# Automated Hotel Review Summarization Using Generative AI: Assessing BART versus T5-small Models

Shawn Oyer
*Drexel University, CS 614*

9/3/2024

## Abstract

The adoption of Generative AI has transformed the travel industry, particularly in enhancing customer experience and operational efficiency. This project investigates the application of AI-driven text summarization models, specifically BART and T5-small, to generate concise summaries of hotel reviews from a large dataset scraped from booking.com. The models were evaluated using ROUGE metrics, which assess the quality of the generated summaries by comparing them to human-created summaries. The results demonstrate that the BART model significantly outperforms the T5-small model across all ROUGE metrics, indicating its superior ability to capture key contextual information while maintaining the structural integrity of the original reviews. This research highlights the potential of advanced text summarization models like BART in summarizing user-generated content, thereby creating more efficient information retrieval and enabling more informed decision-making processes in the travel sector.

## 1 Introduction

The travel industry has been one of the most innovative and fast-evolving sectors in recent years. The industry keeps adopting the latest technology applications to improve customer experience and operational efficiency. In recent years, Generative AI has come to impact the sector by being a game-changer in providing best-in-class solutions that overwhelmingly elevate the respective dimensions of the travel planning process [11]. Consumers now have access to an excessive amount of information which makes it cumbersome to sift through. Automated article summarization using generative AI has emerged as a viable method to extract essential insights and give succinct summaries of articles in this era of information overload. At its core, text summarization seeks to condense lengthy documents or passages into concise summaries while preserving the essential meaning and context. By extracting salient information and discarding redundant or extraneous details, summarization facilitates quicker understanding, decision-making, and information retrieval [2]. This study focuses on the application of AI-driven text summarization models in the travel industry, specifically analyzing their effectiveness in summarizing hotel reviews.

## 2 Background

The integration of AI in the tourism industry offers significant opportunities, as it can enable growth between 7% and 11.6% of total revenue in the sector. This growth can be achieved through various AI applications and several emerging research trends in the field, including areas such as eWOM, service recovery, customer satisfaction, brand/destination image, service quality, big data, netnography, Travel 2.0, Web 2.0, e-tourism, green experience, smart tourism. Challenges such as data complexity, algorithmic bias, financial concerns, and socio-ethical considerations need to be addressed to fully harness the potential of these technologies in tourism [1].

Just in the last year, major travel companies such as TripAdvisor and AirBnb announced new generative AI-driven features that summarize reviews to give travelers a brief, digestible overview of each property [10]. The features also summarize community perspectives on key attributes of quality such as location, atmosphere, or rooms so that guests can get familiar with the property quicker [9].

Generative AI for text summarization marks a significant leap forward in our pursuit of advanced natural language understanding and generation. Leveraging the power of deep learning and transformer architectures, these models offer a versatile and powerful tool for transforming complex textual informa-

tion into concise and meaningful summaries. Open AI is at the forefront of AI research and offers the Large Language Model (LLM), which is a powerful model trained on vast amounts of textual data. There are other models such as Lang Chain, Hugging Face Summarization, BERT, GPT, Gensim Summarization, among others that also have proven successful in vast text summarization and the field is constantly growing [2].

# 3 Methodology

Specifically, this project aims at generating text summarizations for thousands of hotel reviews that incorporate user perspectives on key attributes of quality such as location, atmosphere, hospitality and services using two pre-trained summarization models, BART and T5-small.

## 3.1 Data Source

The data used were sourced from Kaggle and freely available under CC0 1.0 license. The data can be downloaded here: data download [6]

The data consists of a .CSV file with 515,000 customer reviews and ratings of 1493 luxury hotels across Europe scraped from Booking.com in 2017. Here are the 17 fields and descriptions:

- Hotel_Address: Address of hotel.
- Review_Date: Date when reviewer posted the corresponding review
- Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year
- Hotel_Name: Name of Hotel
- Reviewer_Nationality: Nationality of Reviewer
- Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
- Review_Total_Negative_Word_Counts: Total number of words in the negative review
- Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
- Review_Total_Positive_Word_Counts: Total number of words in the positive review
- Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience
- Total_Number_of_Reviews_Reviewer_Has_Given: Number of Reviews the reviewers has given in the past
- Total_Number_of_Reviews: Total number of valid reviews the hotel has
- Tags: Tags reviewer gave the hotel.
- days_since_review: Duration between the review date and scrape date
- Additional_Number_of_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there
- lat: Latitude of the hotel
- lng: Longitude of the hotel

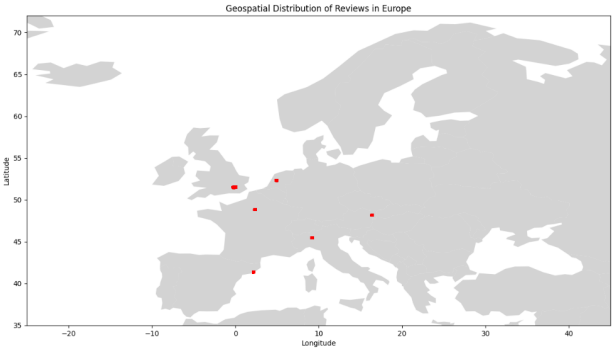



Figure 1: Geospatial Distribution of Hotels in the Dataset

## 3.2 Model and Data Justification

The framework used for this project was the Hugging Face Transformers library, a cutting-edge open-source framework that provides easy-to-use, production-ready implementations of the latest and greatest transformer-based models. These models, trained on vast amounts of textual data, have demonstrated remarkable performance in a wide range of NLP tasks, including text summarization. Some key advantages of using this library is its modular design, which allows users to easily swap out different models or fine-tune them to better suit their specific needs. In addition, the transformers library provides efficient batch processing capabilities,

model caching/deployment, evaluation/metrics, distributed training/inference and customization/fine-tuning that makes this library more sophisticated and scalable than its counterparts [7].

This specific dataset was used for the project due to being openly sourced, and containing a rich amount of positive and negative reviews for a sufficient amount of hotels to be able to test the summarization model on. It also contained many attributes that allowed for quality Exploratory Data Analysis (EDA). In conducting research for other datasets, these requirements were very difficult to attain in openly sourced datasets.

Figure 1 provides a clear overview of the project's workflow. It starts with a comprehensive dataset comprising 515,000 positive and 515,000 negative reviews. The data undergoes a cleaning process, including tokenization and punctuation removal, before being fed into a model. The model organizes the reviews by hotel, and the resulting chunks are passed to a summarizer combinator, which generates review responses. Finally, these responses are evaluated using specific metrics to determine the model's performance.

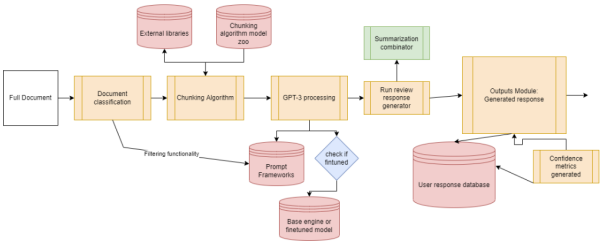



Figure 2: Abstractive Text Summarization Model [8]

### 3.3 Procedure

This was the procedure used for this project:

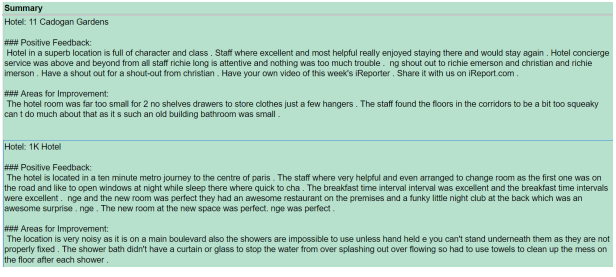
1. Load .CSV file
2. Clean the data by removing unwanted values/characters and pronouns
3. Use the pre-trained Summarization pipeline called BART (bidirectional encoder and Autoregressive decoder), provided by Hugging Face to create text summarizations for the 10 most recent reviews per hotel using max chunk size of 1000, max length of 150 and min length of 50
4. Generate a new .CSV file containing positive summarization reviews and areas for improvement summarization reviews for each hotel
5. Fine-tune hyperparameters with a pre-trained model and tokenizer named t5-small provided by Hugging Face using the top 10 most recent reviews as the Source Text and the positive/areas for improvement summarization reviews as the Target Text to train the model
6. Evaluate the models using ROUGE metrics

## 4 Experiments and Results

### 4.1 Model Output Examples

BART was selected as the pre-trained model due to being particularly effective when used for text generation (e.g. summarization, translation) as part of Hugging Face Summarization pipeline [3]. Using the BART pre-trained model with no fine-tuned parameters, positive feedback and areas for improvement were generated based on the 10 most recent user reviews for each hotel. The decision to use only the most recent 10 reviews was due to memory limitations in Google Colab, which caused crashes when processing all available reviews, so the study focused on this subset.

Figure 3 illustrates an example output with two review summaries. BART effectively categorized feedback into positive aspects and areas needing improvement, while avoiding the use of pronouns, which is appropriate for a review summary. For instance, the positive feedback for 11 Cadogan Gardens Hotel included, "The hotel, in a superb location, is full of character and class. The staff were excellent and most helpful; I really enjoyed staying there and would stay again." Conversely, the areas for improvement highlighted, "The hotel room was far too small for two, with no shelves or drawers to store clothes—just a few hangers. The staff found the floors in the corridors to be a bit too squeaky, which is hard to address in such an old building. The bathroom was small."


**Summary**

Hotel: 11 Cadogan Gardens

### Positive Feedback:
Hotel in a superb location is full of character and class . Staff where excellent and most helpful really enjoyed staying there and would stay again . Hotel concierge service was above and beyond from all staff richie long is attentive and nothing was too much trouble . ng shout out to richie emerson and christian and richie imerson . Have a shout out for a shout-out from christian . Have your own video of this week's iReporter . Share it with us on iReport.com .

### Areas for Improvement:
The hotel room was far too small for 2 no shelves drawers to store clothes just a few hangers . The staff found the floors in the corridors to be a bit too squeaky can t do much about that as it s such an old building bathroom was small .

Hotel: 1K Hotel

### Positive Feedback:
The hotel is located in a ten minute metro journey to the centre of paris . The staff where very helpful and even arranged to change room as the first one was on the road and like to open windows at night while sleep there where quick to cha . The breakfast time interval interval was excellent and the breakfast time intervals were excellent . nge and the new room was perfect they had an awesome restaurant on the premises and a funky little night club at the back which was an awesome surprise . nge . The new room at the new space was perfect. nge was perfect .

### Areas for Improvement:
The location is very noisy as it is on a main boulevard also the showers are impossible to use unless hand held e you can't stand underneath them as they are not properly fixed . The shower bath didn't have a curtain or glass to stop the water from over splashing out over flowing so had to use towels to clean up the mess on the floor after each shower .

Figure 3: Example Output of pre-trained Summarization Pipeline

After generating new hotel summaries, the next

objective was to fine-tune a summarization model to better evaluate the summaries' metrics. The T5-small model, a variant of the T5 transformer specifically designed for text summarization, was chosen for this task and adapted to produce concise and coherent summaries of input text [4]. The dataset was split into 80% for training and 20 for testing, with the original reviews serving as the Source Text and the generated summaries from BART as the Target Text.

The fine-tuning process involved the following parameters: a learning rate of 2e-5, a per-device training batch size of 2, a per-device evaluation batch size of 2, gradient accumulation steps set to 4, weight decay set to 0.01, a save total limit of 3, 3 training epochs, and the use of fp16 precision with the predict_with_generate option enabled.

To assess the model's performance, testing examples were created, allowing users to input two reviews and receive a generated summary based on those reviews. For instance, given the input reviews "The hotel had all of the intricate artwork displayed and breakfast was extraordinary and reasonably priced. The staff were super helpful," and "The hotel was fantastic! It was in a perfect location near the city centre," the model generated the following summary: "The staff were super helpful. The hotel was in a perfect location near the city centre. The hotel had all of the intricate artwork displayed and breakfast was extraordinary and reasonably priced. The staff were super helpful. The hotel was fantastic!" While the model generally performed well in summarizing the reviews, it tended to repeat content to reach the maximum character limit instead of stopping once the summary was complete.

## 4.2 Evaluation Metrics

This section discusses the evaluation metrics performed in this project. ROUGE was used as the evaluation metric to assess the models ability to generate summaries from text reviews. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. ROUGE-1: Measures the overlap of unigrams (single words) between the generated summary and the reference summary and is useful for measuring the basic content overlap, focusing on the presence of individual words.. ROUGE-2: Measures the overlap of bigrams (two consecutive words) between the generated summary and the reference summary and provides more insight into the fluency and coherence of the summary, as it considers pairs of words. ROUGE-L (Longest Common Subsequence) measures the longest common subsequence (LCS) between the generated summary and the reference summary and focuses on capturing the longest sequence of words that appear in both the generated summary and the reference summary in the same order. It is more sensitive to the order of words and thus better reflects the structure of the summary. Higher scores indicate better performance, with more overlap between the generated and reference summaries [5].

Table 1: F1 scores - ROUGE Metrics Comparison

| Metric | Type | Low | Mid | High |
|---|---|---|---|---|
| **ROUGE-1** | BART | 0.1400 | 0.1441 | 0.1487 |
| | t5-small | 0.0335 | 0.0396 | 0.0464 |
| **ROUGE-2** | BART | 0.0167 | 0.0176 | 0.0187 |
| | t5-small | 0.0074 | 0.0096 | 0.0122 |
| **ROUGE-L** | BART | 0.0861 | 0.0883 | 0.0908 |
| | t5-small | 0.0258 | 0.0310 | 0.0361 |
| **ROUGE-Lsum** | BART | 0.0862 | 0.0883 | 0.0906 |
| | t5-small | 0.0265 | 0.0311 | 0.0360 |

Table 1 summarizes the F1 scores for each metric across three different quality levels: Low, Mid, and High. For the ROUGE-1 metric, BART significantly outperformed T5-small across all quality levels. BART achieved F1 scores of 0.1400, 0.1441, and 0.1487 for the Low, Mid, and High levels, respectively. In contrast, T5-small recorded much lower scores of 0.0335, 0.0396, and 0.0464 for the same levels. This indicates that BART is more effective in capturing key words from the source text.

For the ROUGE-2 metric, the BART model achieved F1 scores of 0.0167, 0.0176, and 0.0187 across the Low, Mid, and High levels, respectively. T5-small, on the other hand, achieved lower scores of 0.0074, 0.0096, and 0.0122. The gap between the models indicates that BART is better at preserving contextual information in the form of two-word sequences.

For the ROUGE-L and ROUGE-L sum metrics, BART again outperformed T5-small, with ROUGE-L F1 scores of 0.0861, 0.0883, and 0.0908 across the Low, Mid, and High levels, respectively. T5-small recorded significantly lower scores of 0.0258, 0.0310, and 0.0361. The ROUGE-Lsum scores followed a similar pattern, with BART achieving 0.0862, 0.0883, and 0.0906, compared to T5-small's 0.0265, 0.0311, and 0.0360.

In looking at the heatmaps for both BART and T5 (Figures 4 and 5), they follow a similar pattern of ROUGE 1 having the highest F1 scores, ROUGE 2 having the lowest F1 scores and Rouge L and Rouge L sum right in the middle. In addition, Figure 6 displays the comparison of F1 scores for the evaluated metrics which clearly shows BART as the more robust model.

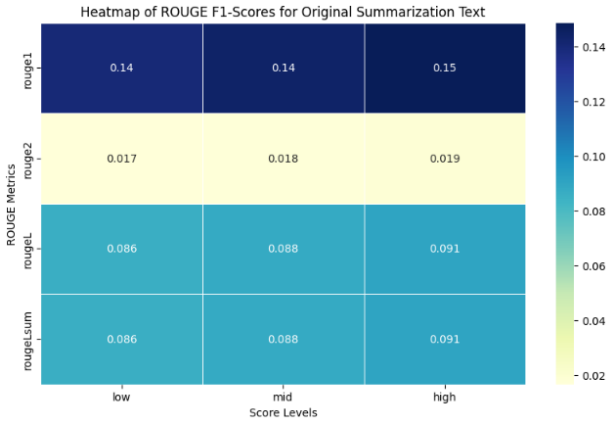



Figure 4: BART Model Heatmap of ROUGE F1 Scores

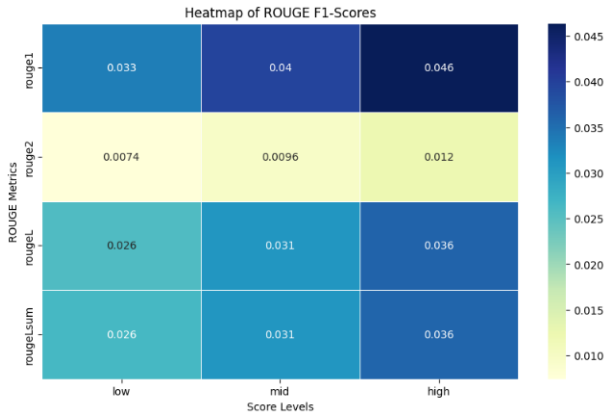



Figure 5: T5-Small Model Heatmap of ROUGE F1 Scores

## 5 Discussion

Overall, the results clearly indicate that the BART model consistently outperforms the T5-small model across all ROUGE metrics and quality levels. BART's higher scores in ROUGE-1 and ROUGE-2 suggest that it is more adept at capturing both individual words and word pairs from the source text, resulting in more accurate and contextually relevant

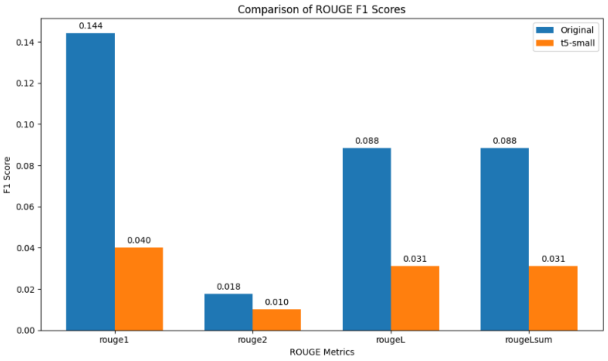



Figure 6: F1 ROUGE Metrics Comparison

summaries. Furthermore, BART's superior performance in ROUGE-L and ROUGE-L sum demonstrates its ability to preserve the structure and order of the original text, making it a more dynamic model for generating accurate text summaries. The significant performance gap between the models emphasizes the advantage of using a more powerful transformer model like BART for tasks involving text summarization.

## 6 Conclusion

This project underscores the effectiveness of generative AI text summarization in processing large amounts of user-generated reviews within the travel industry. In comparing the BART and T5-small models, BART consistently outperforms T5-small across all metrics, particularly in capturing key content and maintaining the structural integrity of the original text. These findings suggest that BART's more advanced transformer architecture makes it a more dynamic tool for summarization tasks, offering significant benefits in providing quality and accurate summaries. Enhancing customer experience and operational efficiency in the travel industry is paramount in this rapidly growing sector, as AI-driven summarization can streamline the decision-making process for travelers by providing quick access to relevant information. Future work could explore the scalability of these models in real-time applications and further fine-tuning to adapt to various domains within the travel sector.

## References

[1] Garcia-Madurga, Miguel A., and Ana J. Grillo-Mendez. "Artificial Intelligence in the Tourism Industry: An Overview of Reviews." MDPI, vol.

13, no. 8, 2023, p. 13. Retrieved from url. Accessed 30 August 2024.

[2] Gopalakrishnan, Karthika. "TEXT SUMMARIZATION USING GENERATIVE AI: A CASE STUDY IN BANKING INDUSTRY." iaeme, 2024. Retrieved from url. Accessed 30 August 2024.

[3] Hugging Face. "facebook/bart-large-cnn · Hugging Face." Hugging Face, 18 January 2024. Retrieved from url. Accessed 3 September 2024.

[4] Hugging Face. "Text Summarization" Hugging Face, 2024. Retrieved from url. Accessed 3 September 2024.

[5] Lin, Chin Y. "ROUGE: A Package for Automatic Evaluation of Summaries." ACL Anthology, 2004. Retrieved from url. Accessed 3 September 2024.

[6] Liu, Jiashen. "515K Hotel Reviews Data in Europe." Kaggle 2017. Retrieved from url. Accessed 30 August 2024.

[7] Oluwafemidiakhoa. "Unlocking the Power of Text Summarization: A Comprehensive Guide to Building a Production-Ready Solution with Hugging Face Transformers." Medium, 8 April 2024. Retrieved from url. Accessed 30 August 2024.

[8] Payne, Matt. "State of the Art GPT-3 Summarizer For Any Size Document or Format." Width.ai, 21 August 2023. Retrieved from url. Accessed 30 August 2024.

[9] Tiernan, Kirstie. "Airbnb has used artificial intelligence & machine learning to disrupt the marketplace to become the fastest-growing hotelier provider in recent years." BDO USA, 28 August 2023. Retrieved from url. Accessed 30 August 2024.

[10] TripAdvisor. "Tripadvisor Launches AI-Powered Hotel Review Summaries." Hospitality Net, 25 October 2023, Retrieved from url. Accessed 30 August 2024.

[11] Veluru, Chandra S. "Transforming Travel Planning: The Impact of Generative AI on Itinerary Optimization, Cost Efficiency and User Experience." Journal of Artificial Intelligence & Cloud Computing, vol. 2(4), no. 2754-6659, 2023, pp. 1-8. Research Gate. Retrieved from url. Accessed 30 August 2024.