Comparative Analysis of Hotel Reviews: A Case Study of New York Marriott Marquis and Competitors

Ali El Masri

**Abstract**

This study offers a detailed comparative analysis of hotel reviews for the New York Marriott Marquis and its top five competitors in New York City. By detecting languages, applying topic modeling, conducting sentiment analysis, and visualizing various metrics, we aim to derive actionable insights. The primary goal is to provide recommendations that can ultimately enhance customer satisfaction and maintain a competitive edge. Of importance, this dataset is not a definitive representation of the hotel’s operations, but instead, an example for educational purposes, which is derived from a limited set of reviews from 2012.

**Introduction**

Third-party booking sites such as TripAdvisor and Booking.com offer hotels metrics to compare their performance against competitors. However, these reports often lack deeper insights into guest reviews. The goal of this study is to connect that gap by analyzing guest reviews of the New York Marriott Marquis and its competitors, thus influencing advanced Natural Language Processing (NLP) techniques. Through this analysis, the goal is to uncover trends and patterns that can inform better business decisions and strategies for improving guest experiences.

**Methodology**

The methodology encompasses several steps: data collection, cleaning, preprocessing, topic modeling, sentiment analysis, and comparative analysis. We employed BERTopic for topic modeling, utilizing transformer-based embeddings from Huggingface. For sentiment analysis, we used DistilBERT, a smaller, faster, and lighter version of BERT (Bidirectional Encoder Representations from Transformers).

According to Egger and Yu (2022), “compared to other NLP techniques, BERTopic works exceptionally well with pretrained embeddings due to its approach of splitting clustering documents and using c-TF-IDF to extract topic representations. BERTopic supports various topic modeling variations such as guided topic modeling, dynamic topic modeling, or class-based topic modeling. Its main strength lies in performing well on multiple aspects of the topic modeling domain, whereas other techniques typically excel in only one aspect. Additionally, BERTopic allows researchers to reduce the number of topics after training the model, enabling a realistic number of topics based on the actual data.” Therefore, we selected BERTopic as our topic modeling algorithm, leveraging its state-of-the-art capabilities to extract coherent topics from textual data.

**Data Collection**

Our dataset includes 878,561 reviews from 4,333 hotels, all sourced from TripAdvisor in JSON format from Carnegie Mellon University’s website. We refined this dataset to approximately 200,000 English reviews, and after further filtering and modeling, we focused on around 10,000 reviews specific to the New York Marriott Marquis and its main five competitors for our visualizations.

**Data Cleaning and Preprocessing**

1. **Language Detection:** Using the langdetect library, we filtered out non-English reviews, retaining only those in English.
2. **Topic Modeling:**
   * **Models Used:**
     + **SentenceTransformer:** "all-MiniLM-L6-v2"
     + **Embedding Model:** We used the SentenceTransformer model "all-MiniLM-L6-v2" to create dense vector representations of the text.
     + **UMAP:** UMAP(n\_neighbors=15, n\_components=5, min\_dist=0.0, metric='cosine', random\_state=42)
     + **HDBSCAN:** HDBSCAN(min\_cluster\_size=150, metric='euclidean', cluster\_selection\_method='eom', prediction\_data=True)
     + **CountVectorizer:** CountVectorizer(stop\_words="english", min\_df=2, ngram\_range=(1, 2))
     + **Maximal Marginal Relevance:** MMR(diversity=0.3)
     + **Topic Representation:** We used KeyBERTInspired and Maximal Marginal Relevance for topic representation to ensure diversity and relevance in the key terms representing each topic.
   * **Dimensionality Reduction:** We applied the UMAP algorithm to reduce the dimensionality of the embeddings, making the clustering process more efficient.
   * **Clustering:** We used HDBSCAN to cluster the reduced embeddings into topic clusters.
   * **Vectorization:** The CountVectorizer with specific parameters (stop words, minimum document frequency, n-gram range) was used to vectorize the text.
   * **Outlier Reduction:** We employed the topic\_model.reduce\_outliers method to reduce noise and ensure more coherent topics.

**Topic Modeling Results**

The topic modeling process helped us identify and visualize the prevalent topics discussed by guests.

A colorful blotches with text

Description automatically generated with medium confidenceFigure 1: Documents and Topics

The scatter plot shows the distribution of documents (reviews) across different topics, projected onto a two-dimensional manifold using UMAP and clustered by color using HDBSCAN. Each color represents a unique topic, and the clustering of points indicates how frequently topics are discussed together. This visualization helps us understand the diversity and concentration of discussions in the reviews.

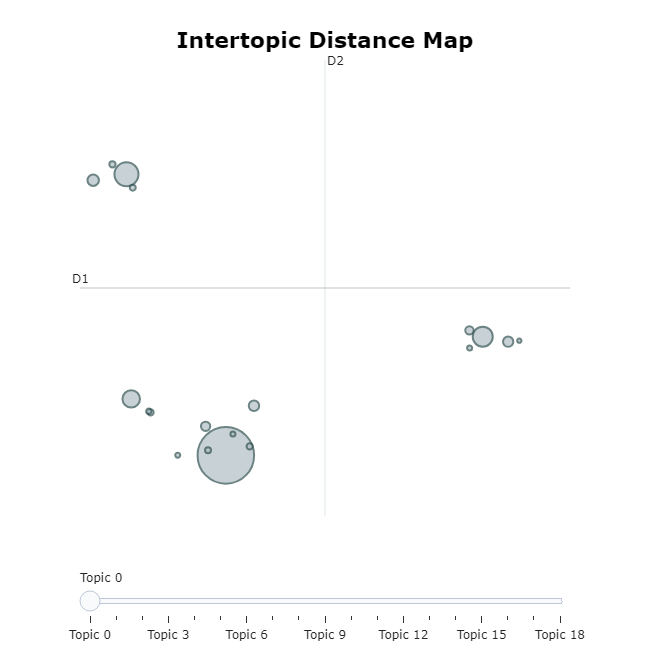


Figure 2: Intertopic Distance Map

The intertopic distance map provides an overview of the distances between topics. Topics closer to each other are more similar, while those farther apart are more distinct. The size of the circles represents the prevalence of the topics. This map is crucial for identifying which topics are more central and which are more peripheral in the discussions.

A graph with text and numbers

Description automatically generated

Figure 3: Topics per Class

The bar chart illustrates the frequency of different topics per class (review category). It highlights which topics are most commonly discussed in different review segments. Understanding the distribution of topics across classes helps us tailor specific strategies for different aspects of the hotel’s service.

A diagram of a clustering graph

Description automatically generated

Figure 4: Hierarchical Clustering

This dendrogram represents the hierarchical clustering of topics derived from the reviews. Each branch represents a topic, and the length of the branches indicates the distance or dissimilarity between topics. Hierarchical clustering helps us visualize the nested structure of topics, revealing how topics group together at various levels of granularity.

**Brief Summary for Clustering and Topic Modeling Results:** Our topic modeling analysis reveals that guest reviews encompass a wide range of themes, from the overall hotel experience to specific issues like room conditions and service quality. The hierarchical clustering visualization provides a detailed view of how these topics are related, showing the hierarchical relationships and groupings within the topics. This helps us understand the overall structure and connections between different themes discussed by guests.

**Sentiment Analysis**

For sentiment analysis, we used the distilbert-base-uncased-finetuned-sst-2-english model.

* **Model Description:** This model is a fine-tuned checkpoint of DistilBERT-base-uncased, fine-tuned on SST-2. This model reaches an accuracy of 91.3 on the dev set (for comparison, BERT bert-base-uncased version reaches an accuracy of 92.7). Grootendorst, M. ([2022](#_top))

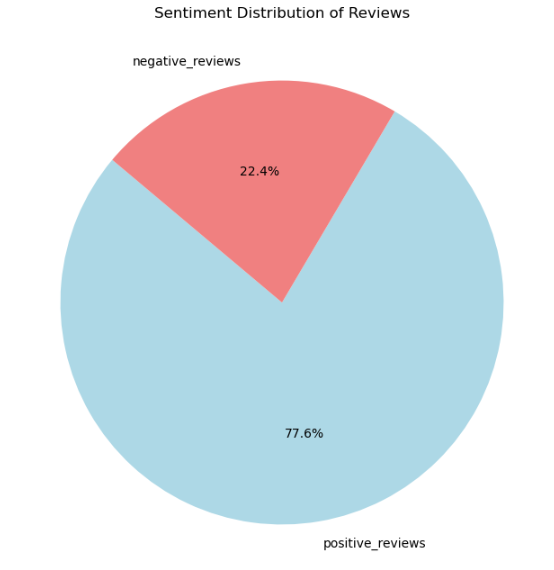
****

Figure 5: Sentiment Distribution of All Reviews

We analyzed the sentiment distribution of all reviews, providing insights into the overall sentiment trends.

**Analysis and Results**

1. **Comparative Rating Analysis:** We compared the average ratings for different dimensions (service, cleanliness, location, rooms, sleep quality, value) between New York Marriott Marquis and its competitors.

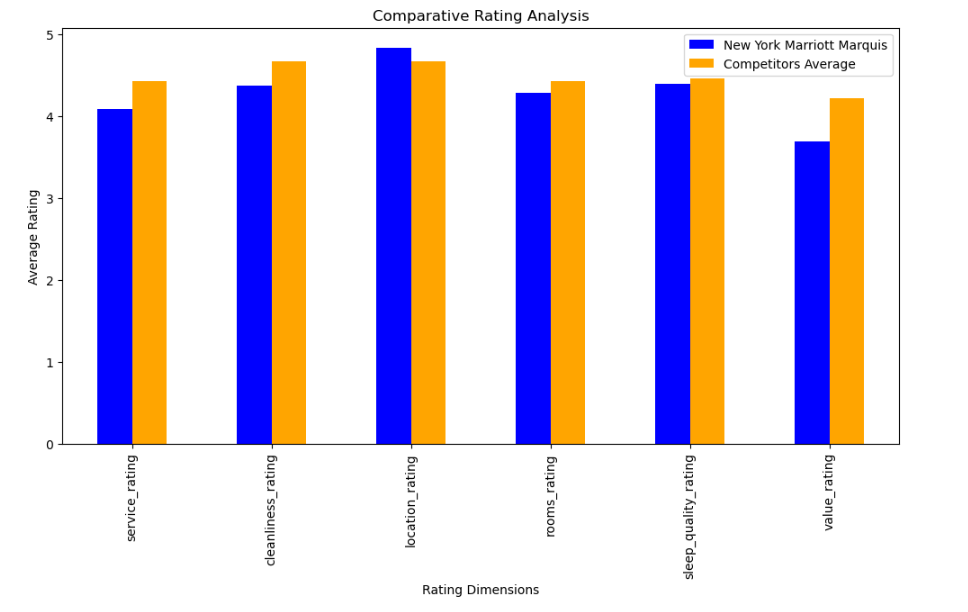


Figure 6: Average Rating Comparison

1. **Sentiment Comparison:** We compared the sentiment distribution (positive vs. negative) for New York Marriott Marquis with that of its competitors.

A graph of different colored squares

Description automatically generated

Figure 7: Sentiment Distribution

1. **Topic Distribution Analysis:** We analyzed the distribution of topics discussed in reviews for New York Marriott Marquis and compared it to that of its competitors.

A graph with different colored bars

Description automatically generated

Figure 8: Topic Distribution Comparison

1. **Time Series Analysis:** We compared the trends in sentiment and overall ratings over time for New York Marriott Marquis with its competitors.

A graph with blue and orange lines

Description automatically generatedFigure 9: Sentiment Trend Over Time

**S.W.O.T Discussion**

**Strengths**

* **Location:** The New York Marriott Marquis excels in this area over its competitors, which we should emphasize in marketing campaigns.
* **Positive Sentiment Trends:** There is a higher increase in positive sentiment trends over negative reviews.

**Weaknesses**

* **Low-Scoring Areas:** Improvements are needed in most areas in which the hotel scores lower than its competitors.
* **Negative Sentiment Trends:** Addressing periods of increased negative sentiment is crucial for maintaining a positive reputation.

**Opportunities**

* **Customer Feedback Loop:** Implementing a robust feedback loop can continuously gather guest feedback and make improvements.
* **New Services and Amenities:** Introducing new services based on positive feedback can enhance the guest experience.

**Threats**

* **Competitor Performance:** Regular benchmarking against competitors is necessary to stay competitive.
* **Proactive Issue Resolution:** Quickly addressing and resolving negative feedback can prevent long-term reputation damage.

**Conclusion**

Our study provides a detailed comparative analysis of hotel reviews for New York Marriott Marquis and its competitors. The insights derived can help improve customer satisfaction and maintain a competitive edge. Future work can focus on incorporating demographic analysis and expanding the scope of analysis to include more competitors.

**References**

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7, 886498. <https://doi.org/10.3389/fsoc.2022.886498>

Grootendorst, M. P. (2022). *Bertopic¶*. BERTopic. https://maartengr.github.io/BERTopic/index.html#citation

**Appendix**

**Supporting Documentation**

GitHub Link for Additional Visualizations and Analysis Code: <https://github.com/elmasri-ali>

**Data Source**

<https://www.cs.cmu.edu/~jiweil/html/hotel-review.html>

**Models Reference Links**

* **UMAP and HDBSCAN:** <https://umap-learn.readthedocs.io/en/latest/clustering.html>
* **SentenceTransformer Embedding model:** "all-MiniLM-L6-v2"
  + <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
  + <https://www.sbert.net/>
* **Topic Modeling:** BERTopic
  + <https://maartengr.github.io/BERTopic/index.html>
* **Text Classification Model:** DistilBERT base uncased finetuned SST-2
  + <https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english#model-details>

**Questions and Answers (Milestone 4)**

1. **What were the main challenges in data preprocessing?** Ensuring accurate language detection and handling missing or incomplete data were the main challenges.
2. **How was the topic modeling approach selected?** We chose BERTopic for its ability to handle large datasets and generate interpretable and meaningful topics.
3. **What insights were gained from sentiment analysis?** The analysis revealed periods of positive and negative sentiment trends, helping us identify areas for improvement.
4. **How can the hotel leverage the findings from this study?** By focusing on strengths, addressing weaknesses, and introducing new services based on positive topic feedback.
5. **What future work is suggested?** Incorporating demographic analysis and expanding the scope to include more competitors and larger dataset.
6. **How reliable are the sentiment analysis results?** The use of a fine-tuned BERT model ensures high reliability in sentiment classification with an accuracy of 91%.
7. **What impact do the identified topics have on guest satisfaction?** Understanding frequently reviewed topics can help the hotel focus on areas that matter most to guests.
8. **How often should the hotel conduct such analyses?** Regular analysis (e.g., quarterly) is recommended to stay updated on guest feedback and market trends.
9. **What other NLP techniques can be used for such analyses?** Techniques like word embeddings, LDA, and advanced transformers can be explored for deeper insights.
10. **How can the hotel improve its response to negative feedback?** By implementing a faster and more efficient feedback resolution system we can significantly improve guest satisfaction.