Comprehensive NLP Analysis of Airbnb Reviews for Improved Decision-Making

George Washington University

Final Term Project

Vishal Bakshi
Department of Data Science
George Washington University
Vishal.bakshi@gwu.edu

Kusum Sai Chowdary Sannapaneni
Department of Data Science
George Washington University
kusumsaichowdary.sannapaneni@gwu.edu

Abstract

This project explores methods to enhance sentiment and aspect-based analysis of Airbnb guest reviews using advanced Natural Language Processing (NLP) techniques. By employing transfer learning, aspect extraction, and summarization, we aim to improve the efficiency of feedback interpretation. Our approach utilizes transformer-based models, including BERT and T5, for sentiment classification and summarization, alongside Word2Vec and KMeans clustering for aspect-based sentiment analysis (ABSA). The study evaluates the influence of aspect granularity, dataset diversity, and pre-trained models on overall performance. Results are expected to demonstrate that combining transformer-based models with aspect-specific features enhances predictive accuracy and provides actionable insights for hosts. This project contributes to the field of text analysis by offering a framework to automate and refine the interpretation of large-scale user-generated content for short-term rental platforms.

1 Key Information to include

- **Mentor:** Prof. Ning Rui
- **Team Contributions:** Both Kusum Sai Chowdary and Vishal Bakshi contributed equally to the report, research, and development of the .py files. The initial text preprocessing phase was a collaborative effort, with both team members conducting extensive research on various methodologies. Phase I was led by Kusum, supported by regular discussions and input from Vishal. In Phase II, both members worked together, with Vishal focusing on SpaCy techniques like dependency parsing and Named Entity Recognition (NER), while Kusum concentrated on Word2Vec embeddings and K-Means clustering. Phase III was led by Vishal, with ongoing research discussions and input from Kusum. For the visualizations, Kusum and Vishal worked closely to create and refine the plots, ensuring the presentation flowed well and provided clear, explainable insights.

2 Introduction

The rapid growth of platforms like Airbnb has generated vast user reviews offering insights into guest satisfaction, property conditions, and host performance. However, analyzing these reviews manually is inefficient. This report leverages Natural Language Processing (NLP) techniques to automate insights extraction through (1) sentiment analysis, (2) aspect-based sentiment analysis (ABSA), and (3) summarization. These methods aim to help hosts, guests, and platform administrators improve guest experiences and address recurring issues effectively.

3 Dataset Description

The analysis is based on the NYC Airbnb Reviews 2021 dataset, which contains approximately 17,444 reviews. Each entry in the dataset includes the following key attributes:

- **listing_id**: A unique identifier for each Airbnb property.
- **url**: The URL of the Airbnb listing page.
- **review_posted_date**: The month and year when the review was posted.
- **review**: The textual content of the guest's feedback.

For this project, we focused exclusively on English-language reviews. After applying filtering and preprocessing steps, we obtained a consistent and clean dataset tailored for Natural Language Processing (NLP) tasks. The dataset's diverse feedback on aspects such as location, cleanliness, and host interaction makes it highly suitable for sentiment analysis, aspect-based sentiment analysis, and summarization.

4 Description of NLP Models and Algorithms

4.1 Sentimental Analysis

We employed a two-staged approach

- **Initial Labelling Using a Pre-trained Model**

  We utilized a pre-trained sentiment analysis pipeline based on DistilBERT from the Hugging Face Transformers library to assign initial sentiment labels ("POSITIVE" or "NEGATIVE") to each review.

- **Fine-Tuning BERT**

  Using the initial sentiment labels, we fine-tuned a BERT model (bert-base-uncased) for binary sentiment classification. Leveraging BERT's transformer architecture, which utilizes self-attention mechanisms, enabled the model to effectively capture the contextual nuances of language, making it highly suitable for sentiment analysis tasks.

4.2 Aspect-Based Sentiment Analysis (ASBA)

To determine the aspects of the stay that influenced sentiment

- **Word2Vec for Embeddings**

  We trained a Word2Vec model using the tokenized reviews from the dataset to generate vector representations for each word. This process involves learning numerical embeddings that capture the semantic relationships and contextual meaning of words based on their co-occurrence patterns within the reviews. These embeddings are essential for identifying patterns and similarities among words, which can be used for downstream tasks such as clustering and aspect extraction.

- **Noun Extraction and Clustering**

  To identify the key topics or aspects guests discuss (like cleanliness, location, or price), we focused on nouns in the reviews. Nouns are words that name things (e.g., "location," "host"). We extracted these nouns and used a method called KMeans clustering to group similar words together. For example, words like "transportation," "subway," and "location" might fall into one group, which we labeled as "Location and Transportation." This grouping helped us organize the feedback into meaningful categories.

- **Aspect Assignment to Reviews**

  Once we had these groups (or aspects), we looked at each review to see which group of nouns it mentioned. For example, if a review included words related to "cleanliness" or "location," we tagged it with those aspects. This allowed us to connect specific feedback to the relevant aspects of the stay, enabling a detailed analysis of what guests liked or disliked about particular features of their experience.
  This process helps us break down the reviews into actionable insights, making it easier to understand which areas are praised or need improvement.

4.3 Text Summarization using T5 Architecture

To generate concise and meaningful summaries of guest reviews, we utilized a T5 model (t5-large). T5 is a transformer-based model that treats all Natural Language Processing (NLP) tasks as text-to-text problems. This means that it can take input text (e.g., a review) and generate output text (e.g., a summary) in a structured and coherent way.

For reviews that were too long, we divided the text into smaller sections, or "chunks," and used the T5 model to summarize each chunk individually. These intermediate summaries were then combined to create a single cohesive summary for the entire review.

The final summaries provide an easy-to-digest overview of the key themes discussed in the reviews, such as location, host quality, neighborhood atmosphere, and cleanliness. This process ensures that the most important aspects of guest feedback are captured in a concise and readable format, making it easier for hosts to quickly identify strengths and areas for improvement in their listings.

5 Experimental Setup

5.1 Data preprocessing

- **Language Detection**

  Used the langdetect library to filter and retain only English-language reviews.

- **Text Cleaning**

  Expanded contractions, removed URLs, emojis (converted to descriptive text), mentions, hashtags, and special characters for cleaner analysis.

- **Tokenization and Lemmatization**

  Employed NLTK to tokenize and lemmatize words, eliminating common stopwords while preserving negations and intensifiers to ensure sentiment accuracy.

5.2 Train-Validation Split

To prepare the data for model training and evaluation, we divided it into two parts: 80% for training the model and 20% for validating its performance. During this process, we ensured that the split maintained the balance between positive and negative sentiment labels. This approach, called stratified splitting, helps prevent any bias in the data distribution and ensures that both classes are proportionately represented in both the training and validation sets.

5.3 Model Training

- **Sentiment Analysis**

  We fine-tuned a BERT model for sentiment classification over two training cycles, known as epochs. During each epoch, we closely monitored evaluation metrics, such as accuracy and F1-score, to ensure the model was learning effectively and to prevent overfitting.

- **Aspect Extraction**

  For extracting key aspects from the reviews, we trained a Word2Vec model with a vector size of 100 on the processed text data. This allowed us to generate word embeddings that captured semantic relationships between words. From these embeddings, we extracted nouns as representative terms for various aspects (e.g., location, cleanliness). Using KMeans clustering, we grouped the extracted nouns into 15 meaningful categories (or aspects) based on their semantic similarity.

- **Summarization**

  To create concise summaries of reviews, we used a T5 model in inference mode. This involved processing large sets of reviews for individual listings and generating text summaries that highlighted key points such as location quality, host interactions, and cleanliness.

5.4 Framework and Hardware

All the experiments were conducted using Python, leveraging libraries such as Hugging Face Transformers (for BERT and T5), NLTK (for text preprocessing), Gensim (for Word2Vec), and scikit-learn (for clustering with KMeans). When available, GPU acceleration was utilized to speed up model training and inference, given the computational intensity of transformer-based models like BERT and T5.

6 Hyperparameter Tuning and Overfitting Prevention

6.1 Hyperparameters

- **Learning Rate**

  We experimented with different learning rates, including 2e-5 and 3e-5, to optimize the model's performance. A learning rate of 2e-5 was selected because it provided stable training and consistently good results, minimizing the risk of divergence or underperformance.

- **Batch Size**

  The batch size, which determines the number of samples processed together during training, was set to 16. This value was chosen to balance GPU memory limitations and ensure efficient training without compromising performance.

- **Number of Epochs**

  The model was trained for 2-3 epochs. Training was stopped early when evaluation metrics (such as accuracy or F1-score) plateaued, ensuring that the model did not overfit the training data.

- **Regularization**

  Regularization techniques were implemented to improve generalization and prevent overfitting. This included applying weight decay to the model's parameters and carefully selecting the maximum sequence length (max_seq_length) to optimize input size while avoiding unnecessary computational overhead. These hyperparameter choices ensured that the model was trained efficiently and produced reliable results.

6.2 Overfitting Prevention

- Early Stopping: Training was halted early when the validation F1-score stopped improving, preventing overfitting to the training data.
- Loss Monitoring: Both training loss and validation loss were closely monitored to ensure the model's performance was balanced and not over-optimized for the training set.
- Data Augmentation: Although data augmentation techniques were considered to enhance model robustness, they were not implemented due to the strong initial performance of the model.

7 Results

7.1 Sentiment Distribution

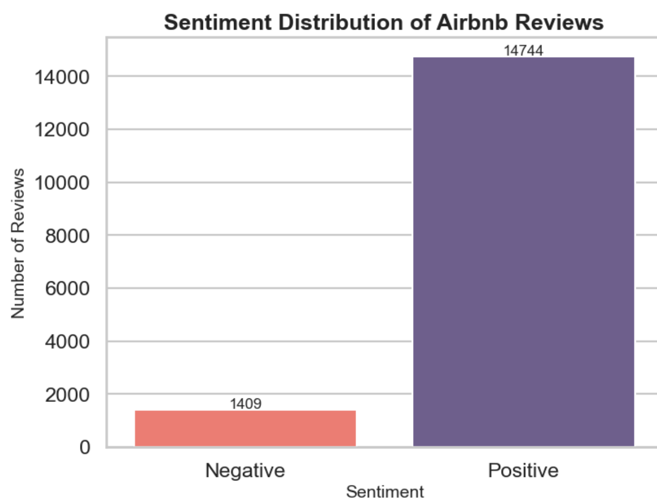

Fig 1. Bar Chart Negative vs Positive Sentiment Analysis

A bar chart comparing the number of positive and negative reviews highlights a strong skew toward positive sentiment, indicating that guests typically have favorable experiences.

Interpretation: Positive reviews greatly outnumber negative ones, reflecting high overall guest satisfaction. This imbalance may also suggest that satisfied customers are more likely to leave feedback.

7.2 Yearly Sentiment Trend

We created a line chart to visualize the number of positive and negative reviews over time. The chart shows that positive reviews consistently dominate throughout the years. While negative reviews are present, their numbers remain relatively low and stable, indicating a consistent trend of guest satisfaction over time. This suggests that the overall quality of experiences provided has been maintained, with fewer fluctuations in negative feedback.

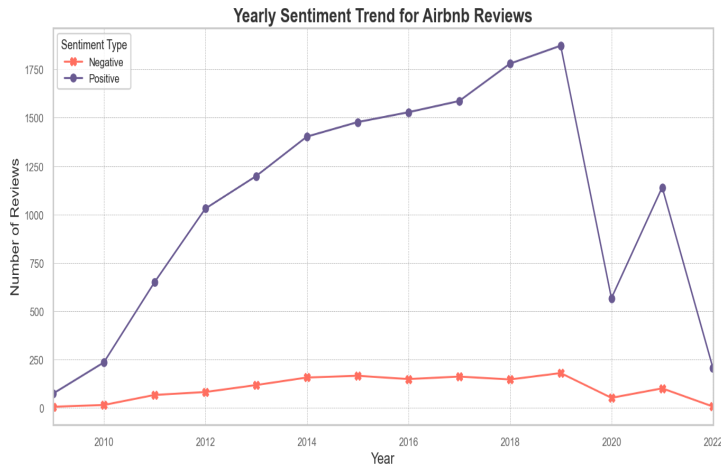

Fig 2. Yearly Sentiment Trend for Airbnb Reviews

Interpretation: Positive sentiment has either increased over time or remained consistently high, which may reflect improvements in service quality or more selective guest acceptance criteria by hosts. On the other hand, while negative sentiment is present, it does not exhibit significant spikes, indicating stable quality standards and a relatively low occurrence of unsatisfactory experiences. This overall trend suggests a steady maintenance of guest satisfaction.

7.3 Top Bi-grams in Positive Reviews

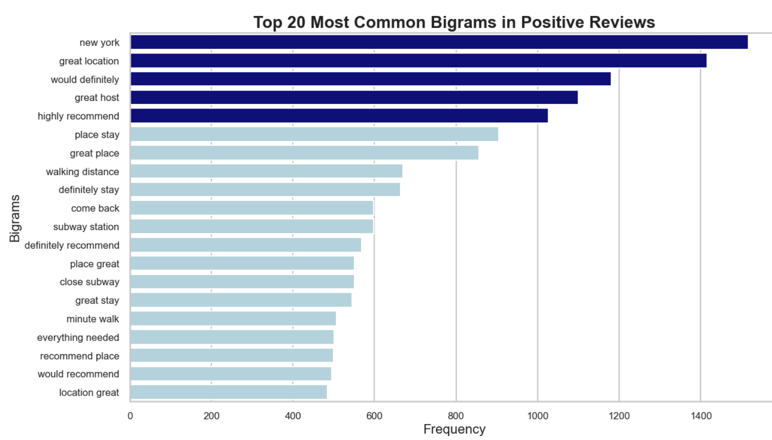

Fig 3. Top 20 Most Common Bi-grams in Positive Reviews

Common bigrams such as “great location,” “great host,” and “highly recommend” frequently appear in positive reviews.

Interpretation: These bigrams indicate that guests often highlight the importance of location and the host’s role in their overall satisfaction. Phrases like “definitely recommend” reflect strong endorsement behavior, suggesting that guests are likely to recommend the property to others based on their positive experiences. This insight underscores the significance of a good host-guest relationship and the property’s strategic location in driving favorable feedback.

7.4 Top Bi-grams in Negative Reviews

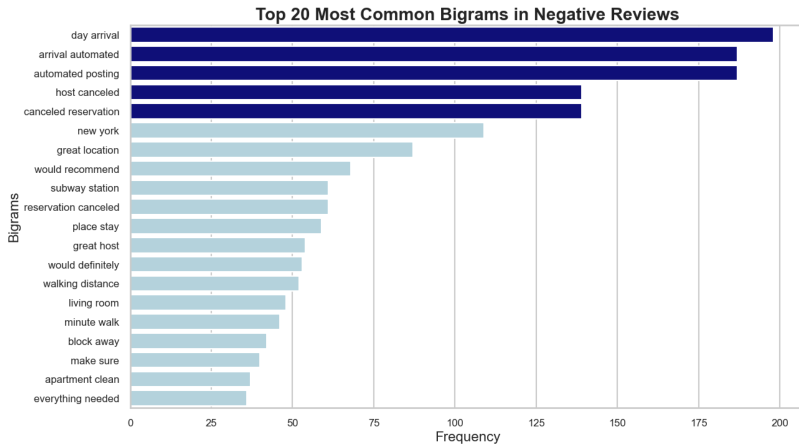

Fig 4. Top 20 Most Common Bi-grams in Positive Reviews

Frequently occurring negative bigrams, such as "host cancelled" and "canceled reservation," highlight dissatisfaction primarily related to booking issues and host reliability.

Interpretation: These bigrams suggest that common negative feedback revolves around cancellations and unmet expectations, which are key areas of concern for guests. Addressing these issues, such as improving booking reliability and setting clear expectations, can help hosts enhance guest satisfaction and reduce negative reviews. This insight emphasizes the importance of consistent communication and reliability in fostering a positive guest experience.

7.5 Word Clouds

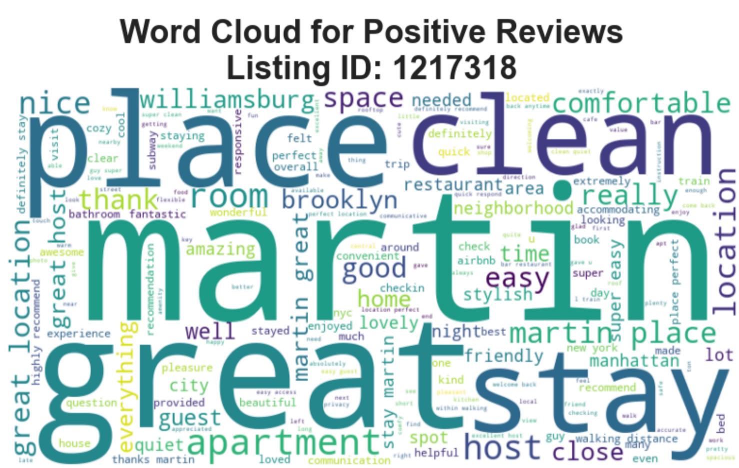

Fig 5. Word Cloud for Positive Reviews

The word cloud for positive reviews highlights frequently used words such as "great," "clean," "comfortable," and mentions of the host's name, indicating recurring themes in guest feedback.

Interpretation: The dominant words in the word cloud suggest that key factors driving guest satisfaction include cleanliness, comfort, and the personal connection or responsiveness of the host. Word clouds offer a quick and intuitive visual representation of what matters most to guests, making it easier for hosts to identify strengths and prioritize these aspects to enhance the overall guest experience.

7.6 Aspect-Based Sentiment (Heatmap)

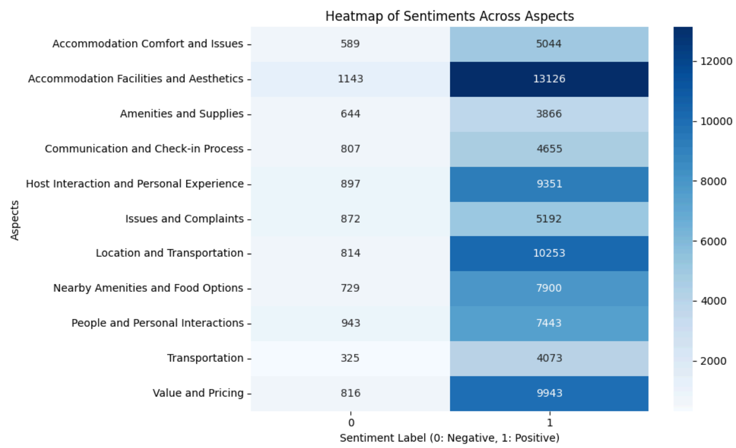

Fig 6. Heatmaps of Sentiments Across Aspects

A heatmap illustrates the relationship between each identified aspect (e.g., "Location and Transportation," "Value and Pricing," "Amenities and Supplies") and the associated positive and negative sentiments in guest reviews.

Interpretation: The heatmap reveals that most aspects, such as "Location" and "Accommodation Facilities," are strongly linked to positive sentiment, indicating guest satisfaction in these areas. However, certain aspects, like "Issues and Complaints," are more frequently associated with negative sentiment. These insights provide hosts with actionable feedback, highlighting strengths to maintain and specific areas, such as resolving guest complaints, that require attention to improve the overall guest experience.

7.7 Radar Chart for a Highly Reviewed

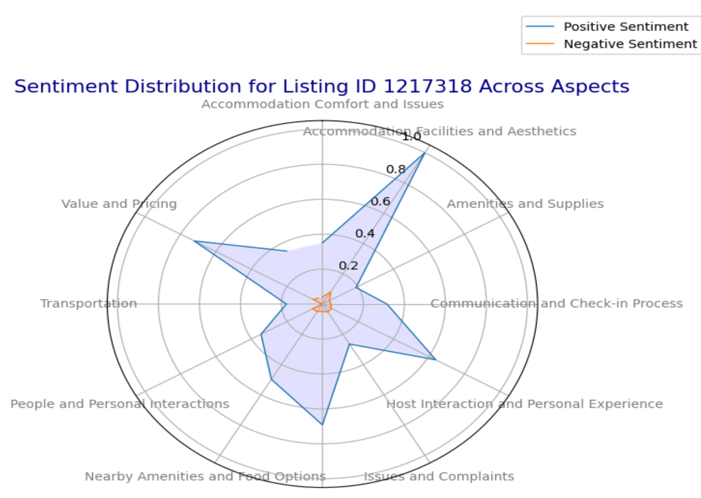

Fig 7. Sentiment Distribution for Listing ID 1217318 Across Aspects

A heatmap illustrates the relationship between each identified aspect (e.g., “Location and Transportation,” “Value and Pricing,” “Amenities and Supplies”) and the associated positive and negative sentiments in guest reviews.

Interpretation: Aspects that extend further along the positive axis of the radar chart, such as “Accommodation Facilities” and “Host Interaction,” represent strengths that guests appreciate. Conversely, aspects where negative sentiment is more pronounced, such as “Value and Pricing,” indicate areas that require improvement. This visualization provides a comprehensive view of the listing’s performance across different aspects, helping hosts identify their strengths and focus on addressing specific areas of concern to enhance guest satisfaction.

7.8 Cumulative Evolution of an Aspect Over Time

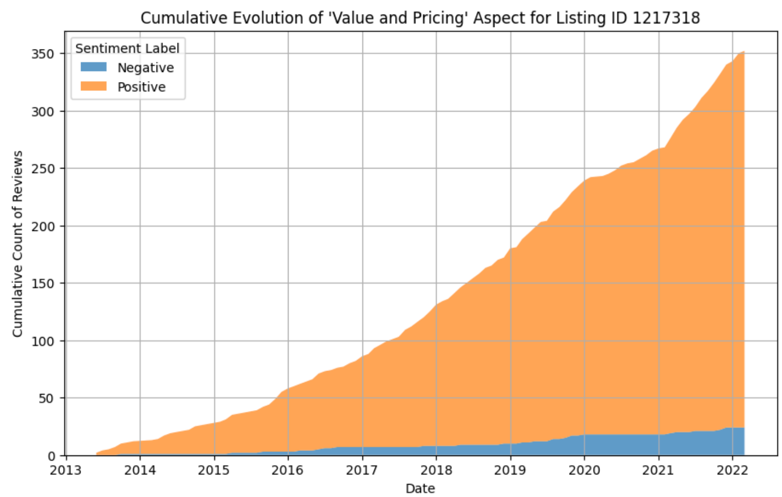

Fig 8. Stacked Area Chart of Cumulative Positive and Negative Reviews Over Time for "Value and Pricing"

A stacked area plot illustrates the cumulative growth of positive and negative sentiments for the “Value and Pricing” aspect. The chart shows that positive sentiment steadily increases over time, while negative sentiment grows at a slower and more stable rate.

Interpretation: The consistent growth in positive sentiment suggests that guests increasingly recognize and appreciate the value offered by the listing. Meanwhile, the relatively low and stable growth of negative sentiment indicates that while some concerns exist, they are not escalating significantly. This trend reflects an overall improvement or sustained satisfaction in the value proposition, highlighting the importance of maintaining competitive pricing and addressing any recurring issues to enhance guest experiences further.

7.9 Summarization Examples

We employed the T5 model to condense hundreds of reviews into a single, concise paragraph for each listing. These summaries emphasized key positive aspects, such as the host’s responsiveness, the convenience of the location, and the overall cleanliness of the property.

Example: *Final Summary for Listing ID 10452*

Angela is a lovely host. She keeps the place clean, responds to messages quickly, and goes out of her way to make sure her guests have everything they need. For $50/night, it's not a terrible value. The location is good, with plenty of stores, bars, and a train station really close that can get you to Manhattan in 15 minutes. I spent one month at Angela's and to be honest, it was great. She is always available and ready to help you in any matter.

Interpretation: Summaries enable guests to quickly assess the strengths and weaknesses of a listing without the need to read through every review. For hosts, these summaries provide valuable insights into how their property is collectively perceived, helping them identify areas of excellence and opportunities for improvement.

8 Summary and Conclusions

We successfully implemented a comprehensive NLP pipeline to derive meaningful insights from Airbnb reviews

- **Sentiment Analysis**

  The sentiment analysis model achieved high accuracy and F1 scores in distinguishing between positive and negative sentiments. It effectively captured the prevalent positive sentiment across reviews while also identifying subtle negative feedback trends. This reliable classification provides a solid foundation for understanding overall guest satisfaction levels.

- **Aspect-Based Sentiment Analysis**

  By extracting nouns from the reviews, grouping them into meaningful aspects using clustering techniques, and linking sentiments to these aspects, we obtained detailed insights. This method pinpointed specific areas of guest satisfaction, such as location, cleanliness, and host interactions, as well as common frustrations like reservation issues and noise disturbances. These fine-grained insights allow hosts to focus on strengths and address recurring challenges effectively.

- **Summarization**

  The T5 model-generated summaries provided concise, high-level overviews of guest feedback, highlighting key themes such as host responsiveness, property cleanliness, and location convenience. These summaries save time for future guests by offering a quick snapshot of the listing's overall quality. For hosts, they serve as an efficient tool for understanding guest sentiment and identifying areas for improvement without the need to read through every review.

9 What We Learned

- Airbnb guests generally provide positive feedback.
- Positive discussions focus on aspects like location, hosts, and facilities.
- Negative feedback, though less frequent, often involves cancellations and reservation issues.
- Summarization tools effectively condense large volumes of reviews into actionable insights.

10 Future Work

- Address class imbalance by utilizing advanced techniques or increasing the number of negative samples.
- Explore more advanced aspect identification methods, such as dependency parsing or topic modeling.
- Further fine-tune the summarization model using domain-specific data for more detailed and nuanced summaries.
- Fake Review detection using Isolationforest
- Topic modelling

References

- **Dataset:** NYC Airbnb Reviews 2021 (Kaggle)
- **BERT:** Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers. *NAACL-HLT.*
- **T5:** Raffel, C. et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR.*
- **Word2Vec:** Mikolov, T. et al. (2013). Distributed Representations of Words and Phrases. *NeurIPS.*
- **Hugging Face Transformers:** https://huggingface.co/transformers/
- **NLTK:** https://www.nltk.org/
- **Langdetect:** https://pypi.org/project/langdetect/
- **Contractions:** https://pypi.org/project/contractions/