30538 Final Project:

Zhuohao Yang & Yue Wang & Gabrielle Pan

2024-12-01

Group Member: Yue Wang, yuew3@uchicago.edu, Aaronnn0912 Gabrielle Pan, gpan@uchicago.edu, ddbb421 Zhuohao Yang, zhuohao@uchicago.edu, 00ikaros

1. Introduction

This project investigates reviews from Las Vegas, aiming to understand the sentiment trends, thematic content, and geographic distribution of public opinion. The analysis combines natural language processing (NLP) techniques, sentiment analysis, and geospatial visualization to derive actionable insights.

2. Research Question

How do public sentiments and thematic trends in reviews evolve over time, and what are their geographic patterns in Las Vegas? The analysis seeks to answer: What are the prevailing themes in reviews? How do sentiments fluctuate temporally? How are sentiments distributed geographically?

Data Cleanning

Structured information was extracted from review and location tag datasets using Python libraries, including: • pandas: For data manipulation and structuring. • re: For regular expression-based text extraction. Incomplete, non-English, and improperly formatted records were removed to ensure data quality. Non-standard formats in text and timestamps were standardized to create a clean dataset for further processing.

Import Dataset

Label Dataset Filtering

Data Preprocessing

Natural Language Processing (NLP) techniques were employed to preprocess the text: • langdetect: Detected and filtered out non-English reviews. • NLTK: Removed noise such as stopwords, tokenized the text, and standardized the format through lemmatization using WordNetLemmatizer. These steps ensured that the data was normalized and ready for sentiment analysis and topic modeling.

Stopwords

Setup and Train the BERTopic Modeling

The BERTopic model was applied to uncover key themes in the reviews: • Embedding Model: all-mpnet-base-v2, a Transformer-based pre-trained model, was used to generate semantic embeddings for the reviews. • Clustering: HDBSCAN was employed to group reviews into meaningful topics. • Temporal Analysis: pandas was used to analyze how topic distributions changed over time.

Trainning

Topics

Retrieve words for each topic

Cluster Similarity Cosine

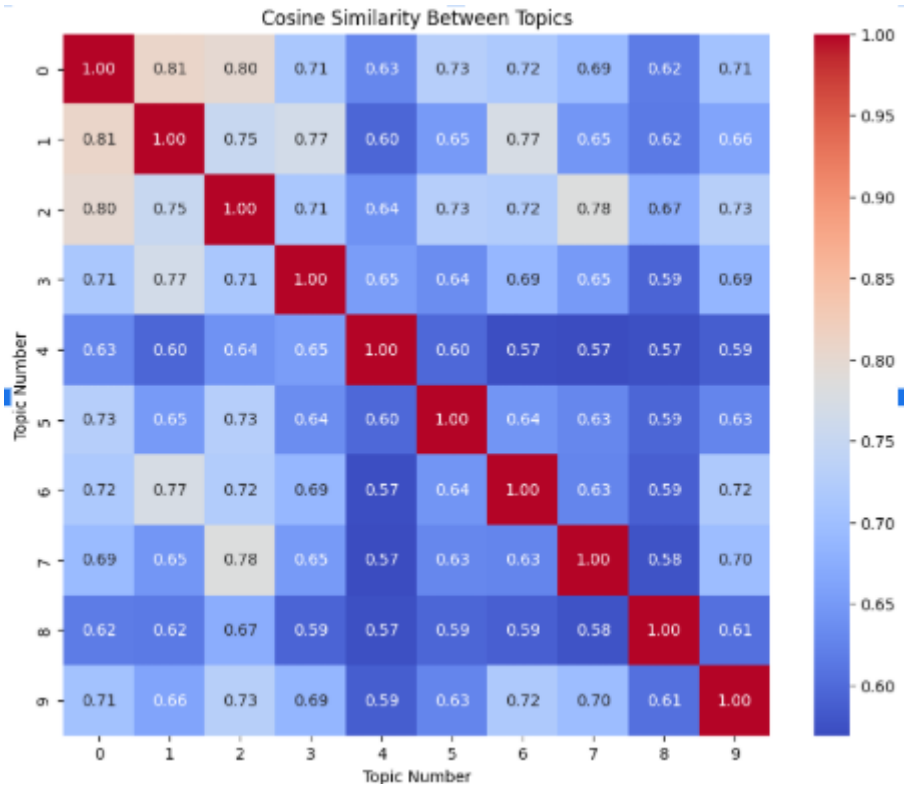


Figure 1: cosine similarity

This heatmap shows the cosine similarity between topics from the BERTopic model, with values closer to 1 indicating stronger thematic overlap. For instance, Topic 0 and Topic 1 (similarity 0.81) suggest shared themes, while lower scores (e.g., Topic 4 and Topic 7 at 0.57) indicate distinct topics. This analysis supports content categorization by grouping similar topics, uncovers thematic connections like shared emotional patterns between Topic 5 and Topic 0, and informs business strategies, such as aligning marketing for related topics like dining (Topic 2) and professional services (Topic 1).

Sentiment Analysis

The VADER sentiment analysis tool quantified user emotions in the reviews: • Sentiment scores ranging from -1 (negative) to 1 (positive) were assigned to each review. • These scores were used to identify temporal trends in sentiment, providing insights into users' emotional responses over time.

Assign the Sentiment Score

- Since the performance of TextBlob was too bad, therefore, we chose VADER as hour sentiment analysis library.

Data Visualization

Altair: Created daily and monthly sentiment trend charts to visualize temporal variations in user emotions. GeoPandas & Matplotlib: Mapped the spatial distribution of sentiment, highlighting areas with positive and negative user feedback. Polynomial Regression: Modeled the nonlinear temporal trends in topic probabilities, enabling long-term pattern predictions.

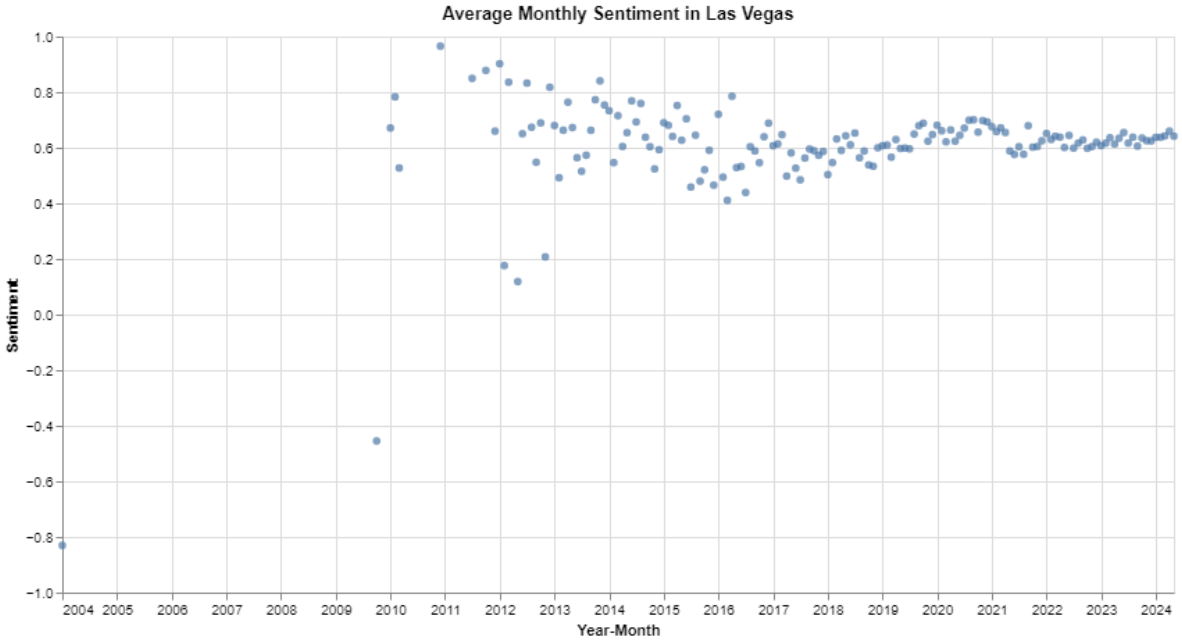


Figure 2: Monthly Chart

Combine with the Topics Probabilities

Polynomial Multi-Regression

Prepare Topic Probabilities Data

Perform Polynominal Regression for Each Topic

Visualize Topic Probability Trends

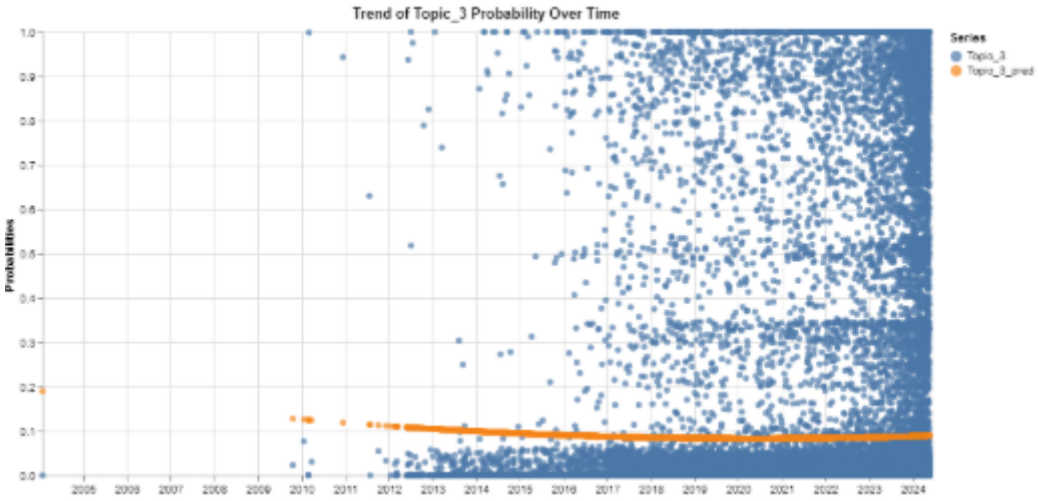


Figure 3: topic_3 Probabilities Over Time

The plots display the daily average sentiment over time and monthly, showing significant variability with a wide range of sentiment scores. This broader spread reflects the influence of specific events or daily visitor experiences, causing fluctuations in sentiment. Despite the variability, most scores cluster around positive sentiment, suggesting that, overall, visitor experiences tended to lean positive even on days with notable fluctuations.

Lon & Lat Geopandas

Prepare Data with Lat/Lon Coordinates

Visualize Sentiment Scores on a Map with Folium

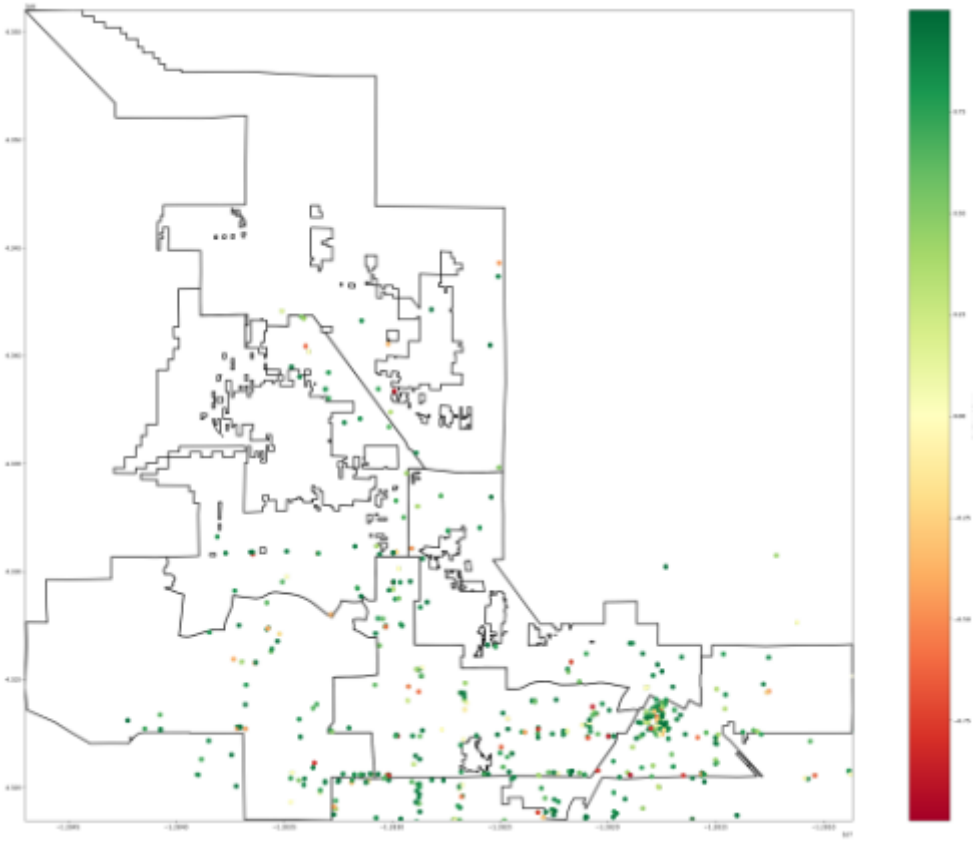
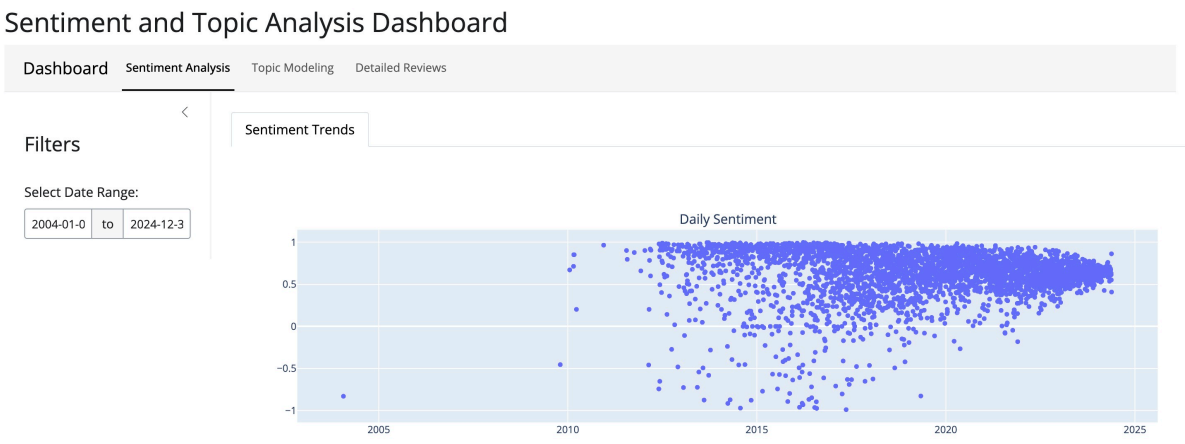


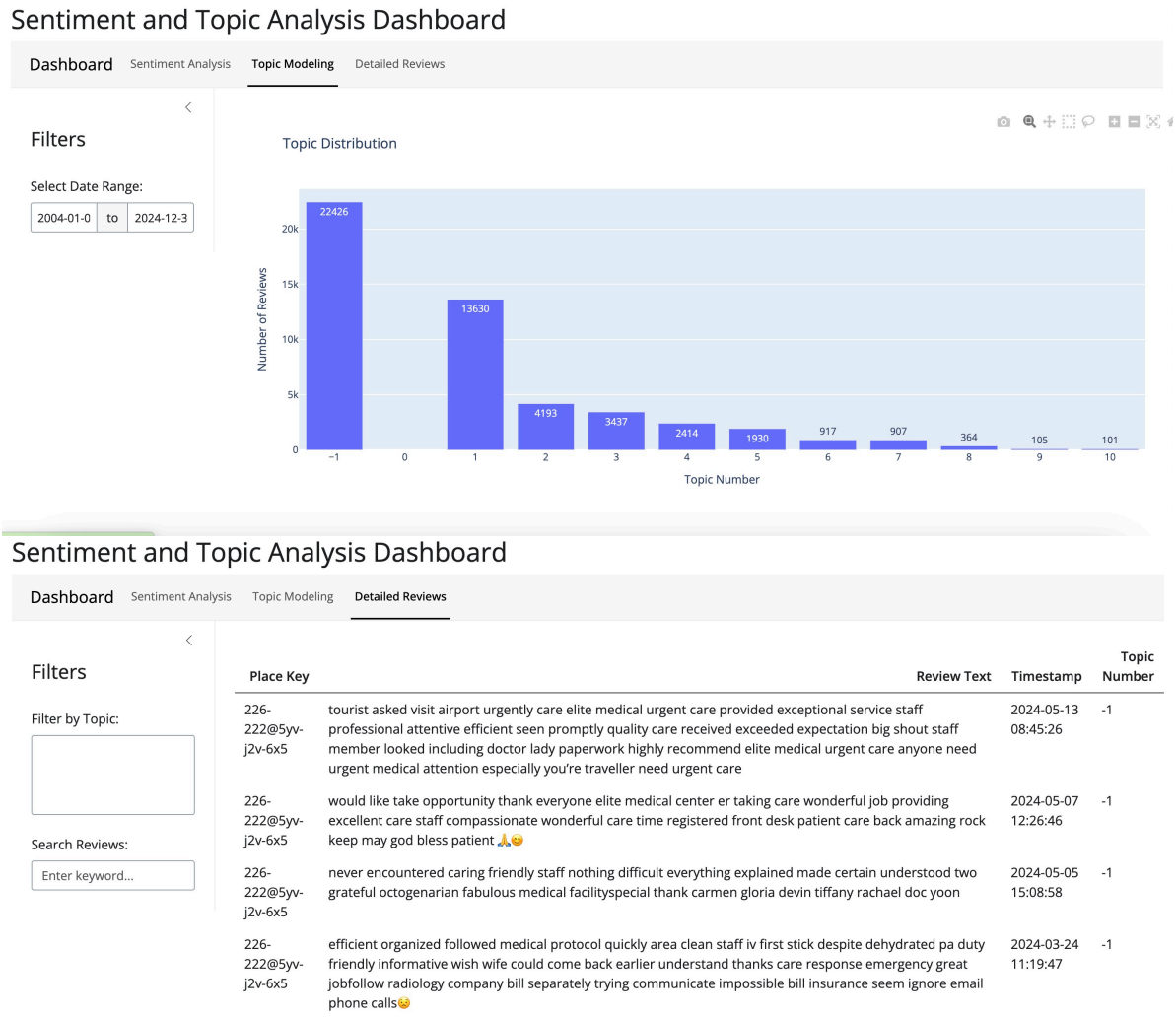
Figure 4: Geo

This map visualizes the spatial distribution of sentiment scores across different areas of Las Vegas. Areas with higher densities of points likely correspond to hubs of activity, such as popular business centers or recreational zones. Green points represent positive sentiment, while red points indicate negative sentiment. Notably, clusters of red points are observed in certain high-density areas, suggesting potential issues such as noise or overcrowding that might negatively impact visitor experiences in these regions. These findings highlight the importance of addressing localized challenges in high-traffic areas to improve overall user satisfaction. The map underscores the need for spatially targeted interventions to balance activity levels and enhance sentiment in less positively perceived areas.

4. Difficulties Encountered VADER's Limitations: • VADER struggled with complex reviews, such as those containing sarcasm or implicit sentiment. For example, reviews like "Not bad at all" could be misclassified, reducing the accuracy of sentiment scores. Geographical Data Issues: • Some reviews lacked latitude and longitude data, limiting the scope of geospatial analysis. • Inconsistent coordinate systems between datasets required extra cleaning and standardization, reducing the coverage of spatial sentiment maps. Computational Constraints: • BERTopic's reliance on Transformer-based embeddings (e.g., all-mpnet-base-v2) was computationally intensive, especially for large datasets. • Polynomial regression for multiple topics and high-order terms further increased computational demands, slowing down the modeling process.

5. Shiny App The app just stored the sentiments over time, topics distribution and the specific reviews that have been assigned to existing topics. From the sidebar, you can choose the time range, selected topics and specific texts.





6. Policy Implications This analysis provides valuable insights for policymakers and businesses: Tourism Development: Positive sentiment trends highlight successful attractions, while negative areas signal improvement opportunities. Public Service Optimization: Identifying geographic sentiment disparities can guide resource allocation for parks and other amenities. Event Impact Assessment: Monitoring sentiment spikes around events helps evaluate their success and public reception.

7. Conclusion This project showcases a comprehensive framework for analyzing reviews through NLP, sentiment scoring, and geospatial methods. By uncovering themes, temporal patterns, and geographic insights, the findings contribute to improving public services and business strategies in Las Vegas.