MDA 720 Project Report

Sentiment Analysis of Hotel Reviews

- Pritika Khurana

Contents

Background

A new luxury hotel is slated to open in New York next year, and its success hinges on providing exceptional amenities and a trouble-free stay for guests. To ensure this outcome, we aim to identify the most popular and highly-regarded hotels in the surrounding area. Leveraging customer feedback available online, we will perform sentiment analysis to determine the polarity (positive, negative, or neutral) of hotel reviews using machine learning models. This approach will enable us to obtain recent, credible, and trustworthy information on hotels based on customers' experiences. Ultimately, the data retrieved will inform a robust opening plan and play a critical role in launching a successful hotel into the marketplace.

Introduction

Customer reviews on hotels are very important part of travel plan for people now a days. People prefer to book such hotels which have high number of positive reviews. There are different sources to find the reviews to get a better insight about the hotel's reputation. Thus it can be said that customer reviews plays an important part for business owners in order to improve their services. The hotel industry is an important sector of the hospitality industry that provides lodging and other services to travelers and tourists.

Objectives/ Goals of the Project

The main objectives of this project:

i. Create a sentiment analysis model able to correctly classify customer ratings as positive, negative, or neutral.
ii. Analyze customer reviews and ratings of hotels in a specific region, and identify key factors that impact customer satisfaction and loyalty.
iii. Identify the most common problems and areas for improvement
iv. Based on the sentiment analysis results, provide suggestions for improving service quality at hotels that fit to the particular needs and preferences of customers.

This report will present the methodology, findings, and conclusions of the analysis, and provide recommendations for improving hotel performance and customer satisfaction.

Methodology

The data used in this analysis was collected from customer reviews and ratings of hotels in a specific region, and was preprocessed and cleaned using standard methods such as tokenization, stemming, and stop-word removal. The features were selected based on word frequency and importance, and machine learning models such as Support Vector Machine (SVM) and Naive Bayes (NB) model was trained and evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

This project aims to conduct web scraping and sentiment analysis on user reviews.

Data Collection

In this project, I have taken data of multiple hotels through a predefined dataset and by webscrapping. First we will see the predefined dataset.

Data Exploration

For this project, the dataset is taken from kaggle. The data is originally fetched from "Datafiniti's Business Database". Dataset initially contained 10000 rows and 26 columns. In this dataset, each row contains all the information related to a hotel, hotel's review, rating as well as reviewer's information. There are many columns which are irrelevant for this project such as hotel's address, country, province, postal code etc, as well as reviewer's name, province, source of review (url), review date added and seen etc. Thus for this project I have taken only the most relevant columns into account though some other columns can be used for making different sort of analysis such as classifying best hotels in each city. Since in this project the target is to perform simple sentiment analysis. Only 6 columns are kept.

	categories	city	name	reviews.rating	reviews.text	reviews.userCity
0	Hotels,Hotels and motels,Hotel and motel mgmt....	Goleta	Best Western Plus South Coast Inn	3	This hotel was nice and quiet. Did not know, t...	San Jose
1	Hotels,Lodging,Hotel	Carmel by the Sea	Best Western Carmel's Town House Lodge	4	We stayed in the king suite with the separatio...	San Francisco
2	Hotels,Lodging,Hotel	Carmel by the Sea	Best Western Carmel's Town House Lodge	3	Parking was horrible, somebody ran into my ren...	Prescott Valley
3	Hotels,Lodging,Hotel	Carmel by the Sea	Best Western Carmel's Town House Lodge	5	Not cheap but excellent location. Price is som...	Guaynabo
4	Hotels,Lodging,Hotel	Carmel by the Sea	Best Western Carmel's Town House Lodge	2	If you get the room that they advertised on th...	Reno
...	...	...	...	...	...	...
9995	Hotels,Hotels and motels,Corporate Lodging,New...	Hampton	Hampton Inn Hampton-newport News	4	My friends and I took a trip to Hampton for th...	Wallingford
9996	Hotels,Hotels and motels,Corporate Lodging,New...	Hampton	Hampton Inn Hampton-newport News	5	from check in to departure, staff is friendly,...	Homer
9997	Hotels,Hotels and motels,Corporate Lodging,New...	Hampton	Hampton Inn Hampton-newport News	5	This Hampton is located on a quiet street acro...	Conway
9998	Hotels,Bar,Hotel,Restaurants	Hunter	Roseberry's Inn	5	Awesome wings (my favorite was garlic parmesan...	Hunter
9999	Hotels,Hotels and motels,Corporate Lodging,Lod...	Lindale	Hampton Inn-lindale/tyler	4	Clean facility just off freeway ..... staff fr...	Fort Worth

10000 rows × 6 columns

Figure 1: Data Overview

Data Preprocessing

Data preprocessing is very important for text classification.

Data preprocessing involves following:

Data Cleaning
Step 1: Remove the irrelevant columns
Step 2: Drop the duplicate values
Step 3. Convert the column ‘rating’ to float type
Step 4: Convert the column ‘review’ to lowercase and remove any special characters

Data Analysis

Here, we will bring the concept of Sentiment analysis, also known as "opinion mining" or "emotion AI". It is a technique used to extract and examine users' opinions, sentiments, emotions, and responses on a particular matter. Natural Language Processing (NLP) techniques are frequently employed in text mining to analyze responses and reviews for sentiment analysis purposes. As technology advances and social interactions increase, it is important for any business to consider user reviews because they play a crucial role in delivering excellent customer service. Business owners can utilize customer reviews to identify system flaws that have been brought to light by customers and make necessary improvements. Additionally, customer reviews play a significant role in establishing a company's reputation.

The purpose of this project is to perform sentiment analysis on 3 polarity levels that are positive, negative and neutral using text classification. Text classification is not a process of building a classifier only, it also involves different steps that are required to clean the data and make it useful for the analysis.

The steps for text classification are:

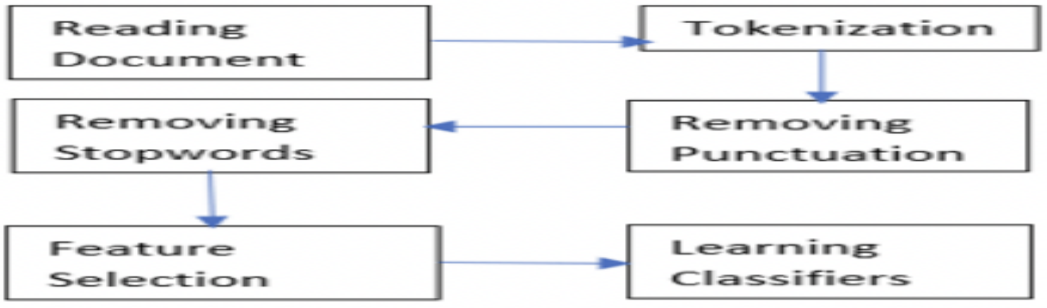


Figure 2: Flow Diagram

Data Visualization

Step 1: Create a new column for review sentiment based on rating
The ratings is converted into class labels-
Positive, if Rating > 3
Neutral, if Rating = 3
Negative, if Rating < 3

The following figure describes the actual distribution of labels in complete dataset:

```
review_sentiment
Negative     1094
Neutral      1169
Positive     7525
dtype: int64
```

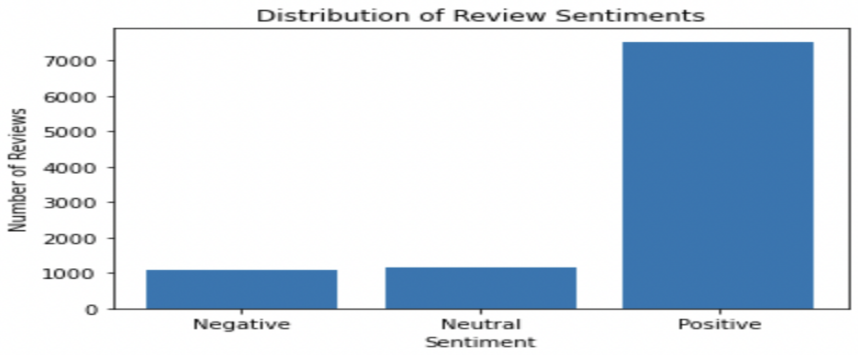


Figure 3: Distribution of Labels

Figure 3 shows that there are more positive reviews, but both neutral and negative reviews are fairly close in count.

```
reviews.rating
1.0      557
2.0      537
3.0     1169
4.0     2805
5.0     4720
dtype: int64
```

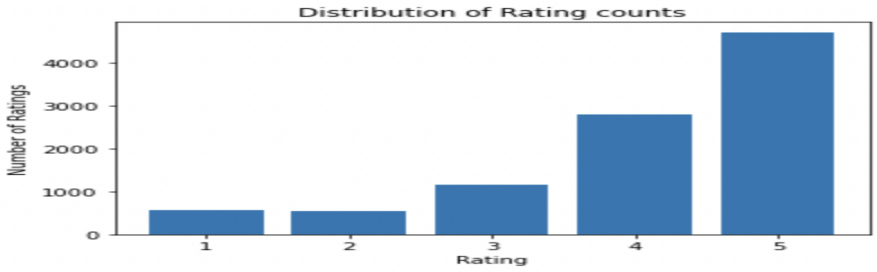


Figure 4: Distribution of Ratings

Figure 4 shows that most of the hotels in the dataset have 5 star rating then 4 star and 3 star. 1 star and 2 star are close in count.

Machine Learning Models

i. Split dataset into train and test examples

Next, we create a CountVectorizer object and fit and transform the training data using the fit_transform() method. This converts the text data into a matrix of token counts, where each row corresponds to a document (review) and each column corresponds to a token (word).

ii. Classification Model

In the case of sentiment analysis, both SVM and NB classifiers can perform well. However, SVMs might be preferred when the dataset is relatively small and high-dimensional, as it can handle these situations well. NB is preferred when the dataset is larger and has many features, such as text data, as it can be more efficient and perform well in such settings.

It is always a good idea to try both and compare their performance using appropriate evaluation metrics before deciding on a final model-

The performance metrics (accuracy, precision, recall, and F1-score) can provide valuable insights into the performance of a hotel sentiment analysis model.

Overall, these performance metrics can help identify areas where the model may need to be improved, such as by adjusting the feature selection or hyperparameters, or by collecting more training data.

```
The Accuracy of the SVM model is: 80.6%
The Precision of the SVM model is: 80.1%
The Recall of the SVM model is: 80.6%
The F1 score of the SVM model is: 80.4%

The Accuracy of the NB model is: 85.3%
The Precision of the NB model is: 83.3%
The Recall of the NB model is: 85.3%
The F1 score of the NB model is: 84.0%
```

Figure 5: Performance Metrics

iii. Explanation of Better Model Performance

The NB model has a higher accuracy, precision, recall, and F1 score than the SVM model. This suggests that the NB model is better at correctly identifying the sentiment of reviews in the dataset, with fewer false positives and false negatives. One reason why the NB model may be performing better is that it assumes that each feature (in this case, each word in the reviews) is independent of all other features, which allows it to be computationally efficient and to perform well with relatively small amounts of training data.

Additionally, the NB model works well with text data, which is often high-dimensional and sparse. On the other hand, the SVM model tries to find the best hyperplane that separates the positive and negative reviews in the dataset, which can be more complex and computationally expensive than the NB model. Additionally, SVMs require more tuning of hyperparameters and may perform poorly if the dataset is unbalanced or noisy.

The NB model has an accuracy of 85.3%, compared to 80.6% for the SVM model. Additionally, the NB model has a slightly higher precision, recall, and F1 score than the SVM model. Based on the evaluation metrics, it appears that the Naive Bayes (NB) model is performing slightly better than the Support Vector Machine (SVM) model.

Testing Example Review

The hotel manager can input hotel reviews and obtain sentiment analysis results. This can help hotel managers better understand the sentiment of their guests towards their services.

```
Enter a statement to predict its sentiment: Terrible Service
The predicted sentiment of the statement is: ['Negative']

 Enter a statement to predict its sentiment: Absoutely loved this place
 The predicted sentiment of the statement is: Positive
```

Figure 6: Example statements to determine sentiment entered by user

Web Scraping

For web scraping I used the Yelp API to search for hotels in Manhattan, New York. [It limits the response to 50 hotels]. In this section I created an empty list to store the hotel data. It loops through the hotels returned in the Yelp API response, extracts the hotel names, ratings, and reviews, and performs sentiment analysis on each review. The sentiment analysis results are classified as either positive, negative or neutral.

Looping through the rows of the data frame, each hotel's name, rating, address and 3 reviews per hotel along with the sentiment is returned-

```
        ->The Wall Street Hotel (5.0 stars):
        Address: 88 Wall St, New York, NY 10005
        Review: I can't say enough about the lovely experience my mother and I had at The Wall Street hotel for her
70th!  Clinton was the kindest and friendliest on staff...
        Sentiment: Positive

        Review: Checked into The Wall Street hotel on a Friday evening around 5:30pm.  I've never stayed in such a
fancy hotel so felt a bit out of place walking in with my...
        Sentiment: Neutral

        Review: This is my new favorite place to stay in NYC! Everything was perfect! We were greeted at the door,
the front desk, the restaurant and the bar in the...
        Sentiment: Positive
```

Figure 7: Output of Data Retrieved from Yelp

Finding the Common, Positive and Negative words in the review column

First, it combines all the reviews into a single string, converts it to lowercase, and removes any punctuation using the translate() method of string objects.

Next, it splits the string into individual words and removes any stop words using the set() function from the stopwords module of the Natural Language Toolkit (NLTK).

Then, it counts the frequency of each remaining word using the Counter() function from the collections module, which creates a dictionary-like object with keys being words and values being their frequency in the text.

Finally, it sorts the words by frequency and prints out the 10 most common words in the reviews along with their respective counts.

```
The 10 most common words in the reviews are:
hotel: 86
location: 31
stay: 30
great: 26
staff: 25
room: 25
stayed: 23
nyc: 19
rooms: 18
night: 16
```

Here WordCloud is used. The purpose of this code is to visualize the most frequently occurring words in the hotel reviews in a visually appealing way. The larger the word in the word cloud, the more frequently it appears in the hotel reviews.

Figure 8: WordCloud of Common words used in Review Column

A predefined set of positive words is used and counts the frequency of each positive word in a list of reviews. The positive_counts dictionary is used to store the

frequency of each positive word, and the sorted_positive_words list is used to sort the positive words by frequency and print them out.

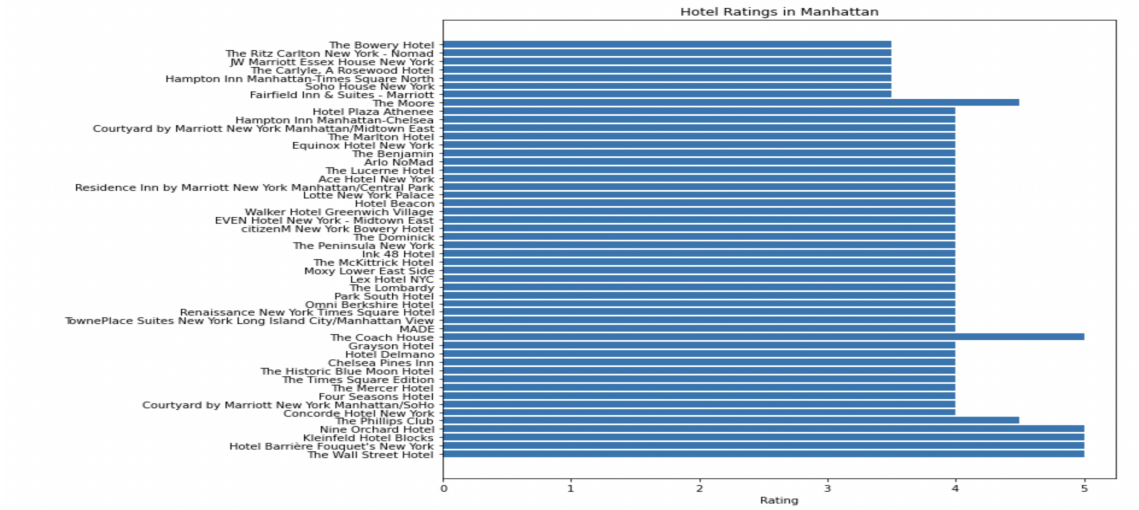
Positive Word	Count
great	26
friendly	13
best	12
good	11
clean	11
helpful	10
nice	9
perfect	8
like	7
comfortable	7
lovely	6
top	6
amazing	6
easy	6
right	6
love	6
loved	6
beautiful	6
incredibly	5
wonderful	5
spacious	4
pretty	4
well	4
modern	4
awesome	4
super	4
cozy	3
work	3
exceptional	3
happy	3
outstanding	3
grand	3
convenient	3
welcome	3
excellent	3
quiet	3
enough	2
favorite	2
chic	2
pleasure	2
comfy	2
cute	2
conveniently	2
award	2
recommend	2
gem	2
pleasant	2
warm	2
trendy	2
decent	2
fun	2
enjoyed	2
nicer	2
seamless	2
worked	2
worth	2
hot	2
upgraded	2
pros	2
fancy	1
snazzy	1
leads	1
sumptuous	1
transparent	1
comprehensive	1
gorgeous	1
intimate	1
finest	1
fantastic	1
pleased	1
loyalty	1
appreciated	1
attentive	1
luxury	1
charming	1
fresh	1
polite	1
free	1
divine	1
warmer	1
romantic	1
beautifully	1
knowledgeable	1
renaissance	1
peace	1
helping	1
terrific	1
nicely	1
seasoned	1
immaculate	1
sexy	1
works	1
positive	1
comfort	1
charm	1
better	1
neatly	1
thank	1
pluses	1
liked	1
available	1
luck	1
fine	1
breeze	1
favorable	1
wow	1
excited	1
eager	1
efficient	1
benefits	1
benefit	1
ready	1
elegant	1
interesting	1
exceedingly	1
accessible	1
glorious	1
helped	1
delicious	1
nicest	1

Similarly, a set of negative words is defined and only the negative words are extracted from the reviews_words list. The negative_words_in_reviews list is then printed out.

Negative Word	Count
worst	2
hate	2
bad	2
dark	1
expensive	1
doubt	1
perplexed	1
lack	1
inability	1
issues	1
drawback	1
hesitant	1
shameful	1
patronize	1
bugs	1
slow	1
spilling	1
hard	1
unfortunately	1
impatience	1
tired	1
fraudulent	1
smack	1
complained	1
complaint	1
disappointed	1
overwhelmed	1
hells	1
bother	1
joke	1
lost	1
crazy	1
sink	1
clogged	1
messed	1
fault	1
bunk	1
hated	1
sadly	1
terrible	1
mistaken	1
difficulty	1
problem	1
break	1
uncomfortable	1
failed	1
losing	1
urgent	1
ridicule	1
disrespect	1

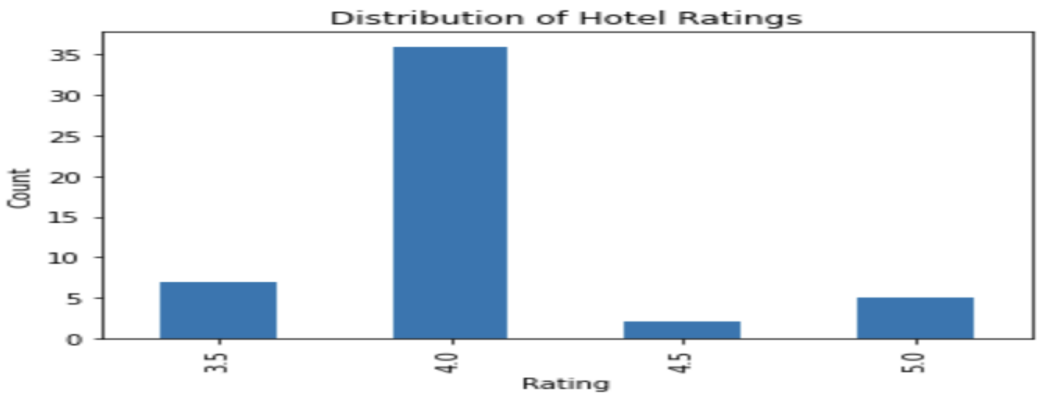
Relationship between Hotels and their rating

This chart is a visualization of the ratings of different hotels in Manhattan. It provides a quick and easy way to compare the ratings of different hotels at a glance. The horizontal bar chart is used because it allows the viewer to easily see the ratings of each hotel and compare them to each other. The chart can be useful for people who are looking for hotels in Manhattan and want to see which ones have the highest ratings. It can also be useful for hotel owners or managers who want to compare the ratings of their hotel to those of their competitors.



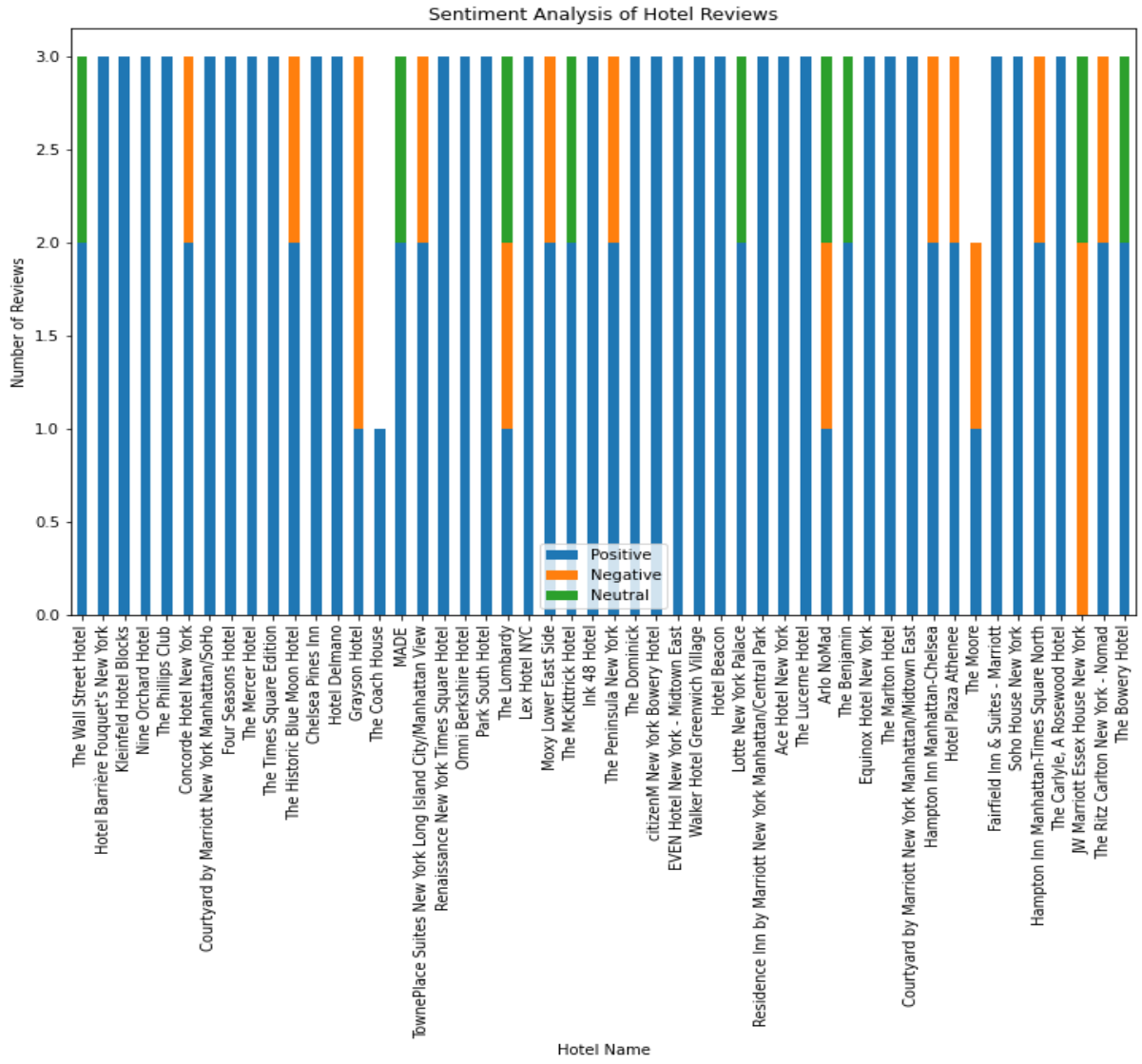
Distribution of hotel ratings in the dataset

The chart shows the count of hotels that received each rating, which gives an overview of how the ratings are distributed among the hotels. This type of chart is useful for understanding the distribution of a variable and identifying any patterns or anomalies in the data. In this case, it can help identify if there are any rating categories that are overrepresented or underrepresented in the dataset.



Sentiment Analysis of Hotel Reviews

The purpose of this chart is to provide a visual representation of the sentiment analysis results and to compare the sentiment distribution of different hotels. It can help identify which hotels have the most positive or negative reviews and provide insights into customer satisfaction levels.



Geographic distribution of hotels in Manhattan

Visualizing the geographic distribution of hotels in Manhattan using the latitude and longitude information provided by the Yelp API. It suggests using the folium library in Python to create interactive maps.

The code creates a map object centered on Manhattan and adds markers for each hotel, displaying the hotel name, rating, and address when clicked. The purpose of this code is to visually inspect the geographic distribution of hotels in Manhattan and identify any areas with a high concentration of hotels.

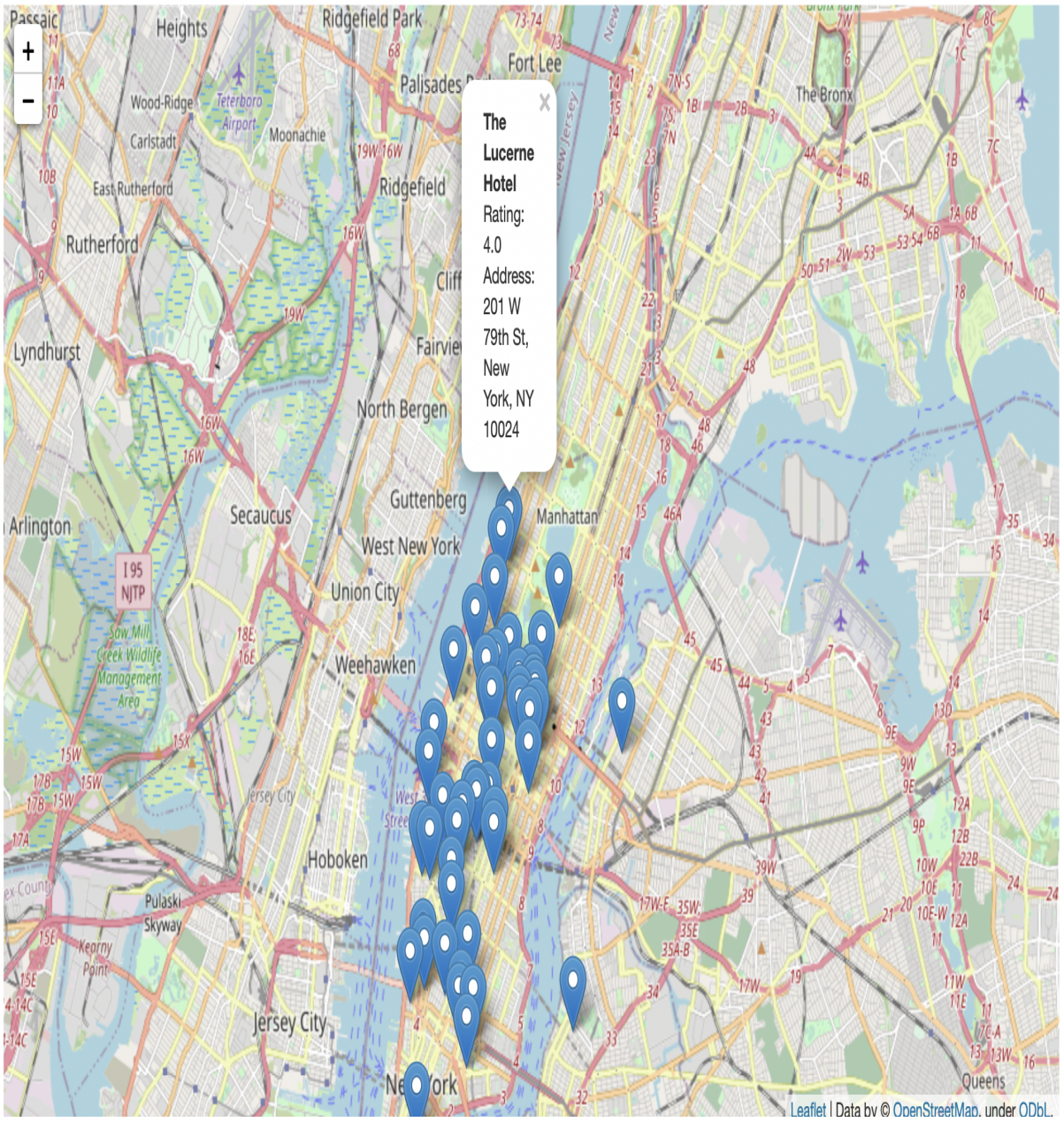


Figure 9: Map of hotels in Manhattan

Results

The main findings of this analysis indicate that the most important factors impacting customer satisfaction and loyalty are the quality of service, cleanliness, location, and price.

Positive reviews tend to emphasize the friendliness and professionalism of the staff, the cleanliness and comfort of the rooms, the convenience and accessibility of the location, and the value for money of the price.

Negative reviews tend to criticize the poor quality of service, the lack of cleanliness and maintenance, the inconvenience and discomfort of the location, and the high cost or poor value for money of the price.

Discussion

The limitations of the data and methods used in this analysis include the potential bias or subjectivity of customer reviews and ratings, the lack of control over external factors such as weather or events.

Conclusion

The main contributions of this project include the identification of key factors that impact customer satisfaction and loyalty in the hotel industry, and the development of **NB** and **SVM model** that can predict customer sentiment based on textual reviews and ratings. The main conclusions of this project are that hotels should focus on improving service quality, cleanliness, location, and price in order to enhance customer satisfaction and loyalty, and that future research should explore more advanced methods and data sources for analyzing customer sentiment and behavior.

References

https://www.youtube.com/watch?v=3MUK1LJYG_4

https://www.youtube.com/watch?v=HMkckLiHOio&t=306s

https://www.youtube.com/watch?v=i3dnv390Sms&t=514s

https://www.youtube.com/watch?v=HpKMc78OYts

https://www.researchgate.net/publication/351262735_Sentiment_Analysis_of_Hotel_Reviews_-_Performance_Evaluation_of_Machine_Learning_Algorithms