account for nuanced language, such as sarcasm, may yield more accurate insights into user interactions within these communities. For future research, one could dive deeper into integrating location-based data and advanced NLP techniques to better capture these underlying dynamics.

Materials and Methods

The Yelp dataset, which was retrieved from https://www.yelp.com/dataset, including 6,990,280 reviews from 1,987,897 users, was filtered to include users with exactly 20 reviews (r = 20). Subsets for r = 18–23 were used for comparative analysis. A network was constructed with nodes representing users and edges formed by co-reviewing businesses. The dataset is split in two, first file is called *review.json* and has the variables (review_id, user_id, business_id, stars, date, text, useful, funny, cool). The other is called *business.json* and has the variables (business_id, name, adressm city, state, postal code, latitude, longitude, stars, review_count, is_open, attributes, categories, hours)
The Louvain method was applied for community detection, and modularity was calculated. Modularity was based on communities made by a partition from the inbuilt Python algorithm community_louvain for all the subsets.
Sentiment scores (1–10) were derived from a wordlist called Data_Set_S1.txt from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026752, and analyzed per user and individually. Statistical analysis, included one-way ANOVA to test subsets.
All analyses were conducted in eaxplainer notebook Python using NetworkX, Matplotlib, Pandas, ForceAtlas and other related libraries. This notebook can be found here: https://github.com/kristineAA/02805YelpProject/blob/main/output/Project_2.ipynb

References

1. A complete guide to successful Customer Analysis in 2024, November 2024, IMD, https://www.imd.org/blog/marketing/customer-analysis-marketing-plan/
2. Barabási, Albert-László. Network Science Chapter 9
3. Kumar, Prem & Verma, Puneet & Singh, Anurag. (2018). A Study of Epidemic Spreading and Rumor Spreading over Complex Networks. 10.1007/978-981-13-2348-5_11. E. van Sebille, M. Doblin, Data from “Drift in ocean currents impacts intergenerational microbial exposure to temperature.” Figshare. Available at https://dx.doi.org/10.6084/m9.figshare.3178534.v2. Deposited 15 April 2016.
4. Calkins, Keith G, (2005), Applied Statistics – Lesson 5 Correlations Coefficients, Andrews University, https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm