

Hotel Reviews Sentiment Analysis

Minwoo Choi

University of Virginia

BS Data Science



Project Introduction

Why hotel reviews?

My older brother is planning a trip to Europe to explore PhD opportunities. Recognizing his intense interview preparations, I wished to help him find a stay at the best hotel. However, I noticed the reviews were often contradictory, mixed with positive and negative comments. I wanted to use data to cut through the noise and provide him with reliable, evidence-based recommendations.

Why this dataset?

The dataset has over 500,000 reviews, meaning it is advantageous in generating statistically significant insights. It also consists of textual data, both positive and negative reviews, along with structured numerical insights like reviewer scores and word counts.

Questions I hope to answer

1. How do customer sentiments align with numerical reviewer scores?
2. Can we quantify guest satisfaction into a simple score?
3. What factors make a hotel highly rated versus poorly rated?



Project Contents



01

Data Cleaning

02

**Exploratory
Data Analysis (EDA)**

03

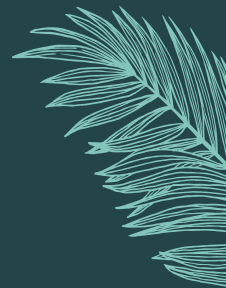
Sentiment Analysis

04

Validation

05

Limitations



Dataset Overview

This dataset was imported from [Kaggle](#) and owned by [Booking.com](#). It consists of 515,000 reviews for 1,493 different hotels across Europe.

* The columns displayed on the right are the relevant one for this project; the rest were discarded during the cleaning process

Column
Hotel_Name
Average_Score
Total_Number_of_Reviews
Reviewer_Score
Positive_Review
Review_Total_Positive_ Word_Counts
Negative_Review
Review_Total_Negative_ Word_Counts



Data Cleaning and Preparation

- Removed irrelevant columns and dropped rows with missing or duplicated values

```
columns_to_drop = ['Hotel_Address', 'Additional_Number_of_Scoring', 'Total_Number_of_Reviews_Reviewer_Has_Given',  
                  'days_since_review', 'lat', 'lng', 'Tags', 'Review_Date', 'Reviewer_Nationality']  
data = data.drop(columns=columns_to_drop).dropna().drop_duplicates()
```


- Applied cleaning functions to remove extra whitespace, special characters, and converted all text to lowercase for consistency

```
def clean_text(text):  
    text = re.sub(r'\s+', ' ', text)  
    text = re.sub(r'^\w\s', '', text)  
    return text.strip()  
  
filtered_data[['Positive_Review', 'Negative_Review']] = filtered_data[['Positive_Review', 'Negative_Review']].applymap(clean_text)  
filtered_data[['Positive_Review', 'Negative_Review']] = filtered_data[['Positive_Review', 'Negative_Review']].applymap(str.lower)
```

- Filtered hotels with at least 1000 reviews and removed rows containing irrelevant placeholders like “Nothing” and “Other”

```
hotels_to_keep = hotel_counts[hotel_counts >= 1000].index  
filtered_data = data[data['Hotel_Name'].isin(hotels_to_keep)]
```

```
placeholders = ["Nothing", "Other"]  
filtered_data = filtered_data[~filtered_data['Negative_Review'].isin(placeholders) & ~filtered_data['Positive_Review'].isin(placeholders)]
```



EDA 1 - Validating the Data Coverage

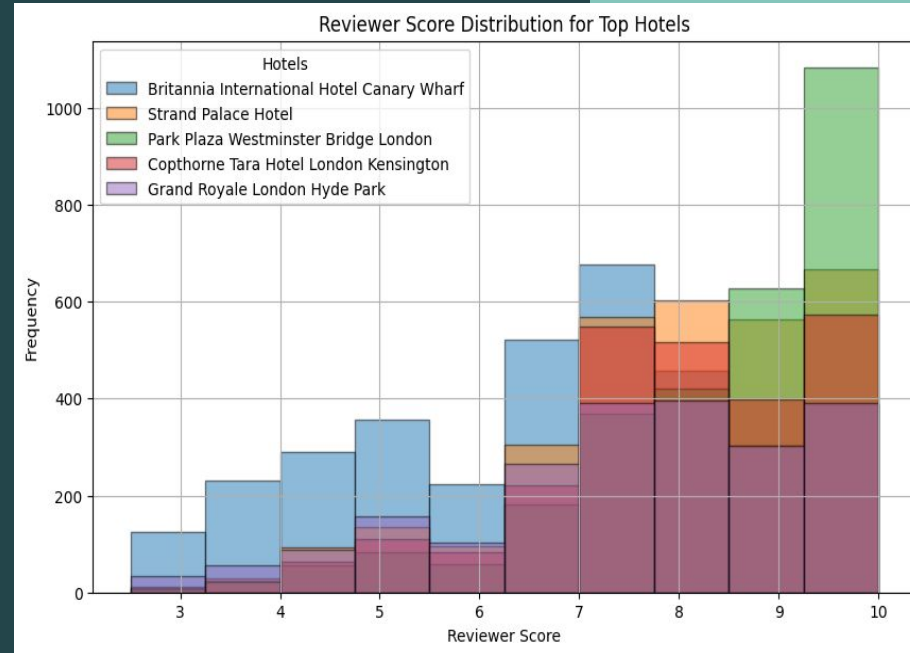
Review Trends Over Time

- The graph ensures the dataset spans a sufficiently broad time period to capture diverse customer experiences
- The graph identifies whether the dataset is dominated by older reviews or if it includes recent, up-to-date feedback



EDA 2 – Analyzing Rating Distributions for Top Hotels

- The overlaying histograms displays which hotels have more diverse or higher / lower ratings
- The graph gives us insights into what we can anticipate in future sentiment analysis
- Results suggest that, on average, review scores tend to be higher than the middle score (> 5)



EDA 3 - Hotel Rankings by Average Score

Top 5 Highest-rated Hotels

Intercontinental London The O2	9.372028
Hotel Esther a	9.121815
citizenM Tower of London	9.025530
citizenM London Bankside	8.988213
Mondrian London	8.969209

Top 5 Lowest-rated Hotels

Mercure Paris Terminus Nord	7.276791
Hilton London Metropole	7.249845
The Cumberland A Guoman Hotel	7.221242
Shaftesbury Premier London Paddington	7.094387
Britannia International Hotel Canary Wharf	6.771916

```
hotel_rankings = data.groupby('Hotel_Name')['Reviewer_Score'].mean().sort_values(ascending=False)
```

The average reviewer score ranges significantly, from 6.77 to 9.37, indicating a clear differentiation in customer experiences and perceived quality among hotels. Furthermore, we can later compare these to the results of the sentiment analysis to verify that they align.

Sentiment Analysis - Calculating Sentiment Polarity

```
data['Positive_Sentiment'] = data['Positive_Review'].apply(lambda x: TextBlob(x).sentiment.polarity)
data['Negative_Sentiment'] = data['Negative_Review'].apply(lambda x: TextBlob(x).sentiment.polarity)
```

The two lines of code calculate sentiment polarity for both positive and negative reviews, meaning they provide us a quantitative measure of how positive or negative the review text is. The score ranges from -1 to 1 (negative to positive)

Example Output:

```
the front desk and entrance was very nice and ... 0.780000
terrible food worst hilton i have ever stayed in miserable depressing rooms -0.9
```



Sentiment Analysis - Calculating Hotel Scores out of 100

```
hotel_sentiments['Hotel_Score'] = (hotel_sentiments['Positive_Sentiment'] - abs(hotel_sentiments['Negative_Sentiment'])) * 50 + 50
# Ensuring that the scores are within 0-100 bounds
hotel_sentiments['Hotel_Score'] = hotel_sentiments['Hotel_Score'].clip(0, 100)
```

Absolute value was used for negative sentiments because their polarity ranges from -1 to 0 while the polarity of positive comments range from 0 to 1

Final Results: Top 10 Highest-Rated & Top 10 Lowest-Rated Hotels and their Sentiment Analysis Scores

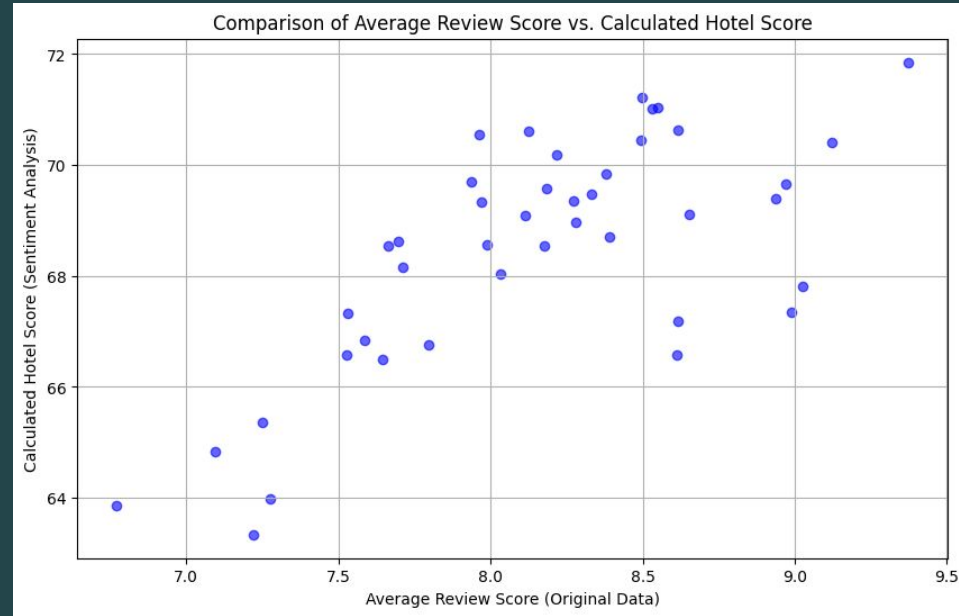
Hotel_Name	Hotel_Score
Intercontinental London The O2	71.853144
DoubleTree by Hilton Hotel London Tower of London	71.219423
St James Court A Taj Hotel London	71.041401
The Tower A Guoman Hotel	71.022560
Hilton London Wembley	70.628802
Park Plaza London Riverbank	70.610807
DoubleTree by Hilton London Docklands Riverside	70.547576
Park Plaza Westminster Bridge London	70.438930
Hotel Esther a	70.402132
Park Plaza County Hall London	70.178690

Hotel_Name	Hotel_Score
The Park Grand London Paddington	66.836044
Norfolk Towers Paddington	66.767927
Millennium Gloucester Hotel London	66.572460
Best Western Premier Hotel Couture	66.566658
Holiday Inn London Kensington	66.502474
Hilton London Metropole	65.367433
Shaftesbury Premier London Paddington	64.827316
Mercure Paris Terminus Nord	63.987580
Britannia International Hotel Canary Wharf	63.860788
The Cumberland A Guoman Hotel	63.338895

Validation 1 - Effectiveness of sentiment analysis in evaluating hotel quality

The general upward trend suggests that hotels with higher average review scores also tend to receive higher sentiment-based scores. This validates the sentiment analysis approach, indicating that the calculated scores align with the actual user ratings.

However, if some points significantly deviate from the trend, it may indicate instances where sentiment analysis does not fully capture the nuances of the written reviews. This could suggest that some hotels may have received high ratings despite negative textual feedback.



Validation 2 - Pearson correlation between Sentiment Analysis Score and Average Reviewer Score

```
from scipy.stats import pearsonr
corr, p_value = pearsonr(comparison['Hotel_Score'], comparison['Avg_Reviewer_Score'])
print(f"Correlation: {corr}, P-value: {p_value}")
```

```
Correlation: 0.7026306005621268, P-value: 2.1367435972484216e-07
```

The positive and high correlation tells us that hotels with higher average scores also have higher sentiment-derived scores

- It effectively captures the general opinion, expressed in texts, and translates it into a meaningful numerical score
- However, since it's not a perfect correlation (1.0), there are some discrepancies, meaning that sentiment analysis does not perfectly predict the average review score

The p-value is extremely small (2.14×10^{-7}), meaning there is strong statistical significance in the correlation

- The observed correlation is very unlikely to have occurred by random chance
- It further supports sentiment analysis as a valid way of estimating hotel ratings



Limitations

Contextual Nuance

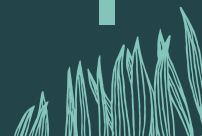
The sentiment analysis model (TextBlob) may not accurately capture sarcasm, humor, or subtle language nuances. For example, a review like "The hotel was amazing... if you enjoy rude staff and dirty rooms." might be misinterpreted as positive.

Discrepancies Between Text Reviews and Numerical Scores

Reviewers may give high scores but leave critical comments (e.g., "Loved the location but the service was terrible" with an 8/10 rating).

Data and Model Generalization Issues

The hotel score formula (based on sentiment polarity) is an assumption and may not reflect real-world rating scales accurately. A more advanced model such as a deep learning sentiment analysis could improve precision.



If I Had More Time, I would further investigate...

Sarcasm and Contextual Sentiment Analysis:

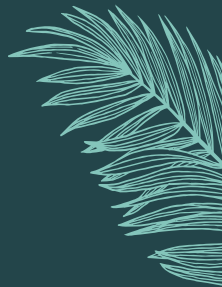
Since TextBlob struggles with sarcasm and nuanced language, I would explore using more advanced NLP techniques, such as a fine-tuned transformer model like BERT, to better capture sentiment within context. A side-by-side comparison of sentiment scores from TextBlob vs. BERT would be insightful.

Multi-Language Sentiment Bias:

True multinational hotels receive guests from all over the world, meaning the reviews would be in multiple languages. I would investigate whether sentiment scores vary based on language differences and whether certain phrases are misinterpreted due to translation issues.

Topic Modeling on Reviews

Using NLP techniques like Latent Dirichlet Allocation (LDA), I would extract the most common themes in positive and negative reviews. This would provide deeper insights into what factors contribute most to customer satisfaction and dissatisfaction, such as cleanliness, staff behavior, or amenities.





Minwoo Choi

University of Virginia



About the Author

Hello! I am a second year at the University Of Virginia studying Data Science with concentrations on Data Engineering and Artificial Intelligence. I hope you enjoyed my work and I look forward to connecting with you! Feel free to reach out!

qce2dp@virginia.edu

[GITHUB](#)

[LinkedIN](#)



Github Project Files

- Raw / Processed Dataset
 - Scripts
 - Graphs