Final Report - Geospatial Analysis for Yelp User Behaviour

Mengyuan Li, Siying Liu, Xianyi Nie, Hao Hu, Yaxin Xue, Xiaocheng Chen (All team members have contributed a similar amount of effort)

1 INTRODUCTION

The rapid growth of online review platforms, such as Yelp, has transformed the way customers evaluate and choose businesses. Yelp as one of the most popular online review platforms provides a wealth of business information and user-generated content, including ratings, tips, and reviews. Among these, customer reviews hold significant value, as they not only convey customer experiences but also reflect public sentiment towards businesses. Yelp data provides a unique opportunity to explore the patterns and trends in user behavior across different regions, which can have important implications for businesses and policymakers.

In this project, we focus on four US cities: Carpinteria, Philadelphia, Santa Barbara, and Tucson. We chose these cities due to their diverse geographic, demographic, and economic characteristics, which can provide a comprehensive understanding of user behavior in different urban environments. Our primary goal is to explore and visualize the geospatial patterns and sentiment of the reviews, as well as to examine user behavior based on their reviews.

2 PROBLEM DEFINITION

Geospatial analysis provides a powerful tool for visualizing and analyzing spatial patterns and trends. [9] By mapping Yelp data, researchers can identify clusters of businesses and users and examine how they vary across different regions.[2] This can provide insights into the factors that influence user behavior and the competitiveness of different business sectors.

As for the reason for success, this approach can inform a wide range of stakeholders, including businesses, policymakers, and consumers. By identifying regional differences in user behavior and the local business environment, companies can tailor their marketing strategies and services to better meet the needs of their customers. Policymakers can also use this information to support economic development and promote business growth in different regions. [5]

Given a dataset of Yelp reviews and business information, our goal is to analyze user behavior, business characteristics, and their relationships across four US cities. We aim to answer the following questions:

- What are the sentiment patterns in reviews across different cities?
- How do user behavior and business characteristics correlate in these cities?
- Can we identify clusters of similar businesses and user behaviors?
- How does user behavior vary across different geographic locations?

3 LITERATURE SURVEY

In contemporary business practices, when decision-makers decide where to locate or expand their business, it is necessary to conduct extensive analysis, such as browsing various locally based social media, government records, and customer surveys to collect data, then analyzing it using Geographic Information Systems (GIS) software [1, 6]. This often leads to increased costs in terms of time and financial resources.

In this context, Yelp, a large public review site, presents an invaluable source of data for business decision-makers to gain a deeper understanding of their target market. However, most current analyses on Yelp datasets have focused on detecting fraudulent reviews [3] and developing recommendation algorithms for users [4, 7], with limited research on the combination of Yelp review data and spatial geographic distribution. Sun, Y. and Paule, J.D.G. used the "AMOEBA" spatially constrained clustering algorithm to analyze the spatial patterns of commercial places in Phoenix, USA based on Yelp data, but only considered the star ratings of the places [8].

4 PROPOSED METHOD

The proposed method aims to conduct geospatial analysis on Yelp user behavior based on Yelp data. The analysis will be carried out by processing and analyzing a large dataset of Yelp user reviews, ratings, and other relevant data in order to derive insights into Yelp user behavior in different geospatial locations. The proposed method involves several steps, including data collection,

data cleaning, data preprocessing, and geospatial analysis.

4.1 Intuition

Our approach combines sentiment analysis, correlation analysis, clustering, and geospatial analysis to provide a comprehensive understanding of user behavior and business landscape in the selected cities. In terms of innovations, the proposed method exhibits the following ideas:

1. Integration of Yelp data with geospatial analysis: The proposed method integrates Yelp data with geospatial analysis techniques to derive insights into Yelp user behavior across different geospatial locations. This approach helps to provide more accurate and detailed information about Yelp user behavior, as it takes into account the specific geospatial context in which user behavior occurs.

2. Combination of sentiment analysis and geospatial analysis: By integrating natural language processing and geospatial techniques, this method offers a novel way to uncover location-based insights and trends related to review sentiments and the keywords driving those sentiments.

3. Application of cluster analysis for spatial segmentation: The proposed method applies cluster analysis techniques to segment geospatial locations based on Yelp user behavior. This approach helps to identify patterns and trends in Yelp user behavior across different geospatial segments, which can then be used to inform business decisions and marketing strategies.

4. Incorporation of visualizations for data analysis: The proposed method incorporates various visualizations such as heat maps, scatter plots, and choropleth maps to visualize the results of the geospatial analysis. These visualizations help to convey the insights derived from the analysis in a more intuitive and understandable manner.

4.2 Dataset

We used three dataframes for our analysis:

- reviews: Contains information on Yelp reviews, including review_id, user_id, business_id, stars, useful, funny, cool, text, and date.
- businesses: Includes details about Yelp businesses, such as business_id, name, address, city, state, postal_code, latitude, longitude, stars, review_count, is_open, attributes, categories, and hours.
- users: Provides information about Yelp users, including user_id, name, review_count, yelping_since, useful, funny, cool, elite, friends, fans, average_stars, and various types of compliments.

4.3 Approaches

1. Data collection and preprocessing: Gather a dataset of reviews with location data (e.g., latitude and longitude, postal codes) and user ratings. Clean and preprocess the data, including normalizing text and labeling sentiment based on user ratings.

2. Location-based sentiment analysis: Perform sentiment analysis to identify keywords or phrases that influence review polarity using techniques like TF-IDF. Associate the discovered keywords or phrases with their corresponding location data, and aggregate the data by location to calculate keyword/phrase frequency and average sentiment.

3. Geospatial visualization: Create interactive maps using geospatial visualization libraries to display the distribution of keywords or phrases and their associated sentiment across the geographical area of interest. Visualizations may include choropleth maps, heatmaps, or point maps with color-coded markers based on sentiment.

4. Spatial clustering: Apply spatial clustering techniques to identify clusters or hotspots of similar sentiment and keyword usage. Analyze these clusters to uncover distinct patterns and preferences among different regions.

By implementing the functions, we can provide a comprehensive understanding of the spatial distribution of review sentiment and the keywords contributing to the polarity of reviews. This approach can be particularly useful for businesses looking to improve their products or services, policymakers seeking to understand public opinion, or researchers exploring regional preferences and trends.

Overall, the proposed method provides a comprehensive and innovative approach to analyzing Yelp user behavior using geospatial analysis techniques. By leveraging Yelp data and applying advanced analytical techniques, the proposed method can help businesses gain a deeper understanding of Yelp user behavior and make more informed decisions.

4.4 Challenges

During the analysis, we encountered several challenges:

- Large dataset: Due to the size of the dataset, we had to optimize our code for faster execution and limit the reviews data size.
- Data manipulation: We needed to handle missing values, merge dataframes, and preprocess text data to extract meaningful information.
- Visualizations: Creating informative visualizations required the use of third-party libraries like folium and adding progress bars for better viewer experience.

5 EXPERIMENTS / EVALUATION

5.1 Testbed Description

Our testbed consists of Yelp datasets containing reviews and business information for four selected US cities: Carpinteria, Philadelphia, Santa Barbara, and Tucson. The datasets include various attributes, such as review text, user and business IDs, ratings, geographic coordinates, and business categories. To evaluate the results of our analysis, we employed both qualitative and quantitative methods. Our evaluation approach focused on addressing the research questions we formulated at the beginning of the project.

5.2 Experiment Objectives

Our evaluation was based on the insights and patterns we were able to derive from the data. We also considered the limitations of our analysis, such as potential unexplored factors or biases in the dataset, while interpreting the results. The evaluation process helped us draw conclusions and discuss the implications of our findings, as well as identify possible future extensions to the project.

5.3 Observations

Sentiment distribution across cities: We compared the sentiment scores obtained through sentiment analysis in a visual manner, such as using histograms or bar charts. This allowed us to assess the variation in sentiment across cities and identify any significant differences between them.

Relationships between user behavior and business attributes: We computed correlation coefficients between various user behavior attributes and business attributes. We evaluated the strength and direction of these relationships by analyzing the correlation matrix and visualizing the results using heatmaps.

Clustering of businesses with similar user behavior patterns: We used KMeans clustering to group businesses based on their geographical location and sentiment scores. To evaluate the quality of the clusters, we examined the cluster sizes, distribution of businesses within clusters, and the consistency of user behavior patterns within each cluster. Additionally, we visualized the clusters on a map to assess the spatial distribution of businesses belonging to the same cluster.

Geographic influence on user behavior patterns: We performed geospatial analysis using Folium to create interactive maps that visualize user behavior patterns in the selected cities. We evaluated the results by observing the spatial distribution of user behaviors and sentiments, identifying areas with distinct patterns, and comparing these patterns across different cities.

6 CONCLUSION AND DISCUSSION

Our experiments revealed distinct sentiment patterns and correlations between user behavior and business characteristics in the four cities. We successfully identified clusters of similar businesses and user behaviors, highlighting areas with similar patterns of sentiment and user activity. Moreover, the geospatial analysis provided valuable insights into the geographic distribution of user behavior patterns.

In this study, we utilize the publicly accessible dataset available on Yelp.com, which encompasses comprehensive information about businesses located in major cities across various states and provinces, including Pennsylvania (PA), Florida (FL), Tennessee (TN), Indiana (IN), Missouri (MO), Louisiana (LA), Arizona (AZ), New Jersey (NJ), Nevada (NV), Alberta (AB), California (CA), Idaho (ID), Delaware (DE), and Illinois (IL). Figure 1 presents the distribution of the number of businesses in each of these cities, offering valuable insights into the geographical landscape of the data.

Figure 2 and Figure 3 show the top 10 most popular businesses across the US and the top 10 cities with the highest number of businesses.

Figure 4 and Figure 5 show the most positive and most negative words for the sentiment analysis.

In order to conduct a more in-depth geospatial analysis, we have chosen to focus on specific cities, starting

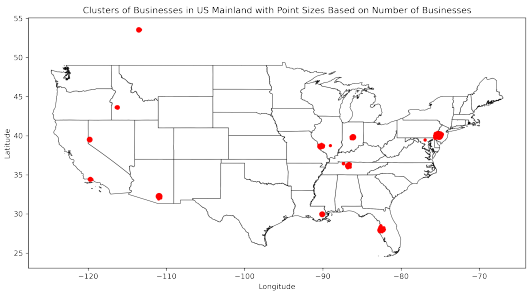

Figure 1: Number of Businesses in US Mainland by Major Cities

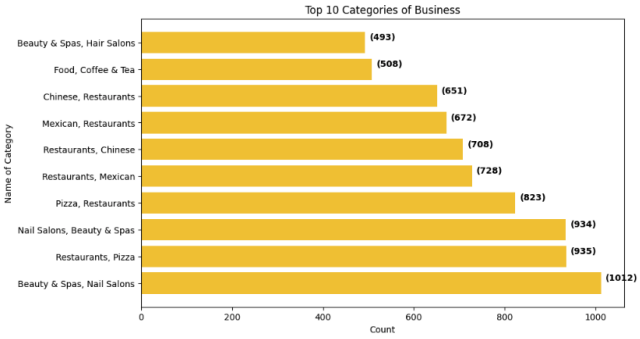

Figure 2: Top10 Business Types

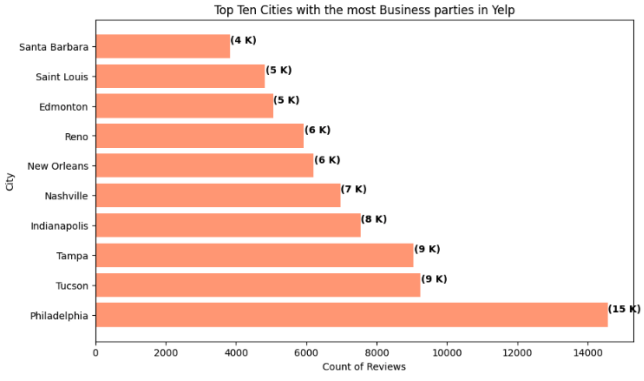

Figure 3: Top10 Cities With Business

with the popular small town of Carpinteria in California. Figure 6 illustrates the distribution of businesses within Carpinteria,with color-coded markers representing their ratings. The diverse range of ratings suggests various levels of customer satisfaction within

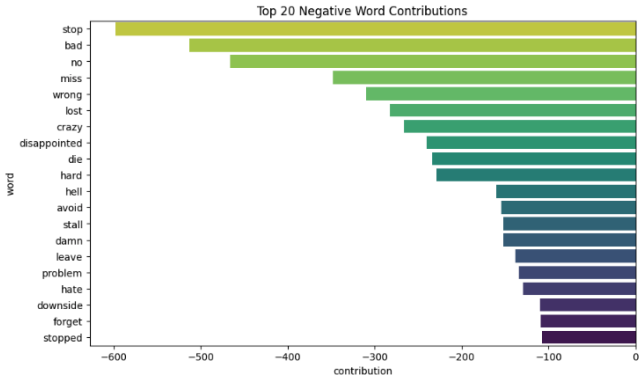

Figure 4: Top20 Negative Words

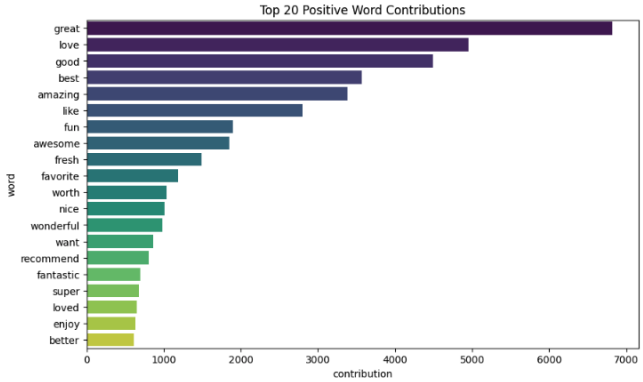

Figure 5: Top20 Positive Words

the town.We observe that a considerable number of businesses are strategically situated near Highway 101, which serves as the primary traffic artery in the area. It is also noteworthy that, despite the high concentration of businesses along this route, there is a diverse range of ratings (indicated by different colors, with green representing higher ratings and red signifying lower ratings).

Additionally, we investigated the geospatial clustering of businesses based on their perceived "expensiveness" as mentioned in customer reviews. Figure 7 displays the results, with red denoting the most expensive businesses and green representing the least expensive ones. We find that the priciest businesses are also predominantly clustered around Highway 101.

This analysis provides valuable insights for business owners, particularly those considering offering luxury services. They must weigh the trade-offs between competing with other high-end service providers in the vicinity of Highway 101 or relocating to a slightly more

remote location. In the latter case, they must also take into account whether there is a sufficient customer base to support their luxury services. Furthermore, this analysis aids consumers in identifying businesses that align with their spending preferences.

Figure 6: Location and Ratings of Businesses in Carpinteria

Figure 7: Most Expensive Businesses in Carpinteria

Sentiment analysis was also performed on Yelp reviews from Santa Barbara, Philadelphia, and Tucson, and the results were presented in Figures 8, 9, and 10. These figures illustrate the distribution of positive and negative sentiments in each city, with blue indicating negative sentiments and red indicating positive sentiments. Based on the results presented in the figures, it is evident that positive comments greatly outnumber negative comments, and their spatial distribution is heavily influenced by the city layout. Additionally, while negative and positive comments are typically scattered, there are regions in the cities where almost all comments are positive. This finding indicates that the sentiment of Yelp reviews varies significantly across different geographic locations, and businesses operating in these cities need to pay close attention to the sentiment of their reviews to tailor their strategies effectively.

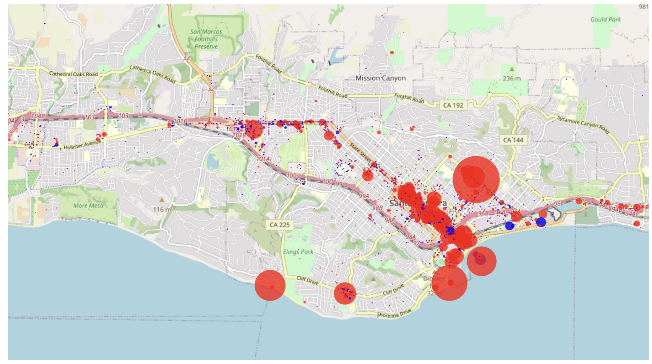

Figure 8: Santa Barbara Business Sentiment Map

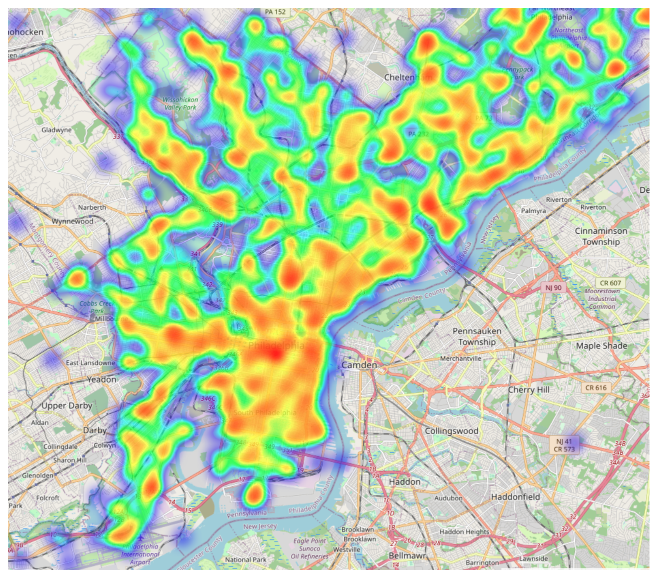

Figure 9: Sentiment Heat Map Philidelphia

Figure 11 and 12 show the top 10 positive and negative words for restaurants in Carpinteria and Santa Barbara. It's evident that there are both similarities and differences in user behavior and preferences between the two locations. Both locations emphasize the importance of delicious food and exceptional customer service. However, reviewers in Carpinteria tend to value the freshness and portion sizes of the dishes, while those in Santa Barbara prioritize the overall dining experience and culinary innovation. To enhance their establishments, business owners in each location should

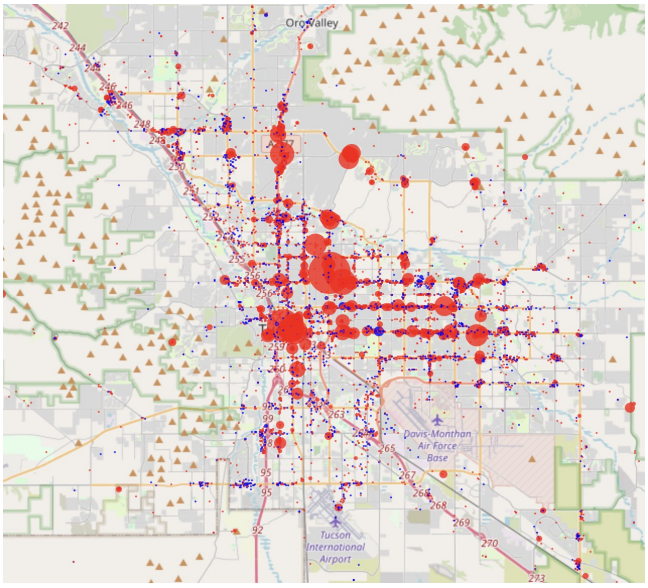

Figure 10: Tucson Business Sentiment Map

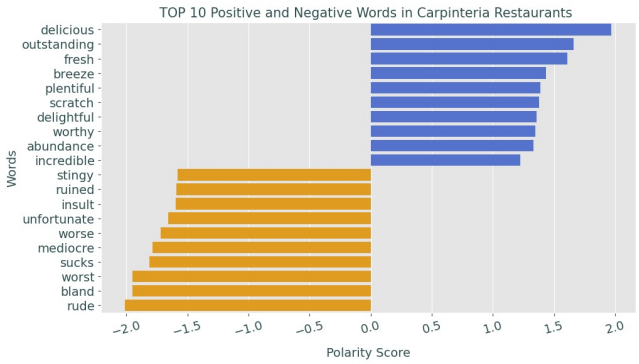

Figure 11: Top 10 positive and negative words for restaurants in Carpinteria

take these preferences and feedback into account. By understanding these unique regional preferences and incorporating them into their offerings, business owners can effectively cater to the tastes of their respective communities and create a more satisfying and successful dining experience for their patrons.

Furthermore, after filtering the reviews within these cities, all relevant comments were rated on a scale from negative one to positive one, revealing a shift from negative to positive sentiment. The insights gained from sentiment analysis can be leveraged to inform business strategies in the hospitality industry. By analyzing the sentiment of Yelp reviews, businesses can gain a better understanding of their customer's experiences

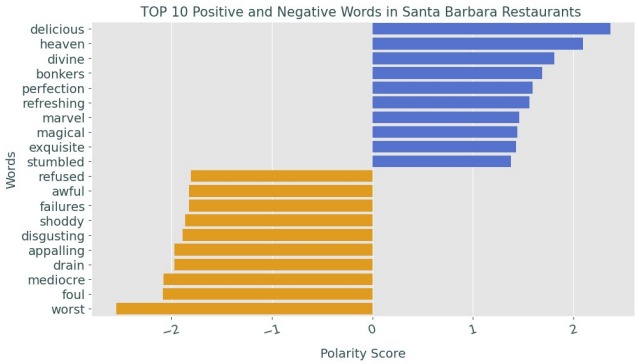

Figure 12: Top 10 positive and negative words for restaurants in Santa Barbara

and perceptions, identify areas for improvement, and take action to enhance their overall reputation. Thus, the findings of this analysis have practical implications for businesses in the hospitality industry seeking to improve their presence and customer satisfaction.

However, our study has some limitations, such as the reliance on Yelp data, which may not accurately represent the entire population's preferences. Additionally, our analysis only considered a subset of cities in the US, and future work could expand the scope to include other cities and countries. Despite these limitations, our findings can be valuable as they shed light on user behavior in different urban environments and the relationships between user behavior, business characteristics, and geographic locations.

7 EFFORT

All team members have contributed a similar amount of effort.

REFERENCES

[1] Michael A. Erskine, Dawn G. Gregg, Jahangir Karimi, and Judy E. Scott. 2014. Business Decision-Making Using Geospatial Data: A Research Framework and Literature Review. *Axioms* 3, 1 (2014), 10–30. https://doi.org/10.3390/axioms3010010

[2] James Huang, Stephanie Rogers, and Eunkwang Joo. 2014. Improving restaurants by extracting subtopics from yelp reviews. *iConference 2014 (Social Media Expo)* (2014).

[3] Michael Luca and Georgios Zervas. 2016. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science* 62, 12 (2016), 3412–3427. https://doi.org/10.1287/mnsc.2015.2304

[4] Yi Luo, Liang Rebecca Tang, Eojina Kim, and Xi Wang. 2020. Finding the reviews on yelp that actually matter to me: Innovative approach of improving recommender systems. *International Journal of Hospitality Management* 91 (2020), 102697.

[5] Shah Jahan Miah, Huy Quan Vu, John Gammack, and Michael McGrath. 2017. A big data analytics method for tourist behaviour analysis. *Information & Management* 54, 6 (2017), 771–785.

[6] James B. Pick. 2005. *Geographic Information Systems in Business*. Idea Group Inc.

[7] Sumedh Sawant and Gina Pai. 2013. Yelp food recommendation system.

[8] Y. Sun and J.D.G. Paule. 2017. Spatial analysis of users-generated ratings of yelp venues. *Open Geospatial Data, Software and Standards* 2, 5 (2017). https://doi.org/10.1186/s40965-017-0020-9

[9] Zheng Xiang, Bing Pan, Rob Law, and Daniel R Fesenmaier. 2010. Assessing the visibility of destination marketing organizations in Google: A case study of convention and visitor bureau websites in the United States. *Journal of Travel & Tourism Marketing* 27, 7 (2010), 694–707.