

Navigating the Noise of Online Hotel Reviews

Amit Halbreich	20817393	amithalbreich	Amit.Halbreich@mail.huji.ac.il
Omer Mushlion	208271197	omer_mushlion	Omer.Mushlion@mail.huji.ac.il
Omri Marom	319133666	omrimar	Omri.Marom@mail.huji.ac.il


1. Problem Description


When planning a vacation and searching for accommodations, many of us turn to read user reviews on websites like booking.com, sifting through thousands of reviews. But how meaningful are these reviews? And what can we conclude from so many contradicting reviews?


Online hotel reviews have been the focus of research across various disciplines, including their impact on consumers choice and hotel performance, text analysis and summarization of reviews, and the motivations behind leaving reviews. Using multiple topics covered during the course, this works presents the **topic-indicativeness scores** ([3.1](#)), a metric that provides nuanced and semantically rich summary of user reviews; Then, we will show how this metric can be used in order to rerank hotels to improve recommendations based on user preferences ([3.2](#)), and to extract concise but meaningful insights from a large amount of reviews ([3.3](#)).

2. Data

We implemented a JavaScript tool to scrape user reviews from booking.com and packaged it into a Google Chrome extension for easy use. When running the script on the reviews page of a hotel on booking.com, it automatically sifts through all reviews pages and scrapes the reviews into a .csv file. Please refer to [Appendix I](#) for the instructions for running this scraping script with a demo video.

For each hotel, we get a .csv file with the following scheme: 

We used this tool to extract the reviews of multiple accommodations with various overall ratings and accommodations types, as displayed below: 

The following fun bonus plot presents the locations of the hotels scrapped for this project on a world map: 

Since we were looking for hotels with at least 1,000 English reviews, this map portrays the apparent bias towards the popular travel destinations in the English-speaking western world, especially among those that tend to use Booking.com. Initially, we tried to follow the list of the

largest cities around the world and pick one hotel in each city, but we found out that many of those cities are either not popular travel destinations (like Karachi); reside in countries banned from Booking.com (like Tehran or Moscow); apparently do not use Booking.com much (like Beijing); or have very few English reviews, with the vast majority in Portuguese or Spanish (like São Paulo). We also wanted to pick at least one hotel from each continent, but there are no accommodations in Antarctica on Booking.com 🤖❄️

3. Solution

3.1. Topic-Indicativeness scores of online hotel reviews

We define an online hotel review as *indicative* if it (a) repeats a point made by a *significant* portion of other reviews for the same hotel in the same sentiment and (b) does not contradict a *significant* portion of other reviews raising the same point in the opposite sentiment. For example, if only two reviews out of 200 mentions the cleanliness of the room, the conclusion about room cleanliness remains inconclusive; and if 40 reviews state that the rooms are clean while 50 reviews state that they are dirty, the matter is inconclusive as well.

In order to do so, we need to perform three main text processing tasks: extract the top topics discussed in all reviews (3.1.1), analyse the sentiment of reviews into positive and negative reviews (3.1.2), and classify each review into one or more topics (3.1.3).

3.1.1. Topic extraction from reviews

One solid, low-hanging approach to extract topics from the reviews is using an off-the-shelf topic extraction model like Bertopic. However, we chose to use tools mentioned in class:

1. **Preprocess reviews text:** tokenize to sentences, remove stop words, lemmatization.
2. **Encode each word using its TF-IDF score:** represent each word with its TF-IDF score with respect to each “document” (review sentence).
3. **Dimensionality reduction:** apply SVD decomposition to the TF-IDF matrix to reduce its dimensions while preserving as much information as possible.

The following plot presents the scores of top 25 words, filtered to include only nouns: (plot)

Based on this ranking, we grouped the top-scored words into the following 5 topics: **location**, **staff**, **food and beverages**, **room amenities** and **hotel amenities**. We can also observe the dominance of these five topics in the following word clouds of the negative and positive reviews: (plot)

3.1.2. Sentiment analysis of reviews

The scheme of our data already provides sentiment analysis for the reviews (separating them into positive and negative), and although it might be a bit noisy, since this is not the main topic of this

project we will use this sentiment information without any additional refinement.

3.1.3. Classification of reviews by topics

For this task we chose to use a zero-shot text classification model with multi-label support, since many reviews discuss several topics even within a single sentence (for example: "*Staff and room cleanliness were excellent*"). We initially experimented with Facebook's `bart-large-mnli`, and later decided to use `GPT-4o` to speed up the prediction process. Please refer to [Appendix II](#) for the prompt we used for this purpose. The classification is currently binary, but exploring the use of continuous values between 0 and 1 could be an interesting direction for future work.

3.1.4. Topic-Indicativeness results

We mentioned two factors in the indicativeness measure we suggested. First, the following histograms present how often different topics are mentioned in reviews across various hotels: **(plot)**

We can observe some variability in the proportion of reviews discussing different topics. This can be attributed to several factors: not all topics are relevant to every hotel (for example, budget hotels may not offer food services); there may be overlap between topics (e.g., hotel breakfast related to hotel amenities); and a notable portion of reviews are generic and do not discuss any specific topic (for example, "*Everything was good*" or "*Had a great time*"). Overall, the data shows that a substantial portion of reviews discuss these topics, confirming they are a solid basis for key topic extraction. Therefore, we will not weight the topic frequencies in our topic-indicativeness scores, although this could be explored in future work to further improve the precision and quality of the scores.

The following histograms present the balance between positive and negative reviews for each topic. This ratio is mapped to range $[-1, 1]$, where -1 indicates only negative reviews; 1 indicates only positive reviews; and 0 indicates an equal number of positive and negative reviews, calculated using the following formula:

$$\text{Sentiment Ratio} = (\# \text{ Positive Reviews} - \# \text{ Negative Reviews}) / (\# \text{ Total reviews for the topic})$$

We will use these ratios as the topic-indicativeness scores. **(plot)**

We can observe several patterns here: the sentiment ratios for "Room amenities" and "Hotel amenities" are approximately centred around 0, indicating that these topics are more debatable, with guests having mixed opinions, and people should approach reading a small subset of reviews on these topics with more caution. In contrast, "Staff" and "Location" are centred around a strong positive value, suggesting that people are more likely to agree on these topics and are generally pleased with the staff and location of the hotels. It's somewhat surprising that "Food and beverages" is also centred around a positive value (though lower than "Staff" and "Location"), as people are often quite particular about their food.

For the evaluation of these scores, we will look at the correlation between the sentiment ratios and the overall ratings of the hotels, as shown in the following plot. We want the sentiment ratios to be somewhat correlative with the overall rating, to be indicative of the hotel's rating, but not too correlative, as they should be more fine-grained than the overall rating. (plot)

Examining the Pearson correlation between sentiment ratios and overall hotel ratings, we see positive correlations with four topics, except for “Location” (which does not show any significant correlation, due to the consistently high sentiment towards location observed in the previous plot). The positive correlation of the first four topics suggests that our scores do indeed capture information about the hotel qualities, and is indicative for this.

3.2. Re-rank candidate hotels based on topic-indicativeness scores

We present a re-ranking or tiebreaker approach to recommend hotels to users based on the calculated topic-indicativensness results as defined in the previous section (3.1):

1. The topic-indicativensness scores serve as the item profiles for the candidate hotels.
2. We gather ratings semi-explicitly from the users by asking them to rate the importance of each of the 5 topics mentioned above (3.1.1) from 0 to 5 to their travel experience (rather than asking them to score the hotels themselves, as in pure explicit rating gathering). We use this rating as the user-profile for the recommendation purposes.
3. We calculate a weighted score of each candidate hotel based on these user ratings and the topic-indicativensness scores, to recommend hotels that better match the travel preferences of the users.

Our approach can be seen as a middle ground between collaborative recommendation approach (as it considers the ratings of other users as part of the topic-indicativensness scores calculation) and content-based recommendation (as it considers the preferences of the user itself).

Similar to the content-based recommendation approach, our re-ranking method, based on the topics importance scores, can recommend hotels to users with unique taste (for example: a user who chooses a hotel based solely on the breakfast), and it is also self-explanatory, due to the simplicity and clarity of the topic-indicativensness scores.

We will demonstrate our method as a second, re-ranking step applied to hotels and break a tie between 5 hotels in London, all with an 8.0 overall average rating (as of the data retrieval date 29/08/24). The following plot present the topic-indicativensness scores of these hotels: (plot)

Now, consider the following three different travel characters and notice how each of them is recommended a different hotel using our approach: (plot)

The above character descriptions were written with the help of GPT-4o, Please refer to [Appendix II](#) for the prompt we used for this purpose.

3.3. Extracting indicative small subset of reviews

We propose an approach to identify and extract a small, meaningful subset of reviews for each hotel that will capture the topic-indicativeness scores, so that by reading only these selected reviews, users can get a good approximation of the overall impression of the hotel.

While text summarization is a well-established research area in NLP, we aim to extract a concise yet comprehensive representation of the overall reviews using graph-based techniques:

1. For each hotel, we initialise a graph where each node represents a user review as a whole (both the positive and negative comments).
2. We calculate the “topic-sentiment vector” for each node, a binary vector of size 10 (5 topics, as defined in [3.1.1](#), and 2 sentiments) that have 1 in the i ’th coordinate iff the user review discusses the i ’th (topic, sentiment) pair.
3. Edges are added between any two nodes that share at least one common (topic, sentiment) pair.
4. The weight of each edge is determined by the normalised dot product between the “topic-sentiment vectors” of the two nodes.
5. We run PageRank algorithm on this graph to identify the most representative reviews of the overall impression of the hotel. We claim that the 10 top-scored reviews serve as a concise and indicative summary of the hotel reviews.

For evaluation, we calculated the “estimation error” of the topic-indicativeness scores using the 10 top-scored reviews, and compared it to a random subset of reviews of size 10 (averaged over multiple iterations). As presented in the following plot, this small subset of top-scored reviews provides a better indicative approximation of the topic-indicativeness scores, providing users with a concise impression of the hotel capturing the conflicting reviews and debatable points: (plot)

Conclusion

This work deals with a challenge many of us encounter in real life: trying to obtain a conclusive impression from too many online reviews, many times with conflicting information. Apart from providing a concise, interpretable, and self-explanatory method to extract the essence of the content of the reviews (which could be easily added to such websites!), we showed several interesting and valuable applications of this measurement. As mentioned along this report, there is room for further refinement and improvement.

We hope you’ll think about these things when booking your next vacation using booking.com 😊
(Disclaimer: This work is in no way connected with, sponsored by (unfortunately...), or endorsed by booking.com or any of its affiliates.)