USC University of Southern California

Project Report - DSO 560

Team: Sentiment Synthesizers

SENTIMENT ANALYSIS ON AIRBNB REVIEWS

Avantika Gargya, Basava Satish Chandra Velagapudi, Krisha Gandhi, Kenneth Fung, Sai Bharadwaj Kalyandurg

Index

1. Introduction
2. Dataset Overview
3. Data Preprocessing
   - 3.1. Handling Missing Values
   - 3.2. Language Detection
   - 3.3. Language Distribution & Filtering
   - 3.4. Language Translation for Sample
   - 3.5. Text Preprocessing for Sentiment Analysis
4. Sentimental Analysis
5. Feature & Key Themes Extraction
6. Results
   - 6.1. Top Features - Sentiment Category
   - 6.2. Top Features - N Grams
   - 6.3. Top Features - Descending Order of TF-IDF Score
   - 6.4. LDA for Topic Modeling
   - 6.5. Theme Extraction across Unique Listings
7. Project Limitations
8. Appendix
   - 8.1. Listing wise Visual Analysis
   - 8.2. Top Tri-Gram Features by Borough
   - 8.3. LDA Analysis

Colab link:

NLP_Project_file_Sentiment Synthesizers_AIRBNB_Group 4.ipynb

1. Introduction

Problem Statement: *Advanced NLP-Driven Insights from Airbnb Reviews for Neighborhood & Listing Centric Recommendations*

Modern businesses rely on reviews and comments to refine their services (hotels, restaurants etc) and enhance their offerings. For this project, we leveraged Airbnb's extensive multilingual property data, focusing on two neighborhoods in New York (Brooklyn and Manhattan). Although models can compute sentiment scores, these same reviews can offer hosts a more nuanced view of guest experiences, neighborhood dynamics, and property-specific details. Our aim is to employ text analytics to transform reviews into actionable insights, guiding neighborhood and listing-centric recommendations that benefit both guests and hosts.

Our goals are twofold: first, to build a pipeline that cleans and translates the raw data; second, to apply sophisticated text analytics that reveal deeper patterns. During this process, we encountered several challenges, including translating multilingual content, context-specific text cleaning, and dealing with missing data.

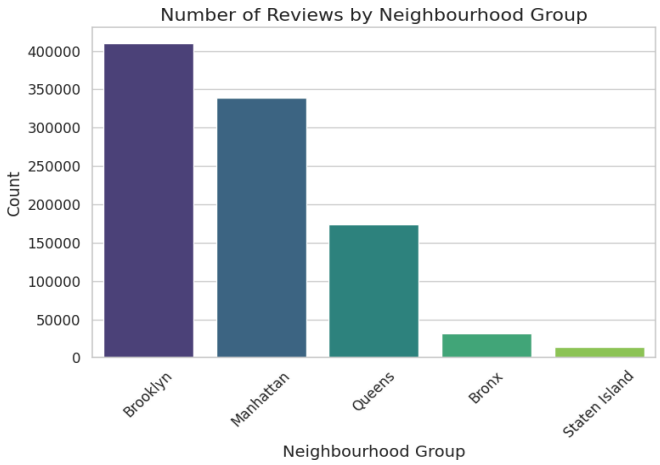
The workflow starts with Data Ingestion, where we consolidate multiple CSV files - encompassing reviews, listings, and neighborhood information - into a single dataset. We then rectify missing values and ensure consistency among reviews. We then perform sampling, language detection via langdetect and translation via google translator, this is then followed by tokenization, stopword removal, and lemmatization. This comprehensive approach yields a cleaned corpus for advanced text analytics. We also tested a perplexity-based fake review detector as a potential pipeline addition but it incorrectly labeled legitimate reviews, jeopardizing valuable insights.

Next, we apply Sentiment Analysis using TextBlob, to calculate polarity and subjectivity scores classifying each review as Positive, Negative, or Neutral. Finally, Theme Extraction via TF-IDF and n-grams subsequently identifies core themes in each classification (from proximity to public transit to overall cleanliness) enabling more targeted recommendations for hosts. This analysis is done for both neighborhood and listing levels.

2. Dataset Overview

We used Airbnb's official data available at: https://insideairbnb.com/get-the-data/
In terms of scope, we selected New York City for our project which had quarterly data available up to January 3, 2025. New York City has 5 boroughs - Manhattan, Brooklyn, Bronx, Staten Island and Queens. Each of these boroughs have a distinct character and resident demographics.

1. Manhattan - As the center of business, wealth and culture in NYC, and as the home of Wall Street, Broadway and tourist attractions, the footfall is more of business travellers.



2. Brooklyn – This neighbourhood is considered "trendy, artsy, and gentrified. Gentrification has recently increased property prices, it's popular for those looking for a more 'local' experience.

3. Queens – The most diverse, family-oriented, and underrated borough famous for housing immigrant communities. It is less popular for tourists, but more popular for long-term travellers.

4. The Bronx – Bronx is historically associated with poverty, crime, and urban decay (especially in the '70s and '80s). It is not as popular for short-term tourists due to safety perceptions.

5. Staten Island – This suburb is often considered **the most d**isconnected from NYC's fast-paced lifestyle.

Due to the distinct characteristics and the number of available reviews for Brooklyn and Manhattan, we used these as a focus area for our project.

The dataset is composed of **4 files - listings.csv** (storing host information, neighbourhood info, description etc., **reviews.csv, neighbourhoods.csv and neighbourhoods.geojson.** We tailored our analysis towards analysing the reviews for each neighbourhood, and reviews broken down by listings. Overall, there are about **37.7K listings and 969K reviews** overall.

3. Dataset Pre-processing

3.1. Merging the Dataset

For ease in processing, we merged the reviews.csv file with some relevant information from the listings dataset such as price, room_type, neighbourhood_group. We also removed reviewer information such as name and reviewer id as we felt this was not essential to the analysis.

3.2. Handling Missing Values

Out of our essential variables, only two of them contained missing values - reviews & price:

- Reviews (235 rows): we dropped these rows because it is crucial to proceeding
- Price (19K rows): we imputed with the median price for a listing for that neighbourhood

3.3. Neighbourhood Division

Next we divided the data into separate dataframes for the 5 boroughs - Manhattan, Brooklyn, Queens, Bronx & Staten Island. For the purpose of analysis, we selected Brooklyn and Manhattan, as these boroughs have the highest number of listings and reviews. Additionally, they share similar urban characteristics, making comparative analysis more meaningful.

3.4. Language Detection

Language detection plays a crucial role in our project, as Airbnb reviews come from a diverse set of guests worldwide, leading to a multilingual dataset. Applying the ***langdetect library*** across all reviews, we identified ~30 different languages, with English, Spanish & French being the most prevalent. Our final dataset consisted of: Brooklyn: ~389k reviews | Manhattan: ~317k reviews

3.5. Language Distribution and Filtering

To streamline our dataset, we performed a distribution analysis of detected languages and applied a filtering criterion. Only languages with over 5,000 reviews were retained, which further refined our dataset to ***English***, ***Spanish***, and ***French*** while maintaining computational feasibility. To enable a granular analysis of guest sentiment at the listing level, we created a structured subsample with an equal number of reviews per listing to accommodate the computational constraints. This approach allows hosts to identify key sentiment trends across their properties.

Sampling Criteria:

- We filtered out listings with fewer than 100 reviews to ensure statistical reliability
- 1111 high-review listings were identified in Brooklyn and 790 in Manhattan
- To create a balanced dataset that is computationally affordable, we sampled 46 reviews per listing, resulting in ~51k reviews for analysis of Brooklyn data. Similarly, we sampled 65 reviews per listing, resulting in ~51k reviews for analysis of Manhattan data

This structured subsample allows us to conduct listing-level sentiment analysis, identifying recurring themes and patterns in guest experiences across different properties.

3.6. Language Translation for Sample

Translation of non-English reviews was necessary to ensure uniform sentiment analysis. We utilized the ***Google Translator API*** to translate French and Spanish reviews into English. This step ensured that all reviews were in a common language, allowing us to conduct sentiment analysis without linguistic inconsistencies. To ensure efficiency we applied the translation function only to non-English reviews. The process was automated using vectorized operations to handle large-scale text translation. Error handling mechanisms were implemented.

3.7. Text Preprocessing for Sentiment Analysis

Text preprocessing is a crucial step in Natural Language Processing(NLP) to clean & standardize textual data, removing unnecessary elements while retaining meaningful information. In this project, we preprocess the translated English reviews to enhance the quality of data for sentiment analysis and thematic exploration.

3.7.1. Text Preprocessing Steps

1. Setup and Initialization

- Lemmatization: We initialize WordNetLemmatizer to reduce words to their root form (e.g., "running" → "run")
- Dictionary Validation: We use multiple English dictionaries (US, UK, Canadian, Australian, New Zealand, and South African variants) to validate words
- Stopword Management: We load standard English stopwords from NLTK but retain essential stopwords like "not," "no," and "against" to preserve sentiment meaning

2. Text Cleaning and Normalization

- Convert to lowercase: Convert text to lowercase to ensure uniformity
- Tokenization: Split text into individual words
- Stopword Removal: Remove the above updated unimportant stopwords
- Apostrophe Normalization: Standardize apostrophes and merge possessive forms
- Punctuation Removal: Retain only letters, digits, spaces, apostrophes to maintain clarity

3. Token Filtering and Lemmatization

- Validate words using the English dictionaries & retain only those recognized in at least one variant
- Apply lemmatization to convert words to their base forms, reducing dimensionality while preserving meaning

4. Reconstructing the Processed Text

After filtering & lemmatization, tokens are rejoined into structured text for further NLP processing. This ensures clean text maintaining semantic integrity while eliminating noise.

3.7.2. Implementation on Brooklyn & Manhattan Reviews

After defining the preprocessing function, we apply it separately to the Brooklyn and Manhattan datasets. The final cleaned dataset is now ready for sentiment analysis and further NLP modeling. This structured approach ensures that only meaningful text is retained, enhancing the quality of insights derived from the data through sentimental analysis.

4. Sentiment Analysis

Now that we had the cleaned reviews column in our dataframe, we proceeded to perform sentiment analysis using the Textblob library. Sentiment analysis is the process of determining the emotion of a given text whether it is positive or negative, or neutral. The sentiment function of TextBlob helps us understand the underlying emotion of a text by returning two properties, polarity, and subjectivity. Polarity returns a float value that is in the range of [-1,1] where 1 means a positive statement and -1 means a negative statement. Subjectivity is also a float value that is in the range of [0,1]. 0 means a personal opinion and 1 means factual information.

To categorize reviews with an appropriate sentiment label, we defined the following ranges: If the polarity score is <0, we tagged it as a negative review and if >=0, it was classified as positive. For subjectivity score, we defined the following ranges: If <0.3, labelled as objective, if <=0.7, mixed, else subjective.

We calculated sentiment scores and assigned labels at individual review level across Manhattan and Brooklyn for a neighbourhood analysis- so what patterns do we see across listings in the two neighbourhoods as well as at listings' level, where we provided mean scores over unique listings. This would help hosts understand that on average, what is the underlying sentiment expressed by guests through reviews pertaining to their property.

5. Feature and Key Themes Extraction

Next, we decided to take this further so that the stakeholders (hosts and guests) would not only gain insight on the average sentiment but also be able to identify key themes/features that stood out from the reviews.
To do this, we tried the TF IDF (unigram) as well as N-grams feature extraction approach, first across both neighbourhoods (Manhattan and Brooklyn), where we regrouped the sub-datasets into positive and negative reviews based on the labels assigned in the sentiment analysis step, and then taking all the reviews in these dfs as the corpus, we identified "key features" for positive and negative reviews and sorted them in descending order based on the TF IDF scores.
Then for the negative reviews, we additionally did **Latent Dirichlet Allocation (LDA) modelling.** Given the key themes that were highlighted across negative reviews, this method allowed us to group features to help hosts identify broader topics where improvements can be made across all listings in Brooklyn/Manhattan, such as improvement in amenities and comfort.

Next, we performed key theme extraction after grouping by listings as well, so taking all reviews for a unique listing as a mini-corpus. For each listing, we extract key phrases using TF-IDF to highlight guest experiences, summarizing reviews for quick insights. This helps hosts understand what guests love and what needs improvement. By identifying unique selling points like

"amazing rooftop view" or *"quiet and cozy"*, hosts can optimize their listing descriptions to attract more bookings or charge higher prices. At the same time, spotting recurring complaints such as *"slow Wi-Fi"* or *"complicated check-in"* allows them to address issues and improve guest satisfaction. This data-driven approach helps hosts refine pricing, stay competitive, and set clear expectations, ultimately leading to better reviews, higher ratings, and increased bookings.

For Airbnb guests, TF-IDF analysis makes reviews more useful by summarizing common feedback, allowing them to quickly identify key phrases without reading hundreds of reviews. This helps guests find listings that match their needs, avoid common pitfalls like *"noisy street"* or *"small bathroom"*, and trust reviews based on real guest experiences rather than just star ratings. By making booking decisions more informed and efficient, this analysis ensures guests have a smoother and more satisfying stay.

6. Results

Guests are generally satisfied with the accommodations rented in Brooklyn and Manhattan as the reviews in the 2 boroughs are overwhelmingly positive, with average sentiment scores of 0.4141 and 0.4086, respectively, on a scale of -1 to 1. The counts of review sentiment are as follows.

sentiment_label	count_brooklyn	count_manhattan
Positive	50548	50502
Negative	558	848

Table 1. Counts of review sentiment

Using the sentiment scores, we can retrieve the average scores for each listing. For example, the sentiment score for the Brooklyn listing with ID 6848 is 0.3623 and that for the Manhattan listing with ID 6990 is 0.404; both are categorized as 'Positive' (Appendix I).

After analyzing the processed reviews, we identified the 100 most relevant 3-word phrases with relevance weightings for positive and negative reviews respectively using word clouds (Appendix II). Across all listings in Brooklyn and Manhattan, the most relevant phrase for positive reviews is 'in great location' and its average TF-IDF scores are 0.0136 and 0.0179, respectively. As for negative reviews, the most relevant phrases are 'within walking distance' and 'worst experience ever'.

To help hosts understand the features that drive average sentiment scores and to help guests select the best listing, we analyze at listing level as well. For the listing with ID 6848, the most

relevant features are ‘great location’, ‘comprehensive amenities’, and ‘comfort’. The host can include these key features in their description to attract more guests, allowing guests to quickly determine if the listing meets their requirements.

The LDA model identified 5 main themes using bi-grams from negative reviews across listings in the 2 neighborhoods (Appendix III): location, check-in experience, bathroom, cleanliness, and maintenance. If Airbnb hosts could provide a smooth check-in process and maintain a clean environment, especially for bathrooms, they would address most guests’ concerns and reduce the likelihood of getting negative reviews. Compared to listings in Brooklyn, hosts with Manhattan listings should pay more attention to providing better customer service to guests

7. Limitations

Our analysis could be improved by using the entire dataset, as we only randomly selected 46 and 65 reviews from each listing in Brooklyn and Manhattan due to limited time and computational resource constraints. In addition, our analysis would be more accurate if we could identify and filter out fake reviews without existing labels. Additionally, the insights could be more comprehensive by identifying patterns in sentiments and topics from reviews based on the cultural backgrounds of guests, which would help hosts to cater to specific needs. Moreover, we could analyze reviews from specific times of the year to help hosts prepare in advance for a better hospitality experience, increasing the likelihood of guests returning to Airbnb.

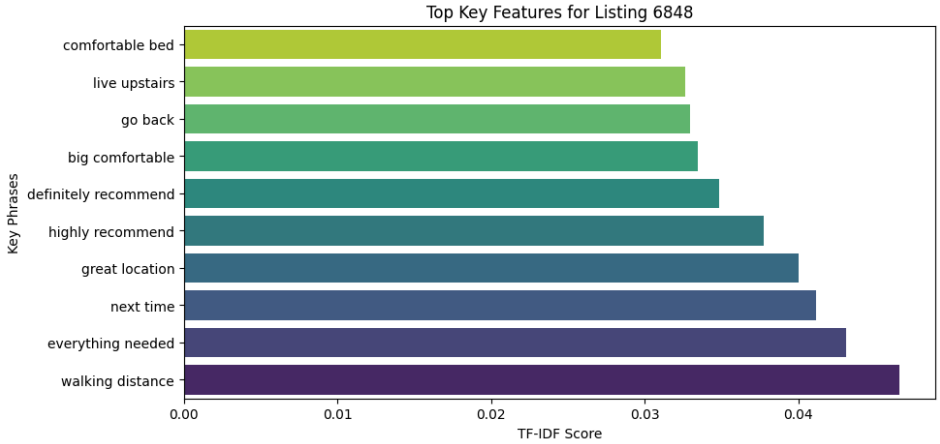
CONTRIBUTIONS TABLE -

Avantika Gargya	Project Skeleton setup, Dataset pre-processing, LDA, EDA, visualizations, wordcloud, Iterations of overall code with parameter tuning
Sai Bharadwaj Kalyandurg	Fake News Classifier, Data Pre-Processing, LDA, troubleshoot coding issues
Ki On Kenneth Fung	Project idea alignment. Price-sentiment label correlation graph, Iterations of overall code with edits, troubleshoot coding issues
Krisha Gandhi	Project Concept Planning, Literature Review, Sampling iterations, Brooklyn Code File (Neighbourhood+Listings Analysis), Sentiment Analysis, Theme Modelling (TF-IDF, N-grams), WordCloud+TF-IDF Bar Chart Visualizations, Code iterations for optimizing output
Basava Satish Chandra Velagapudi	Project Concept Refinement, Data Pre-processing, Manhattan Code (Language Translation, Sentiment Analysis, TF-IDF, N-grams), Dictionary Optimization, Stopword Tuning, Code Debugging & Output Enhancement

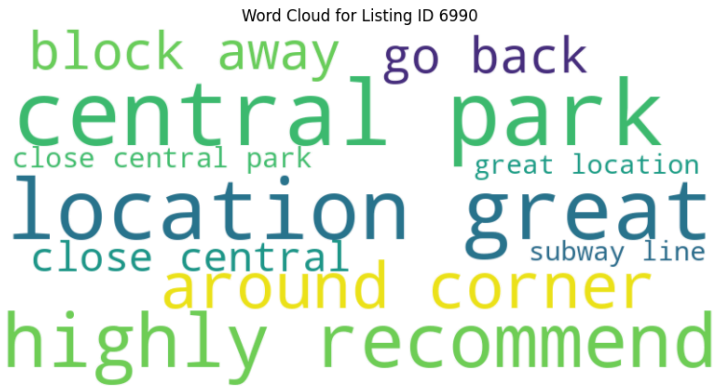
8. Appendix

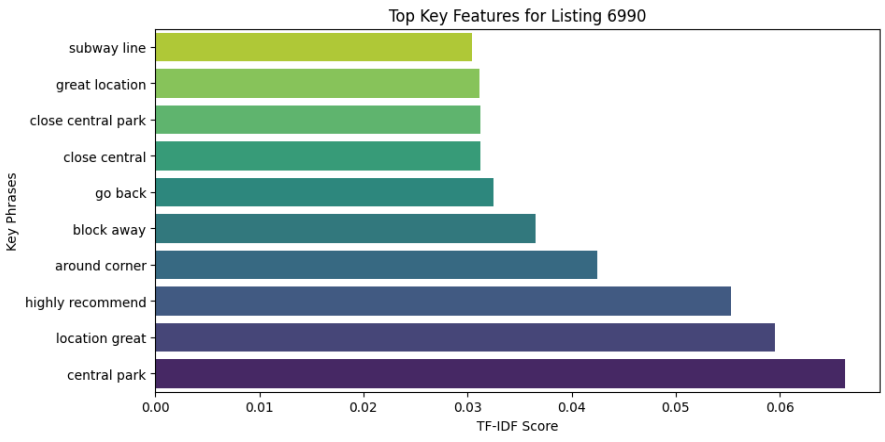
8.1. Appendix I

- Word Cloud for listing with ID = 6848:



- Word Cloud for listing with ID = 6990:





8.2. Appendix II:

- Top 100 tri-gram features by each borough's **positive** reviews, in tables with weightings and word clouds

Tri-gram phrases for Brooklyn listings	
feature	**mean_tfidf**
in great location	0.013591
within walking distance	0.010521
home away home	0.007245
highly recommend staying	0.006475
definitely book again	0.006214
...	...
above beyond make	0.001562
out way make	0.001528
definitely again next	0.001497
location easy access	0.001435
restaurant within walking	0.001406

Tri-gram phrases for Manhattan listings	
feature	**mean_tfidf**
in great location	0.017892
within walking distance	0.010157
great location great	0.006831
close central park	0.006270
easy check in	0.005751
...	...
clean in great	0.001703
distance central park	0.001695
walk subway station	0.001654
sure everything needed	0.001641
few minute walk	0.001585

Word Cloud for **Brooklyn** listings:

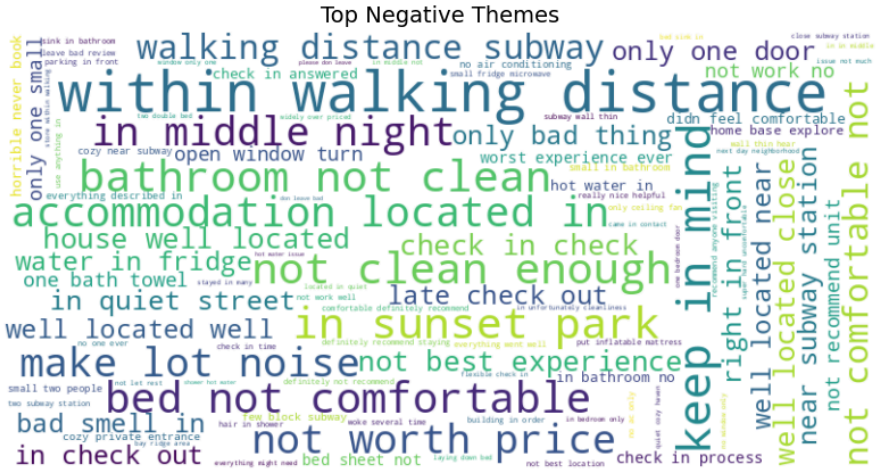
Word Cloud for **Manhattan** listings:

- Top 100 tri-gram features by each borough's **negative** reviews, in tables with weightings and word clouds

Tri-gram phrases for Brooklyn listings	
feature	**mean_tfidf**
within walking distance	0.009175
keep in mind	0.005817
in sunset park	0.005376
bathroom not clean	0.005376
not clean enough	0.005376
...	...
came in contact	0.001493
two double bed	0.001493
don leave bad	0.001307
hot water issue	0.001203
shower hot water	0.001203

Tri-gram phrases for Manhattan listings	
feature	**mean_tfidf**
worst experience ever	0.010547
not worth price	0.006691
within walking distance	0.005896
few block away	0.004717
front desk staff	0.004427
...	...
late check out	0.004312
walk time square	0.004099
not worth money	0.004092
not disturb sign	0.003636
not good experience	0.003538

Word Cloud for **Brooklyn** listings:



Word Cloud for **Manhattan** listings:

8.3. Appendix III:

- **LDA Topic Modelling for Brooklyn listings' negative reviews:**
  - **Topic 1**: ['located in', 'not best', 'walk subway', 'not comfortable', 'close subway', 'first night', 'few day', 'not clean', 'late night', 'check in']
  - **Topic 2**: ['in life', 'in quiet', 'not clean', 'worst experience', 'air conditioning', 'in front', 'check in', 'in bathroom', 'well located', 'not recommend']
  - **Topic 3**: ['in kitchen', 'check out', 'not work', 'subway station', 'near subway', 'walking distance', 'not sure', 'check in', 'in bathroom', 'not clean']
- **LDA Topic Modelling for Manhattan listings' negative reviews:**
  - **Topic 1**: ['not recommend', 'subway station', 'in bathroom', 'in front', 'in kitchen', 'not work', 'time square', 'not good', 'front desk', 'check in']
  - **Topic 2**: ['no one', 'few block', 'come back', 'late night', 'didn work', 'front desk', 'in bathroom', 'check out', 'well located', 'check in']
  - **Topic 3**: ['next door', 'not even', 'worst experience', 'in middle', 'not recommend', 'in building', 'location great', 'hot water', 'not clean', 'in bathroom']