Sequence analysis in text mining: how does it work among reviews?

An extended research project report submitted to the University of Manchester for the degree of MSC. DATA SCIENCE in the School of Social Sciences

2024

Student ID: 11349153

Table of Contents

WORD COUNT – 6982/7500

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In today's world, social media plays a crucial role in customer experience management. For service providers like Airbnb, evaluating customer feedback is essential for improving service quality and enhancing customer satisfaction. The primary objective of this research is to trace the shifts in sentiments when a customer normally reviews a listing, that is, whether they tend to start with a positive tone and then shift towards complaints/negative or vice-versa. Specifically, this research involves listings and host data from Airbnb, based in Manchester, which can yield targeted insights for service improvement in this region.

By analysing the order, frequency, and sentiment score of aspects, the research identifies key drivers of positive and negative reviews and assesses their impact on the overall customer experience. Findings show that certain aspects, like amenities and environment, are frequently mentioned with varying sentiment polarity, highlighting areas where improvements are needed to enhance guest satisfaction. Additionally, the study offers insights into how the sequence of aspect mentions within a review can influence the overall sentiment and perception of a listing. These results provide actionable insights for Airbnb hosts to prioritize improvements in certain areas, ultimately aiming to improve guest satisfaction and maintain high service standards in a competitive market.

Keywords- *NLP, SENTIMENT ANALYSIS, ASPECT DETECTION, AIRBNB*

ACKNOWLEDGEMENT

I would like to express my gratitude to my supervisor, Dr. Azar Shahgholian, for her unwavering support, constant encouragement, and excellent guidance throughout this project. I am fortunate to have attended her meetings and gained valuable knowledge from her.

I also extend my special thanks to Dr. Riza Batista Navarro, my Text Mining Professor, whose lectures on Natural Language Processing (NLP) have been immensely beneficial.

Finally, I am profoundly grateful to my family back in India for their continuous support in every aspect while I was abroad.

DECLARATION

No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

INTELLECTUAL PROPERTY STATEMENT

i. The author of this extended research project report (including any appendices and/or schedules to this report) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this report, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has entered into. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks, and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the report, for example graphs and tables ("Reproductions"), which may be described in this report, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this report, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see https://documents.manchester.ac.uk/display.aspx?DocID=24420), in any relevant dissertation restriction declarations deposited in the University Library, The University Library's regulations (see https://www.library.manchester.ac.uk/about/regulations/) and in The University's Guidance for the Presentation of dissertations.

1.INTRODUCTION

1.1 Background

Artificial Intelligence (AI) is transforming the whole world, and travel and tourism industry is not an exception. It has brought significant advancements in terms of operational efficiency, customer experience, and service optimization. Within the bigger picture of AI, **sentiment analysis** and **aspect detection** are powerful tools for understanding customer opinions and feedback. **Sentiment analysis** involves the identification and categorization of sentiments expressed in text, typically classifying them as positive, negative, or neutral. **Aspect detection**, on the other hand, is an advanced approach to that which helps in identifying specific elements (or aspects) of a product or service mentioned in the text, such as "location" or "cleanliness" in a review, and analysing the sentiment associated with each aspect. Together, these techniques enable businesses to gain deeper insights into customer satisfaction and identify areas for improvement.

Moreover, the evolution of the Internet has led to the creation of new business models, the attraction of foreign capital through tourist offerings, and the reinvention of traditional models. The hospitality sector is a critical revenue base, and within the sharing economy, Airbnb stands out as a digital platform connecting guests with hosts, facilitating accommodation without owning any property. This platform introduces a new way of lodging, offering tourists personalized experiences, interaction with locals, and access to unique accommodations (Yiannakou, et al., 2022)

Airbnb stands out as a testament to how AI and digital platforms are reshaping the tourism industry. The platform's success is not only due to its innovative business model but also its effective use of AI to utilize vast amounts of data generated via user interactions. These capabilities enable businesses to maximize profit, reduce costs, and create targeted marketing campaigns that resonate more effectively with consumers. Below figure shows how AI can prove to be useful with social media data.

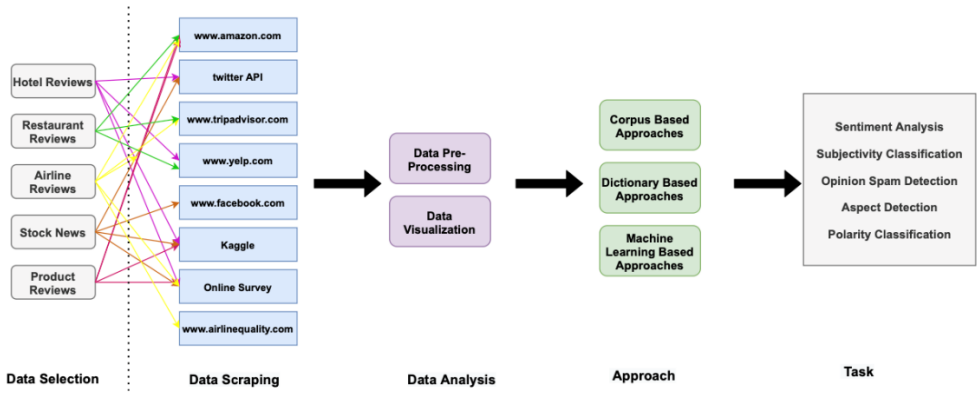


Figure 1 Use of AI in social media data (Wankhade., 2022)

Now that we know the key applications of AI in this industry, it's time to understand how these innovations are implemented and the working principles behind them. One such application of AI and data is discussed in detail in this report. Here, we have analysed customer feedback data from the insideAirbnb website specifically for the city of Manchester and drawn inferences from it. This includes uncovering hidden patterns from the data and using those insights to help businesses evaluate their current market scenario and suggest steps for future improvements. All of these were done using NLP (Natural Language Processing) techniques like text mining, aspect-based sentiment analysis and sequence mining. By utilizing these, businesses can automatically process and interpret customer reviews, social media comments, and other textual data to extract meaningful insights. Understanding the sequence of comments is also crucial because it can reveal deeper insights about customer experience. For example, a review that begins with positive comments about the location but ends with negative remarks about the cleanliness might indicate a potential area for improvement for the host. This research is based on the broader context of customer experience management, a field that has gained significant value in recent years. In the context of Airbnb, the customer experience starts from the booking process to the stay itself and finally to the post-stay review. Each of these stages provide opportunities for the host and the platform to enhance the overall experience.

Although the integration of AI in hospitality offers multiple benefits, there are some drawbacks as well. Issues such as data privacy, lack of significant investment, and potential displacement of jobs needs serious consideration. To be precise, the industry must ensure that data and AI systems are used responsibly, and ethical considerations are kept in mind.

1.2 SHAPE AND SCOPE OF THE STUDY

1.2.1 Purpose of the Study

Customer reviews are an important component of the service industry, particularly on platforms like Airbnb, where they serve as a powerful social statistic tool. Reviews shared on platforms like this significantly influence potential customers' decisions and help shape their expectations. By going through reviews from previous guests, customers can get an idea about the quality of service offered. Positive reviews can greatly enhance a host's reputation, making their listing more lucrative to future guests (Luca, 2016). On the other hand, negative reviews may deter potential customers, leading to fewer bookings (Chevalier, 2006). These dynamics display the importance of maintaining high service standards to ensure favourable reviews, which can be crucial in such a competitive market.

Customer reviews also highlight areas that need improvement. Constructive criticism helps hosts to identify and address issues such as location, environment, and communication, which might lead to enhanced guest experiences and, ultimately, improve reviews in the future (Levy, 2013). Reviews also help in a listing's visibility and search rankings on Airbnb. Listings with higher ratings and more positive reviews are often prioritized in search results, creating a feedback loop where good reviews lead to more bookings, which, in turn, generate more reviews (Anderson, 2012). Additionally, reviews give a realistic overview to future guests by providing detailed insights into the strengths and weaknesses of a listing (Vermeulen, 2009). Those insights allow potential guests to make informed decisions based on their specific preferences, whether that be location, amenities, or host interaction.

The aim of this research is to explore and comprehend such dynamics within customer reviews, with a particular focus on the sequence of emotional expressions and their connection to various aspects of Airbnb listings. This research will bridge the gap between existing literature, which is mostly based on sentiment analysis, opinion mining, and predicting future bookings based on reviews (Pang, 2008). These aspects will be discussed in detail in the literature review section. However, unlike other works, this research is not solely based on predicting or calculating sentiments but rather focuses on analysing aspects and sentiments combined, along with the sequence involved. This research builds on existing methodologies, integrates certain

features from those approaches, and emphasizes aspect priority and sentiments involved to provide further tailored analysis (Liu, 2012).

1.2.2 Issues Addressed

The research will address the following broad questions:

Sentiment Sequence Patterns (RQ1): Is there a discernible pattern in the way reviewers express their emotions over the course of their reviews? For instance, do reviews typically start with positive comments and then shift to negative ones, or is there a different common sequence?

Aspect Prioritization (RQ2): Can the sequence analysis of review comments reveal the priority or importance of various aspects of Airbnb listings (e.g., cleanliness, location, host interaction) as perceived by customers?

In addition to these broad questions, the research will also involve various listing-level analytics to provide a comprehensive overview of customer feedback. These analyses include:

- Filtering Based on Positive and Negative Reviews: Identifying listings with predominantly positive or negative reviews to understand the underlying reasons.
- Identifying Most Used Positive Words: Analysing common positive words to highlight aspects that customers appreciate the most.
- Relation of Reviews with Respect to Price and Availability: Investigating how review sentiments correlate with the price and availability of listings.
- Visualisations: Creating visual representations of the data to better understand and communicate findings.

These questions and analyses will be explored using advanced text mining and data analysis techniques, including aspect-based sentiment analysis, to extract meaningful insights from the review data. The findings of this research will contribute to a better understanding of customer experiences and lead to better decision making by hosts for improving Airbnb services and customer satisfaction.

1.3 Structure

This report is systematically designed to explore and analyse customer reviews within the Airbnb platform. It begins with a brief introduction followed by a comprehensive Literature Review section, which provides an overview of natural language processing techniques used in the context of customer reviews and delves deeper into existing methodologies for sentiment analysis, and examines previous studies, particularly within the hospitality industry, to establish a foundation for the research. Next up, the Data section offers a detailed description of the Airbnb review dataset, including its source and key attributes involved, then the pre-processing steps like handling missing values, removing stopwords and text normalization that were essential for preparing the data for analysis. The Methodology section describes about the advanced text mining techniques available, and the methodologies used here, including natural language processing (NLP) and machine learning algorithms, alongside a description of the technique, like VADER, NLTK and opinion mining. The Results section presents the findings from the sentiment and aspect analysis, including implications from this research. The report concludes with a Conclusion section that summarizes the main findings and suggests scope for future research. This structured approach ensures a comprehensive examination of the research questions and provides clear, actionable insights for both academic and practical applications.

2. LITERATURE REVIEW

In the field of customer service management, the analysis of online reviews has become quite popular because of its ability to understand and improve service quality. This section highlights the current methodologies and approaches in the field of aspect detection and sentiment analysis, with a particular focus on their application to online reviews, like those on Airbnb.

2.1 Existing Methodologies

The existing body of research on aspect detection and sentiment analysis employs various techniques to extract meaningful insights from large volumes of textual data. Before digging deep into the methodologies lets first look at what those terminologies mean.

Sentiment Analysis

Sentiment Analysis is the process of analysing textual data and then infer the emotional tone of the data whether its positive, negative, or neutral. In today's world, companies have access to data more so than ever, be it social media data like tweets or reviews, customer emails, conversation with chatbots, all of those if analysed properly can lead to better decision making by the business.

Review ID	Review Text	Sentiment
1	"The hosts were great."	Positive
2	"The kitchen was in a poor condition."	Negative
3	"I love the hotel but not the neighbourhood."	Neutral

Table 1 Sentiment Analysis example

Traditional methods of sentiment analysis include NLP techniques like Lexicon based approaches which is further subdivided into dictionary based and corpus-based approaches. The dictionary-based approach depends on opinion mining, and then searches the dictionary of their synonyms and antonyms. The corpus-based approach begins with a seed list of opinion words, and then finds other opinion words in a large corpus to help in finding opinion words with context specific orientations (Medhat et al., 2014). More advanced approaches include machine learning and deep learning models to analyse sentiment and detect aspects within

reviews. Some of the approaches used for sentiment analysis have been shown in the figure below.

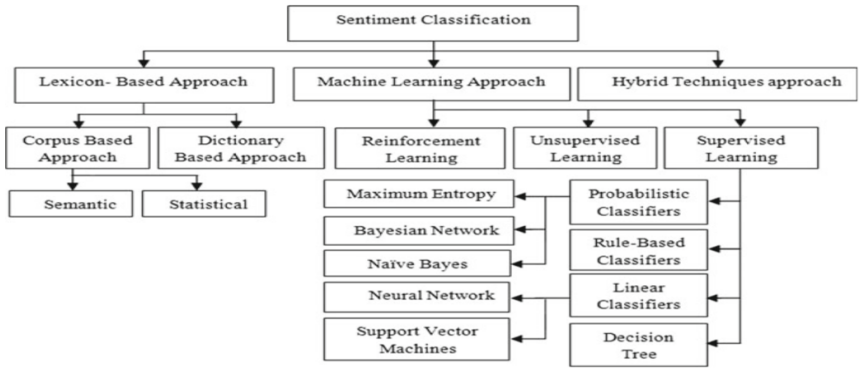


Figure 2 Sentiment Classification Techniques (Aqlan, 2019)

As depicted in above figure, one common approach for sentiment analysis is the use of supervised machine learning algorithms, where models are trained on labelled datasets to classify the sentiment of reviews whether positive, negative, or neutral. These models make use of features like word n-grams, part-of-speech tags, and syntactic dependencies. Popular algorithms in this domain include Support Vector Machines (SVM), Naive Bayes, and various neural network architectures. Most of the prior work in sentiment analysis (Blitzer et al., 2007, Choi et al., 2005, Narayanan et al., 2009) view sentiment classification as a text classification problem where an annotated text data with sentiments labelled is required for training the classifier (He et al., 2011). Various pre-trained models can also be used for sentiment analysis tasks like VADER, TEXTBLOB etc. We will talk more about VADER in the study design section.

Aspect-based sentiment analysis (ABSA) is another critical area of research. It is a hybrid approach. ABSA aims to identify specific aspects or features mentioned in reviews and determine the sentiment expressed towards each aspect. Techniques such as conditional random fields (CRFs) and recurrent neural networks (RNNs) can be used to improve the accuracy of aspect detection and sentiment classification. (Ezzameli et al., 2023)

Sequence Mining

Sequence mining is also an advanced area of research within the field of text mining and sentiment analysis that focuses on uncovering patterns in the order of words or aspects within textual data, such as reviews. This technique is particularly useful for understanding the relative

importance of different aspects in user-generated content (Rashidi, 2014). Methods such as sequential pattern mining (SPM) and hidden Markov models (HMMs) are commonly used for this.

Link Analysis

Link analysis involves studying the relationships between different entities mentioned in the reviews. This include analysing how different aspects are connected to each other and to the overall sentiment of the review. Techniques such as network analysis and graph-based methods are commonly used.

Aspect Detection

Aspect detection involves identifying specific components or attributes of a product or service mentioned in customer reviews. Various techniques, such as unsupervised learning models using cosine similarity scores and assigning clusters (Ghadery et al., 2018), Latent Dirichlet Allocation (LDA) use bag of words and supervised learning models are used for aspect detection. These models aim to automatically extract key aspects, such as location, cleanliness, and service, neighbourhood from the text. However, it doesn't work well with imbalanced data, its computationally intensive and overfitting might occur.

Sentence. Despite the flaws in the script development, I enjoyed this film, especially the ending, the compelling love stories, and the message it conveys to the audience.	
Sub-tasks	**Output**
Aspect Term Extraction	script, film, love stories, message
Aspect Term Polarity	{script: Negative}, {film: Positive}, {love stories: Positive}, {message: Positive}
Aspect Category Detection	Screenplay, Movie, Story, Content
Aspect Category Polarity	{Screenplay: Negative}, {Movie: Positive}, {Story: Positive}, {Content: Positive}

Table 2 Aspect Detection example (Vazan et al., 2022)

Although each approach comes with its own pros and cons, the ultimate choice of approach depends on the outstanding task, available data resources, computational power, and domain knowledge.

2.2 Recent Works

Some of the previous research works in this topic include (Ordenes et al., 2014) who applied a linguistics-based technique for customer feedback analysis, highlighting the significance of linguistic features in understanding customer sentiments. They demonstrated the effectiveness

of various text mining techniques in extracting meaningful insights from customer reviews. X. (Fu et al., 2013) explored a sophisticated approach to sentiment analysis, particularly focusing on Chinese online social reviews. They used an integrated approach of topic modelling and HowNet semantic lexicon which is domain independent and was effective in capturing complex Chinese reviews. (Kranzbuhler et al., 2017) researched customer experience at different levels and the importance of understanding customer feedback on a granular level. The framework covered analysis from individual reviews to broader trends, providing a comprehensive view of customer feedback. Similarly, (Xia et al., 2015) presented an approach for aspect-based sentiment analysis, combining machine learning techniques with natural language processing to effectively extract and categorize sentiments related to different service aspects.

Further contributing to this field, **(**Cambria et al., 2017) proposed a novel sentiment analysis approach that combined general knowledge with deep learning techniques, enhancing the accuracy of sentiment detection and offering tailored insights into customer opinions. (Pontiki et al. 2016) introduced the Semeval-2016 task on aspect-based sentiment analysis similarly like (Xia et al., 2015) but offering a relative overview of various methods and datasets used for aspect and sentiment extraction also providing insights into the performance of different approaches. (Medhat et al., 2014) also provided an overall review of sentiment analysis techniques, including aspect-based sentiment analysis, discussing the strengths and weaknesses of various methods, and suggested limitations and future research scope.

2.2.1 Evaluation

The performance metrics for the reviewed studies vary widely based on the methodologies and datasets used. (Pontiki et al. 2016) in the Semeval-2016 task showed accuracies from 65-85%, with top systems achieving F1 scores around 0.75 to 0.85. (Kranzbuhler et al. 2017) did not report any specific accuracy scores since it focused on the multilevel nature of customer experience research. (Xia et al. 2015) reported accuracy between 75-85%, with F1 scores between 0.75 to 0.85 for aspect-specific sentiments. (Ordenes et al. 2014) utilized linguistic cues for sentiment analysis, typically achieving accuracy between 70-80%, with detailed precision, recall, and F1 scores depending on the data. (Medhat et al. 2014) provided a survey with sentiment analysis techniques showing accuracy typically between 70-85%, with varying precision, recall, and F1 scores whereas (Cambria et al. 2017) reported accuracy often exceeding 85% using a combination of common-sense knowledge and deep learning, with high

precision, recall, and F1 scores frequently above 0.85. These metrics reflect the advancements and types of work done in the field of sentiment analysis for customer experience management.

2.3 Proposed Approach

The approach used in this report builds upon existing methodologies as discussed in section 2.2 and incorporating several novel aspects that aim to address the research objectives as stated in section 1.2. Unlike previous studies that often focus on either sentiment analysis or aspect-based sentiment analysis, this methodology integrates both to provide a more comprehensive understanding of customer feedback. Advanced NLP techniques were employed to improve the accuracy of sentiment detection and aspect extraction.

Sequence Mining

One of the unique features of this research is the application of sequence mining techniques to analyse the shift in sentiments in reviews. Based on the research done by Ordenes et al. (2014), who emphasized the importance of sequence in understanding consumer narratives, we utilized Sequential Pattern Mining to uncover patterns in how guests structure their feedback. This approach allowed us to identify repetitive sequences of aspect mentions and understand how these sequences impact the overall sentiment of the review. For example, from the table 3 below, it's evident that reviews follow a similar sequence starting positively then shifting to negative or neutral.

Review Text	Aspect Sequence	Sentiment Sequence
"The location is perfect, very close to the metro. However, the cleanliness could be improved. Overall, a good stay."	Location → Cleanliness → Overall Experience	Positive → Negative → Positive
"Great host, very responsive. The neighborhood was quiet, but the bed was uncomfortable."	Host → Neighborhood → Amenities	Positive → Positive → Negative
"The apartment was spotless and well-decorated. The check-in process was smooth, but the Wi-Fi \	Cleanliness → Decor → Check-in → Amenities	Positive → Positive → Positive → Negative
"The view from the balcony was stunning! However, the noise from the street was disturbing."	View → Noise	Positive → Negative

Table 3 Sequence Mining

Aspect-Based Sentiment Analysis (ABSA)

Although previous studies like the one by (Pontiki et al. 2016), did demonstrate the utility of Aspect-Based Sentiment Analysis (ABSA) in understanding customer sentiments at a deeper

level, our approach extends this by integrating ABSA with sequence mining techniques. This integration allows us to not only identify the sentiments associated with specific aspects (such as location, environment, and worth) but also helped in analysing the sequence in which these aspects are mentioned within reviews. Also, we used rule-based approach for finding the sentiments associated with each aspect. This involved pos-tagging the reviews and then applying custom pattern searches on that, to uncover attached sentiments.

Sentiment Analysis

To ensure accuracy and efficiency in sentiment analysis, we have used the VADER (Valence Aware Dictionary and sentiment Reasoner) tool, which is well-suited for analysing short, informal text like customer reviews. VADER assigns sentiment scores based on a pre-built lexicon and aggregates these to produce an overall compound sentiment score, ranging from -1 (most negative) to +1 (most positive), while also capturing the intensity of sentiments. (Hutto et al., 2014) The classification as positive/negative/neutral is done based on a threshold value. Its sensitivity to punctuation, capitalization, and slang makes it particularly effective for this type of research where people express their opinions freely on social media. We applied VADER to classify the sentiment of each review, accurately capturing both the overall tone and specific texts that could influence a guest's experience. Additionally, we fine-tuned VADER by incorporating domain-specific terms related to Airbnb reviews, ensuring that sentiment analysis remained not only accurate but contextually relevant. This tailored approach resulted in a robust sentiment analysis model, effectively supporting the broader objectives of our research.

Review: This is the worst service I have ever experienced.
Sentiment: {'neg': 0.584, 'neu': 0.416, 'pos': 0.0, 'compound': -0.8423}

Table 4 VADER example

Regional Analysis

This research also involves segregating the data by region and time, allowing us to analyse regional differences in sentiments and how these change over time. By comparing aspect frequency and sentiment trends across different regions, we can provide tailored

recommendations for hosts in specific areas. Additionally, by examining changes in sentiment over time, we can identify evolving trends and shifts in guest expectations, which can inform future strategies for both hosts and the Airbnb platform.

Custom Lexicon Based Approach

In addition to using VADER for sentiment analysis, we developed a custom dictionary tailored specifically to the context of Airbnb reviews using most common words from reviews. We also applied POS tagging and then specified language rules to better capture the aspect sentiments. These lexicons include domain-specific terms related to hospitality, travel, and accommodation, which are not always properly captured by general-purpose sentiment analysis tools. By incorporating these custom lexicons, the precision of our sentiment analysis was enhanced, particularly in identifying and categorizing sentiments that are specific to the Airbnb context. LIWC, a text analysis tool, was also a paid alternative available for this task. Below, I have shown an example of the grammar rule I have used in this research for aspect sentiment matching.

grammar = r""" NP: {<JJ><NN.*>+}: {<NN.*>+<JJ>}

Noun followed by Adjective Adjective followed by Noun

Visual Analytics

To facilitate the interpretation and significance of our findings, we employed various data visualisation techniques like heatmaps, trend lines, and comparative bar charts to display the relationships between different aspects, sentiments, and review outcomes. This has been discussed in detail in the Results section.

3.RESEARCH DESIGN

3.1 Aim

The primary aim of this research was to analyse the huge data that we got from the inside Airbnb website, which included over 188,757 reviews provided by the customers. Starting from cleaning the data since most of the reviews were unstructured to visualizing it to discover underlying patterns and finally decoding the sentiment transitions along with the top aspects involved was the main part of this extended research project. Also, our aim was to keep data as secure as possible by anonymizing it.

3.2 Objectives

Sentiment Analysis: Determine the sentiment expressed in reviews to understand customer experience.
Sequence Analysis: Analyse the position of aspects within the reviews to determine their relative importance and capture the sentiment shifts in reviews.
Aspect Detection: Identify specific aspects or features mentioned in Airbnb reviews, such as cleanliness, location, host interaction, and amenities.
Frequency Analysis: Calculate the frequency of each identified aspect to highlight the most discussed features in the reviews.
Insights derived: Provide Airbnb hosts with clear, prioritized aspects that require attention and improvement based on customer feedback.

3.3 Research Overview

To achieve these objectives, a structured research design was used, starting with data collection from Airbnb reviews and listings to analysing the aspects involved. Natural Language Processing (NLP) techniques were applied throughout the research as discussed in the proposed approach section 2.3. Main workflow of this research can be segregated into following steps as shown in figure 3.

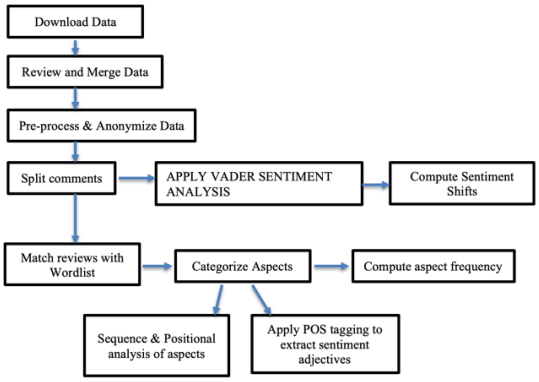


Figure 3 Research Workflow

Data Collection and Pre-processing: A comprehensive dataset of Airbnb reviews was collected, which includes various metadata such as review text, date, and listing ID. The pre-processing steps are discussed in detail in section 4.4. This thorough pre-processing ensured data integrity and consistency of the input data for further analysis.

Sentiment Analysis: Sentiment analysis is performed on each aspect as well as the review comments to determine whether the sentiments expressed by the reviewers are positive, negative, or neutral. For this purpose, we used the VADER sentiment analysis tool. VADER methodology has been discussed in the earlier sections. **In this research, we chose not to use a machine learning approach for sentiment analysis. This decision is because the study is focused on analysing the reviews and the emotions conveyed rather than predicting sentiments. Also, sentiment-labelled data for Airbnb reviews was not available, and manually labelling so many reviews could introduce bias, as there is no universally accepted gold standard for sentiment labelling in this context.** Hence, using VADER allowed for a quick and consistent analysis of sentiments expressed in the reviews. By combining rule-based aspect detection with sentiment analysis, this methodology offers a comprehensive view of how different aspects of Airbnb listings influence customer experience.

Sequence Analysis: Sequential pattern mining has been used to analyse the sentiment shift patterns in a review. It is also used for positional analysis of aspects mentioned in the reviews by calculating percentage of aspect positions in beginning/middle/end. Based on the results obtained from this, we could state that reviewers normally start reviews with a positive tone and then progress towards complaints, if any. Refer to table 3 for example. Using positional analysis, we can also determine aspect priority, as aspects mentioned first have much higher priority than those mentioned later.

Aspect Detection: For aspect detection, a rule-based pattern matching approach has been utilized, which involves creating a detailed Excel file that maps specific keywords and phrases to different aspect of the Airbnb listings. (see table 5) This file includes categories such as amenities, outcome, worth and neighbourhood, among others. By matching these predefined aspects with the reviews, we can easily identify which aspects are being discussed in each review. By manually labelling the aspect categories and their corresponding keywords, this method ensures that context of an aspect is accurately captured. Frequencies of mentioned aspects were calculated by creating binary dummy variables indicating presence and absence of aspects in each review and then adding up the counts. This helped us in addressing RQ2. Other than that, we also applied POS tagging on the tokenized noun phrases and then certain grammar rules were applied on those tokens to capture the adjectives or sentiment associated with each of those aspects. An example for that is shown in section 2.3.

Amenities	Environment	Neighbourhood	Worth	Outcome	Recommend
bedroom	train station	Salford	excellent	definitely stay	definitely recommend
kitchen	pub	Oldham	lovely	stay here again	highly recommend
bathroom	festival	Moss side	nice	stay there again	won't recommend

Table 5 Dictionary for aspect categorization

The above table gives an example of the custom dictionary file which was created for this research. "Amenities" refers to the amenities provided in that property, "Environment" refers to the surrounding area whereas "Neighbourhood" specifies the place of the property, "worth" describes the experience of the customer, "Outcome" describes the overall review of the stay and lastly "Recommend" reflects how likely the customer is willing to recommend the property to others.

Results: Finally, various data visualisations were used to provide a descriptive analysis of the data, aiding in addressing the research questions and revealing additional key insights. Since this research did not rely on prediction-based modelling, and, we didn't have pre-labelled data to compare with, so, traditional metrics such as accuracy scores and other benchmark tests were not applicable in this case.

4. DATA

4.1 Source

Airbnb does not allow users to download data directly from their website, instead they are scattered across various sources or news outlets. Access to such data is available on third party websites like InsideAirbnb (available online at http://insideairbnb.com/about) from where data has been downloaded for this research. I have used the data specifically for Greater Manchester and the research is based upon that. Below is the representation of the data available on InsideAirbnb.

Greater Manchester, England, United Kingdom

26 June, 2024 (Explore)

Country/City	File Name	Description
Greater Manchester	listings.csv.gz	Detailed Listings data
Greater Manchester	calendar.csv.gz	Detailed Calendar Data
Greater Manchester	reviews.csv.gz	Detailed Review Data
Greater Manchester	listings.csv	Summary information and metrics for listings in Greater Manchester (good for visualisations).
Greater Manchester	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
Greater Manchester	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
Greater Manchester	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

Figure 4 Greater Manchester Dataset from InsideAirbnb

Although, there were multiple files available for download on the website but for research purpose, only the listings and reviews file have been used, since the other datasets don't help much in meeting the objectives of this research.

4.2 MetaData

As mentioned in the above section two files have been used for this research, which are reviews.csv and lsitings.csv. Brief overview of both the datasets have been provided below.

Listings

The Listings Dataset consists of 6,136 entries with 18 columns, providing detailed information about each Airbnb listing. Key columns include `id`, which serves as a unique identifier for each listing, and `name`, which denotes the listing's name. The `host_id` and `host_name` columns identify the host associated with the listing. Geographic information is provided through `neighbourhood_group`, `neighbourhood`, `latitude`, and `longitude`, specifying the

location of the property. The dataset also includes `room_type`, indicating the type of accommodation (e.g., Entire home/apt, Private room), and `price`, detailing the cost per night. Additionally, `minimum_nights` shows the required minimum stay, while `number_of_reviews` counts the total reviews a listing has received. The `last_review` column records the date of the most recent review, with `reviews_per_month` providing an average review frequency. Other important columns include `calculated_host_listings_count`, which counts the number of listings managed by the host, `availability_365`, which shows the number of days the listing is available for booking in a year, and `number_of_reviews_ltm`, indicating the number of reviews in the last 12 months. The `license` column, although mostly empty, is intended to capture license information for the listings. An overview of the dataset has been provided below.

	listing_id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_rev
0	157612	Loft in Salford · ★4.92 · 2 bedrooms · 2 beds ...	757016	Margaret	Salford	Salford District	53.501530	-2.262490	Entire home/apt	42.0	2	
1	283495	Home in Middleton · ★5.0 · 1 bedroom · 1 bed ·...	1476718	Alison	Rochdale	Rochdale District	53.562710	-2.218240	Private room	75.0	100	
2	310742	Rental unit in Manchester · ★4.66 · 1 bedroom ...	1603652	Francisca	Manchester	Ancoats and Clayton	53.484110	-2.229190	Private room	34.0	180	
3	332580	Condo in Manchester · ★4.86 · 1 bedroom · 1 be...	1694961	Manchester	Manchester	City Centre	53.480170	-2.232850	Private room	50.0	2	
4	411843	Rental unit in Manchester · ★4.76 · 2 bedrooms...	2046430	Tom	Manchester	Hulme	53.468518	-2.244034	Entire home/apt	100.0	2	

Figure 5 Listings Dataset

Reviews

The Reviews Dataset contains 188,757 entries across 6 columns, focusing on customer feedback related to the listings. The `listing_id` column links each review to its corresponding listing, while `id` provides a unique identifier for each review. The `date` column records the date of the review. The dataset also includes `reviewer_id` and `reviewer_name`, which identify the individual who left the review. Finally, the `comments` column captures the full text of the review, offering qualitative insights into the guest's experience. Together, these datasets provide a comprehensive overview of Airbnb listings and the feedback they receive, enabling detailed analysis of host performance, guest satisfaction, and property characteristics.

The data is a complete representation of the Manchester Airbnb host base. It includes reviews from various locations and property types in and around Manchester.

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	157612	919313	13/02/2012	1378688	Kristin	Margaret and her husband were the perfect host...
1	157612	922493	14/02/2012	1724861	Katy	Margaret and Tom are warm, welcoming and incre...
2	157612	1244776	07/05/2012	2284316	Ian	The place was great, and the photographs give ...
3	157612	1486412	15/06/2012	1440146	Tim	Super Place, Margaret and Tom are Lovely peopl...
4	157612	1538944	22/06/2012	2640396	Sherry	Margaret was such a great host and was extreme...
5	157612	1600081	01/07/2012	2062280	Mary Ellyn	It was such a pleasure to stay in Margaret's l...
6	157612	1609117	02/07/2012	2177275	Neil	Margaret and Tom were both very friendly and h...
7	157612	1621213	03/07/2012	2426728	Natalie	Great rooms; clean and cosy with breakfast and...
8	157612	1844197	30/07/2012	3027783	Anja	Everything was just great. We stayed one night...
9	157612	1990044	14/08/2012	377437	Anthony	10 out of 10!\r The accommodation is clean...

Figure 6 Reviews Dataset

4.3 Ethics

While working with social media data, it is very important to keep data ethics into consideration. Since this research involves personal identifiers like name and ID so we had to anonymize data properly so that its unidentifiable by backtracking and to ensure data privacy. For achieving that, columns like reviewer name were removed since it won't be that useful in our analysis. Other than that, the host names were replaced by pseudonyms to anonymize the data.

```
Host to Pseudonym Mapping:
Margaret -> Host_1
My-Places -> Host_2
Joey -> Host_3
Francisca -> Host_4
Fiona -> Host_5
Soraya & Shahin -> Host_6
Nick -> Host_7
Laurence -> Host_8
John -> Host_9
Anne Phil -> Host_10
```

Figure 7 Host name anonymization

4.4 Pre-processing

The pre-processing steps began with importing essential libraries like `pandas`, `numpy`, `re`, `nltk`, `langdetect`, and `unidecode`, followed by downloading NLTK's 'punkt' tokenizer for text processing. Then we expanded contractions in the review text to ensure consistency and clarity. This step involves replacing contractions like "can't" with "cannot" and "won't" with "will not," making the text more uniform and easier to analyse. Then the reviews were tokenized into sentences. Following this, the dataset was filtered to include only reviews from 2019 onwards and comments with less than 30 characters were removed ensuring that the analysis focuses on more recent and meaningful data. Additionally, the text in comments were converted to lowercase to maintain uniformity, non-English reviews were removed, and non-ASCII characters were converted to ASCII to standardize the text further.

The next stage involved merging the review data with additional information from the listings dataset, such as host details, neighbourhood, room type, and availability. This merger added further contextual info to the reviews. After merging, a series of text cleaning steps were applied. Unnecessary columns like license were removed, and records with host names containing more than one name were dropped. Common textual replacements were made, such as converting "she" or "he" to "host" and standardizing various location names. Additionally, stopwords were removed, punctuation and apostrophes stripped out, and any HTML tags or unnecessary characters were cleaned from the text. This thorough pre-processing ensured that the data was well-prepared for the subsequent analysis, with clean and standardized text which can accurately reflect the content of the reviews.

	comments	comments_clean
0	tom and margaret are a very well bred couple. ...	tom margaret well bred couple . drove london h...
1	well equipped converted attic space, perfect f...	well equipped converted attic space , perfect ...
2	margaret and her husband are extremely welcomi...	margaret husband extremely welcoming attentive...
3	ideal location with great host, perfect for a ...	ideal location great host , perfect work trip ...
4	we had a wonderful stay at margaret's place. t...	wonderful stay margaret 's place . little home...

Figure 8 Pre-processed comments

4.5 Data Summary

The dataset used in this analysis comprised of a collection of customer reviews, which provides valuable insights into customer experiences with Airbnb listings. A total of 125253 reviews were present after pre-processing. From the figure below, we can see a rise in reviews over the years and a dip during the covid period presumably, because of lockdown, which saw a halt in tourism business.

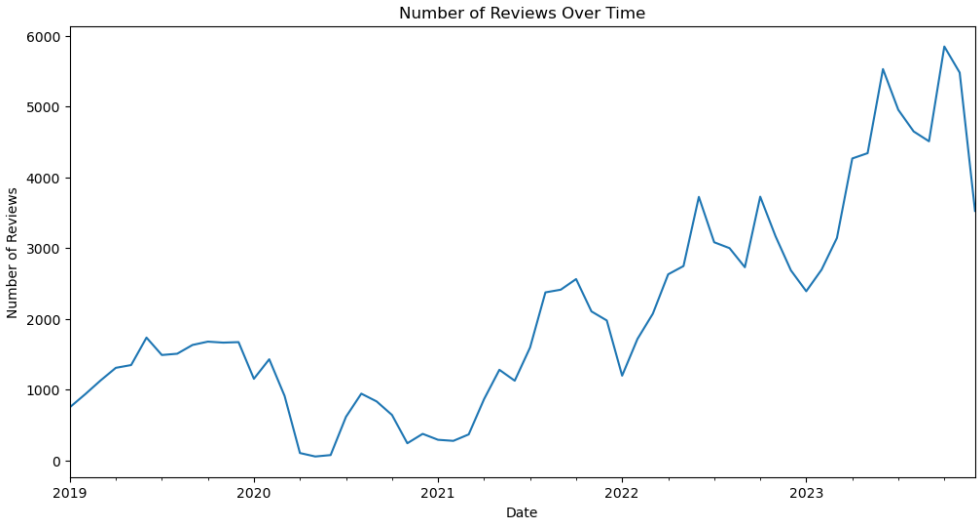


Figure 9 Reviews per year.

Other statistics include the average review length, which was 225.36 characters which clearly indicates that reviewers prefer to keep their reviews short mostly.

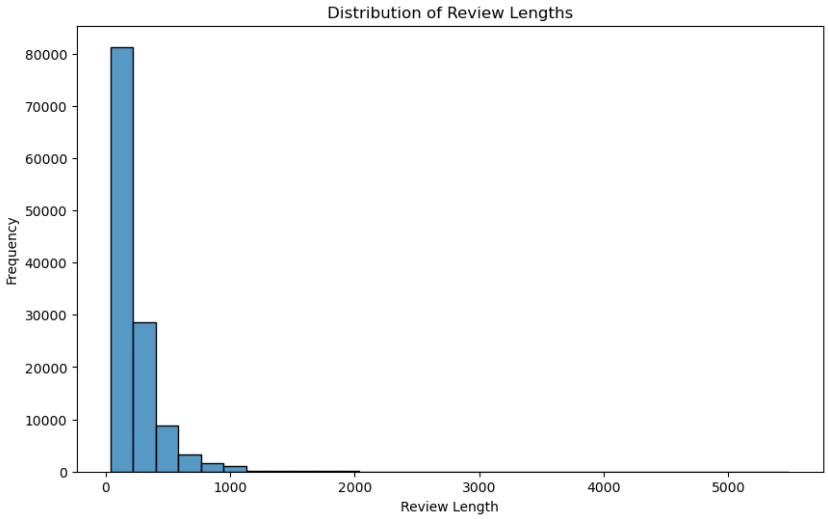


Figure 10 Distribution of Review Lengths

Most used words in reviews

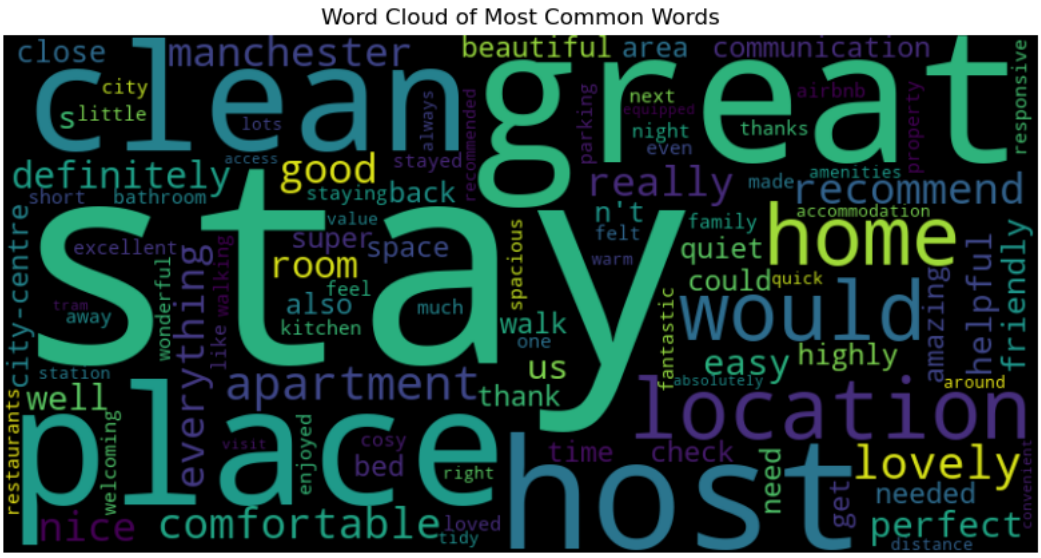


Figure 11Most common words in reviews

The word cloud in above figure shows the most frequently mentioned words in the reviews. Bigger words appear more frequently in the text, indicating the topics and sentiments that are most expressed by guests. Key words like "stay," "location," "home," "clean," "great," and "host" are prominent, suggesting that these are significant factors in driving guest experiences. Positive adjectives like "comfortable," "lovely," "recommend," and "definitely" also appear frequently, reflecting dominance of positive sentiments over negative in the reviews. The emphasis on words related to location, cleanliness, and the host indicates that these aspects are critical from a guest's perspective.

Price vs neighbourhood

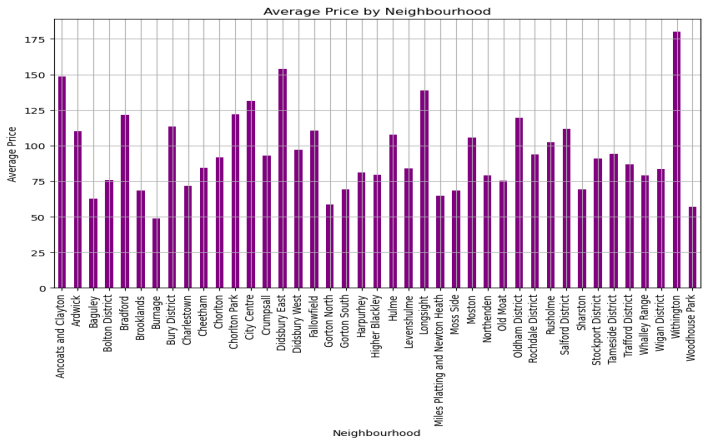


Figure 12 Price vs Neighbourhood

From the above figure, we can get an idea about the price distribution in different regions. We can see varying average prices across different regions, among which, Withington reported the highest average price for listings in that area and Burnage had the lowest average price. This could be due to factors like proximity to attractions, quality of accommodations, or overall ambiance.

Some other statistics have been shown below.

Total Hosts	Total Listings	
		Neighbourhood with the least listings: Baguley (9 listings)
1571	4777	Neighbourhood with the most listings: Salford District (965 listings)

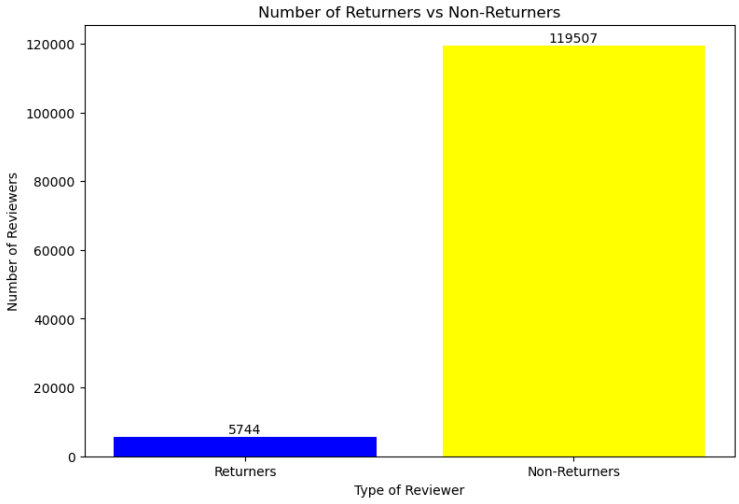


Figure 13 Count of returners vs non-returners

5.RESULTS & ANALYSIS

5.1 Findings

In this research, we conducted a comprehensive sentiment and sequence analysis of Airbnb reviews to meet the objectives mentioned in section 3.2. The analysis leveraged various machine learning and natural language processing techniques to fulfil the objectives.

Sentiment Analysis

DISTRIBUTION

Before digging deeper into sentiment analysis, first, let's look at the distribution of sentiments across the reviews used in this research.

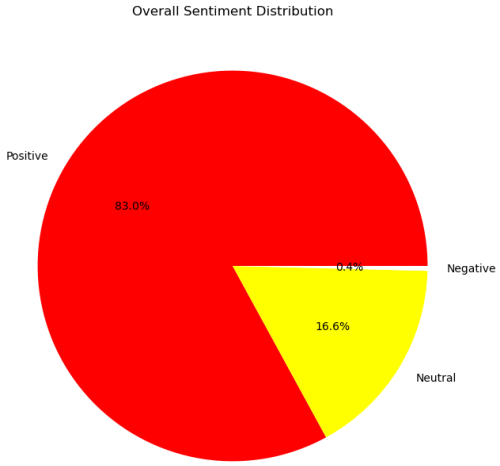


Figure 14 Sentiment Distribution

Clearly, we can see that the dataset is highly imbalanced with over 83% reviews being positive. This reflects on the quality of service and overall guest experience provided by Airbnb hosts. In this case, we can't do anything to fix this data imbalance since these are actual reviews posted by people and resampling this data will lead to potential loss of information from the data. However, the small presence of negative and neutral reviews suggests there are still scope for improvement which can be explored further.

TRANSITION(RQ1)

	Sentiment Sequence	Frequency
0	Positive → Positive	74317
1	Single Sentiment	22151
2	Positive → Neutral	14863
3	Neutral → Positive	7073
4	Positive → Negative	2971
5	Neutral → Neutral	1564
6	Negative → Positive	1244
7	Neutral → Negative	405
8	Negative → Neutral	383
9	Negative → Negative	276

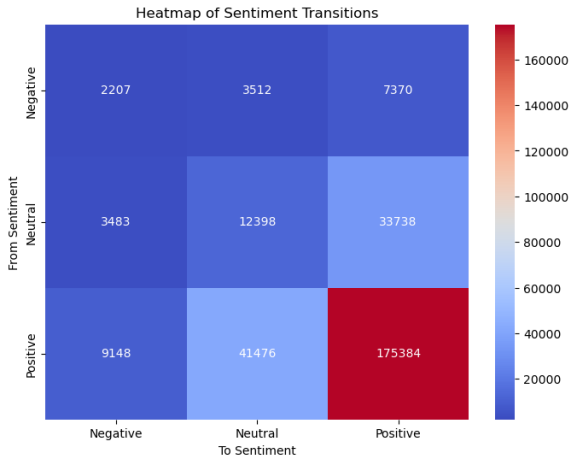


Figure 15 Sentiment Transitions

The above visualisations reveal a strong tendency for reviewers to remain consistent with their reviews when they start on a positive note, however, there were some instances when they did eventually shift towards negative parts of the review. Presence of transitions from negative to positive are also visible suggesting that they didn't like certain aspects of their stay but overall had a good experience. Additionally, there is a significant positive shift from neutral to positive sentiment, suggesting that many reviews, while starting neutrally, ended on a favourable note. These insights highlight potential areas of strength for businesses, as aspects of their service or product are turning neutral experiences into positive ones. These sentiment transitions can be quite helpful for hosts to strategically focus on enhancing customer experiences, especially by identifying and amplifying the factors that lead to a positive conclusion in reviews.

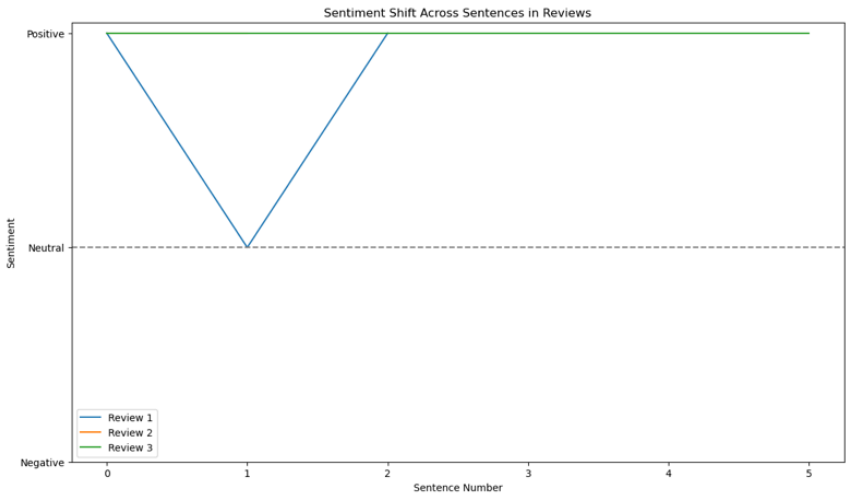


Figure 16 Sentiment shifts in a review

ASPECT PRIORITY (RQ2)

Priority/Importance of aspects were determined based on the frequency each aspect, where high frequency suggests it is of high importance from customer's perspective while shaping reviews.

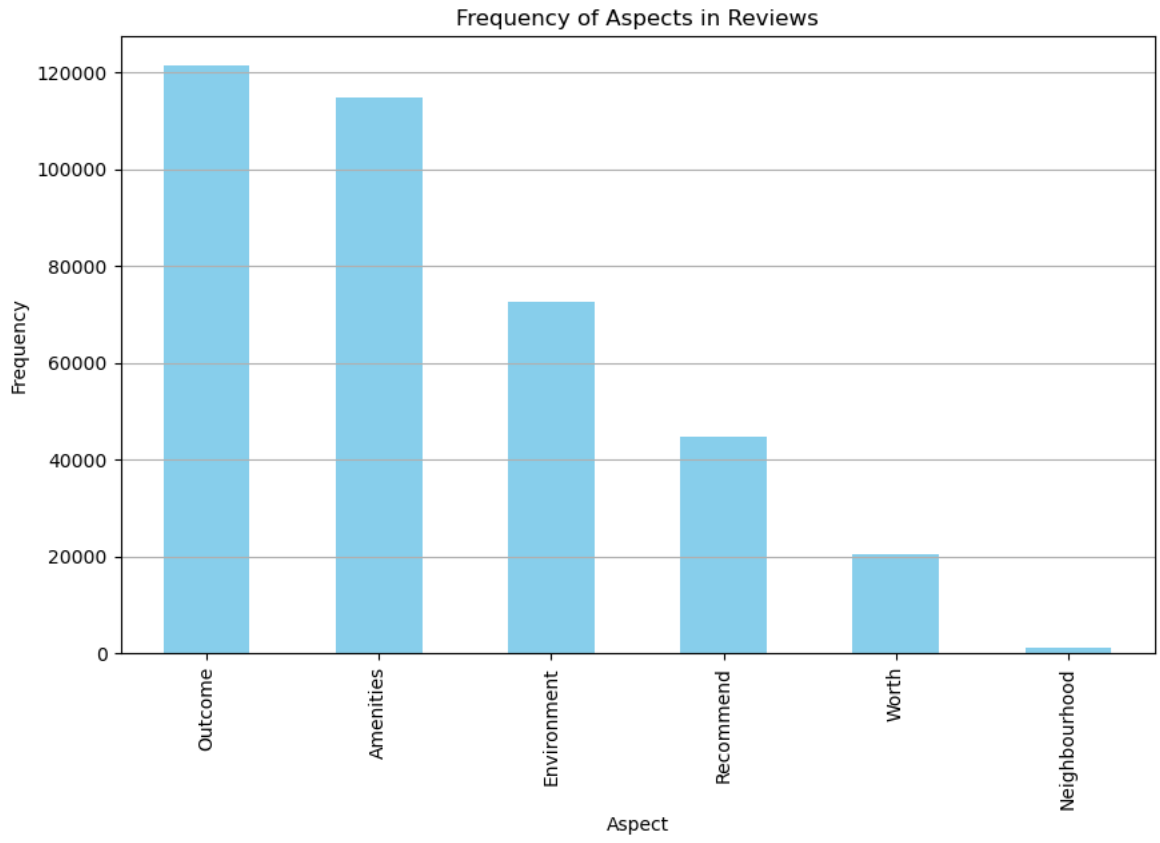


Figure 17 Frequency of mentioned aspects.

Clearly, we can see from the above figure that reviewers have mostly talked about overall outcome in their reviews, followed by the amenities provided in that listing and then environment and finally neighbourhood count wise. Based on this, we can derive that most guests tend to provide an overall evaluation of the place in their reviews, often highlighting the specific amenities they liked or disliked. After visualising the frequency of aspects overall, the next part was to do a deeper level sequence analysis of aspect positions in reviews to see which aspect they talk about first, which gives us a clear picture about aspect priority as well.

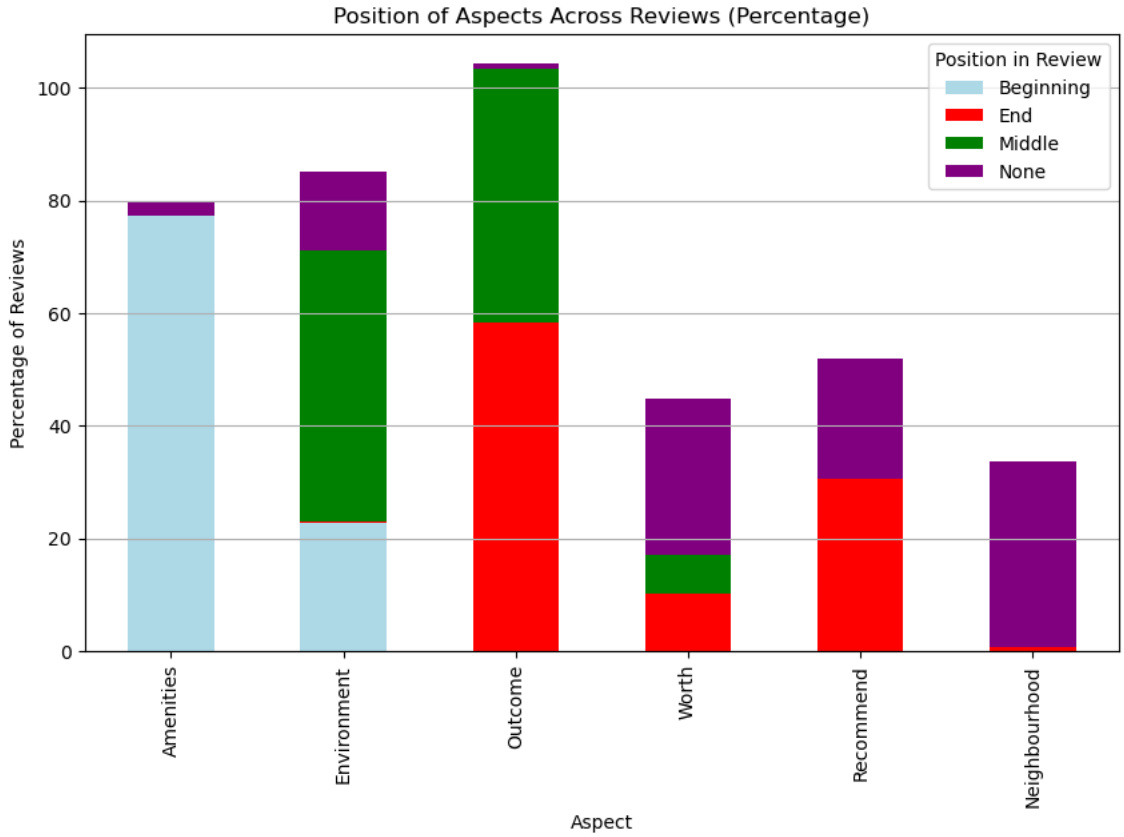


Figure 18 Aspect position in reviews.

Figure 18 highlights that reviewers tend to shape their reviews in a clear narrative progression, beginning with immediate, tangible features like amenities, which are frequently mentioned at the start of reviews, depicting their importance to the overall experience. Environment often follows in the middle, suggesting that after addressing the amenities, reviewers shift focus to describe their surroundings like nearby places of interest. Lastly, aspects such as outcome and worth are commonly discussed towards the end, indicating that these are part of the reviewer's overall experience of the stay and final judgment. Recommendations typically conclude the review, reflecting their willingness to recommend the property to others if they are satisfied. Notably, some reviews omit mentioning neighbourhood and worth altogether, implying these aspects are not that relevant to all reviewers. This pattern shows a focus on immediate features, followed by a broader reflection on the experience, and concluding with an overall judgment.

By delving deeper into the amenities mentioned at the beginning of the reviews, we can gain a clearer understanding of which amenities are most important to guests and how these influence their likelihood of returning to the same listing.

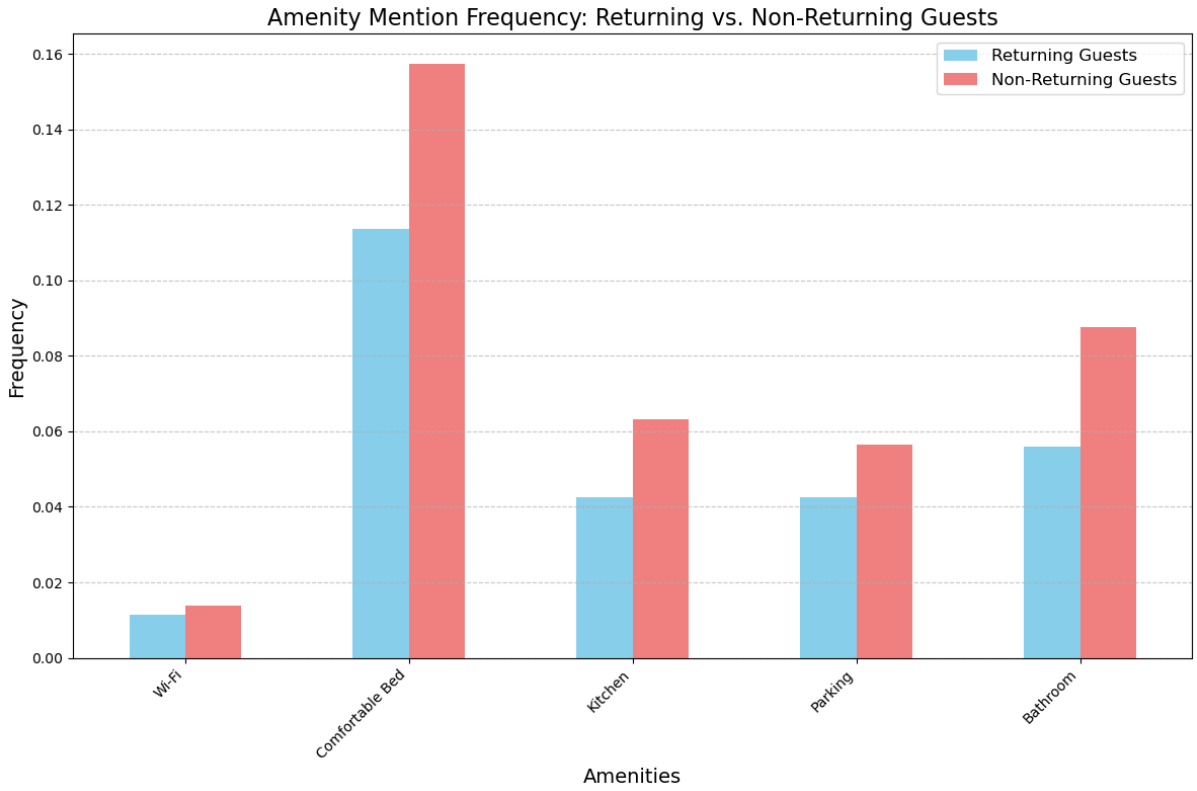


Figure 19 Customer retention based on amenities.

From the above figure, we can see a clear comparison between guests who have returned and those who haven't. Amenities like "Comfortable Bed" and "Bathroom" are mentioned more frequently by non-returning guests. This could signify areas where improvements could potentially increase the likelihood of guest return. On the other hand, "Wi-Fi" and "Parking" show similar mention frequencies across both returning and non-returning guests, suggesting they are standard expectations but may not significantly influence the decision to return.

5.2 Implications

The sequence analysis showed a distinct pattern in how reviewers structured their feedback. Positive comments about the host and location were typically mentioned at the beginning of the reviews, while negative feedback, if any, was mentioned towards the end. This pattern suggests that reviewers prefer to start with a positive tone and then eventually shift to other sentiments, possibly to balance their criticism & appear fair in their evaluation. Possible implications from this research can be: -

1. **Customer Experience:**

The high percentage of positive reviews indicate that Airbnb hosts are generally providing satisfactory services, but the presence of negative and neutral reviews highlights areas for improvement. By focusing on the aspects that drive negative sentiments, such as certain amenities or environmental factors, hosts can enhance the overall guest experience.

2. **Focus on Key Amenities:**

The analysis reveals that amenities like "Comfortable Bed" and "Bathroom" are frequently mentioned by non-returning guests. This suggests that these amenities are not up to guest expectations, leading to lower return rates. Hosts should consider investing in the quality of these amenities, as improvements in these areas could enhance guest satisfaction and encourage repeat bookings.

3. **Marketing:**

The finding that guests often start their reviews by discussing immediate, tangible features such as amenities suggests that these features are most important for guests. Hosts can leverage this by highlighting these aspects during marketing campaign and during guest interactions, ensuring that these elements meet guest expectations from the outset.

4. **Utilizing Sentiment Transitions:**

The sentiment transitions show that guests who begin their reviews on a positive note tend to remain positive, while those who start neutral often shift towards positivity. This suggests that the host is the key, and creating positive first impression will definitely help them in the long run resulting in improved customer retention.

5. **Tailored Services:**

The sequence analysis of aspect mentions in reviews reveals that guests prioritize discussing amenities first, followed by the environment, and then overall outcomes and recommendations. This suggests that while guests appreciate a wholesome experience, the amenities of the stay are most immediately important. Hosts can tailor their services and communication strategies

to emphasize these aspects, ensuring that they meet the high expectations that guests have for these features.

6. **Customer Retention Strategies:**

The analysis of returning guests provides insights into which amenities are most likely to influence return rates. By focusing on highly mentioned amenities by non-returners, hosts can formulate strategies to improve guest retention. This might include upgrading bedding quality, improving bathroom facilities, or addressing any recurring issues that guests have complained about.

7. **Broader Implications:**

The patterns observed in this research are not limited to Airbnb and can be applied broadly across the hospitality industry. Other accommodation providers can apply these techniques to understand guest priorities and improve service quality. The emphasis on aspects like comfort, cleanliness, and amenities highlight universal factors that influence guest satisfaction and loyalty.

6.CONCLUSION

Gathering all the insights

This research aimed to uncover patterns in Airbnb reviews to understand the sequence of emotions and aspects which are key to customer experience. Through detailed text mining and sequence analysis, we identified key themes and their implications for both research and practice. To sum it up, we can say that customers mostly had a positive experience during their stay with some exceptions where certain customers showed a strong dislike towards certain amenities. There were no complaints regarding neighbourhood. In cases where customers started with a negative sentiment, they tried to balance their review by putting an overall review of the place in a positive tone, potentially to appear fair and less judgemental. Other than that, aspects like cleanliness, host's behaviour, and value for money were highly mentioned early in reviews. This reflects their value to the overall guest experience. Hosts and platform managers should work on factors driving negative sentiments to improve customer retention and limit negative reviews.

Scope for Research

The findings are an addition to the existing body of literature on customer experience in online reviews. This research proves that reviews have a structured sequence mostly, which, upon analysing we found out that it generally starts with a positive impression. Advanced sequence analysis techniques can be employed on top of this to uncover more complex hidden patterns from reviews. There's also scope of applying this analysis on a large-scale data instead of just focussing on Manchester. This might bring out some demographic or socio-economic factors influencing sentiments which was missing in this research.

Limitations

The main limitation of this research was the imbalanced dataset, with most reviews being positive. Non-English reviews were removed due to language constraints, missing potential insights from diverse cultural perspectives. Incorporating multilingual sentiment analysis could address this. Additionally, the lack of a sentiment-labelled dataset led to reliance on VADER, which, although efficient, struggles with complex reviews, sarcasm, and contextual

understanding. For aspect detection, creating a custom dictionary was challenging due to the large volume of reviews, leading to a focus on the most common terms, which limited the scope of pattern matching.

REFERENCES

- Yiannakou, A., Apostolou, A., Birou-Athanasiou, V., Papagiannakis, A. and Vitopoulou, A., 2022. Branding places through experiential tourism: A survey on the features of the experiential product and enterprises in Greek regions. *Tourism and Hospitality*, 3(2), pp.435-450. Available at: https://doi.org/10.3390/tourhosp3020028.
- Luca, M., 2016. Reviews, reputation, and revenue: The case of Yelp.com. *Harvard Business School Working Paper*, No. 12-016, March 2016.
- Chevalier, J.A. and Mayzlin, D., 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), pp.345-354. Available at: https://doi.org/10.1509/jmkr.43.3.345.
- Garfield, B. and Levy, D., 2013. *Can't buy me like: How authentic customer connections drive superior results*. New York: Penguin.
- Anderson, C., 2012. *The impact of social media on lodging performance*.
- Vermeulen, I.E. and Seegers, D., 2009. Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, 30(1), pp.123-127.
- Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), pp.1-135. Available at: http://dx.doi.org/10.1561/1500000011.
- Liu, B. and Zhang, L., 2012. A survey of opinion mining and sentiment analysis. In: C. Aggarwal and C. Zhai, eds. *Mining Text Data*. Boston, MA: Springer, pp.415-463. Available at: https://doi.org/10.1007/978-1-4614-3223-4_13.
- Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), pp.1093-1113. Available at: https://doi.org/10.1016/j.asej.2014.04.011.
- Blitzer, J., Dredze, M. and Pereira, F., 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In: *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp.440-447.
- Choi, Y., Cardie, C., Riloff, E. and Patwardhan, S., 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In: *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp.355-362.

- Narayanan, R., Liu, B. and Choudhary, A., 2009. Sentiment analysis of conditional sentences. In: *Proceedings of the 2009 conference on empirical methods in natural language processing*, pp.180-189.
- He, Y. and Zhou, D., 2011. Self-training from labeled features for sentiment analysis. *Information Processing & Management*, 47(4), pp.606-616. Available at: https://doi.org/10.1016/j.ipm.2010.11.003.
- Ezzameli, K. and Mahersia, H., 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99, p.101847.
- Saffari, R.M. and Rashidi, H., 2015. A new framework based on learning automata for user community detection in social networks. *International Journal of Computer Science Issues (IJCSI)*, 12(2), p.118.
- Ghadery, E., Movahedi, S., Faili, H. and Shakery, A., 2018. An unsupervised approach for aspect category detection using soft cosine similarity measure. *arXiv preprint arXiv:1812.03361*.
- Ordenes, F.V., Theodoulidis, B., Burton, J., Gruber, T. and Zaki, M., 2014. Analyzing customer experience feedback using text mining: A linguistics-based approach. *Journal of Service Research*, 17(3), pp.278-295.
- Fu, X., Guo, L., Guo, Y. and Wang, Z., 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. *Knowledge-Based Systems*, 37, pp.186-195. Available at: https://doi.org/10.1016/j.knosys.2012.08.003.
- Kranzbuehler, A.M., Kleijnen, M., Morgan, R. and Teerling, M., 2018. The multilevel nature of customer experience research: An integrative review and research agenda. *International Journal of Management Reviews*, 20(4), pp.433-456. Available at: https://doi.org/10.1111/ijmr.12140.
- Xia, R., Xu, F., Zong, C., Li, Q., Qi, Y. and Li, T., 2015. Dual sentiment analysis: Considering two sides of one review. *IEEE Transactions on Knowledge and Data Engineering*, 27(8), pp.2123-2137.
- Cambria, E., Poria, S., Gelbukh, A. and Thelwall, M., 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6), pp.74-80. Available at: https://doi.org/10.1109/MIS.2017.4531228.
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M. and Eryiğit,

G., 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California. Association for Computational Linguistics, pp.19-30.

APPENDIX

CODE

The entire code has been uploaded on GITHUB, accessible via link below https://github.com/Abhinandan305/BM-11349153-ERP

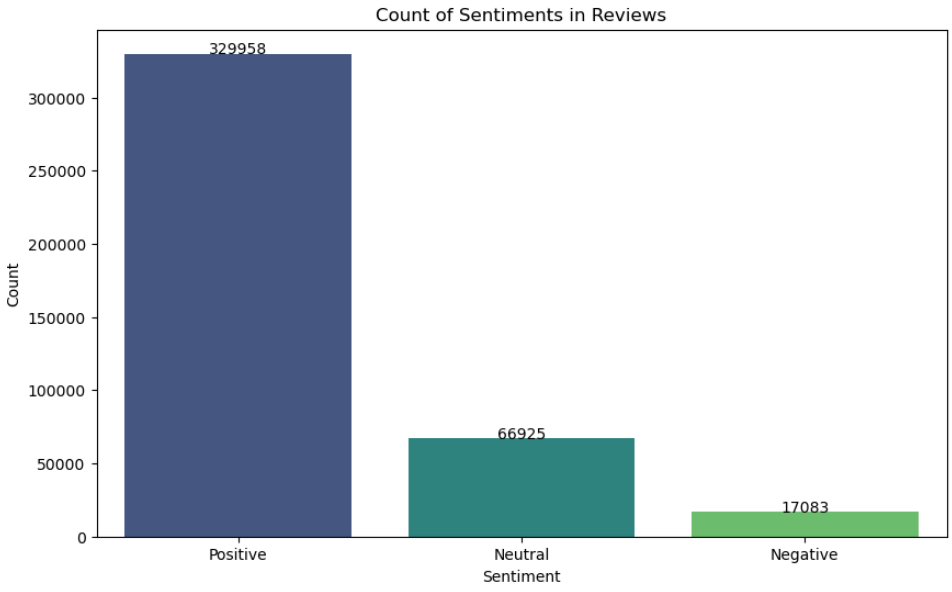
DATA

This review data of Manchester was downloaded from an open-source website https://insideairbnb.com/get-the-data/

Refer to additional materials for more information regarding code and data.

ADDITIONAL ILLUSTRATIONS

1.Count of sentiment transitions

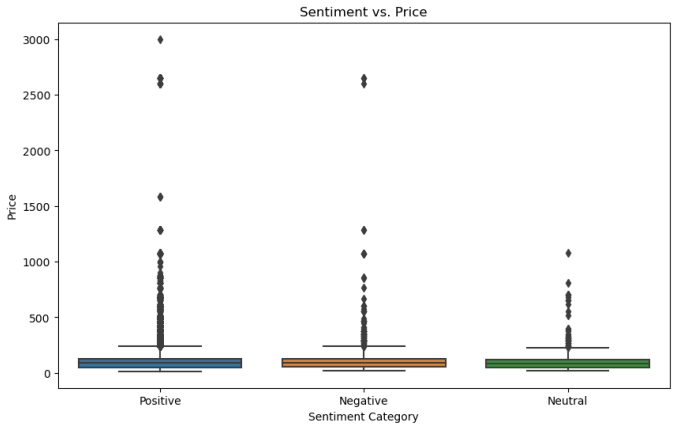


The overwhelming number of "positive" sentiments in figure 17 suggests that data is highly imbalanced. This may be because we removed reviews of other language and we also removed too outdated reviews for our analysis. On the other hand, it also indicates that most of the hosts are doing well with regards to service standards. Few negative sentiments indicate that there is still room for improvement.

2.Most common keywords in positive reviews

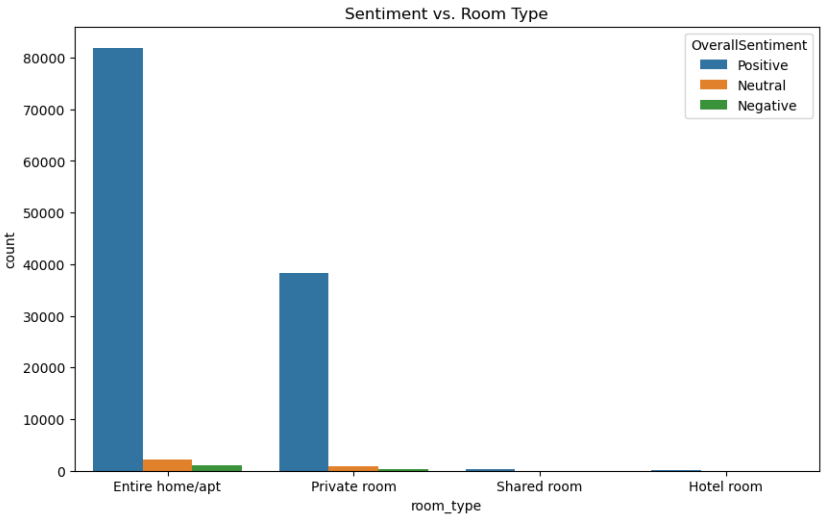
This word cloud suggests that Airbnb guests frequently mention the overall quality of the place and their positive experiences with the stay and host. The emphasis on cleanliness, host interaction, and location underscores the importance of these factors in creating a positive guest experience. Words like “highly recommend”, “great host”, “great place” also shows that customers are really satisfied with the services of the host.

3.Sentiment Vs Price



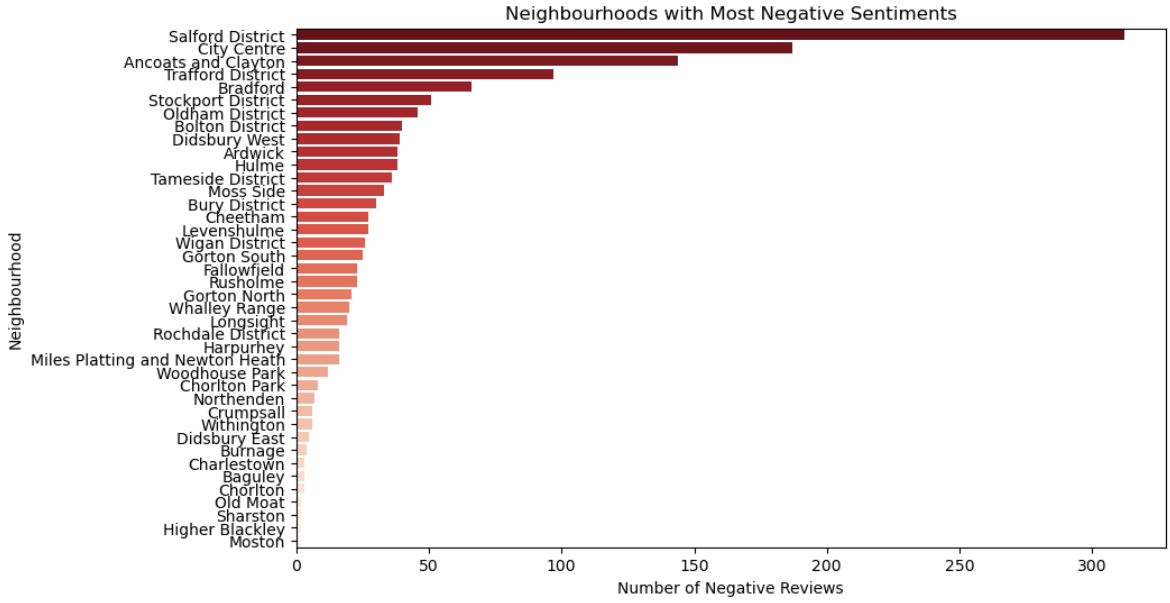
The plot shows that majority of prices are concentrated at the lower end of the scale across all sentiment categories. There are a few listings with significantly high prices that received Positive or Neutral reviews, and some outliers in the Negative sentiment category as well. This indicates that high price do not guarantee positive sentiment.

4.Sentiment distribution by Room type



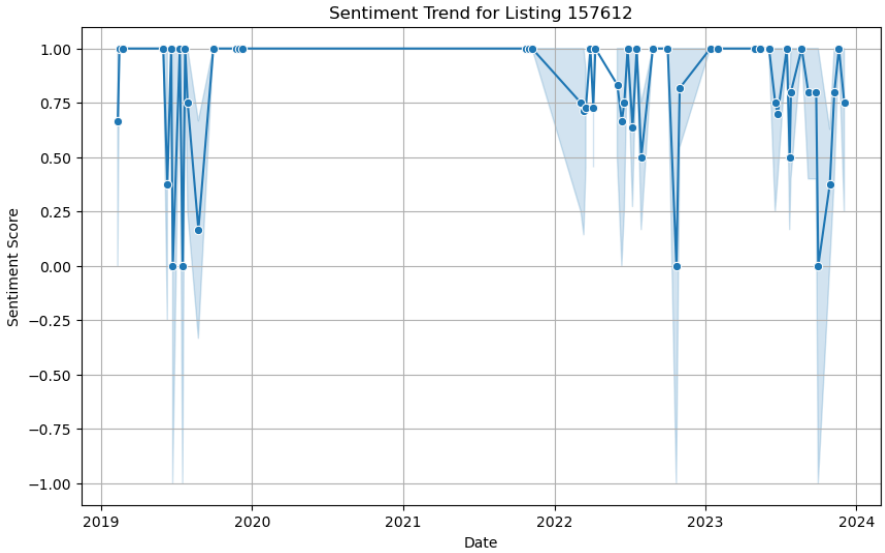
The above figure suggests that entire homes/apartments are the most popular and highly rated type of accommodation on Airbnb, followed by private rooms. Shared rooms and hotel rooms receive far fewer reviews, which may indicate a lower preference among Airbnb guests. The number of positive sentiments across all room types reflects generally high satisfaction among guests.

5.Sentiment distribution by neighbourhood



The bar chart reveals that neighbourhood with most listings (Salford) have more negative reviews while neighbourhood with few listings have fewer negative reviews. This shows which area needs more improvement.

6.Sentiment polarity for a specific listing



The above figure shows the trends in sentiment for a particular listing over the years. It suggests that this listing has generally maintained a very positive tone over the years, but an event around

2023 caused a sharp dip in this trend, which could be due to various factors, such as changes in Airbnb policies, market conditions, or specific events affecting guest experiences during that period.

7. Most used words in negative reviews

This figure suggests there might have been some issues with the bathroom, bed and kitchen as per reviews from former guests which should be focussed upon by the hosts to prevent such experience in future.

REQUIRED PACKAGES

All the required libraries have been mentioned in the requirements.txt file uploaded on Github repository.