Project 2:
Hospitality Sentiment Analysis of Vietnamese Cities Using Reddit Headlines

Team members: Nguyen Tran Chung Vu & Tran Tue Nhi

1. Project Background

a. Overview

Understanding public sentiment about various cities can provide critical insights for local governments, marketers, and policymakers. This project aims to analyze and visualize sentiments expressed in headlines from Reddit posts about different cities in Vietnam. Utilizing the Reddit API, we will extract relevant headlines and apply sentiment analysis using the VADER tool, presenting our findings through an interactive choropleth map integrated into a dashboard.

b. Problem Definition

The primary problem we aim to solve is understanding regional sentiment variations across Vietnam. Public sentiment, especially as expressed on social media platforms, can provide real-time insights into the perceptions and concerns of residents in different cities. Traditional methods of gauging public sentiment, such as letters, calls and polls, are time-consuming, expensive, and often fail to capture the dynamic nature of public opinion [1].

Our project addresses the following key issues:

- **Real-time Sentiment Analysis:** Traditional surveys do not provide real-time insights. Our solution leverages social media data to offer current sentiment analysis
- **Geographic Sentiment Mapping:** Existing sentiment analysis studies often lack a geographic dimension, making it difficult to understand regional variations
- **Dynamic Exploration:** Conventional sentiment analysis reports are static. Our interactive dashboard allows users to explore sentiment data dynamically, adjusting time frames and filters to gain deeper insights.

c. Novelty of Solution

Our solution stands out due to the following innovative aspects:

- **Integration of Sentiment Analysis with Geographic Data Visualization:** Our project provides an intuitive way to understand regional sentiments by combining sentiment analysis with geographic data visualisation. Similar studies often focus on either sentiment analysis or geographic visualization, but rarely both. This integration allows for a more comprehensive analysis for city planners and socio-economic developers [2].
- **Use of Reddit API for Data Collection:** Reddit, with its diverse and active user base, offers a rich real-time data source. Many existing sentiment analysis projects rely on Twitter data, which might not capture the depth and variety of discussions found on Reddit [3]. By using Reddit API, we access a different subset of social media discussions, potentially uncovering unique insights
- **Interactive Dashboard for Dynamic Exploration:** Unlike static reports, our interactive dashboard allows users to explore sentiment data over different time frames and with various filters. This feature enables a more nuanced understanding of how sentiments evolve and vary across different regions and time periods.
- **Application of VADER for Sentiment Analysis:** VADER (Valence Aware Dictionary and Sentiment Reasoner) is specifically optimized for analyzing social media text [4]. Its accuracy in handling the informal and varied language used in Reddit posts enhances the reliability of our sentiment analysis.

2. Justification of Approach

- **Sentiment Analysis:** We chose VADER for sentiment analysis because it is optimized for social media text and provides reliable sentiment scores for short texts, such as headlines. VADER's lexicon and rule-based sentiment analysis make it well-suited for processing the informal language commonly found on Reddit.
- **Visualization Method:** A choropleth map was selected for its effectiveness in displaying geographical data with varying intensities. By using color gradients to represent sentiment scores, we can intuitively convey the emotional landscapes of different cities. Integrating this map into an interactive dashboard allows users to explore sentiment trends over time, adding depth to our analysis.

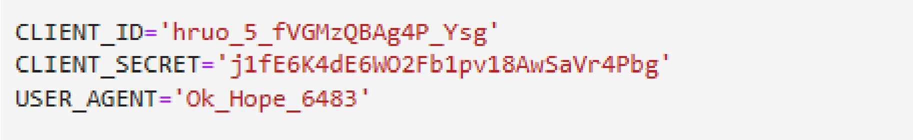
3. Final Product

a. Dataset Processing

In this project, the dataset consists of headlines extracted from Reddit posts about tourism in various cities in Vietnam. The main reason that led us choosing Reddit as the primary data source is because Reddit is a well-known platform that provides real-time feedback as a form of discussions of its users. Moreover, Reddit hosts a wide variety of communities (which are subreddits or subnames) that discuss numerous topics, including local news and events. Last but not least, Reddit's API allows for efficient data extraction, enabling the collection of large volumes of data needed for comprehensive sentiment analysis, while the API service is free-to-use to their users (with a limit of ~1000 headlines/day).

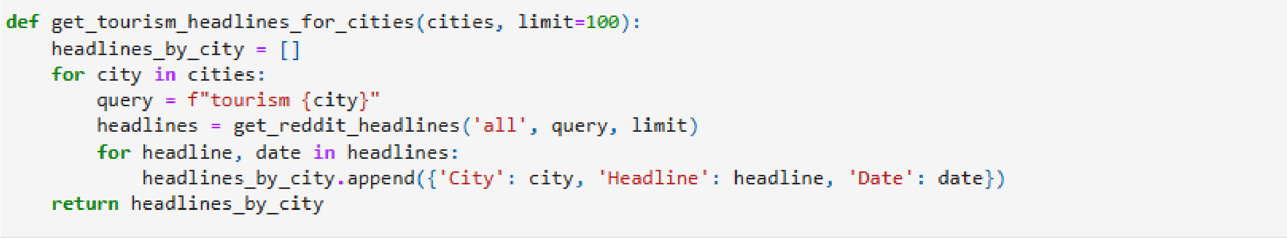
To prepare the dataset for our analysis, the following steps were undertaken:

- Token preparation: In order to use the Reddit API, we were asked to sign up an account on Reddit and then generate a personal token so that the API can access it through that key.

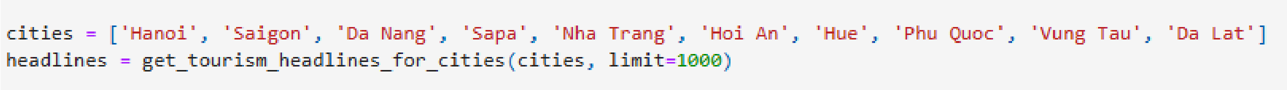


```
CLIENT_ID='hruo_5_fVGMzQBAg4P_Ysg'
CLIENT_SECRET='j1fE6K4dE6WO2Fb1pv18AwSaVr4Pbg'
USER_AGENT='Ok_Hope_6483'
```

- Data extraction: We specify the major cities that tourists normally tend to be interested in and let the API extract the headlines with the subreddit equivalent to the city's name.

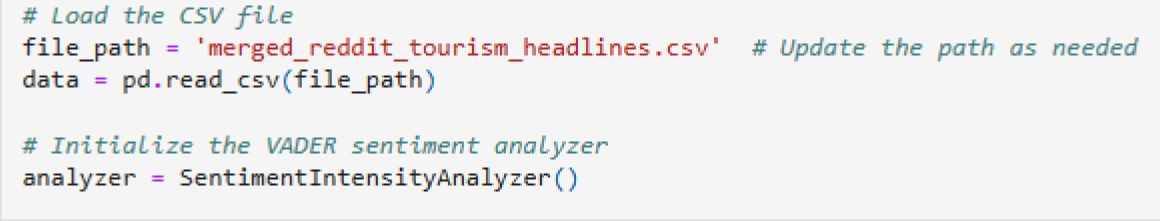


```
def get_tourism_headlines_for_cities(cities, limit=100):
    headlines_by_city = []
    for city in cities:
        query = f"tourism {city}"
        headlines = get_reddit_headlines('all', query, limit)
        for headline, date in headlines:
            headlines_by_city.append({'City': city, 'Headline': headline, 'Date': date})
    return headlines_by_city
```



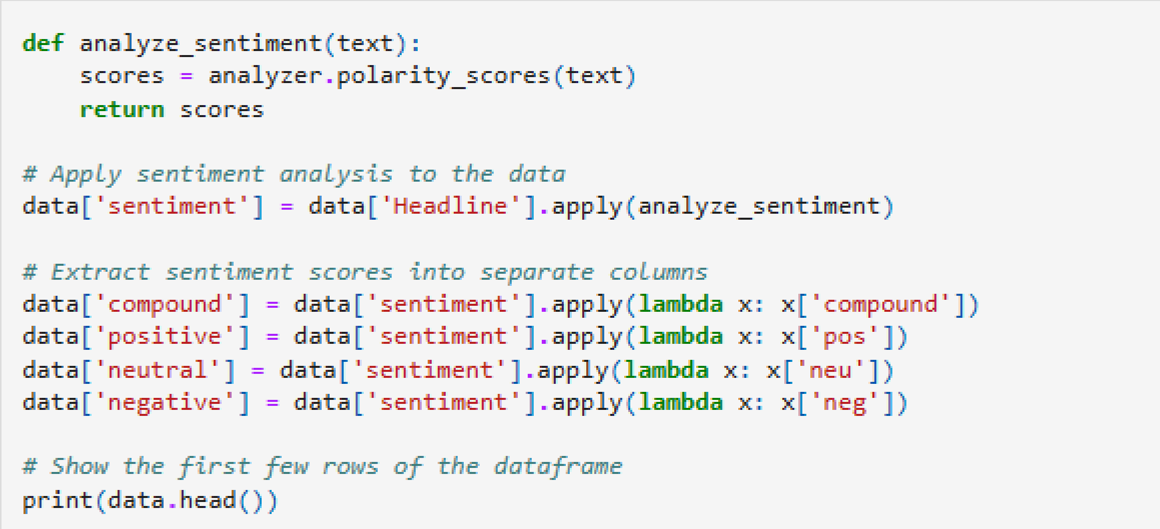
```
cities = ['Hanoi', 'Saigon', 'Da Nang', 'Sapa', 'Nha Trang', 'Hoi An', 'Hue', 'Phu Quoc', 'Vung Tau', 'Da Lat']
headlines = get_tourism_headlines_for_cities(cities, limit=1000)
```

- Sentiment Analysis: The extracted headlines were then processed using the VADER sentiment analysis tool to label. VADER assigns sentiment scores to each headline in 3 different categories which are positive, negative and neutral sentiment.



```
# Load the CSV file
file_path = 'merged_reddit_tourism_headlines.csv'  # Update the path as needed
data = pd.read_csv(file_path)

# Initialize the VADER sentiment analyzer
analyzer = SentimentIntensityAnalyzer()
```



```
def analyze_sentiment(text):
    scores = analyzer.polarity_scores(text)
    return scores

# Apply sentiment analysis to the data
data['sentiment'] = data['Headline'].apply(analyze_sentiment)

# Extract sentiment scores into separate columns
data['compound'] = data['sentiment'].apply(lambda x: x['compound'])
data['positive'] = data['sentiment'].apply(lambda x: x['pos'])
data['neutral'] = data['sentiment'].apply(lambda x: x['neu'])
data['negative'] = data['sentiment'].apply(lambda x: x['neg'])

# Show the first few rows of the dataframe
print(data.head())
```

b. Libraries

The product was developed using several libraries and tools in R:

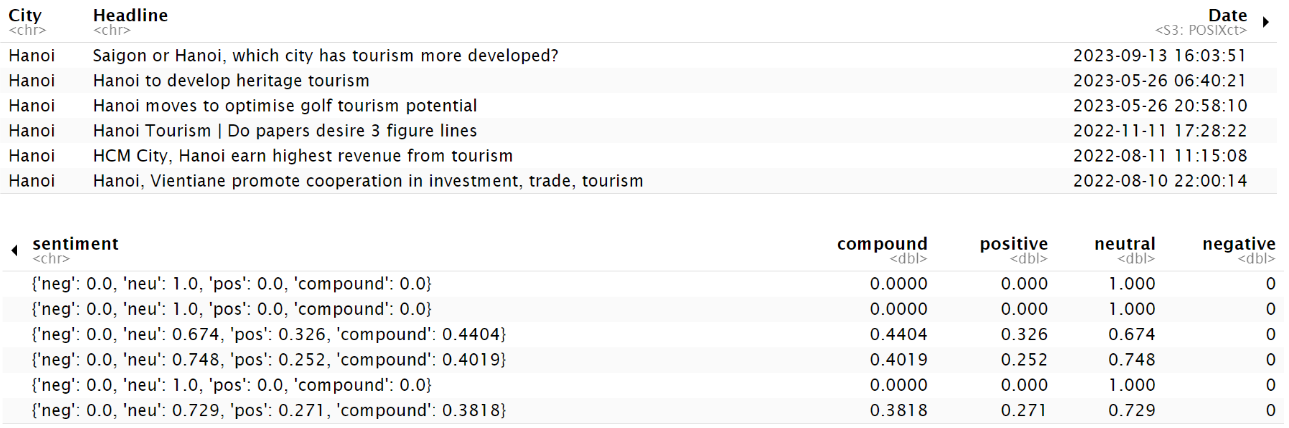
- Data Manipulation: tidyverse, dplyr, readr, lubridate
- Sentiment Analysis: syuzhet, vader, textTinyR, textdata

- Visualization: ggplot2, leaflet, sf, viridis
- Interactive Elements: shiny

These libraries were chosen for their robustness and ease of integration, allowing us to create a comprehensive and interactive tool for sentiment analysis.

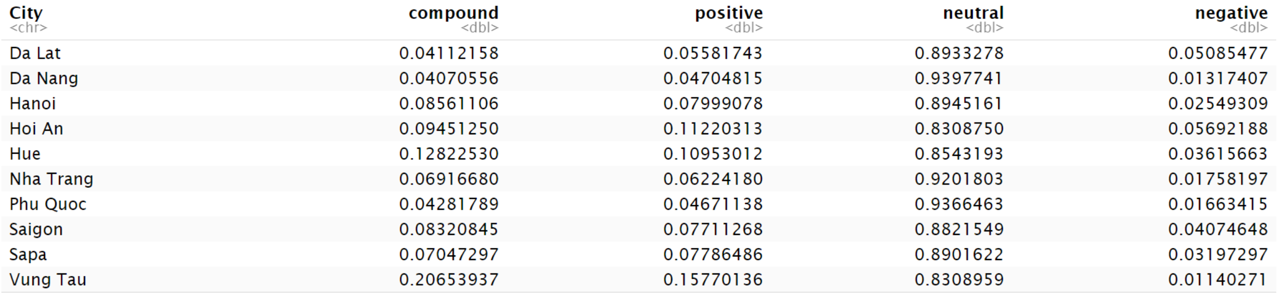
c. Elements of Product

- **Dataset Summary:** The dataset is summarized to provide an overview of its structure and key variables. This includes converting date columns to appropriate formats and aggregating sentiment scores by city.



City <chr>	Headline <chr>	Date <S3: POSIXct>	sentiment <chr>	compound <dbl>	positive <dbl>	neutral <dbl>	negative <dbl>
Hanoi	Saigon or Hanoi, which city has tourism more developed?	2023-09-13 16:03:51	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0.0000	0.000	1.000	0
Hanoi	Hanoi to develop heritage tourism	2023-05-26 06:40:21	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0.0000	0.000	1.000	0
Hanoi	Hanoi moves to optimise golf tourism potential	2023-05-26 20:58:10	{'neg': 0.0, 'neu': 0.674, 'pos': 0.326, 'compound': 0.4404}	0.4404	0.326	0.674	0
Hanoi	Hanoi Tourism | Do papers desire 3 figure lines	2022-11-11 17:28:22	{'neg': 0.0, 'neu': 0.748, 'pos': 0.252, 'compound': 0.4019}	0.4019	0.252	0.748	0
Hanoi	HCM City, Hanoi earn highest revenue from tourism	2022-08-11 11:15:08	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}	0.0000	0.000	1.000	0
Hanoi	Hanoi, Vientiane promote cooperation in investment, trade, tourism	2022-08-10 22:00:14	{'neg': 0.0, 'neu': 0.729, 'pos': 0.271, 'compound': 0.3818}	0.3818	0.271	0.729	0

Table 1: Summary of Dataset



City <chr>	compound <dbl>	positive <dbl>	neutral <dbl>	negative <dbl>
Da Lat	0.04112158	0.05581743	0.8933278	0.05085477
Da Nang	0.04070556	0.04704815	0.9397741	0.01317407
Hanoi	0.08561106	0.07999078	0.8945161	0.02549309
Hoi An	0.09451250	0.11220313	0.8308750	0.05692188
Hue	0.12822530	0.10953012	0.8543193	0.03615663
Nha Trang	0.06916680	0.06224180	0.9201803	0.01758197
Phu Quoc	0.04281789	0.04671138	0.9366463	0.01663415
Saigon	0.08320845	0.07711268	0.8821549	0.04074648
Sapa	0.07047297	0.07786486	0.8901622	0.03197297
Vung Tau	0.20653937	0.15770136	0.8308959	0.01140271

Table 2: Aggregated Sentiment Scores

- **Sentiment Distribution Bar Plot:** A bar plot is created to represent the distribution of sentiment scores (positive, negative, neutral) across all cities. This visualization helps users quickly grasp the overall sentiment landscape.

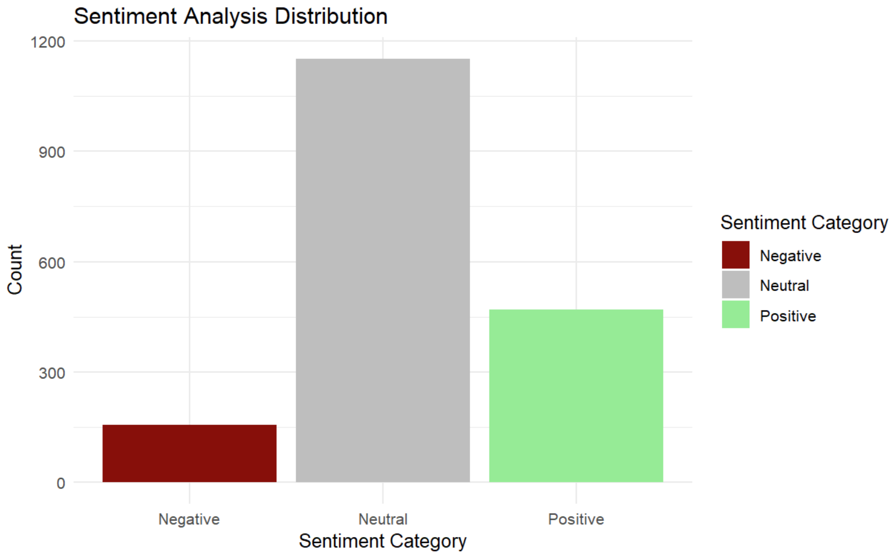


Figure 3: Sentiment Analysis Distribution

- **Sentiment Distribution Across Cities Bar Plot:** A bar plot representing the distribution of sentiment scores in different cities, allowing for a comparative analysis of sentiment across locations.

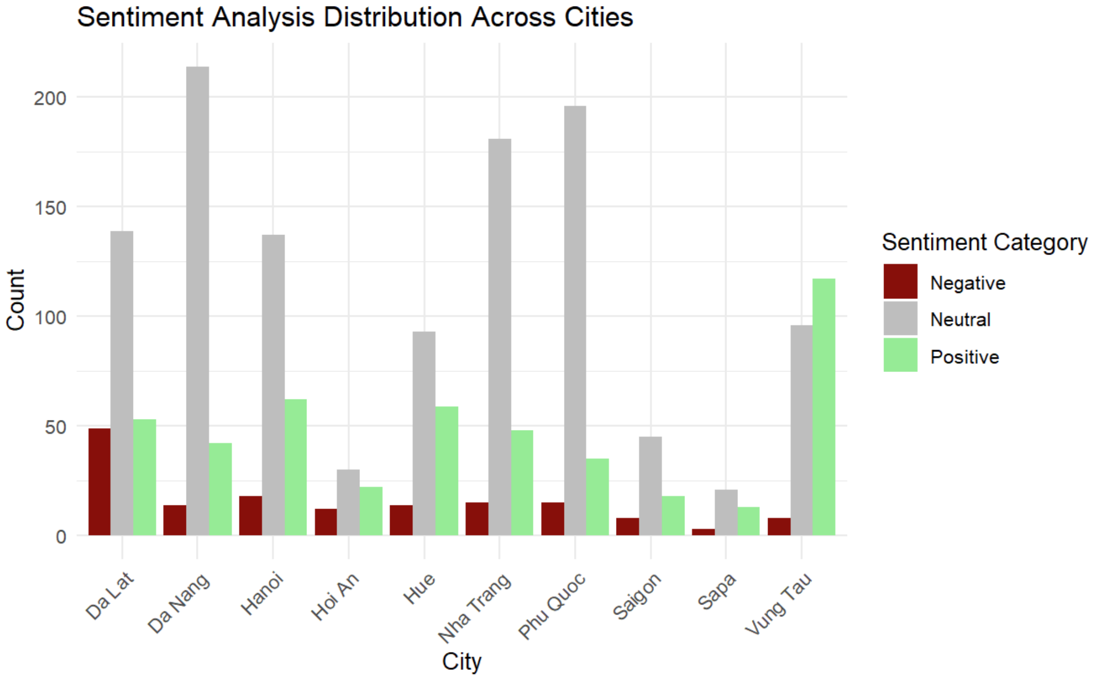


Figure 4: Sentiment Analysis Distribution Across Cities in Vietnam

- **Time Series Plot Comparing Hospitality Sentiment in Different Cities:** A time series plot that represents the counts of each sentiment category over time for various cities, providing insights into temporal sentiment trends.

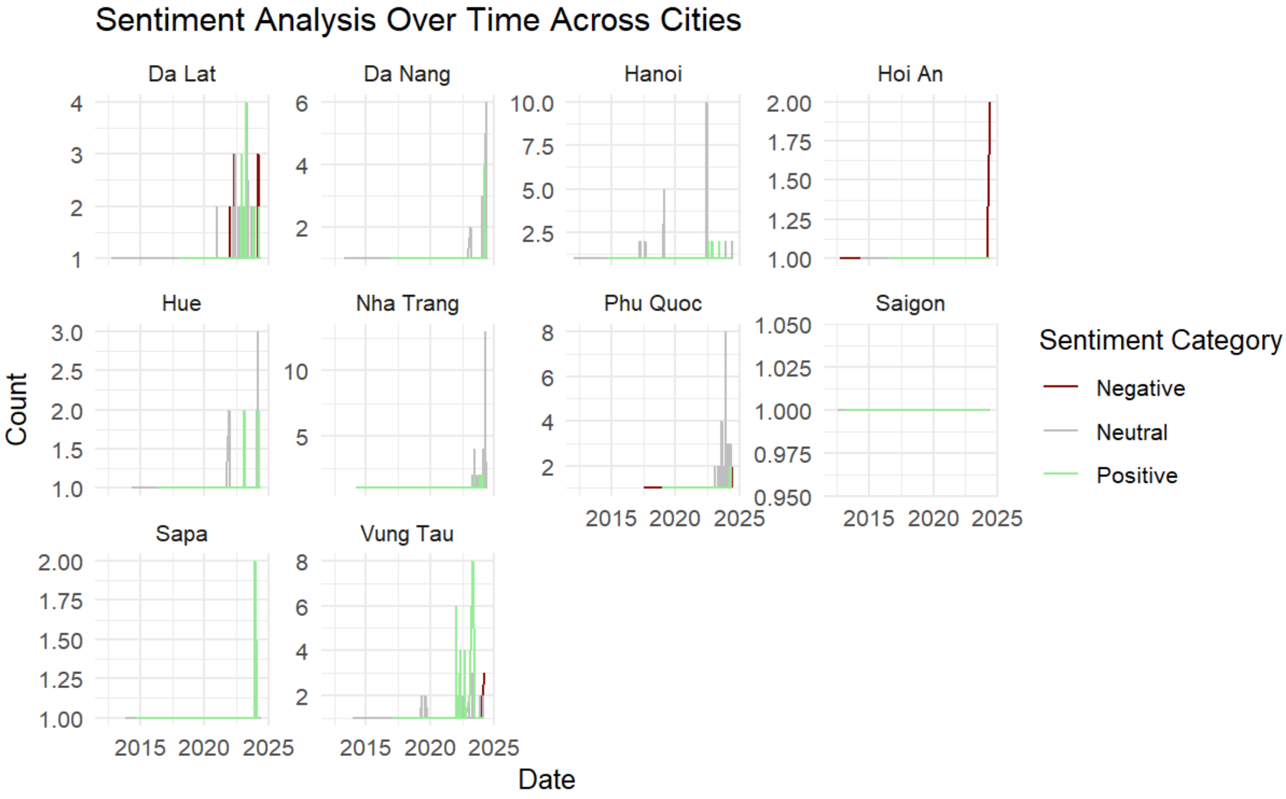


Figure 5: Time Series Plot Comparing Hospitality Sentiment in Different Cities in Vietnam

- **2D Choropleth Map:** A choropleth map using leaflet and sf libraries, displaying average sentiment scores for different cities on a geographical map of Vietnam.

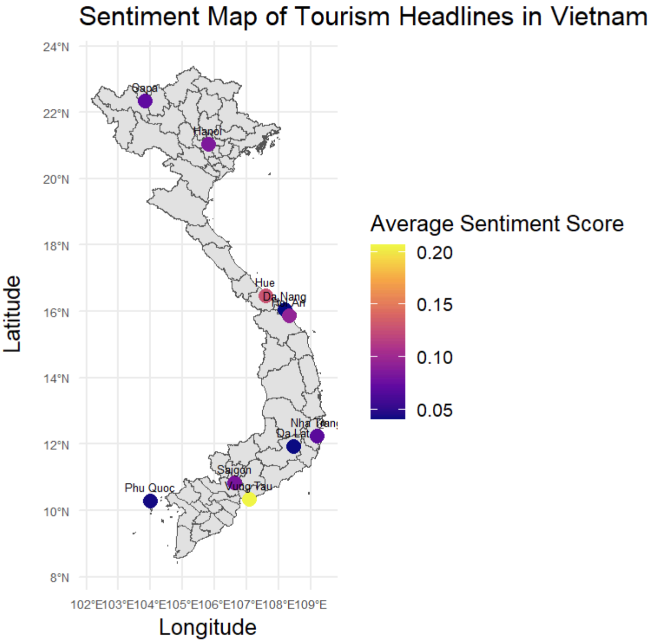
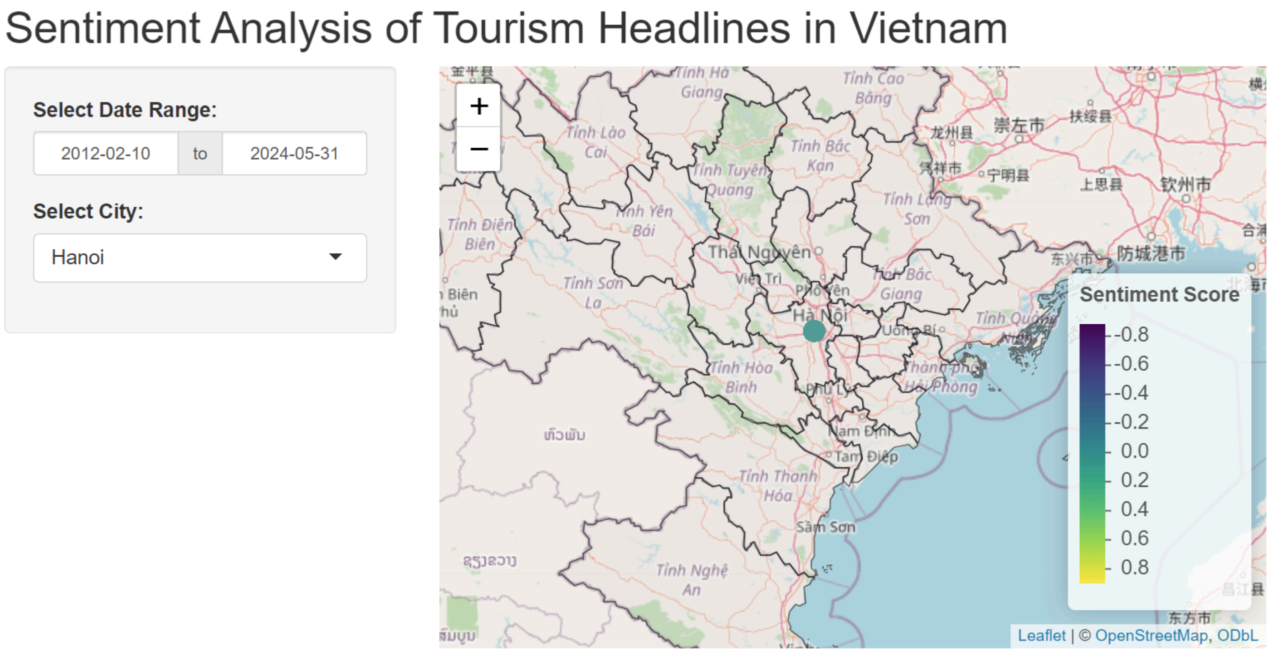


Figure 6: 2D Chropleth Map

- **Advanced 2D Chropleth Map Dashboard:** This enhanced dashboard integrates city sentiment data over time, providing a detailed geographical representation of sentiment trends. Users can select flexible data range along with choosing the city they want to analyze



Sentiment Analysis of Tourism Headlines in Vietnam

Select Date Range:

2012-02-10 to 2024-05-31

Select City:

Hanoi

Figure 7: 3D Choropleth Map Dashboard

4. Discussion

Most sentiments in most cities are neutral, as seen in cities like Hanoi, Da Nang, Nha Trang and Phu Quoc. Positive sentiments are significant in Vung Tau, Sai Gon, Sapa and Hue, while negative sentiments are less frequent across these cities. This indicates a general trend where tourists have neutral or positive experiences, with a relatively small proportion of negative feedback.

Small cities like Vung Tau and Hue generally show more positive sentiments, reflecting their cultural and historical significance. Conversely, famous cities for travelling like Phu Quoc, Da Lat and Hoi An exhibit more negative sentiments, possibly due to a large number of tourists or local issues.

It is noticeable that many cities, such as Da Nang, Vung Tau and Nha Trang, have seen an increase in sentiment counts closer to the present day, reflecting growing tourism discussions or data collection. Positive sentiments appear to be more recent, indicating improving perceptions or experiences in these cities.

However, in some cities like Saigon, Sapa and Hoi An, sentiment data is relatively sparse, suggesting either less tourism coverage or data availability.

5. Limitations

- Data Bias:
  - The dataset is limited to Reddit users, who may not represent the general population.
  - Some cities have more headlines than the other.
- API Restriction:
  - Reddit API rate limits and availability limit data collection.
  - Headlines does not give the whole context of the post.
- Sentiment Analysis Challenges:
  - There exists mixed sentiment in posts
  - It is difficult to interpret neutrals

6. Future Directions

In light of the findings and limitations identified in our analysis of hospitality sentiments across Vietnamese cities using Reddit headlines, several potential future directions can be considered to enhance and expand upon this research:

- **Expand the data source:** to address the data bias inherent in using Reddit as the primary source of data, future research could incorporate data from other social media platforms such as Facebook, Twitter, and Instagram. We can use a Python library that is specific for data crawling such as BeautifulSoup. This would provide a more comprehensive view of public sentiment across different demographics and user bases.
- **Decrease the number of neutral sentiments:** social media posts are usually full of memes and questions that hindered the ability for a more accurate analysis. Hence, decreasing the number of neutral sentiment not only lowers the chance of data bias, but also makes the data more meaningful for further and deeper analyses.
- **Seasonal and Event-based Analysis:** conducting a more granular temporal analysis to identify sentiment trends during specific seasons, holidays will be beneficial for city planners and policy makers.

References

[1] Brooker, R. G., & Schaefer, T. (2015). Methods of measuring public opinion. Public opinion in the 21st century. https://www. uky. edu/AS/PoliSci/Peffley/pdf/473Measuring% 20Public% 20Opinion.
[2] Pino, C., Isaak Kavasidis, & Spampinato, C. (2016). Assessment and visualization of geographically distributed event-related sentiments by mining social networks and news. Consumer Communications and Networking Conference. https://doi.org/10.1109/ccnc.2016.7444806
[3] Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. Social Media + Society, 7(2), 205630512110190. https://doi.org/10.1177/20563051211019004
[4] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550