# The Perfect Getaway

*A text analytical approach to uncover factors of consumer satisfaction from multilingual online reviews of diverse hotel types*

**Author:** Eduard Anghel
**Student number:** 592926

A thesis presented for the degree of MSc Economics & Business
Specialisation Data Science & Marketing Analytics at Erasmus University Rotterdam

**Supervisor:** Dr. A. Tetereva
**Second assessor:** Dr. C. Bellet

July 29, 2022

# Abstract

The advent of the internet brought about a change in the way users share their experiences with their circles. Online reviews are now a significant part of the consumer decision making process, whereby users can quickly judge an experience by aspects others have commented on. This is especially true in the hospitality industry, where consumers have hundreds of options to choose from, facilitated by online travel agencies. With the competition in the industry, hoteliers need to take into account factors of consumer satisfaction to improve the service they offer and make themselves truly unique and stand out from the crowd. This thesis provides a framework for managers to analyse factors of consumer satisfaction from online consumer reviews through the use of novel text analytics methods to extract topics, and further adjust for consumer heterogeneity based on cultural and economic dimensions of the reviewer's home country. Distinct factors of consumer satisfaction are extracted from 5 main hotel types - beach, ski, business, residential and city hotels. Further, multi-lingual reviews are analysed through the use of deep learning translation methods, which has not been seen frequently in the industry. This framework provides a deep look at factors of satisfaction while accounting for differing variables, and yields state-of-the-art accuracies ranging from 77-84% as well as easily interpretable results. This thesis finds that factors of consumer satisfaction do differ across hotel types (for example proximity to the beach for beach hotels, a room with a mountain view for ski hotels and attention to detail for residential stays), while cultural and economic differences also impact satisfaction. It is found that for example, individuals from countries with higher uncertainty avoidance indices are more sensitive to arrival and departure issues and those from countries with lower masculinity tend to focus on aspects that improve relaxation and factors such as decor. These results can be further used by managers to create unique consumer profiles and activate targeted marketing schemes that could save costs and be more effective in the long run. Also, it provides outlets for improving the stay of a consumer, especially if something goes wrong along the way - this could easily increase the rating given for the overall stay.

Keywords: text analytics, LDA, consumer satisfaction, hospitality, cultural dimensions, consumer heterogeneity

# Table of Contents

# List of Tables

# List of Figures

# Section 1: Introduction

Since the early 2000s, the number of people using the internet has increased by nearly ten-fold, with over 3 billion daily users in 2016 (World Bank, 2022). This exponential increase in accessibility to the internet has allowed for the creation of websites and platforms that facilitate our daily lives - from being able to establish communication with friends and family worldwide at the click of a button, to easily finding online reviews of products and experiences. Several studies have shown the importance of word-of-mouth (WOM) communication on consumer behaviour, with it being between seven and nine times as effective as traditional print marketing (Katz & Lazarsfeld, 1955; Day, 1971). Prior to the broad use of the internet, communication was typically limited to a close circle of individuals that one knew well; these persons would thus seek out opinions of those in their circle to make decisions as consumers, which is the principle of WOM communication (Harrison-Walker, 2001). The ease and convenience of online information exchange unlocked the ability for individuals to share their thoughts and opinions on their day-to-day encounters with a global audience in an impersonal environment - known widely as electronic word-of-mouth (eWOM) (Sun, Youn, Wu & Mana, 2006). Traditional WOM and eWOM differ in the mode, speed and reach of which the information travels (Phelps, Lewis, Mobilio, Perry & Raman, 2004). Phelps et al. (2004) argue that eWOM consists of primarily text-based information compared to vocal transmission of traditional WOM, while eWOM also travels much quicker and is able to reach more individuals with significantly less effort. eWOM propagation through text sparked new methods of analysing consumer behaviour - specifically natural language processing and text analytics. These novel methods have been used on user generated content in studies across a plethora of fields ranging from e-commerce to hospitality and medicine (Finch, 1999; Xiang, Schwarts, Gerdes & Uysal, 2015; Cunningham, Tablan, Roberts & Bontcheva, 2013).

In particular, the hospitality industry has been heavily affected by the digitization of our world. The internet brought with it online travel agencies (OTAs) giving consumers an opportunity for users to book their own trips and stays, without having to pay an external agent to do so. McCarthy, Stock & Verma (2010) discovered that the large majority of consumers responding to their survey on hotel purchase decisions prefer online search engines as a method of information collection prior to making a final decision. Thus, consumers are able to access a larger pool of information prior to decision making. An important segment of the information they may access is user generated reviews, which allow them to see what others thought of a particular hotel before booking and reduce uncertainty when booking a vacation (Woodside & King, 2001). These reviews have been shown to drastically affect the behaviour of consumers, with over 60% of consumers finding that reviews are important when making a purchase decision (Smith, 2013). With the popularity of user generated reviews, it becomes important for managers and companies to understand which factors are most important to potential travellers, and adjust their services and offerings accordingly.

In this thesis, we investigate consumer generated reviews using state-of-the-art text analytics and machine learning methods to extract sentiments that are important to consumers in the tourism industry. These novel methods allow for extraction of key information from a large amount of text reviews, greatly improving processing efficiency compared to manually reading each review. We use techniques such as aspect based sentiment analysis to understand how consumers feel regarding different factors of a given hotel, while addressing consumer heterogeneity by accounting for cultural differences. Further, hotels are broken down into five types with the aim of uncovering specific factors of consumer satisfaction that may have been otherwise masked amongst more common aspects.

## 1.1: Research Question

The aim of our research is to discover factors important to consumers when making hotel booking decisions. As such, we will use multilingual user generated reviews obtained from Booking.com of hotels across Europe that are split into different categories based on their amenities and descriptions. Hofstede's cultural dimensions will also be used to address cultural heterogeneity amongst consumers. By doing this, we want to answer the following research question:

*What are the key hotel-based and cultural aspects contributing to consumer satisfaction in different hotel types?*

In order to answer this question, we will use the following methodological approach:

1. Text will be heavily preprocessed to eliminate noise and generate a usable dataset
2. Latent Dirichlet Allocation (LDA) will be used to extract topics from reviews of each hotel type, using different tokenization methods (unigrams, bigrams, trigrams, and k-skip-n-grams)

3. A simple and interpretable classification method (logistic regression) will be used to develop a sentiment analysis model to classify reviews based on previously extracted aspects while also examining the effect on overall satisfaction of the individual's cultural background and economic indicators
4. A logistic regression model will be then run with interactions between cultural dimensions, economic indicators and factors of satisfaction to extract insights into how these combinations amplify or decrease satisfaction
5. A classification method that allows for cohort splits will be used (logistic regression trees) to better analyse direct impacts of cultural and economic indicators on factors of consumer satisfaction

Following preprocessing, topic extraction and classification, factors that are of significance will be analysed and compared across hotel types, and the models evaluated.

### 1.2: Relevance & Contribution

Consumer generated reviews make up a significant portion of a consumer's purchase decision journey as they contain valuable information for prospective consumers. Thus, it is crucial to examine what is meaningful to consumers to adjust products and services accordingly. This thesis will provide a way for managers to efficiently examine user generated content, while also exploring features that consumers deem important to them during their stay. This thesis contributes to existing literature by classifying hotels into different types based on their descriptions and amenities and examining cultural and economic dimensions, so as to discover how consumer heterogeneity contributes to satisfaction while finding unique aspects of satisfaction for certain hotel types. Moreover, this study is one of the first to make use of neural translation models to extract sentiments and topics from hotel reviews in languages other than English. This allows us to better examine cultural heterogeneity amongst consumers that may write reviews in their own language, as opposed to solely in English.

# Section 2: Theoretical Background

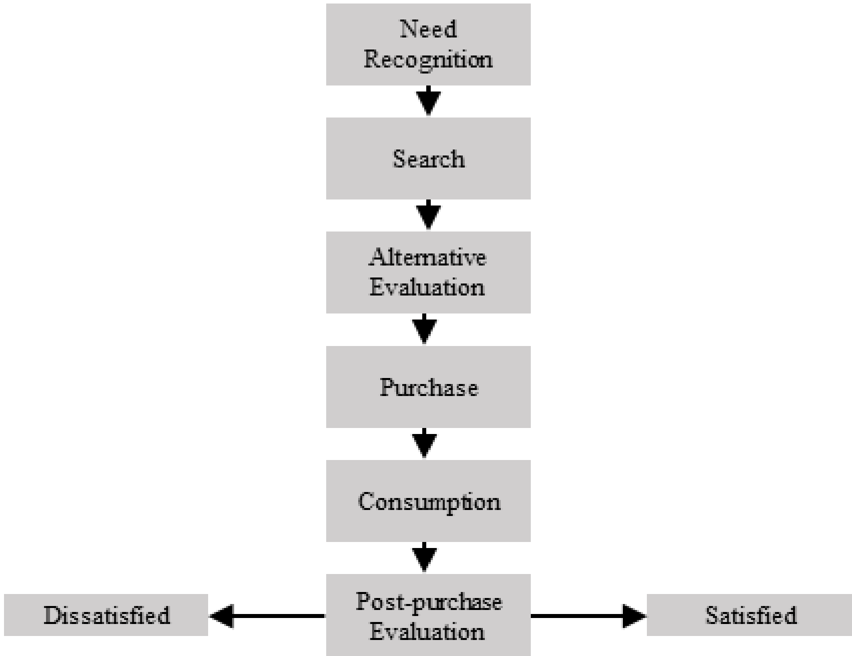
This chapter focuses on analysing literature based on three key aspects. Firstly, an examination into consumer behaviour with its relevance to this thesis is presented. Next, aspect based sentiment analysis is introduced from a literature perspective, and its importance to our thesis explained. Finally, a literature review of existing studies in the hospitality industry is presented.

## 2.1: Consumer Behaviour

When faced with a product, consumers have to undergo a decision making process to determine whether or not they are willing to buy that product. This process relies heavily on the information available to consumers from a variety of different sources (Huber & McCann, 1982; Haubl & Trifts, 2000; O'Brien, 1971). Previously, the sources of information were limited to word-of-mouth (WOM) communication between individuals, and the information that they received directly from the retailer (O'Brien, 1971; Phelps et. al., 2004). Huber & McCann (1982) show clearly that omitting a piece of information, such as an attribute of a product - be it price, or quality - has a significant negative impact on the likelihood to buy of a consumer. Thus, consumers that have transparent access to more information about a product, without any asymmetry between them and the retailer, are more incited to engage in a purchase decision. An especially important factor in consumer decision making is the information received from WOM communication, with some researchers finding that it has a significantly larger impact on purchase decision compared to traditional advertising (O'Brien, 1971).

The advent of the internet allowed for the ability of individuals to communicate anonymously in a one-to-many platform through electronic word-of-mouth (eWOM) (Jalilvand., Esfahani, & Samiei, 2011). Thus, the amount of information available to consumers has drastically increased with billions of users now able to exchange opinions in a simple manner. With the increase in available information, consumers now must be more selective than ever before when parsing their sources before making a purchasing decision. Engel, Blackwell & Miniard (1995) define the consumer decision making process as consisting of six key steps. First, the consumer must recognize that they have a problem which a given product or service can resolve - this step is known as need recognition. Following this, the consumer will search for information to address their needs - this is a crucial step wherein the large amount of information available to users on the internet plays a role. Further, consumers will evaluate their available options before actually making a purchase decision of their chosen alternative. Finally, the last two stages consist of a consumption decision, and a post-purchase alternative evaluation whereby consumers will determine their satisfaction or dissatisfaction with the chosen good. Figure 1 summarises the consumer decision making process as proposed by Engel et al (1995).

**Figure 1:** Consumer decision making process (Figure adapted from Engel et al., 1995)




Consumers make decisions based on the information they gather, and in exchange, they can easily share their opinions on their chosen product, facilitated nowadays by the internet. These user generated reviews give us insight into the last step of the consumer decision making process - the post-purchase evaluation, as we are able to determine whether or not an individual was satisfied, and which factors contributed to that satisfaction.

Engel et al. (1995) state that environmental influences such as culture as well as individual differences also play a part in the consumer decision making process. Thus, consumer heterogeneity clearly has an impact on how a consumer will perceive and evaluate a given product or service. Several researchers have stated that cultural heterogeneity is important when tied to consumer behaviour, as it could change the way they make decisions (Kacen & Lee, 2002; Petersen, Kushwaha & Kumar, 2015). One of the most widely adopted methods for measuring cultural differences is Hofstede's cultural dimensions - these are numerical values of 6 different dimensions relating to the culture of over 100 countries (Hofstede, 2013). These dimensions consist of the power distance index, individualism versus collectivism, masculinity versus femininity, uncertainty avoidance index, long term orientation versus short term orientation and indulgence versus restraint. The power distance index measures whether people in a given society accept that power is distributed unequally, thus a higher value in this index means that the majority of people accept this unequal distribution. A lower value indicates that people demand for equality of power and find inequalities unjustifiable. Individualism versus collectivism measures if members of a society are expected to care for themselves and immediate families only (this leads to a higher value), or if the members work together to look after one another (leading to a lower value). Masculinity versus femininity looks at how competitive a society is, whether the values are attributed with success, achievement, heroism and assertiveness (higher value, more masculine) or whether they are associated with modesty, quality of life and cooperation (lower value, more feminine). The uncertainty avoidance index examines how uncomfortable a certain society is with the unknown and outstanding, higher values typically mean a society is more rigid in its constructs and expects behaviour to be heavily assimilated, while lower values are attributed to more relaxed societies. Long term versus short term orientation measures whether a society is thrifty and pragmatic in terms of change (higher values - long term oriented) or more normative and prefers tradition (lower values - short term oriented). Finally, indulgence versus restraint measures whether members of a society prefer to enjoy the fruits of life, and have fun (higher values - indulgence) or whether they are restrained and must adhere to social norms (lower values - restraint). All of these explanations are adapted from Hofstede's (2013) work that clearly defines each dimension and how they are scored.

These cultural dimensions will be used in this thesis to examine cultural heterogeneity amongst consumers and determine whether their cultures have an impact on how they review certain hotels.

## 2.2: Methodology

Analysing consumer sentiments from text once used to be an arduous, lengthy task with individual reviewers having to comb over the text and manually annotating each review (van Atteveldt, van der Velden & Boukes, 2021). Nowadays, this process can be fully automated through the implementation of various text analytics and natural language processing techniques, known broadly as sentiment analysis. This allows us to examine many consumer reviews efficiently, while also being able to summarise overall sentiments towards a given entity (Hu & Liu, 2004).

Liu (2010) defines three main types of sentiment analysis - document level, sentence level and aspect based. Document level sentiment analysis determines whether an entire text expresses a positive or negative sentiment towards a certain feature. Sentence level determines whether a given sentence within a document contains a positive or negative sentiment, while aspect based sentiment analysis determines whether there is a positive or negative opinion derived from a certain feature that has been discussed. For example, if a reviewer leaves the following review: *'I loved the hotel, the breakfast was splendid. However, the bed was uncomfortable'* document level sentiment analysis would extract that this entire review contains positive sentiments (loved, splendid), but it also contains negative sentiments (uncomfortable). However, we would not know what these sentiments relate to, and since two contradictory sentiments are extracted, we cannot learn anything from this review. Sentence level sentiment analysis would be able to classify the first sentence as positive, and the second as negative, but once again, we are unable to determine what these sentiments relate to. Feature based sentiment analysis would associate the hotel with a positive sentiment (loved), the breakfast with a positive sentiment (splendid) and the bed with a negative sentiment (uncomfortable). Thus, for our research, as we aim to find out which specific features of a hotel contribute to consumer satisfaction, aspect based sentiment analysis will be our preferred methodology.

Aspect based sentiment analysis typically consists of three main steps - identification, classification and aggregation (Tsytsarau & Palpanas, 2011). The identification step typically consists of extracting aspects from the text and the corresponding sentiment, followed by the classification of that sentiment as positive or negative. Finally, all of the corresponding sentiments for a given aspect are aggregated and the final sentiment is summarised. Flavius & Frasincar (2015) present an in-depth survey on aspect-level sentiment analysis, and propose that there are multiple methods to perform aspect detection in a text ranging from frequency-based to more complex methods using machine learning. Frequency based methods rely on extracting single or compound noun pairs that are frequently used in the text that is being analysed. However, this method allows for added noise as some frequently used nouns that have a high frequency may be misinterpreted as aspects. Further, other methods of aspect detection such as syntax-based methods exist but are much harder to implement. Syntax-based methods require relationships between certain word pairs and structures to be predefined, thus an in-depth knowledge of the language is required. One method of aspect detection that seems to be successful and widely used, is the use of unsupervised machine learning to extract key aspects from text (Flavius & Frasincar, 2015). With the use of topic modelling, and more specifically Latent Dirichlet Allocation (LDA), aspects can be automatically extracted from text and their sentiments classified based on the words associated with them (Moghaddam & Ester, 2013; Blei, Ng & Jordan, 2003). Several studies have gathered positive results by using LDA as a tool for aspect extraction, with final sentiment classification accuracies ranging in the 79%-97% range (Lu, Ott, Cardie & Tsou, 2011; Lakkaraju, Bhattacharyya, Bhattacharya & Merugu, 2011; Moghaddam & Ester, 2013).

Following aspect extraction, we will then need to classify our aspects. This is usually done in several ways - either through syntax-based analysis, supervised learning or unsupervised learning (Flavius & Frasincar, 2015). As we are looking at hotel reviews, typically each review is associated with a given score. Thus, we can use a supervised learning method such as logistic regression to classify the aspects we extracted into either positive or negative, by analysing how they affect the reviewer rating overall. This has been done in multiple studies using multiple review sources, such as product and movie reviews, camera reviews and hotel reviews - all with decent F1 scores ranging between 70% - 88% (Li, Han, Huang, Zhu, Xia, Zhang & Yu, 2010; Jin, Ho & Srihari, 2009; Zim, Niepert, Stuckenschmidth & Strube, 2011).

### 2.3: Previous Studies in the Hospitality Industry

Several studies have been done already using text analytics in the hotel industry to analyse a variety of different factors. Table 1 summarises selected studies in the hospitality industry. Across the studies that have been performed, investigating consumer satisfaction is quite prevalent. Berezina, Bilgihan, Cobanoglu & Okumus (2015) examined hotel reviews for Saratosa, Florida and classified sentiments from the texts using more rudimentary methods such as manual word categorization and text link analysis. They found that satisfied customers typically tend to focus more on intangible aspects, such as service received at the hotel, while unsatisfied customers focus more on tangible aspects like the cleanliness of their room. Büschken & Allenby (2016) took this analysis a step further by splitting up hotels into two main categories - upscale hotels in Manhattan and midscale hotels by JFK Airport. This was done to see if these different types of hotels would provide differing aspects that relate to consumer sentiments. They used LDA to model topics from both hotel types, and classified the aspects with their sentiment effects by using logistic regression to model the review ratings. They found that upscale and midscale hotel topics did not differ too significantly, although there were some topics such as noise and smell that were found in midscale hotels but not in upscale hotels. Further, the midscale hotels proximity to JFK Airport yielded unique topics to do with JFK and even the shuttle to the airport. However, the authors did not experiment with different tokenization, and did not account for any heterogeneity across consumers. Guo, Barnes & Jia (2017) performed a similar analysis on hotels amongst 16 different countries, using LDA as the main method to extract aspects. They then performed stepwise regression to classify the sentiments, and created a perceptual map as a managerial tool to summarise important findings. They managed to split up the uncovered aspects into 'controllable' and 'uncontrollable' dimensions - whereby controllable dimensions can be improved by managers, such as staff service & room experience, while uncontrollable factors cannot be adjusted, like location. They also found that the heterogeneity between age & gender of reviewers did impact their overall review score. Further, room experience and communication were found to be important for low end hotels, while homeliness and events were important for high end hotels. Zhang (2019) performed a study on AirBnb consumer satisfaction using LDA to extract topics from over 1 million reviews. Zhang found that negative reviews tend to be more authentic and credible than positive ones, and contain more relevant information in determining whether an AirBnb is actually of quality. Further, it was also found that AirBnb's have unique topics compared to hotels with mentions of AirBnb specific words like 'host' and 'apartment'. However, no sentiment analysis was done following the topic extraction. Finally, Chang, Liu, Xu, Li & Hsu (2020) studied consumer satisfaction in the luxury hotel industry, using Bag-of-Words and doc2vec to extract sentiment aspects from reviews. Further, these aspects were assigned polarity to them based on a given dictionary. Random forest was then used to classify the sentiments. They found that luxury hotels should focus on ensuring staff and cleanliness of utmost quality while the location does have a tremendous impact on the consumer review. Thus, hotels marketed as luxury that are not in a prime location may get heavily docked by consumers.

Other studies done in the hospitality industry include assessing review helpfulness and even examining brand attitude. Chaterjee (2020) investigated the factors determining how helpful user generated hotel reviews are, by conducting sentiment analysis to associate polarity and emotions to a given review, then classifying this review as helpful or not using various machine learning methods. Chaterjee found that higher sentiment content and polarity tend to lead to less helpful reviews, as they may be emotionally charged and highly biased. Low arousal negative emotions tend to lead to more helpful reviews - as they are calmer, and have a natural tone. Further, Chatterjee also noted that machine learning techniques were better at detecting review helpfulness compared to other econometric-based techniques. Ray, Bala & Rana (2021) used user generated hotel reviews to determine brand attitude towards hotels by extracting sentiments and performing part of speech analysis before classifying each review into a positive or negative brand attitude. They found that extracted sentiments have a significant impact on determining a brand attitude, while parts of speech were not useful in the analysis.

As seen in the previous studies, many researchers have chosen to focus on the hospitality industry when performing text analysis. However, there are certain limitations to the previous studies that this paper aims to investigate. Firstly, cultural heterogeneity is not examined and accounted for in any of these studies. Next, these studies only focus on extracting sentiments from a broad group (or groups) of hotels, without further dividing the hotels into buckets based on the facilities and activities they have. As the research by Büschken & Allenby (2016) suggests, examining extracted aspects from different hotel types may give additional insights into what consumers like about a specific hotel. Finally, none of these use multilingual reviews, frequently mentioning that they drop all non-English reviews, thus they cannot extract how international consumers felt about a given hotel. In a world where globalisation is thriving, especially in the hospitality industry where most of the consumers are international, understanding how all consumers feel is important. Thus, discriminating based on language could lead to loss of important information that may be crucial to other consumers.

**Table 1:** Summary of previous studies in the hospitality industry

| Authors | Topic | Data | Method | Findings | Limitations |
|---|---|---|---|---|---|
| Berezina, Bilgihan, Cobanoglu & Okumus, 2015 | Consumer Satisfaction from hotel reviews | 2510 TripAdvisor hotel reviews for Saratosa, Florida | •PASW modeler<br>•Word categorization<br>•Text link analysis | Satisfied customers -> focus on intangible aspects (service)<br><br>Dissatisfied customers -> focus on tangible aspects (cleanliness) | •No sentiment analysis / review rating modelling<br>•Focused on one city |
| Zhang, 2019 | Consumer Satisfaction from Airbnb reviews | 1,026,988 Airbnb reviews in 7 cities in the USA | •LDA | Negative reviews more authentic and credible<br><br>Unique topics on Airbnb compared to hotels | •No sentiment analysis / review rating modelling<br>•Focused on one country |
| Ray, Bala & Rana, 2021 | Consumer Brand Attitude from hotel reviews | 10,000 TripAdvisor hotel reviews | •Sentiment analysis (content, polarity, arousal)<br>•Part of speech analysis<br>•Logistic regression<br>•ANN | Sentiments most important in predicting brand attitude<br><br>Parts-of-speech aspects have no significant impact on brand attitude | •No review rating modelling / consumer satisfaction analysis<br>•No topic modelling / LDA |
| Chaterjee, 2020 | Review helpfulness | 942 TripAdvisor hotel reviews | •Sentiment analysis (content, polarity, arousal)<br>•Logistic regression<br>•ANN<br>•Random forest<br>•SVM | Higher sentiment content and polarity lead to less helpfulness<br><br>Low arousal negative emotion leads to helpfulness<br><br>Machine learning techniques perform marginally better than econometric techniques | •No review rating modelling / consumer satisfaction analysis<br>•No topic modelling / LDA |
| Hu, Zhang, Gao & Bose, 2019 | Consumer Satisfaction from hotel reviews | 27,864 TripAdvisor hotel reviews for New York City | •STM<br>•Topic correlation analysis | Customer complaints for high end hotels mainly related to service; facilities for low end hotels | •No review rating modelling<br>•Focused on one city |
| Guo, Barnes & Jia, 2017 | Consumer Satisfaction from hotel reviews | 266,544 TripAdvisor hotel reviews for 16 countries | •LDA<br>•Stepwise regression<br>•Perceptual mapping | 19 controllable dimensions for consumer satisfaction (checking in/out, resort facilities, communication, homeliness, bathroom, room experience etc.)<br><br>Heterogeneity among demographic groups (age & gender)<br><br>Room experience & communication important for low end hotels, homeliness and events for high-end | •No analysis including different tokenization |
| Chang, Liu, Xu, Li & Hsu, 2020 | Consumer Satisfaction from hotel reviews | 500,000 Booking.com luxury hotel reviews in Europe | •BOW/doc2vec<br>•Sentiment analysis (sentiment polarity – pos/neg)<br>•Random forest | Luxury hotels should focus on staff training, cleanliness and location<br><br>RF efficient and well performing model for sentiment classification | •No topic modelling / LDA<br>•No analysis including different tokenization |
| Büschken & Allenby, 2016 | Consumer Satisfaction from hotel reviews | 3,212 expedia.com upscale hotel reviews in Manhattan<br><br>1,255 expedia.com midscale hotel reviews in Manhattan | •LDA (regular, sentence-constrained, sticky)<br>•Logistic regression | Sentence based topics found to be more distinguished and coherent than word-based analysis<br><br>Upscale topics include – check-in, attractions, recommend, noise and room negative, room positive, location, amenities, staff<br><br>Midscale topics include – noise/smell, recommend, food, service, room/free amenities, front desk, JFK, shuttle | •Focused on one city<br>•No analysis including different tokenization |

## Section 3: Data

The dataset used for this research was scraped from Booking.com, a Dutch multinational online travel agency that travellers can use to reserve their accommodations and travel plans. Booking.com is one of the world's largest online travel agencies (OTAs), with over 28 million listings in over 43 languages, and dominates the OTA market coming in second to Airbnb in terms of market cap (Booking.com, 2022; Statista, 2022). Thus, with such a large presence in the hospitality industry, Booking.com was chosen as an ideal resource to analyse consumer reviews for hotels worldwide. The data was scraped by using a proprietary web crawler, built in python, to extract approximately 500 hotels and their reviews for 11 European countries. Europe was chosen as the market to explore as it hosts the largest number of tourism arrivals annually, with 746 million tourists in 2019 from all over the world (UNWTO, 2022). The final dataset consists of 500 hotels, and 171,299 total consumer reviews.

### 3.1: Data Scraping

Using python, a web crawler was built to navigate through the Booking.com webpages and extract the first 1000 hotel results for each country analysed. The scraper captured the hotel name, the overall hotel rating, the number of reviews, the approximate location of the hotel, and its URL from the results using the name of the country as a search term. Following the generation of the list of the hotels, the crawler scraped through the URLs of each of the hotels and captured more detail about each - the given description, popular facilities, all facilities, the overall location, value for money, staff, comfort, cleanliness, and facilities ratings as well as the accommodation type. This data was merged with the list of the hotels to create our first database consisting of all the hotels to be analysed and generic information about them.

Secondly, the list of hotels was refined by keeping only hotels with > 300 reviews, as too few reviews may add noise and the hotel may not be popular, thus the insights we would be able to generate from these reviews may not be significant. Further, a hotel classification was added based on the description of the hotel, the popular facilities, the hotel type, and the hotel name. Initially, k-means cluster analysis was used to try to extract natural groupings of the hotels based on description and facilities (similarly to how Booking.com allows users to search for specific features in a hotel), however, the clusters were not meaningful - we only saw two main clusters without an interpretable split between them. Thus, for analysis a total of 5 hotel categories were chosen to explore - beach hotels, ski hotels, city hotels, business hotels and residential accommodations (including apartments), derived manually from popular facilities amongst hotels. To ensure consistency among our analysis and to speed up scraping and ease computational load, 100 random hotels in each category were kept for analysis.

Finally, the adjusted list of hotels was used to extract all consumer reviews from each hotel in a separate database. This data consists of the review title, reviewer name, reviewer nationality, review date, stay date, review text and review rating. Further, we kept only hotel reviews that had more than 10 words, since longer reviews tend to give more insights into why the consumer either enjoyed or did not enjoy the hotel (Schindler & Bickart, 2012). Since Booking.com is accessible by consumers worldwide, there are many reviews in our dataset that are not English. As such, for further analysis we decided to study reviews written in other languages, through the means of deep translation. Review language was first detected by using the spaCy language detection pipeline in python - it was chosen due to its state-of-the-art speed and high language detection accuracy of 97% on average (Honnibal & Montani, 2017)**.** The reviews were then translated using the Opus-MT deep learning translation model developed by the Language Technology Research Group at the University of Helsinki, due to its ability to translate entire sentences while preserving context, which is highly important for sentiment analysis (Tiedemann & Thottingal, 2020). These neural translation models are trained on a collection of texts known as OPUS, which consists of over 3.2 billion sentences (Tiedemann, 2016). We translated and kept the reviews of the 6 most common languages in our dataset - Dutch, French, German, Spanish, Swedish and Italian. Our final dataset has 171,299 English (original & translated) reviews, with 26,929 reviews for beach hotels, 14,198 for ski, 37,687 for business, 14,543 for residential and 27,551 for city hotels.

### 3.2: Preprocessing

The review text contains all of the raw reviews that have been scraped from the website - in this thesis we will refer to these texts as the corpus. Before doing any further analysis, it is important that we preprocess the corpus to allow it to be in a digestible format for our models. Firstly, basic preprocessing is done, involving transforming the corpus into all lowercase values since a computer will not easily recognize that ‘Hotel’ and ‘hotel’ are the same word. Contractions were then expanded to get the full meaning of the words, and remnants of HTML web scraping were removed (for example, unicode characters like \r and ._x000D_, which are used in parsing HTML text). Further, all English stop words were removed, to ensure that we do not include high frequency meaningless words in our analysis, such as 'and'. We also kept a set of corpi without removing negations to be used in bigram, trigram, 1-skip-1:2-gram and 1-skip-1:3-gram tokenization. Next, highly frequent words that do not add any meaning were removed - in our case, 'hotel' and 'room' which occurred in almost every review. Finally, punctuation and whitespace were removed, leaving us with a final corpus for each grouping ready for analysis.

### 3.3: Cultural Dimensions

Hofstede's cultural dimensions were obtained from their proprietary website, and consist of data for over 100 countries (Hofstede, 2016). The obtained data was appended to our review dataset by means of matching the reviewer nationality to the country from Hofstede's data. All rows with null values, due to a cultural dimension not being defined for a specific country, were removed so as to not interfere in our model.

### 3.4: Economic Indicators

Economic indicators were included in the modelling to determine whether they have an impact on consumer satisfaction. Two main macroeconomic indicators were chosen - GDP per capita (PPP) and unemployment rate - for further analysis. The reasoning for this choice is that these macroeconomic features best represent the health of an economy - GDP per capita (PPP) allows us to adjust for purchasing power and wealth, while unemployment rate tells us how active individuals are in an economy. Both metrics were obtained from the World Bank (2022), and appended to the dataset based on the reviewer nationality. Since not all countries in our dataset had values for both of these economic metrics available, all entries with a null value were removed prior to modelling.

### 3.5: Descriptive Statistics

Taking a closer look at the data we have extracted, from Figure 2 we see that the review rating variable is on a scale from 1 to 10, with the most observations falling in the 8-10 range. Thus, the dataset is highly imbalanced with only 5% of reviews falling between 1 and 4. To account for this issue, we must balance the training dataset following the train/test split so as to ensure our predictive models can learn about factors of negative consumer satisfaction as well.

**Figure 2:** Distribution of review rating in dataset

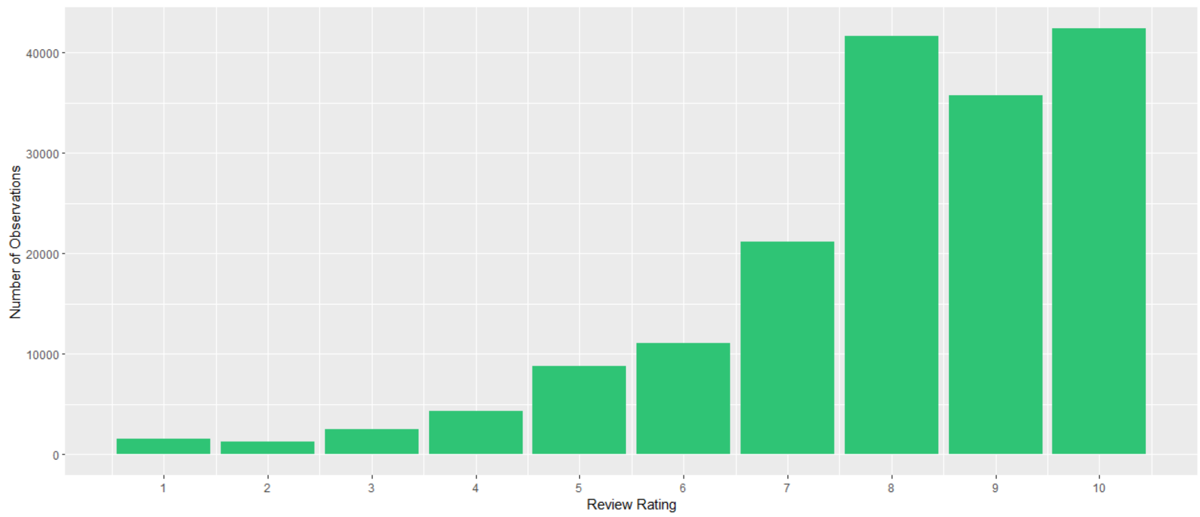
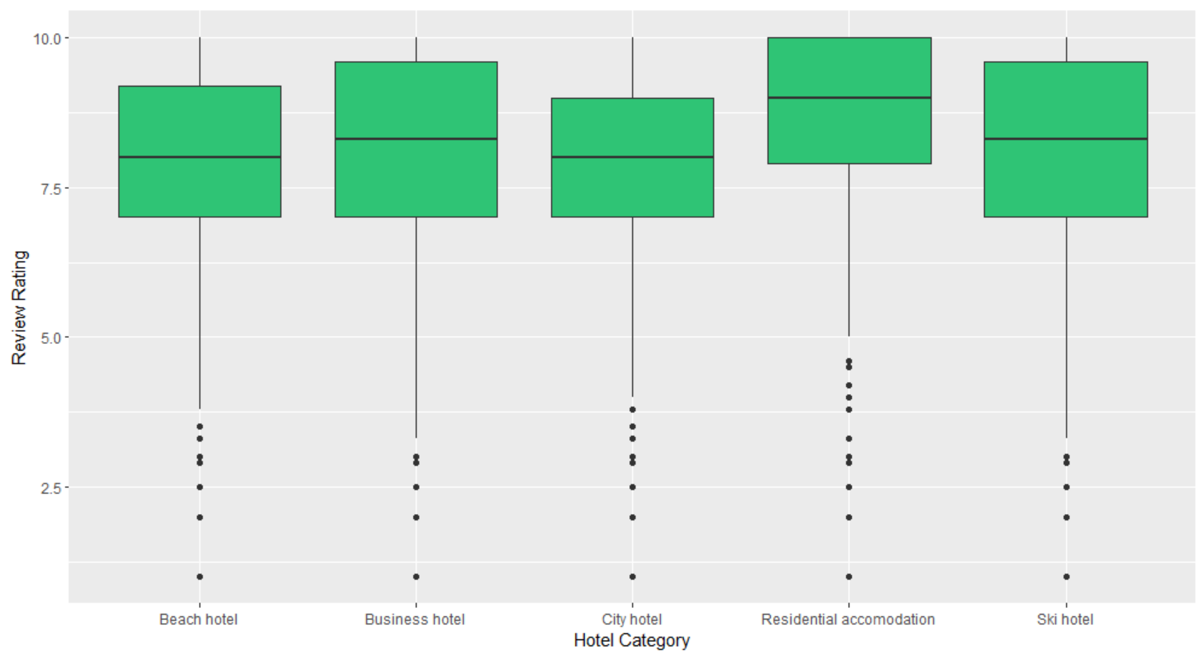



Figure 3 exhibits the distribution of review rating by hotel type. As seen, the average review rating and overall distribution is quite similar across five types, thus no further adjustments are needed prior to analysis.

**Figure 3:** Boxplot of review rating by hotel type




## 3.6: Feature Engineering

To perform sentiment analysis on the prepared corpi, we must first transform our outcome variable - review rating - to a binary format for ease of interpretation and classification. To do this, we have taken all reviews with a rating less than or equal to 4 as 'negative' reviews, and those with a rating greater than or equal to 8 as 'positive' reviews - a new variable named review quality was created with 0 identifying negative reviews and 1 identifying the positive. The reasoning behind this split was that reviews with ratings less than or equal to 4 are likely to be highly negative, especially since consumers tend to rate quite favourably, as we see from our split above. From these highly negative reviews, we aim to extract factors that are severely detrimental to consumer satisfaction, allowing our model to properly classify sentiments based on the reviews. If we had included reviews from ratings 5 to 7, we would likely encounter sentiment classification issues since these reviews are likely to be neutral. These neutral sentiments can be difficult to classify, and thus would hinder model performance. Similarly, we chose the reviews above rating 8 to be our positive ones as they are likely to contain highly positive sentiments, allowing our classifier to learn these easily. Analysing the distribution of the new feature for each dataset shows that there is severe imbalance with a skew towards the positive reviews. This unbalancedness could have an impact when we create our classification models, and as such we decide to upsample our training data to improve negative review prediction in our classification models.

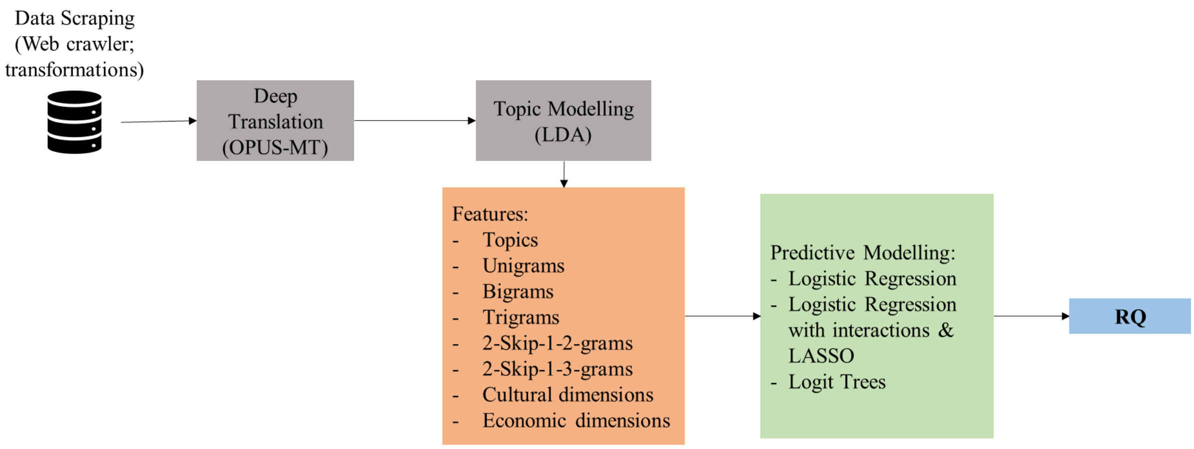
# Section 4: Methodology

This section details the various methods used to perform the analysis in this paper. Firstly, the research framework is presented. Then, latent dirichlet allocation - as a method for topic extraction - will be discussed. Further, the classification methods used in this paper - logistic regression and local logit trees - will be explained. Finally, the evaluation process of our models will be presented.

## 4.1: Research Framework

The research framework consists of five key steps: data scraping, deep translation, topic modelling, feature extraction and predictive modelling. Firstly, data is scraped using a web crawler from Booking.com for all hotels. The data is then transformed and cleaned to be prepared for translation. Reviews are then translated to English using a deep translation model, in preparation for further text analysis. Text preprocessing is then performed as part of this step. Thirdly, LDA is run to extract topics from reviews split among five hotel types. Then, features are taken from the LDA results with different tokenizations, combined with Hofstede's cultural dimensions and two macroeconomic indicators. Finally, three predictive models are run - logistic regression, logistic regression with interactions and LASSO and logit trees - the results of which will be used to answer the research question. This framework is summarised in Figure 4.

**Figure 4:** Summary of Research Framework




## 4.2: Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a topic modelling method commonly used to extract and cluster topics based on raw text. First developed by Blei, Ng & Jordan (2003), LDA is a probabilistic model whereby text can be separated into different topics, based upon a distribution of words occurring in each of these topics. Certain words that co-occur frequently can lead to a topic being extracted, for example if we have multiple reviews that have the words 'walk' and 'beach' occurring frequently, then a topic relating to the distance of the beach from the respective hotel may be discovered. LDA is used in our research to extract topics from each of the different hotel types, compare them amongst each other to see if there are unique topics per type, and finally use the probability of topics in a classification task whereby we can predict their impact on consumer satisfaction.

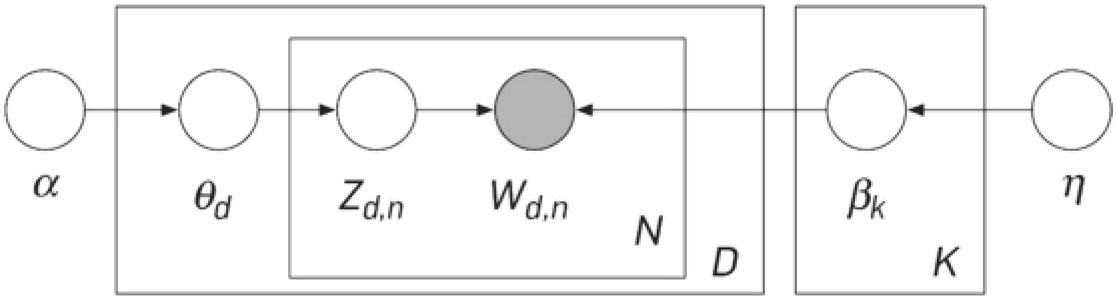
From a technical perspective, LDA is a generative probabilistic model that typically uses a corpus of text as an input and provides a probability distribution vector for each topic extracted over all of the documents in the corpus (Blei et al, 2003). The extracted topics are unobservable without the use of the model, and are thus often referred to as 'hidden' - hence the need for their extraction (Blei & Lafferty, 2009). Further, all of the extracted topics have some probability value for each document, and all topic probability values for a single document must sum up to 1 (Reisenbichler & Reutterer, 2018). To get this output, the model calculates a conditional joint distribution of the topics and the observed variables - or rather, the words within a document (Reisenbichler & Reutterer, 2018). Formally, a topic can be defined as ‘a distribution over a fixed vocabulary’ (Blei, 2012). Figure 5 represents a graphical model of LDA, which shows that LDA can be further subdivided into three main plates - the corpus level (unlabeled exterior plate), document level (denoted by the M plate) and word level (denoted by the N plate). Each of these levels contain their own parameters that need to be defined in the model.

Formally, at the document level, the topic proportions for the *d*th document are denoted by $\theta_d$, which tells us how the topics are distributed across each document (Blei, 2012). Further, at the word level, the topic assignment for a given word n is denoted by $Z_{d,n}$ which tells us the probability that a word belongs to a topic (Blei, 2012). The observed word n for a document d is denoted by $W_{d,n}$ (Blei, 2012). Finally, the distribution over all words for a given topic is denoted by $\beta_k$, which is the probability of words belonging to a topic k. As seen in Figure 5, certain dependencies exist between the different parameters. For example, we see that $Z_{d,n}$ depends upon $\theta_d$, and in turn, $W_{d,n}$ depends upon $Z_{d,n}$ and $\beta_k$ (Blei, 2012). Notice that there are two parameters that are needed to calculate both $\theta_d$ and $\beta_k$ - these are the document topic density α and the topic word density η. To calculate the document topic probability distribution, we must use $\theta_d \sim dirichlet(\alpha)$ and similarly, to calculate the distribution over all words for a given topic, we use $\beta_k \sim dirichlet(\eta)$ (Blei, 2012).

To summarise, LDA assumes a generative process consisting of four steps for each document in a corpus (Blei, 2012):

1) Firstly, we need to choose the words N
2) Secondly, we randomly choose the topic document probabilities $\theta_d \sim dirichlet(\alpha)$
3) Thirdly, for each of the words ($W_{d,n}$), we randomly choose a topic ($Z_{d,n}$) and randomly choose a word from the corresponding vocabulary distribution.

**Figure 5:** Plate notation of LDA model (Blei, 2012)




As part of our topic modelling, the number of topics *k* is selected randomly and the resulting model evaluated based on two metrics - perplexity and coherence.

Perplexity is a measure that allows us to evaluate how well a probability based model can predict a never seen before sample (Blei et al., 2003). More formally, it is the inverse of the geometric mean per-word likelihood (Blei et al., 2003). Mathematically, for a set of M documents

$$Perplexity(M) \;=\; e^{(-log(P(W_d))/(N)}$$

where W represents all of the words in the set of documents, and N represents the total number of words. Thus, perplexity is the exponential of the negative logarithm of the probability of all words in all documents, divided by the total number of words. When comparing models, we select the one with the lowest value of perplexity as that indicates better model performance in predicting on an unseen set of data - minimising perplexity can be seen as maximising the probability, since they have an inverse relationship (Blei et al., 2003).

Coherence is another metric that we use to compare topic models. Coherence allows us to see how similar two words are and how often they co-occur within the same topic (Mimno, Wallach, Talley, Leenders & McCallum, 2011). For example, if a topic is extracted related to the cleanliness of the bathroom, then we would expect the words 'clean' and 'bathroom' to occur multiple times within the same document, and be clustered together in the same topic. Since coherence measures the co-occurrence of words in a topic, a higher value of coherence means that the topics defined by our model are better. Thus, to choose the number of topics *k*, we want to maximise coherence.

## 4.3: Logistic Regression

Logistic regression (logit) is a supervised learning method that is generally used for classification tasks (James, Witten, Hastie, & Tibshirani, 2021). With this model, observations can be classified into their predicted class based on predictor variables specified by the researchers. Logistic regression is a relatively simple, but powerful model that allows relationships between predictors and response to be easily discovered and interpreted. Due to the ease of interpretation and implementation, the logit model was chosen as a baseline for this analysis as it allows us to explore relationships between extracted topics from hotel reviews, cultural values, economic indicators and customer satisfaction quite easily. Mathematically, logistic regression is defined as follows (James et al., 2021)

$$log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

where the probability that a chosen sample belongs to class 1 given the predictors $X_n$ is defined by $P(Y = 1|X)$ and the coefficients associated to these predictors are denoted by $\beta_n$. The logistic regression model outputs a set of coefficients related to the predictor variables that determine the magnitude of the effect each variable has on the log odds of the response. Since the coefficients are numeric, they can be interpreted and compared amongst each other with ease.

## 4.4: Logistic Regression Trees

Logistic regression trees are a combination of simple decision trees and logistic regression, whereby the data is first partitioned based on certain criteria and then a logistic regression model is fit at each node (Landwehr, Hall & Frank, 2005). The combination of these two popular methods gives a final single tree that can be interpreted by the logistic regression results found in each leaf. With these regression results, effects of predictor variables can be quantified and relationships between predictor, response and split variables can be made. The tree follows a process whereby the data is recursively partitioned based upon selected splitting variables, creating binary splits. Each node contains a logistic regression model that is run on that given partition of the data. The benefit of this type of model is that it becomes quite easy to see if certain predictors differ based upon the splits of data they are run on. In the case of this thesis, the logistic regression tree model will be used to identify significant cuts of data based upon cultural and economic factors, whereby logistic regression models will be run at each node consisting of the LDA extracted topics as predictors, and review rating as the response. This method allows the examination of how topics differ in importance based upon differing cultural and economic indicators, allowing the formation of clearer relationships between factors of satisfaction and individual background. Logistic regression by itself provides us with a basic framework that allows the investigation of significant factors of consumer satisfaction, but does not allow us to examine relationships between these factors easily. Interaction effects could be modelled, but due to a large number of predictors, it is not computationally feasible or easily interpretable. Further, these interaction effects would not show the true differences in cohorts consumer preferences as the model would be run on the entire dataset, not on partitions of it.

Formally, the data is recursively partitioned through the performance of a parameter instability test - if there is instability with respect to any splitting variables, the node will be split into two (Zeileis, Hothorn & Hornik, 2008). Zeileis et al. (2008) outline four key steps of the recursive partitioning algorithm. The first step of the algorithm is to fit the model using all observations in the node, and thus estimating the parameter set β, by minimising the respective objective function. In this case, the objective function will be the negative log-likelihood. Secondly, the parameter estimates are assessed by running through the fluctuation test, to determine if any instability exists. If instability is determined, the parameter with the highest instability is chosen as a splitting variable. Thirdly, the split point is then calculated ensuring that the objective function related to the model being run is optimised. Finally, the node is split into two daughter nodes and the algorithm is repeated until a stopping criterion is met.

Mathematically, the test statistic for parameter instability of numeric variables can be defined as follows (Zeileis et al., 2008):

$$\lambda_{supLM}(W_j) = max_{i=1,\dots,n}(i/n * (n - i)/n)^{-1}||W_j(i/n)||_2^2$$

where *i* corresponds to a given observation, *n* relates to the number of observations in the node, and $W_j$ is the partial sum process of the scores obtained from the model that is run. The test statistic is then compared to a p-value generated from comparing the test statistic to asymptotic critical values as defined by Andrews (1993). If the p-value is significant at a chosen level of α, then the split occurs using that splitting variable.

When implementing logistic regression trees in R, the partykit package is used (Hothorn & Zeileis, 2015; Zeileis et al., 2008). Within this package, the logistic regression tree model can be tuned according to several parameters, which are the same as those used by classical decision trees:

1. **minsplit**: the minimum number of observations in a node to be considered for a split
2. **maxdepth:** regulates how many layers the tree may grow

Following tuning, the resulting logistic regression tree model can be plotted and the splits and resulting logit models analysed per node to find any differences amongst cohorts.

## 4.5: Model Evaluation

To evaluate the logistic regression and tree models that we create, three measures will be used to choose the best model: accuracy, F1 score and AIC. Accuracy is a measure that allows us to see how many predictions our model has classified correctly, which can be defined as follows (James et al., 2021):

$$Accuracy = \frac{TP + TN}{\{TP + TN + FP + FN\}}$$

As we have initially unbalanced data, it can be helpful to also examine the F1 score (post balancing) to determine if our model can classify predictions correctly into both the positive and negative classes. The F1 score is defined as the harmonic mean of both precision and recall (James et al., 2021). Mathematically, F1 score is defined as:

$$F1 = 2\,\frac{Precision * Recall}{Precision + Recall}$$

where precision and recall are further defined as:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

thus, precision and recall show the proportion of true positives our model predicts compared to the total predicted positives and compared to the actual positives respectively.

Thirdly, the Akaike Information Criterion (AIC) is used to compare models amongst each other. The AIC determines the complexity of the model, and how well the model predicts, however, when used by itself it is not informative. Thus, the AIC of one model must be compared to other similar models in order to determine the best value. When using AIC, the goal is to choose the model with the lowest value - it is important to note that AIC penalises complex models, thus simpler models may have lower AIC values, but may not be as strong in predicting.

When evaluating a model, all three of these are taken into consideration so as to provide a balanced approach to evaluation and ensure that one-metric bias is minimised.

# Section 5: Results

The results of the multiple analyses are presented in this chapter, in the following order. First, the extraction of topics from the reviews of the five different hotel types is examined. The second section focuses on predictive modelling of review score based on these topics whereby the best model for each hotel type will be chosen. Finally, the results of each of the best models will be interpreted so as to find out which topics and cultural dimensions are important to different consumers.

## 5.1: Topic Modelling

To extract topics from each hotel review, we first created a Document Term Matrix (DTM) for each tokenization method across all hotel types. In total, 25 DTMs were created, 5 for each hotel type consisting of unigrams, bigrams, trigrams, 1-skip-1:2-grams, and 1-skip-1:3-grams. Following the creation of the DTMs using different tokenization methods, we tuned all of the LDA models to find the optimal number of topics ($k$), by using perplexity and coherence as performance measures. A training and validation set split was first created before running the LDA models to ensure our model performs well on unseen data, and that overfitting is avoided as much as possible. To determine the optimal number of topics, we want to maximise validation coherence while minimising perplexity. In cases where coherence increased slightly and perplexity kept decreasing with $k$, we chose the lowest $k$ as we want to keep our final models as interpretable as possible. Figure 6 shows an example of perplexity scores while iterating over $k$ and Figure 7 shows how coherence changes with the same iteration. For this particular example, we chose $k = 22$ as there is a spike in coherence and perplexity is low given that number of topics, but looking further ahead, the coherence does not significantly increase. Thus, 22 topics is the lowest number we can reasonably choose according to the metrics (for simplicity sake). Following tuning of *k,* we run the LDA model with the optimal parameter on each DTM and extract key topics. Figures 8 to 12 show example topics extracted from each hotel type for each different tokenization method.

**Figure 6:** Example of perplexity score iterating over *k*

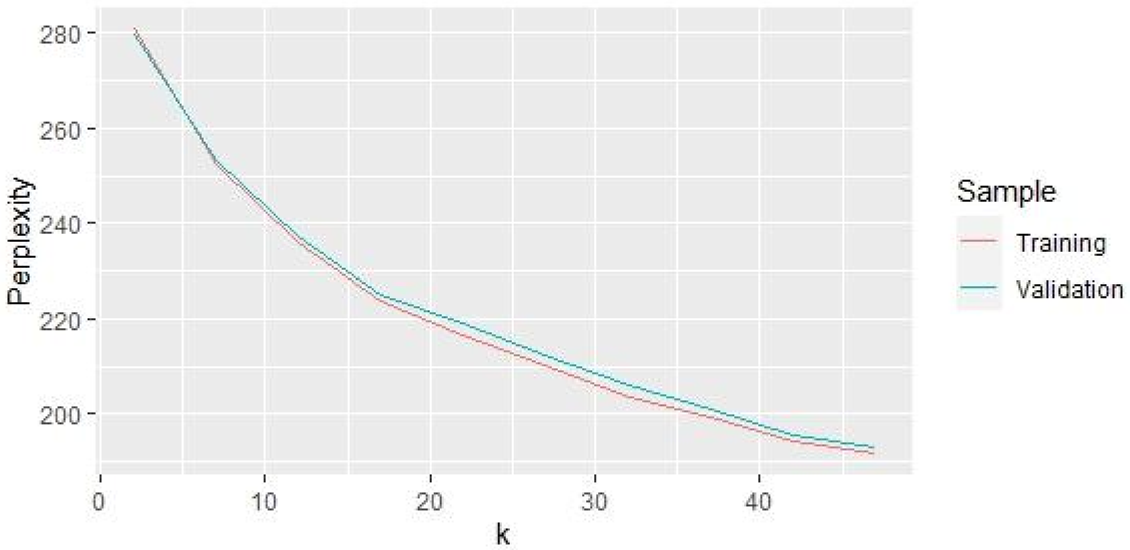
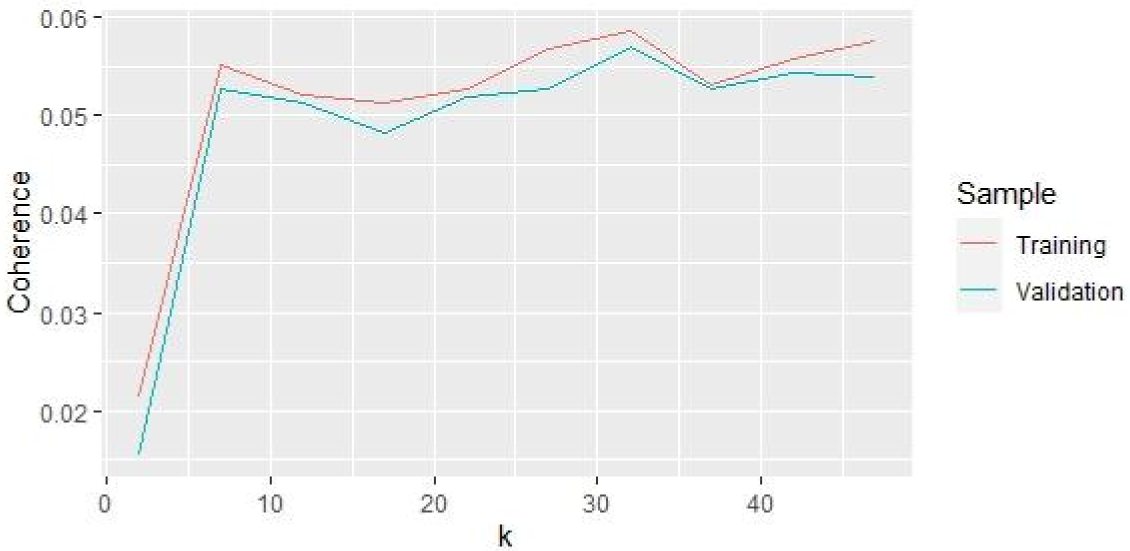
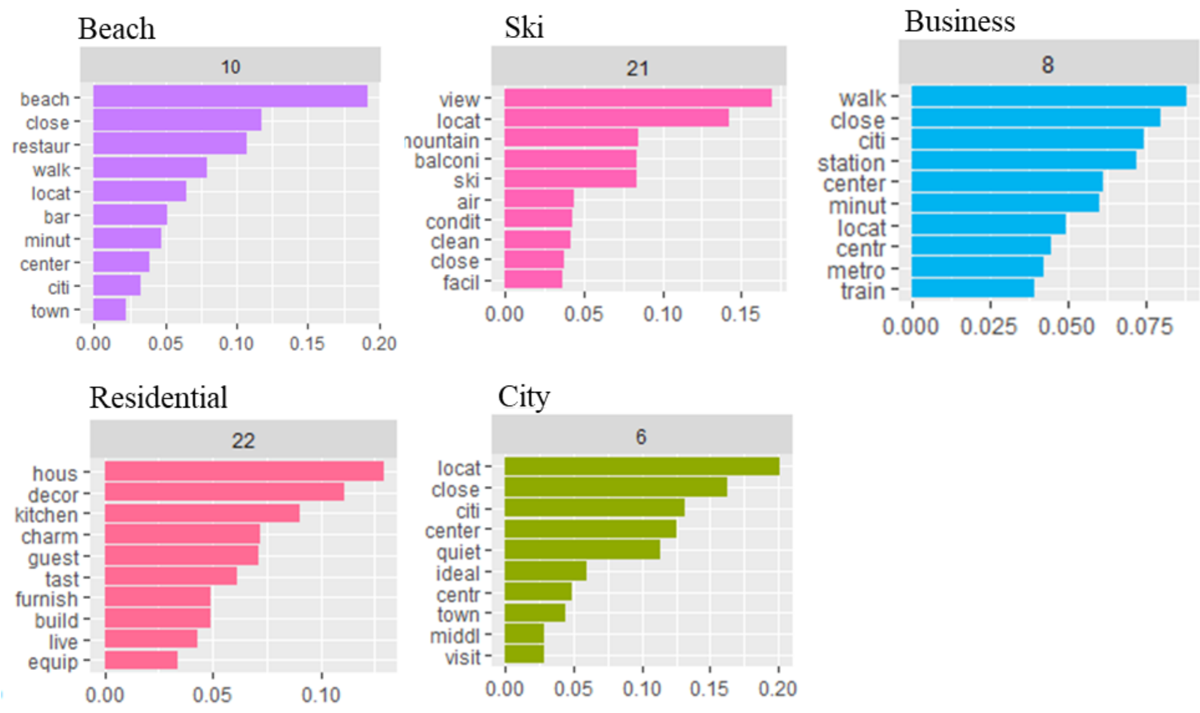



**Figure 7:** Example of coherence score iterating over *k*



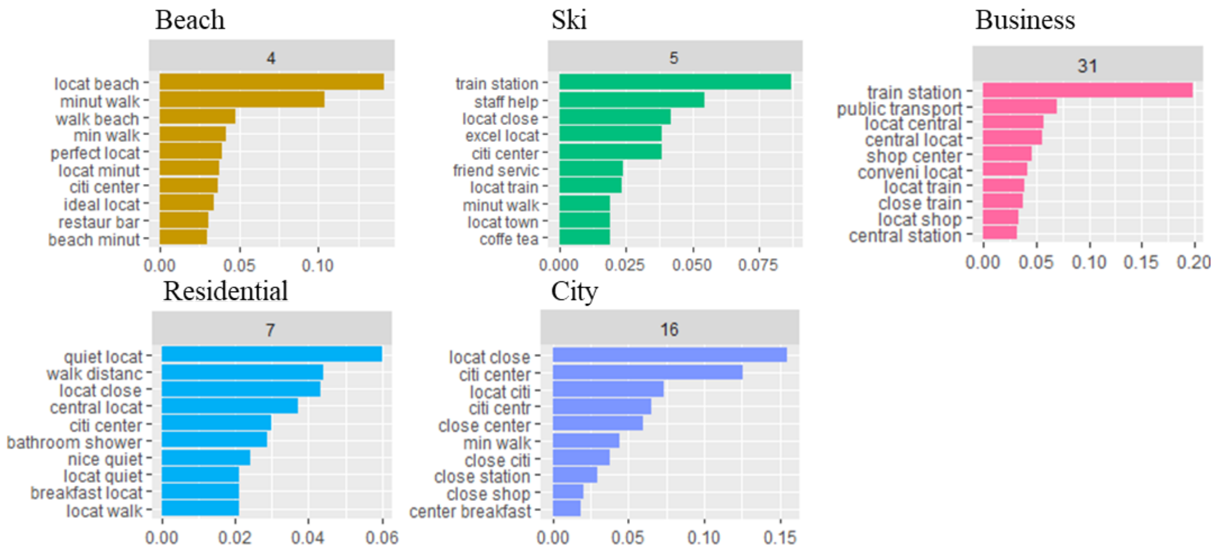

Looking at the extracted topics using unigrams (Figure 8), we can see that topic 10 for Beach hotels contains common words such as 'beach', 'close', and 'walk' - this topic could be interpreted as the proximity of the hotel to the beach, whereby consumers are able to quickly and easily access the beachside. This topic is strictly unique to Beach hotels, and was not extracted for any other hotel types. Similarly, we were able to extract a topic for Ski hotels that includes common words such as 'view', 'locat', 'mountain' and 'ski'. This could be consumers that are talking about ski facilities in the vicinity of the hotel. For business hotels, topic 21 comprises words such as ‘walk’, ‘citi’, ‘station’ and ‘metro’ which can be interpreted as the availability of public transportation near the hotel. This topic may show up for business hotels because people travelling for business purposes may need to use public transport to arrive at conferences or meetings and it is convenient for them to have this near the hotel, whereas leisure travellers may not have a given destination to get to. Residential hotels also show a unique topic with common words like ‘hous’, ‘decor’ and ‘kitchen’. This topic clearly relates to how the owners have decorated the residence, whereby the word ‘hous’ shows up instead of hotel - further, it seems as if these consumers care about the charm and effort that an owner puts into their home. Finally, topic 6 extracted from city hotels consists of ‘locat’, ‘close’, ‘citi’ and ‘center’ which can be interpreted as the proximity of the hotel to the center of the city, an aspect that may be important to travellers looking to stay overnight inside of city limits.

**Figure 8:** Example topics extracted using LDA unigram tokenization



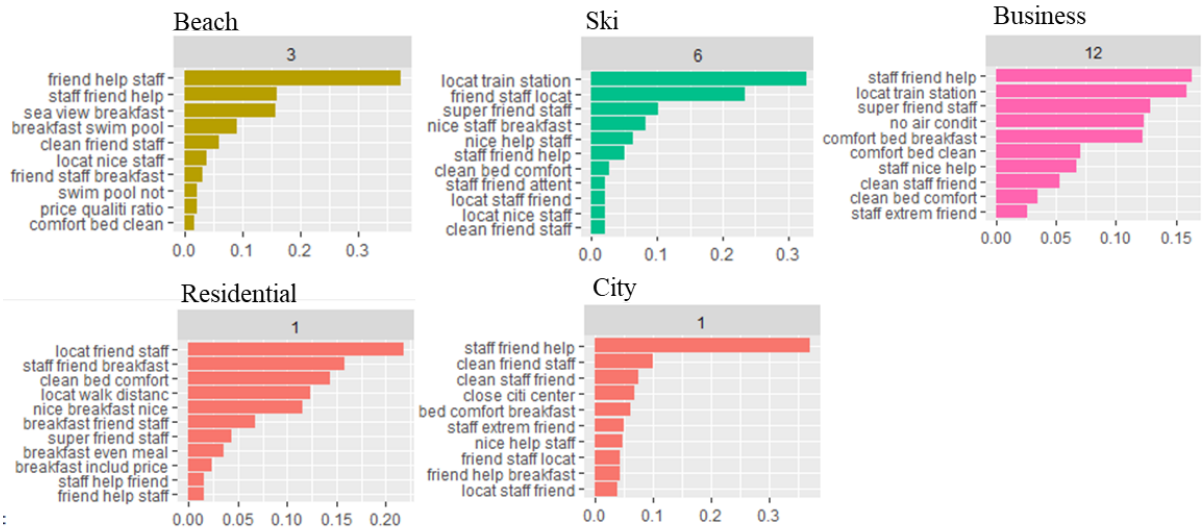

Bigrams were used next to extract topics from each hotel type. Figure 9 summarises example topics that were found. Topic 4 from beach hotels relates to word combinations such as ‘locat beach’ and ‘minut walk’ which could be referring to the proximity of the hotel to the beach. Topic 5 for ski hotels mention ‘train station’ and ‘staff help’, implying that these consumers care about proximity to a train and staff may have been helpful in instructing them how to get to the train from the hotel. This does make sense as skiing resorts tend to be in smaller villages and access to them from metropolitan areas is important. Business hotels have a topic relating to 'train station' and 'public transport', similarly to unigrams, it seems that these consumers care about proximity of their hotel to accessible transport as they tend to be more time conscious. Topic 7 extracted from residential hotels mentions 'quiet locat' which could be attributed to these consumers wanting to get away from the hustle and bustle of the city, which is why they might choose residential accommodation in place of a hotel. Topic 2 of city hotels mentions 'locat close', 'citi center' and 'locat citi', which are location based attributes that consumers may want to see if they are staying in a city; city activities like 'center breakfast' are also mentioned.

**Figure 9:** Example topics extracted using LDA bigram tokenization



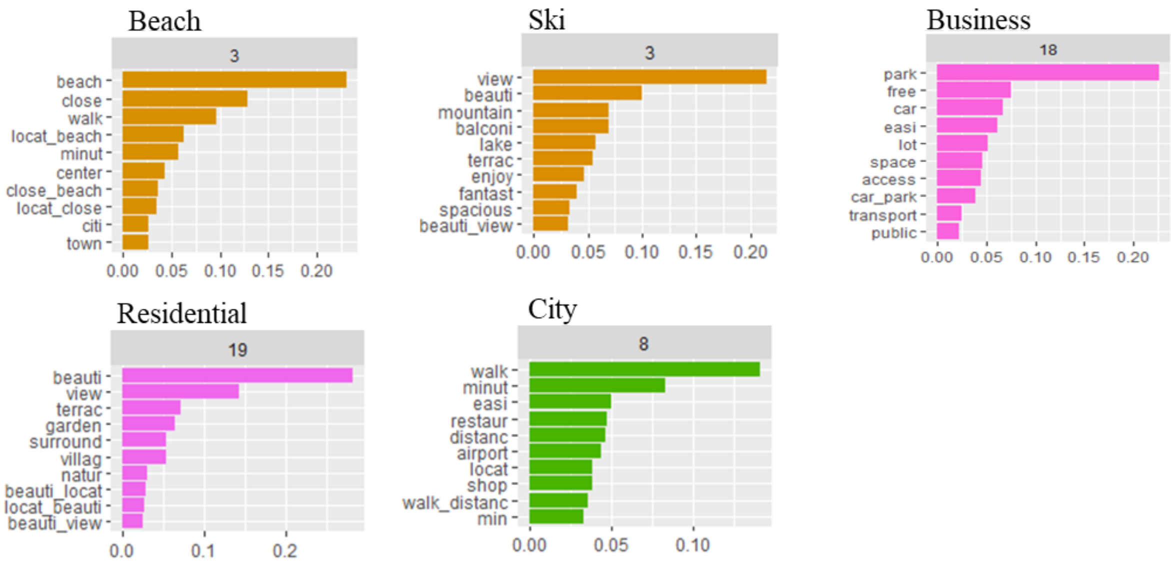

Trigrams are next used to extract topics, albeit with poorer results than both unigrams and bigrams. Figure 10 shows that 'staff friend help' or 'friend help staff' are common trigrams across all hotel types, with very few trigrams being unique. The unique trigrams extracted include 'locat close beach' for beach hotels, 'locat train station' for ski hotels, 'locat minut walk' for business hotels and 'walk citi centr' for city hotels. Trigrams do not seem as informative as the other tokenization methods, and we thus expect low predictive performance from using solely these features in our models.

**Figure 10:** Example topics extracted using LDA trigram tokenization



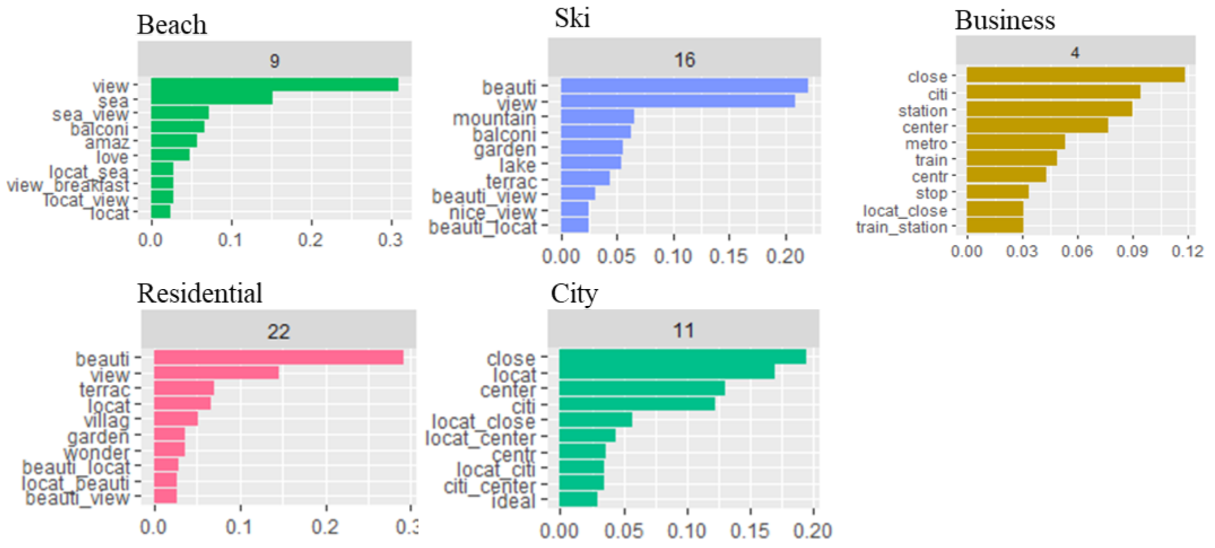

1-skip-1:2-grams were then used as a way to combine both unigrams with bigrams, while also allowing for skipping over words to find common and related word pairs. Figure 11 summarises example topics for each hotel type generated with this tokenization method. We can see that we get similar words and word pairs as we did using unigrams and bigrams, but overall, topics seem to be more specific and coherent with less diversity amongst common words. We see that for example, topic 3 for beach hotels consisted of unigrams such as ‘close’ and ‘walk’ but also bigrams like ‘locat_beach’, thus we know that the consumer is specifically mentioning that the beach is within a walking distance. For ski hotels, we are able to find common words such as ‘lake’ and ‘mountain’ in the same topic as ‘beauti_view’, meaning that a nice view does affect their thoughts about the hotel. Further, business hotel consumers mention a ‘car_park’ and associate it to words like ‘free’. Thus, free parking might be a sought after feature. Extracting sentiments is facilitated with this approach as seen in topic 19 for residential hotels where the bigram ‘locat_beauti’ is found, thus the word ‘beauti’ relates specifically to the location of the house, and this would be a positively classified sentiment. For city hotels a topic relating to words such as ‘walk’ and ‘airport’, ‘shop’ and ‘walk_distanc’ are found referring to the facilities in proximity of the hotel, which might be important for guests.

**Figure 11:** Example topics extracted using LDA 1-skip-1:2-gram tokenization




Finally, the last tokenization method used is 1-skip-1:3-grams whereby we aim to look for unigrams, bigrams, and trigrams while allowing for words to be skipped if a word pair occurs frequently in the corpus. Figure 12 summarises example topics for this tokenization method for each hotel type. Overall it seems that adding trigrams as a possibility has not significantly changed the topics we extract from 1-skip-1:2-grams - with few, if any topics containing a 1-skip-3-gram. The topics we extract are similar to the ones seen in the 1-skip-1:2-gram approach since we have a lack of trigrams. For beach hotels, we see once again a topic relating to ‘sea_view’, while for ski hotels we see that there is a topic related to ‘view’ and ‘mountain’. Business hotels have a topic mentioning ‘locat_close’ and ‘train_station’ similarly to that found in other tokenization approaches. Topic 22 of residential hotels mention ‘garden’ and ‘beauti_locat’, which could mean that guests appreciate the ability to enjoy the outdoors while staying at these accommodations. Finally, city hotels once again mention ‘center’, and ‘locat_citi’ in topic 11.

**Figure 12:** Example topics extracted using LDA 1-skip-1:3-gram tokenization




## 5.2: Predictive Modelling

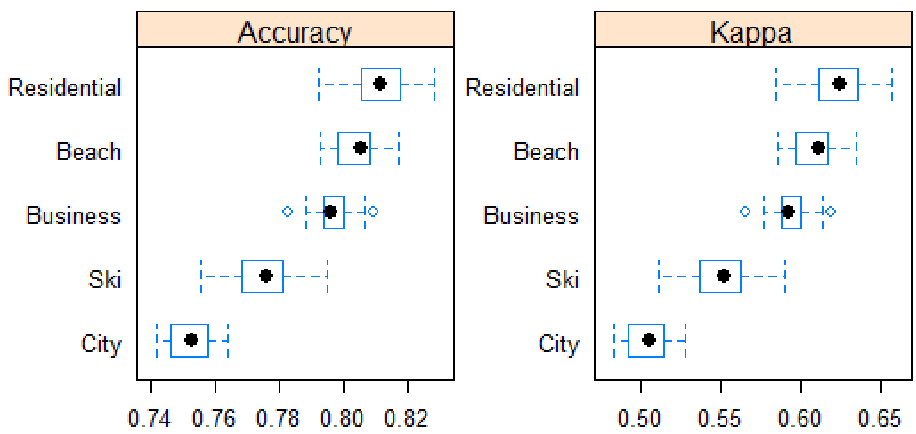
Following topic extraction and distribution matching with a given review, the classification task of reviews based on their rating was done using logistic regression, as a baseline, interpretable classifier. Before logistic regression is done, Hofstede's cultural dimensions were attributed to each review based on the reviewer nationality, so as to address cultural heterogeneity and examine the impact of given cultural dimensions on consumer satisfaction. Further, since our datasets are significantly imbalanced, the training dataset was upsampled on the minority class, so as to teach our models how to predict lower reviews better. Table 2 summarises the results of the logistic regression for each of the 25 feature models, with the accuracy, F1 score and AIC shown. As we can see from the table, the best performing model tends to be the logistic regression using LDA features derived from 1-skip-1:2-gram tokenization, with accuracies ranging from 77.17% - 84.19%. This model also tends to outperform others in terms of F1 score and AIC. We can also see that unigrams performed relatively well, however, using just bigrams and trigrams alone does not yield high predictive power. Adding bigrams and allowing for flexibility of skipping words proved to increase performance in all five hotel types. For further interpretation, we thus choose the 1-skip-1:2-gram logistic regression models based on their higher accuracy, F1 score and low AIC.

**Table 2:** Results of logistic regression by hotel type and tokenization

| Hotel Category | Metric | Unigrams | Bigrams | Trigrams | 2-skip-1-2-gram | 2-skip-1-3-gram |
|---|---|---|---|---|---|---|
| **Beach** | Accuracy | 80.97% | 68.83% | 61.41% | **82.07%** | 81.28% |
| | F1 Score | 88.63% | 80.12% | 75.27% | **89.31%** | 88.09% |
| | AIC | 31,057 | 32,011 | 5,538 | **29,153** | 30,536 |
| **Ski** | Accuracy | 79.96% | 73.13% | 64.53% | **80.27%** | 79.30% |
| | F1 Score | 88.36% | 84.17% | 78.22% | **88.56%** | 87.92% |
| | AIC | 16,850 | 18,888 | 3,025 | **17,459** | 17,814 |
| **Business** | Accuracy | 78.62% | 68.83% | 54.73% | **81.22%** | 80.11% |
| | F1 Score | 87.22% | 80.14% | 69.94% | **88.91%** | 88.71% |
| | AIC | 45,513 | 53,282 | 11,047 | **41,846** | 42,044 |
| **Residential** | Accuracy | 82.48% | 78.17% | 56.78% | **84.19%** | 83.44% |
| | F1 Score | 90.01% | 88.06% | 71.94% | **91.05%** | 90.60% |
| | AIC | 16,900 | 18,546 | 1,576 | **15,313** | 15,790 |
| **City** | Accuracy | 78.23% | 75.96% | 54.91% | **77.17%** | 76.91% |
| | F1 Score | 87.02% | 86.01% | 70.48% | **86.23%** | 85.11% |
| | AIC | 34,515 | 34,250 | 6,846 | **34,975** | 35,127 |

After selecting the use of 2-skip-1-2-grams as our preferred method, 10-fold cross validation was run on the logistic regression model for each hotel type, the results of which are summarised in Figure 13. As can be seen, our models had the highest predictive power for residential, beach and business hotels, while the ski and city hotel models performed slightly worse. However, all models achieved an average accuracy above 75% and an average kappa above 0.5, thus they all perform well on our dataset and can be used for interpretation.

**Figure 13:** Logistic regression performance based on 10-fold cross validation



### 5.3: Interpretation of Logistic Regression Results

Following the training and evaluation of the models, interpretation can take place. The 1-skip-1:2-gram models are selected for interpretation, with Tables 4-8 summarising the coefficients and significance levels obtained from each of these models. All results are ordered by decreasing coefficient magnitude and marked by at least one star if it is significant at the $\alpha = 0.05$ level. Topics are analysed in terms of probability on a scale of 0 to 1 and cultural dimensions are on a scale of 1-100. Economic indicators (GDP per capita PPP and unemployment) are on scales of dollars and percentage (1-100%) respectively. It is also important to note that all interpretation is relative to the reference topic of 'Staff_Help' which is positive in terms of consumer satisfaction - this topic was omitted from the logistic regression models to solve the problem of perfect multicollinearity which occurs with LDA, since all topic probabilities add up to 1. Thus, since the omitted topic is positive, the results need to be carefully interpreted since a negative coefficient does not necessarily mean that the factor is negative with regards to consumer satisfaction, but rather in relation to the omitted topic - which applies solely when interpreting other topics, since this is where the multicollinearity occurred. This would not apply to interpretations of cultural or economic dimensions. As we are removing a multicollinear topic from our regression analysis, the effect of this omitted topic becomes absorbed into the intercept of the model - thus, we can use the intercept and topic context for guidance in determining whether a topic is truly positive or negative.

Table 3 showcases the results for all of the logistic regression models, split by hotel type. For beach hotels, features that contribute to higher consumer satisfaction include being in a beautiful and quiet location and having a clean and spacious room. Further, unique factors that contribute towards positive consumer sentiment include the beach being a close walk away, having access to pool and spa facilities and having a view of the sea (albeit the latter two factors are not statistically significant in our model). Cultural aspects of guests staying at beach hotels that contribute towards a more positive review include individuality, uncertainty avoidance, masculinity and power distance. It could be argued that individuals from cultures meeting these factors may prefer staying at a beach hotel, since they would like to better themselves by relaxing (contributing to individuality), but also they might not like the uncertainty that comes with a busier, more disorganised trip. Further, long term oriented individuals tend to rate lower all else equal, as they prefer to be thrifty and associate better with driving success, than short-term oriented individuals. Factors that drive consumer satisfaction downwards include overall negatives (with the strongest effect) - these range from being associated to not having breakfast available to a staff member not being accommodating. Dirty bathrooms are the next largest cause of concern, with arrival departure issues and nighttime noise also having a significant negative impact on consumer satisfaction.

The analysis for ski hotels shows that the factor contributing most towards positive consumer sentiments is everything being ‘excellent’. Since we eliminated helpful staff from our model to combat multicollinearity, this value is now taken into account in the intercept. Thus, having such a large positive intercept, looking at factors such as loving the decor, being close to ski facilities and having mountain views also have a positive impact on consumer satisfaction, albeit potentially less than helpful staff would.  Culturally, societies with more masculinity, higher power distance and individualism tend to rate ski hotels higher. This does make sense as ski hotels are typically associated with outdoor, rewarding activities that can be done on one’s own, without need for interaction with others. Negative aspects of these hotels include bathroom facility issues, nighttime noise, and arrival and departure issues. Further, price quality issues are common amongst these hotels, possibly because many ski hotels are located in countries such as Switzerland which tend to be more expensive than the rest of Europe, and thus higher quality is expected for the paid price.

Factors that influence satisfaction for business hotels positively include a perfect location and being close to a public transport station. Further, having access to a bar or pool and being close to the centre is important for these guests. Culturally, societies with a lower indulgence versus restraint score tend to rate business hotels lower, this could make sense as these cultures may prefer to indulge in fun activities like those found at ski and beach hotels, and not be in a corporate and socially constricted environment. Further, an interesting result is that cultures that are more masculine tend to rate business hotels lower as well, which is opposite to what we would expect as masculine cultures tend to favour competitiveness and hard work, but they also prefer achievements and rewards - which could be the reason they may not enjoy being away from their personal lives for business. Individualistic cultures also tend to rate these hotels lower, possibly due to the fact that business employees tend to collaborate and work in a collective fashion. Factors that negatively affect satisfaction include lack of gratuities such as free coffee or tea, and non functioning bathroom facilities can be frustrating for these individuals. An interesting result is that these individuals highly disfavour when no coffee, tea or other gratuities are given out, which may all be expected of a business hotel. Economically, reviewers from countries with higher GDP per capita tend to rate more favourably as well as those from countries with higher unemployment - which is contradictory to what we would expect, since higher unemployment typically means a less healthy economy.

The analysis of residential hotels shows that consumers mentioning factors relating to having a perfect host, as well as great attention to detail tend to rate the accommodation higher. These are both unique nuanced topics that are important solely for residential hotels - the word host is used as opposed to the generic ‘staff’, as users staying at residential hotels generally do have to interact with the owner of the residence. Since these are residential accommodations, all finishing touches and decoration are done by the host, thus these users care when the owner puts in effort to paying attention to the details that make their trips lovely. Further, having a terrace, a view or a pool increases consumer satisfaction (but is not statistically significant). Culturally, masculine, long-term oriented and indulgent cultures tend to rate these hotels lower, which could be due to the fact that guests from indulgent cultures may not prefer to perform tasks themselves such as cooking, cleaning and so on; indulgent cultures may like having these tasks done for them, especially while vacationing. Negative factors of satisfaction include mentioning any negatives, as well as nighttime noise, arrival and departure issues, bathroom issues and the comfort of the bed. Economically, unemployment has a significant effect on the ratings of these reviewers, with those coming from countries with high unemployment leaning towards lower ratings.

Finally, city hotels show that factors contributing to higher satisfaction are mentions of everything nice, as well as having a lovely view. Further, having a comfortable bed and being in a great location is important to these consumers. A quieter city location tends to increase satisfaction (but not as much as helpful staff), but so does being a short walk away from the city center. Nearby public transport is important to guests at city hotels, and having a free parking lot can also help in improving satisfaction. Culturally, societies with lower values of indulgence versus restraint tend to prefer city hotels, which does make intuitive sense since these cultures may follow stricter social norms and prefer the business-like aspect of a city. Negative factors of consumer satisfaction include having a dirty bathroom, arrival and departure issues, night time noise, and pool or spa facilities that you need to pay for. City hotels may be more expensive than hotels further outside of a city, thus better quality could be expected from guests if they are paying a higher price. Economic indicators such as GDP per capita and unemployment seem to have an effect on reviewer rating of these hotels, with higher GDP per capita increasing the likelihood of a reviewer rating positively, as well as higher unemployment which is counter intuitive. It could be that countries with higher unemployment may be in the process of urbanisation and as such the guests enjoy the modernity of cities.

**Table 3:** Summary of coefficients of logistic regression by hotel type

**Beach**

| | *Dependent variable:* |
|---|---|
| | review_bi |
| (Intercept) | 4.38$^{***}$ |
| Beauti_Quiet_Locat | 19.65$^{***}$ |
| Excel_Perfect_Locat | 19.21$^{***}$ |
| Clean_Spacious | 14.59$^{***}$ |
| Friend_Staff | 10.02$^{***}$ |
| Breakfast | 4.54$^{***}$ |
| Everything_Nice | 4.05$^{***}$ |
| Beach_Close_Walk | 3.62$^{***}$ |
| Pool_Spa_Facilities | 0.55 |
| uai | 0.02$^{***}$ |
| idv | 0.01$^{***}$ |
| mas | 4.90e-3$^{***}$ |
| ivr | 1.16e-4 |
| GDP_cap | 5.86e-6$^{***}$ |
| pdi | −0.01$^{**}$ |
| unemployment | −0.02$^{***}$ |
| ltowvs | −0.03$^{***}$ |
| View_Sea | −1.04 |
| Breakfast_Buffet | −1.65 |
| Locat_Super | −3.14$^{**}$ |
| Bed_Comfort | −4.97$^{***}$ |
| Restaurant_Bar | −5.12$^{***}$ |
| Park_Lot_Free | −5.72$^{***}$ |
| Servic_Equip | −6.80$^{***}$ |
| Arrival_Departure_Issues | −29.38$^{***}$ |
| Night_Nois | −29.76$^{***}$ |
| Dirty_Bathroom | −30.89$^{***}$ |
| Negatives | −37.93$^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Ski**

| | *Dependent variable:* |
|---|---|
| | review_bi |
| (Intercept) | 17.94$^{***}$ |
| Excellent | 4.96$^{**}$ |
| mas | 0.01$^{***}$ |
| pdi | 0.01$^{***}$ |
| idv | 0.01$^{***}$ |
| ivr | 2.77e-3 |
| GDP_cap | −1.42e-6 |
| uai | −1.78e-4 |
| unemployment | −0.004 |
| ltowvs | −0.01$^{***}$ |
| Love_Decor | −0.71 |
| Locat_Ski | −5.93$^{***}$ |
| Everything_Nice | −6.07$^{***}$ |
| View_Mountain | −6.33$^{***}$ |
| Breakfast_Delicious | −8.32$^{***}$ |
| Clean_Spacious | −9.61$^{***}$ |
| Recommend | −12.99$^{***}$ |
| Locat_Close_Station | −13.98$^{***}$ |
| Meals | −15.76$^{***}$ |
| Locat_Walk_Central | −16.03$^{***}$ |
| Restaurant_Food | −16.23$^{***}$ |
| Staff_Friend | −17.87$^{***}$ |
| Breakfast | −19.58$^{***}$ |
| Bed_Comfort | −20.10$^{***}$ |
| Park_Free | −20.97$^{***}$ |
| Servic_Buffet | −25.65$^{***}$ |
| Price_Quality | −30.74$^{***}$ |
| Arrival_Departure_Issues | −48.83$^{***}$ |
| Night_Nois | −54.94$^{***}$ |
| Bathroom_Facilities | −55.25$^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Business**

| | *Dependent variable:* |
|---|---|
| | review_bi |
| (Intercept) | 13.64$^{***}$ |
| Excellent | 6.42$^{***}$ |
| Locat_Perfect | 2.68$^{**}$ |
| Walk_Station_Minut | 2.45$^{**}$ |
| Everything_Nice | 0.45 |
| Comfort_Bed | 0.05 |
| unemployment | 0.02$^{***}$ |
| uai | 4.83e-3$^{***}$ |
| ltowvs | 1.32e-3 |
| GDP_cap | 8.77e-6$^{***}$ |
| pdi | −3.16e-3 |
| idv | −4.98e-3$^{***}$ |
| ivr | −0.01$^{***}$ |
| mas | −0.01$^{***}$ |
| Bar_Beauti_Pool | −2.33$^{**}$ |
| Clean_Spacious | −2.67$^{**}$ |
| Locat_Close_Center | −3.92$^{***}$ |
| Staff_Friend | −4.17$^{***}$ |
| Breakfast_Price_Quality | −8.50$^{***}$ |
| Staff_Welcoming | −12.99$^{***}$ |
| Park_Free | −14.09$^{***}$ |
| Restaurant_Food | −15.17$^{***}$ |
| Breakfast_Servic | −18.17$^{***}$ |
| Bed_Love_Comfy | −18.38$^{***}$ |
| Arrival_Departure_Issues | −28.66$^{***}$ |
| Bathroom_Facilities | −37.75$^{***}$ |
| Negatives | −39.29$^{***}$ |
| No_Gratuities | −48.98$^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**Residential**

| | *Dependent variable:* |
|---|---|
| | review_bi |
| (Intercept) | 13.86$^{***}$ |
| Perfect_Host | 14.54$^{***}$ |
| Love_Attention_Detail | 8.80$^{***}$ |
| Everything_Nice | 4.31$^{**}$ |
| Excellent | 4.29$^{**}$ |
| Beauti_View_Terrace | 1.05 |
| uai | 0.02$^{***}$ |
| idv | 0.01$^{*}$ |
| GDP_cap | −6.93e-7 |
| mas | −4.37e-3$^{***}$ |
| pdi | −0.01 |
| ltowvs | −0.01$^{**}$ |
| ivr | −0.02$^{***}$ |
| unemployment | −0.05$^{***}$ |
| Breakfast_Delicious | −0.78 |
| Close_Walk | −1.17 |
| Clean_Spacious | −3.70$^{*}$ |
| Pool_Kitchen_Facilities | −4.76$^{**}$ |
| Staff_Friend | −6.84$^{***}$ |
| Breakfast_Fresh | −8.23$^{***}$ |
| Owner_Hous_Decor | −8.70$^{***}$ |
| Locat_Quiet | −12.55$^{***}$ |
| Breakfast | −13.45$^{***}$ |
| Park_Lot_Easy | −13.99$^{***}$ |
| Restaurant_Price | −14.01$^{***}$ |
| Bed_Comfort | −20.74$^{***}$ |
| Bathroom_Facilities | −39.61$^{***}$ |
| Arrival_Departure_Issues | −40.59$^{***}$ |
| Night_Nois | −47.06$^{***}$ |
| Negatives | −47.51$^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

**City**

| | *Dependent variable:* |
|---|---|
| | review_bi |
| (Intercept) | 17.21$^{***}$ |
| Everything_Nice | 3.50$^{**}$ |
| Beauti_View_Terrac | 2.76$^{*}$ |
| unemployment | 0.02$^{***}$ |
| uai | 2.45e-3 |
| pdi | 1.39e-3 |
| mas | 4.35e-4 |
| GDP_cap | 1.07e-5$^{***}$ |
| ltowvs | −7.19e-4 |
| idv | −3.51e-3$^{**}$ |
| ivr | −0.01$^{***}$ |
| Comfort_Bed | −3.04$^{**}$ |
| Breakfast_Fine | −3.23$^{**}$ |
| Excellent | −3.95$^{***}$ |
| Staff_Pleasant | −4.33$^{***}$ |
| Locat_Super | −5.27$^{***}$ |
| Perfect_Recommend | −6.40$^{***}$ |
| Walk_Station_Minut | −7.11$^{***}$ |
| Quiet_Close_Center | −8.66$^{***}$ |
| Friend_Staff | −8.88$^{***}$ |
| Locat_Easi_Access | −13.62$^{***}$ |
| Park_Free | −17.17$^{***}$ |
| Clean | −17.91$^{***}$ |
| Restaurant_Bar | −23.29$^{***}$ |
| Pool_Spa_Facilities_Price | −26.81$^{***}$ |
| Bathroom_Facilities | −28.48$^{***}$ |
| Breakfast_Fresh | −28.57$^{***}$ |
| Breakfast_Service | −28.63$^{***}$ |
| Bed_Bad | −38.54$^{***}$ |
| Dirty_Bathroom | −40.43$^{***}$ |
| Night_Nois | −44.24$^{***}$ |
| Arrival_Departure_Issues | −50.67$^{***}$ |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

## 5.4: Logistic Regression Model & LASSO with Interactions

Logistic regression models with interactions between indicators and extracted topics were created for each hotel type. As these generated a large number of predictors, LASSO regularisation was used to perform variable selection so as to improve interpretability of the models. The performance of these models are summarised by hotel type in Table 4. For the sake of interpretability, only the coefficients greater than 1 or less than -1 are reported, as these are the factors that have the largest impact on consumer satisfaction. Table 5 summarises the coefficients of all the logit interaction models split by hotel type.

As seen from the results, most of the interactions between cultural dimensions and extracted topics do not have a large impact on consumer satisfaction, which is indeed surprising since we have determined that there is indeed a correlation between cultural dimensions and satisfaction. However, it can be noted that interactions between unemployment and certain factors of satisfaction tend to have a higher impact on overall rating - albeit, these interactions are still quite difficult to interpret. For example, in the beach hotel model we see that the interaction between breakfast and unemployment has a negative impact on satisfaction. This could be potentially explained if the hotels charged extra money for breakfast, and thus individuals from countries with higher unemployment may not want to pay extra for this meal. However, we do not know this about all hotels analysed, thus it is quite difficult to draw that conclusion. We do see some interactions with a larger coefficient in the residential hotel model, but interpretability is questionable. However, these models do show that factors specific to a given hotel type such as the beach being a close walk away, having a sea view room, having a mountain view room, being located close to a station, having pool and kitchen facilities and free parking all have strong impacts on consumer satisfaction. Thus, it further strengthens the argument that factors of satisfaction do differ across hotel types, and consumers should be specifically targeted based on the hotel at which they are staying.

**Table 4:** Summary of logit with interactions and LASSO performance

| Hotel Category | Metric | Logit with Interactions and Lasso |
|---|---|---|
| Beach | Accuracy | 83.10% |
| | F1 Score | 89.46% |
| Ski | Accuracy | 81.76% |
| | F1 Score | 88.76% |
| Business | Accuracy | 81.04% |
| | F1 Score | 89.78% |
| Residential | Accuracy | 84.89% |
| | F1 Score | 91.85% |
| City | Accuracy | 77.72% |
| | F1 Score | 87.04% |

**Table 5:** Summary of coefficients of logistic regression with interactions by hotel type

Beach

| | Predictor | Coefficient |
|---|---|---|
| 6 | Beach_Close_Walk | 60.12 |
| 21 | View_Sea | 49.77 |
| 10 | Breakfast | 34.32 |
| 18 | Restaurant_Bar | 31.69 |
| 8 | Bed_Comfort | 15.42 |
| 19 | Park_Lot_Free | 13.74 |
| 14 | Everything_Nice | 5.06 |
| 93 | Beauti_Quiet_Locat:unemployment | 1.05 |
| | ... - COEFFICIENTS LEFT OUT - ... | |
| 69 | Beach_Close_Walk:unemployment | -1.50 |
| 15 | Pool_Spa_Facilities | -2.02 |
| 101 | Breakfast:unemployment | -2.55 |
| 16 | Night_Nois | -2.96 |
| 5 | Excel_Perfect_Locat | -9.63 |
| 3 | Servic_Equip | -10.68 |
| 20 | Clean_Spacious | -20.71 |
| 17 | Locat_Super | -22.41 |
| 7 | Negatives | -39.82 |
| 4 | Breakfast_Buffet | -43.23 |
| 13 | Dirty_Bathroom | -46.94 |

Ski

| | Predictor | Coefficient |
|---|---|---|
| 17 | Breakfast | 95.37 |
| 20 | Locat_Close_Station | 94.45 |
| 12 | Bed_Comfort | 78.89 |
| 23 | Restaurant_Food | 78.19 |
| 9 | Breakfast_Delicious | 77.67 |
| 7 | View_Mountain | 35.49 |
| 2 | Staff_Friend | 5.22 |
| 19 | Recommend | 5.13 |
| 143 | Staff_Help:unemployment | 2.56 |
| 39 | Staff_Friend:unemployment | 1.65 |
| 79 | View_Mountain:unemployment | 1.41 |
| 127 | Clean_Spacious:unemployment | 1.38 |
| 36 | Staff_Friend:ltowvs | 1.22 |
| 103 | Locat_Walk_Central:unemployment | 1.15 |
| | ... - COEFFICIENTS LEFT OUT - ... | |
| 156 | Breakfast:ltowvs | -1.02 |
| 71 | Love_Decor:unemployment | -1.05 |
| 119 | Bed_Comfort:unemployment | -1.50 |
| 191 | Park_Free:unemployment | -1.75 |
| 159 | Breakfast:unemployment | -1.76 |
| 183 | Locat_Close_Station:unemployment | -1.86 |
| 4 | Arrival_Departure_Issues | -8.24 |
| 8 | Servic_Buffet | -8.62 |
| 16 | Night_Nois | -13.46 |
| 18 | Everything_Nice | -26.47 |
| 21 | Park_Free | -27.26 |
| 13 | Clean_Spacious | -32.58 |
| 5 | Excellent | -40.36 |
| 14 | Price_Quality | -63.56 |
| 11 | Bathroom_Facilities | -98.34 |

Business

| | Predictor | Coefficient |
|---|---|---|
| 12 | Bar_Beauti_Pool | 83.24 |
| 15 | Bed_Love_Comfy | 46.74 |
| 5 | Staff_Help | 36.27 |
| 11 | Walk_Station_Minut | 30.55 |
| 8 | Arrival_Departure_Issues | 25.39 |
| 14 | Breakfast_Price_Quality | 17.54 |
| 9 | Everything_Nice | 9.14 |
| 4 | Comfort_Bed | 7.37 |
| 19 | Negatives | 2.18 |
| 45 | Clean_Spacious:unemployment | 1.99 |
| 133 | Breakfast_Price_Quality:unemployment | 1.78 |
| 189 | Staff_Welcoming:unemployment | 1.51 |
| | ... - COEFFICIENTS LEFT OUT - ... | |
| 181 | Restaurant_Food:unemployment | -1.12 |
| 93 | Everything_Nice:unemployment | -1.15 |
| 141 | Bed_Love_Comfy:unemployment | -1.40 |
| 117 | Bar_Beauti_Pool:unemployment | -1.64 |
| 173 | Negatives:unemployment | -1.87 |
| 20 | Restaurant_Food | -5.51 |
| 13 | Locat_Perfect | -8.65 |
| 2 | Bathroom_Facilities | -11.16 |
| 17 | Excellent | -25.42 |
| 3 | Clean_Spacious | -33.56 |
| 21 | Staff_Welcoming | -36.56 |
| 16 | Staff_Friend | -37.50 |
| 7 | Breakfast_Servic | -51.45 |
| 10 | No_Gratuities | -57.94 |

Residential

| | Predictor | Coefficient |
|---|---|---|
| 17 | Pool_Kitchen_Facilities | 108.14 |
| 13 | Breakfast_Delicious | 105.49 |
| 6 | Park_Lot_Easy | 94.02 |
| 4 | Everything_Nice | 87.83 |
| 19 | Excellent | 59.01 |
| 10 | Close_Walk | 45.19 |
| 23 | Love_Attention_Detail | 37.44 |
| 159 | Pool_Kitchen_Facilities:unemployment | 2.20 |
| 144 | Clean_Spacious:pdi | 2.05 |
| 151 | Clean_Spacious:unemployment | 1.13 |
| 167 | Beauti_View_Terrace:unemployment | 1.08 |
| 129 | Arrival_Departure_Issues:idv | 1.03 |
| 192 | Staff_Friend:pdi | 1.00 |
| | ... - COEFFICIENTS LEFT OUT - ... | |
| 152 | Pool_Kitchen_Facilities:pdi | -1.00 |
| 204 | Love_Attention_Detail:ltowvs | -1.06 |
| 103 | Close_Walk:unemployment | -1.19 |
| 147 | Clean_Spacious:uai | -1.25 |
| 120 | Breakfast_Delicious:pdi | -1.54 |
| 55 | Everything_Nice:unemployment | -2.02 |
| 199 | Staff_Friend:unemployment | -3.10 |
| 71 | Park_Lot_Easy:unemployment | -3.19 |
| 18 | Beauti_View_Terrace | -8.95 |
| 11 | Owner_Hous_Decor | -23.54 |
| 14 | Arrival_Departure_Issues | -51.68 |
| 8 | Negatives | -53.43 |
| 16 | Clean_Spacious | -65.95 |
| 20 | Breakfast_Fresh | -74.87 |

City

| | Predictor | Coefficient |
|---|---|---|
| 12 | Quiet_Close_Center | 55.21 |
| 11 | Park_Free | 49.79 |
| 15 | Staff_Help | 46.71 |
| 20 | Comfort_Bed | 31.30 |
| 23 | Locat_Super | 24.95 |
| 22 | Locat_Easi_Access | 22.97 |
| 16 | Excellent | 18.00 |
| 25 | Friend_Staff | 16.16 |
| 5 | Walk_Station_Minut | 15.50 |
| 18 | Perfect_Recommend | 4.78 |
| 73 | Beauti_View_Terrac:unemployment | 2.39 |
| 161 | Breakfast_Service:unemployment | 1.38 |
| 137 | Staff_Pleasant:unemployment | 1.26 |
| 193 | Breakfast_Fresh:unemployment | 1.23 |
| 201 | Locat_Easi_Access:unemployment | 1.09 |
| | ... - COEFFICIENTS LEFT OUT - ... | |
| 145 | Staff_Help:unemployment | -1.07 |
| 153 | Excellent:unemployment | -1.09 |
| 97 | Everything_Nice:unemployment | -1.63 |
| 13 | Bed_Bad | -1.92 |
| 129 | Bed_Bad:unemployment | -2.24 |
| 113 | Park_Free:unemployment | -2.81 |
| 8 | Breakfast_Fine | -10.09 |
| 10 | Dirty_Bathroom | -16.01 |
| 14 | Staff_Pleasant | -18.37 |
| 9 | Everything_Nice | -19.52 |
| 19 | Restaurant_Bar | -22.53 |
| 6 | Beauti_View_Terrac | -29.51 |
| 21 | Breakfast_Fresh | -29.66 |
| 4 | Arrival_Departure_Issues | -37.33 |
| 17 | Breakfast_Service | -42.00 |
| 2 | Bathroom_Facilities | -52.58 |
| 7 | Pool_Spa_Facilities_Price | -58.92 |

## 5.5: Logistic Regression Trees

Following the analysis of reviews on customer satisfaction using two relatively simple models, we then analysed further some potential interactions between cultural dimensions and topics found in reviews while also accounting for economic indicators to get a better picture of how different consumers tend to rate their stays. The reasoning behind this is that the simple logistic regression model with interactions did not give a clear view into how consumer preferences differ based on cultural and economic dimensions - the interaction effects were difficult to interpret and many were not present as high influences to satisfaction. To expand on the previous analyses, we used a logistic regression tree approach and selected the splitting variables to be amongst the cultural and economic indicators.

For beach hotels, we see that the main split occurs on the uncertainty avoidance index, with reviewers from countries with low values giving more importance to the location of the hotel, the availability of a bar and restaurant and having a view of the sea. Negative aspects that these reviewers identify include dirtiness of the bathroom and nighttime noise. Individuals from countries with high uncertainty avoidance index values (>93.5) prefer features such as having a clean and spacious room and gather more satisfaction from the quietness of the location and the presence of pool and spa facilities, which could come as a nice relaxing surprise for these risk averse individuals. Negative aspects include arrival and departure issues and dirty bathrooms - these are highly negative and intuitively it does make sense as these individuals tend to avoid uncertainty. Thus, a trip being ruined by arrival and departure issues might affect them greatly. However, these individuals also like having activities where they can spend more money such as restaurants and bars. Individuals from countries with lower values of the uncertainty avoidance index (<72.5) and low values of masculinity (<16.5) tend to have higher preference for activities such as the pool and spa, and also enjoy a beautiful and quiet location as well as the proximity to the beach and breakfast service. Further, nighttime noise is their biggest disruptor. Individuals from countries with higher masculinity tend to be most affected positively by having a clean and spacious room, as well as a perfect location. Further, any mention of negatives in their review heavily lowers their satisfaction, when compared to individuals from less masculine countries. It seems that reviewers from less masculine countries prefer different activities involved in relaxation, indulging in food and nature, compared to those from more masculine countries, that seem to be heavily influenced by the tangible aspects of the hotel itself. Appendix 1 exhibits the logistic regression tree for beach hotels.

The model on ski hotels shows a main split on unemployment, with reviewers from countries with a higher value (>4.82%) being further split by uncertainty avoidance index. Individuals with a higher uncertainty avoidance index (>56) in general tend to have less of a tolerance for arrival and departure issues, and are more sensitive to the service they receive from the buffet - for example, if they were expecting to have buffet service included and it was not the case, this may significantly impact their trip. Those from countries with a low unemployment (<4.82%) are split again based on masculinity, and those with low masculinity (<27.5) tend to be more satisfied by the decor, and friendly staff. For individuals with higher masculinity, we see that the view of the mountain is more important, as well as having a clean and spacious room and the location close to a station. Also, it seems that the location being close to a station is also a large factor. These individuals are also more sensitive to night time noise. Appendix 2 exhibits the logistic regression tree for ski hotels.

For business hotels, we see a main split on power distance index, with interesting results for countries with a low power distance index (<36.5) and low masculinity (<67). Those with a low value of masculinity tend to give more preference to having access to a bar or pool and having a friendly staff, while those with a higher value tend to care more about having a good price quality for breakfast, and ensuring that they have facilities like free parking and a clean and spacious room if they expected it. We also see that those that are more averse to uncertainty (>89) and a higher power distance index tend to give more importance to practical, usual aspects like having a comfortable bed, nice bathroom facilities and breakfast price quality. Less uncertainty averse individuals tend to care more about the extras that they get from a hotel, such as staff attitude, hotel location and extra amenities. We also see that individuals with lower power distance index tend to prefer interactions with friendly staff, while those with higher indices are not as interested in this. Appendix 3 exhibits the logistic regression tree for business hotels.

Residential hotels have a similar main split on power distance index, and further splits on masculinity. Reviewers from countries with a higher power distance index (>36.5) and a lower masculinity (<42) tend to give more importance to factors like house decor. Appendix 4 exhibits the logistic regression tree for residential hotels.

Finally, city hotels have a main split on power distance index, with further splits on uncertainty avoidance, masculinity and individualism. Individuals from countries with higher power distance indices (>52) and lower uncertainty avoidance (<87) strongly favour when they have a beautiful view or terrace, while those with higher uncertainty avoidance tend to be more affected by factors such as having to pay extra for a restaurant meal. Further, having an easily accessible location is important to them, possibly because they may not have access to a car, also shown by the favouring of the station being a minute walk away. Individuals with lower power distance indices, low masculinity (<57) and low individualism (<52) care more about the location of the hotel, as well as having a nice view and cheap access to spa and pool facilities. Those that are more individualistic prefer indulging in food such as breakfast, and are sensitive to the quality of their meal. Finally, individuals with high masculinity (>57) prefer having nice views and being close to the centre, and are also less affected by pool and spa facilities being pricey. Appendix 5 exhibits the logistic regression tree for city hotels.

### 5.6: Comparison of Model Performances

Hyperparameter tuning of the logistic regression tree model showed that accuracy tended to increase with the increase in the minsize parameter, while maxdepth did not have a significant impact. The accuracy of the models were on average slightly lower than those obtained by the simple and interaction-based logistic regression, which could be due to the fact that sample sizes may have been slightly constrained as the trees underwent multiple splits. However, the accuracies obtained from the logistic regression trees model are all within 2% of those obtained from the full logit model. What the logistic regression tree model adds is a new level of interpretability, due the partitioning of the larger dataset into smaller nodes based on determined important variables. The focus of this thesis was not accuracy, but rather interpretability and each model added a new level of understanding to the question at hand. The simple logistic model exhibited significant coefficients for differing factors of satisfaction amongst different hotel types, while also establishing a correlation between consumer satisfaction and cultural and economic variables. The logistic regression model with interactions allowed us to trim our models and further confirmed that hotel type specific aspects are truly important, while the interactions between factors of satisfaction and indicators were difficult to analyse. Finally, the logit tree model allowed for the exploration of differences amongst cohorts split by cultural and economic indicators, confirming some existence of correlation between indicators and differences in factors of satisfaction. Table 8 below exhibits the accuracy and F1 score of the three models.

**Table 6:** Comparison of logit and logit tree model performance

| Hotel Category | Metric | Logit | Logit with Interactions | Logit Tree |
|---|---|---|---|---|
| Beach | Accuracy | 82.62% | 83.10% | 82.06% |
| | F1 Score | 89.72% | 89.46% | 89.30% |
| Ski | Accuracy | 78.24% | 81.76% | 80.28% |
| | F1 Score | 87.12% | 88.76% | 88.57% |
| Business | Accuracy | 81.52% | 81.04% | 80.78% |
| | F1 Score | 89.09% | 89.78% | 88.61% |
| Residential | Accuracy | 83.72% | 84.89% | 83.79% |
| | F1 Score | 90.76% | 91.85% | 90.82% |
| City | Accuracy | 77.43% | 77.72% | 76.15% |
| | F1 Score | 86.42% | 87.04% | 85.60% |

# Section 6: Conclusion

Gathering insights on consumer satisfaction in the internet age is crucial for a business to stay relevant and continually improve. With consumers having access to thousands of hotel choices at their fingertips, hoteliers must stay savvy on consumer preferences and ensure that their hotel is attractive in the competitive market. While previously opinions and consumer reviews were mainly transmitted by word-of-mouth amongst social circles, individuals can now publicly share their reviews with millions of others at the click of a button. Thus, hoteliers also have access to information that was once harder to find, and it is important to make use of these learnings to improve the brand. The advent of novel text analytical and machine learning methods makes it possible for managers to quickly explore and analyse consumer perspectives and discover factors of satisfaction. As no two consumers are exactly the same, it also becomes important to account for their cultural and economic backgrounds to discover how preferences differ amongst individuals. This information could then be used to create consumer profiles and improve targeted marketing, or attract new clients by appealing to their desires.

This thesis aimed to answer the question "*what are the key hotel-based and cultural aspects contributing to consumer satisfaction in different hotel types?*" by utilising several different machine learning and natural language processing approaches. The findings show that there are indeed differences in consumer preferences and factors of satisfaction across different hotel types - with factors such as proximity to the beach and having a mountain view being examples of these differing preferences. As an example, these factors are important when marketing a hotel, since a consumer could easily be attracted if offered a promotion on a mountain view room. These ‘upgrades’ could even be used to appease upset consumers and increase their overall satisfaction. Clients staying at beach hotels saw an increase in satisfaction if their hotel was located closer to the beach, if the location was quiet and if they had a sea view from their room. Individuals who took a ski hotel preferred staying in a room with a mountain view, and were enamoured by the proximity of the hotel to ski facilities. Business hotel guests preferred amenities such as a bar or pool, and the proximity of the hotel to the centre of a city. However, they strongly disliked if the hotel did not have any gratuities (such as free coffee, tea or wifi). Guests staying in residential accommodation mentioned their ‘host’ as opposed to only ‘staff’, and were thrilled by attention to detail in the home as well as factors such as kitchen facilities. Finally, city hotel guests like to be closer to the city centre, and dislike if spa or pool facilities cost extra. Of course these are only some of the preferences that have a strong effect on consumer satisfaction determined by the reviews that were scraped, in reality, many more preferences exist and could be analysed. However, the key finding is that consumer preferences do differ amongst different hotel types, and it is important for hotels to keep in mind that their guests may be looking for different things based on their stay - for example, a free coffee for a business client is more important than an upgrade to a sea view room.

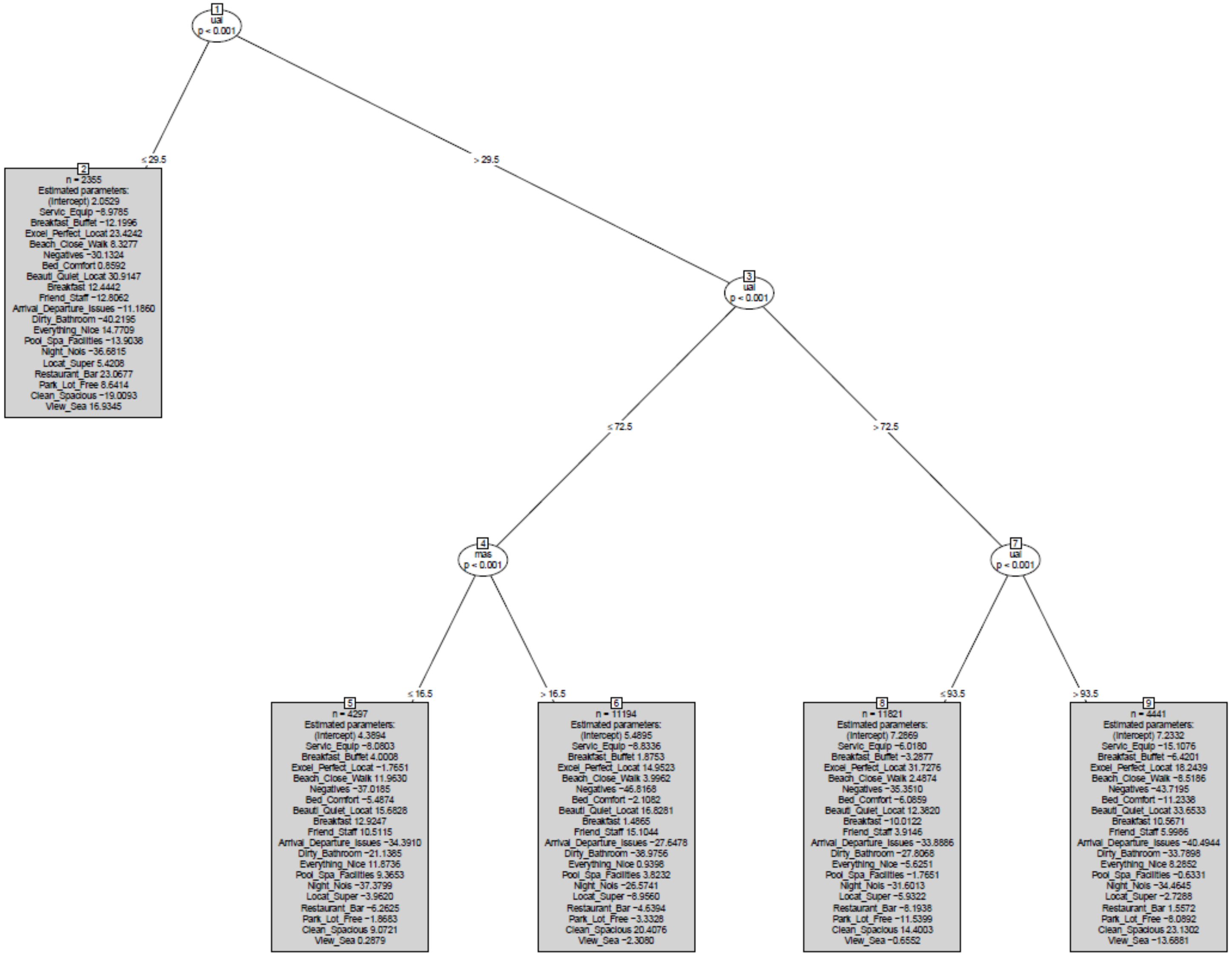
Besides differing factors of consumer satisfaction, cultural and economic dimensions also play a role in preferences. Consumers from less masculine cultures tend to favour features that are more based on relaxation, attention to detail, decor and views - things that are more artistic and creative. Those from more masculine cultures may prefer sport activities such as skiing or hiking. Further, those from countries with high uncertainty avoidance indices detest unforeseen expenses, or when unexpected arrival or departure issues happen. These singular events could ruin an entire trip for these individuals, and it is important to target them and offer extra compensation for the inconveniences. Economic indicators also have an impact on overall consumer satisfaction, whereby guests from countries with higher GDP per capita tend to be more favourable when rating their stay at a hotel. This could be due to the fact that the higher level of wealth may mean there is less scrutiny with how money is spent. Individuals that, for example, are able to afford multiple vacations per year may be more likely to be less critical of their trip than those that are only able to take less, due to their spending power.

From a methodological perspective, this thesis provides a framework for analysing consumer perspectives and how they differ based on certain economic and cultural factors. LDA is used to extract factors of consumer satisfaction from reviews, and logistic regression used to determine the relationship between these factors and overall satisfaction. A novel method, logistic regression trees, is then used to examine differences between cohorts based on predefined splitting variables - in our case, the cultural and economic values. This methodology has not been seen before in text analytical applications and could easily be applied to other fields and use cases to examine cohort differences and discover new insights based on the splits that occur. Without this novel model, we would have not been able to concretely make correlations between cultural and economic dimensions, and the factors of consumer satisfaction that were extracted from the reviews. The logistic model with interactions was quite difficult to interpret, and did not give us a clear picture of how consumer behaviour may differ amongst cultures, but it did confirm that specific factors relating to given hotels are important to consumers. This thesis also shows that incorporating deep translation as a text processing method allows for further exploration of consumer heterogeneity - which is novel and has not been done before in combination with LDA and sentiment analysis. These methodological approaches yielded high accuracy levels of 77-84%, which are in line with previous sentiment classification studies. It is important to note that the goal of this study was not necessarily ensuring a high accuracy, but rather ensuring ease of interpretation. Thus, more sophisticated machine learning models were not examined.

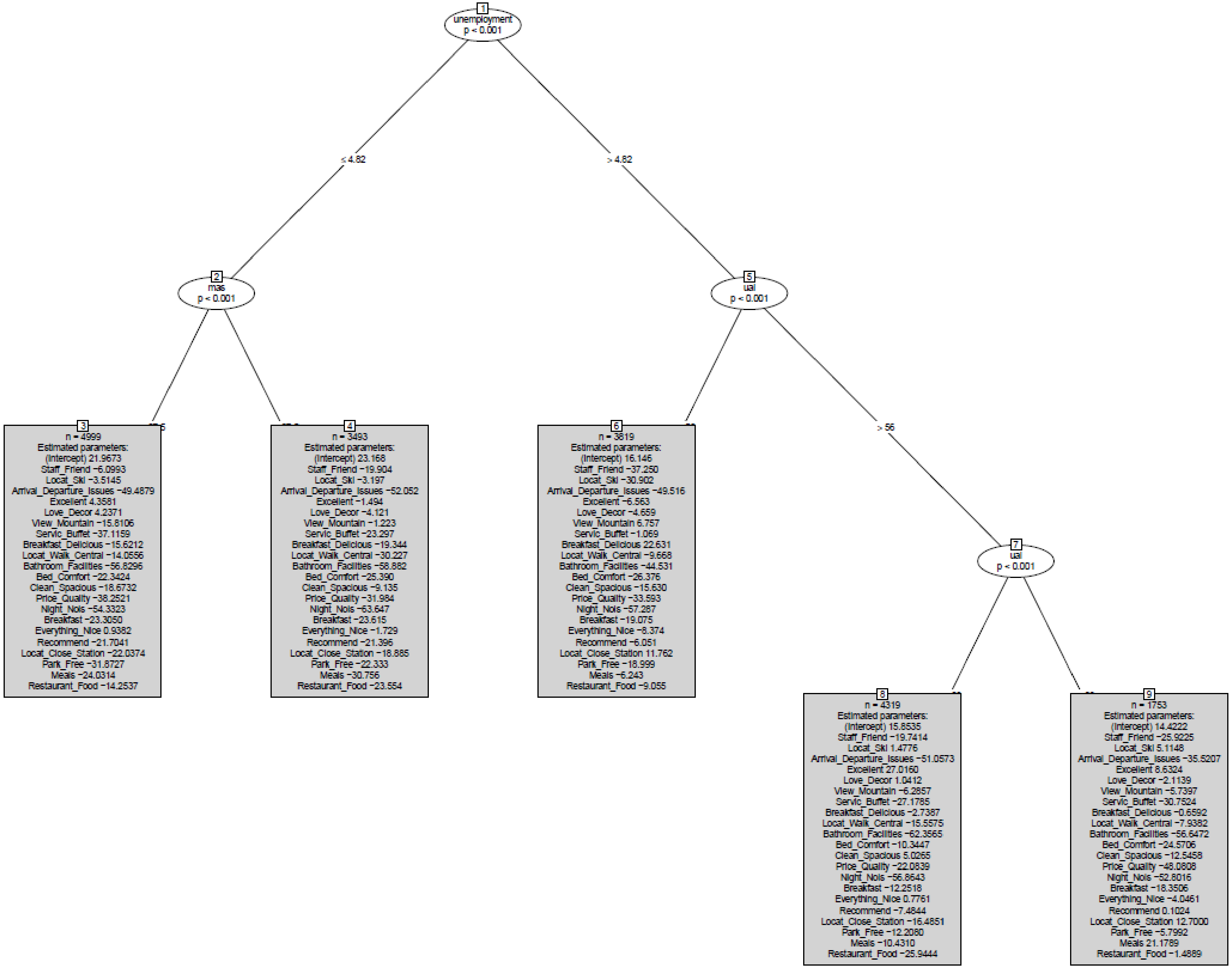
While this thesis provides hoteliers with a method to analyse consumer preferences and proves that they do differ based on hotel type and cultural or economic background, there is more room for improvement in this research space. Further research could be done including more geographical regions outside of Europe to adjust further for consumer preferences. Researchers can also look into taking the obtained heterogeneous preferences and creating different consumer profiles that may be used in recommender systems. For example, tying a reviewer's account information, such as their nationality, to the respective cultural and economic indicators may allow for search engines to provide them with hotels they would enjoy more. Using these extracted consumer preferences to generate tailor made recommendations could improve the match between a hotel and a guest based on what they are searching for. Further, Online Travel Agencies could use this method to classify hotel types and promote those with the respective positive factor of satisfaction. For example, if a user is searching for ski hotels, the OTA could first promote ski hotels that are located in extremely close proximity to ski stations, which could boost a consumer's overall satisfaction. Another limitation that could be further developed upon is the trustworthiness of reviews. Since the reviews were scraped from a website without any manual classification, it is possible that some reviews are fake and written by the company itself, in which case, the average score of the hotel may be skewed heavily. Review trustworthiness could be explored in further research to improve the models and ensure that authentic factors of consumer satisfaction are being captured. More sophisticated classification models such as random forest could also be explored in further studies to examine the impact of such algorithms on the accuracy of sentiment classification, especially in combination with deep translated input text. Finally, other confounding variables could be taken into account in further studies - for example, how preferences may change over time, or including neutral reviews to examine what makes a review balanced.

# Appendices

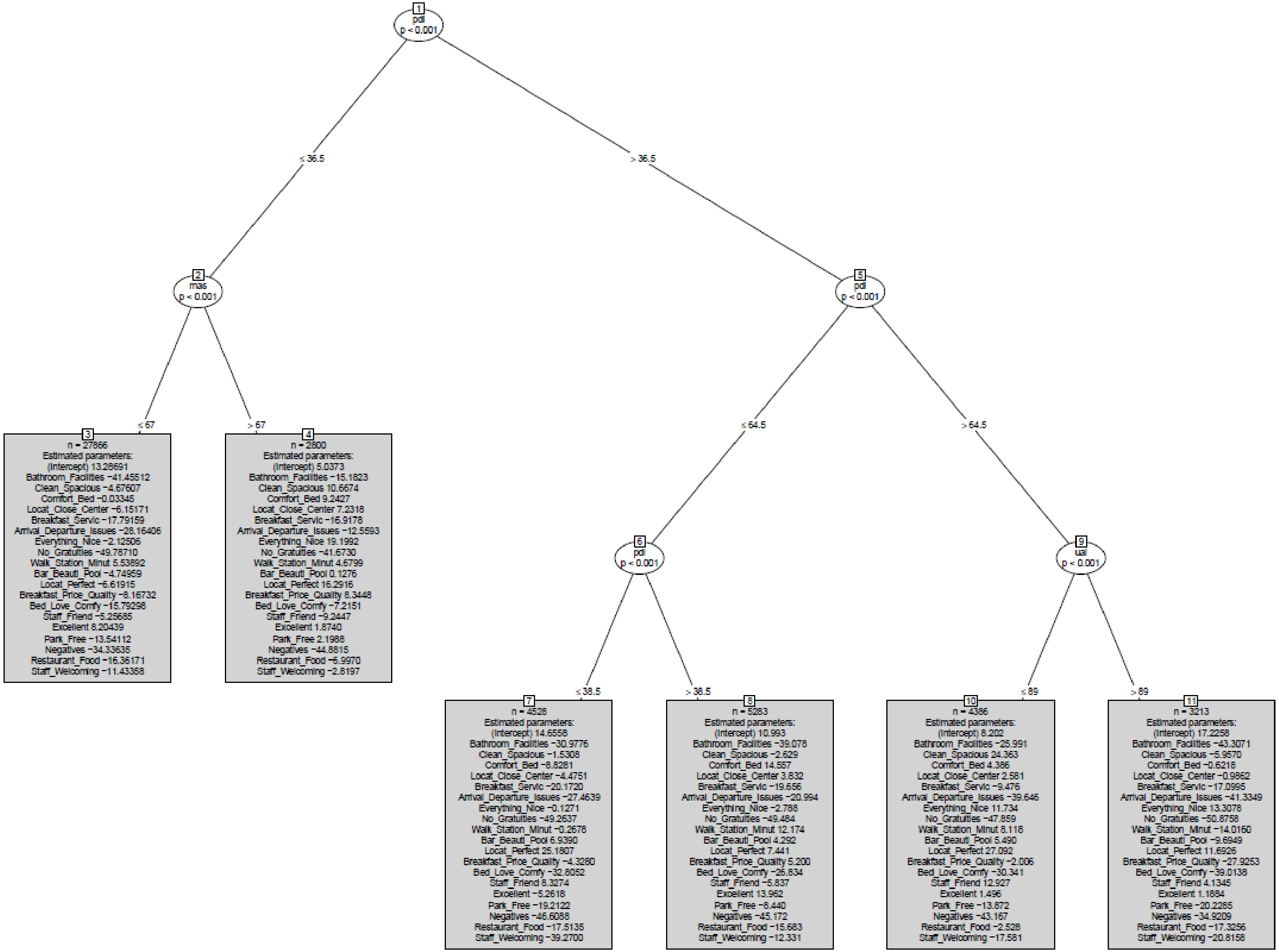
**Appendix 1:** Logistic regression tree for beach hotels



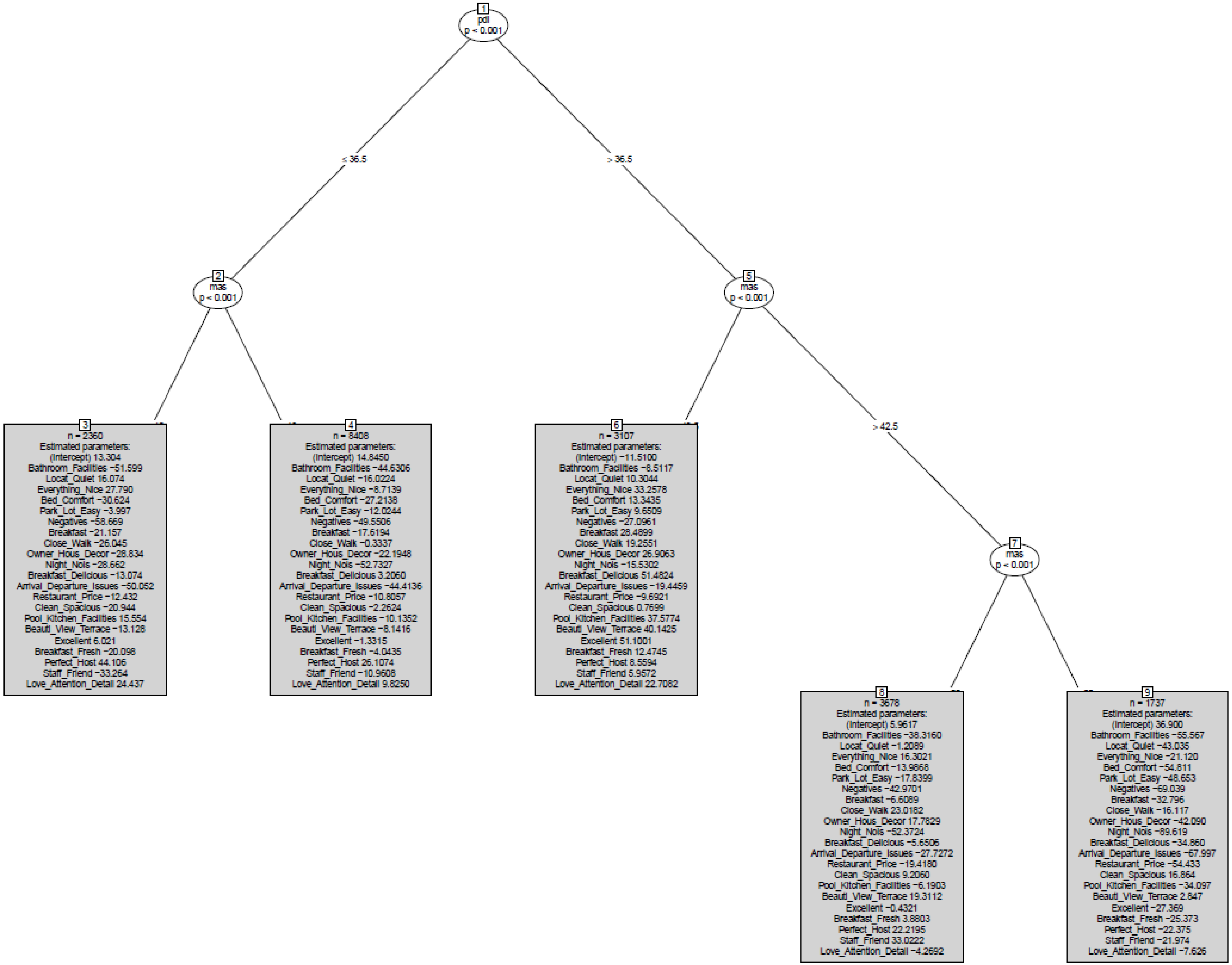
**Appendix 2:** Logistic regression tree for ski hotels



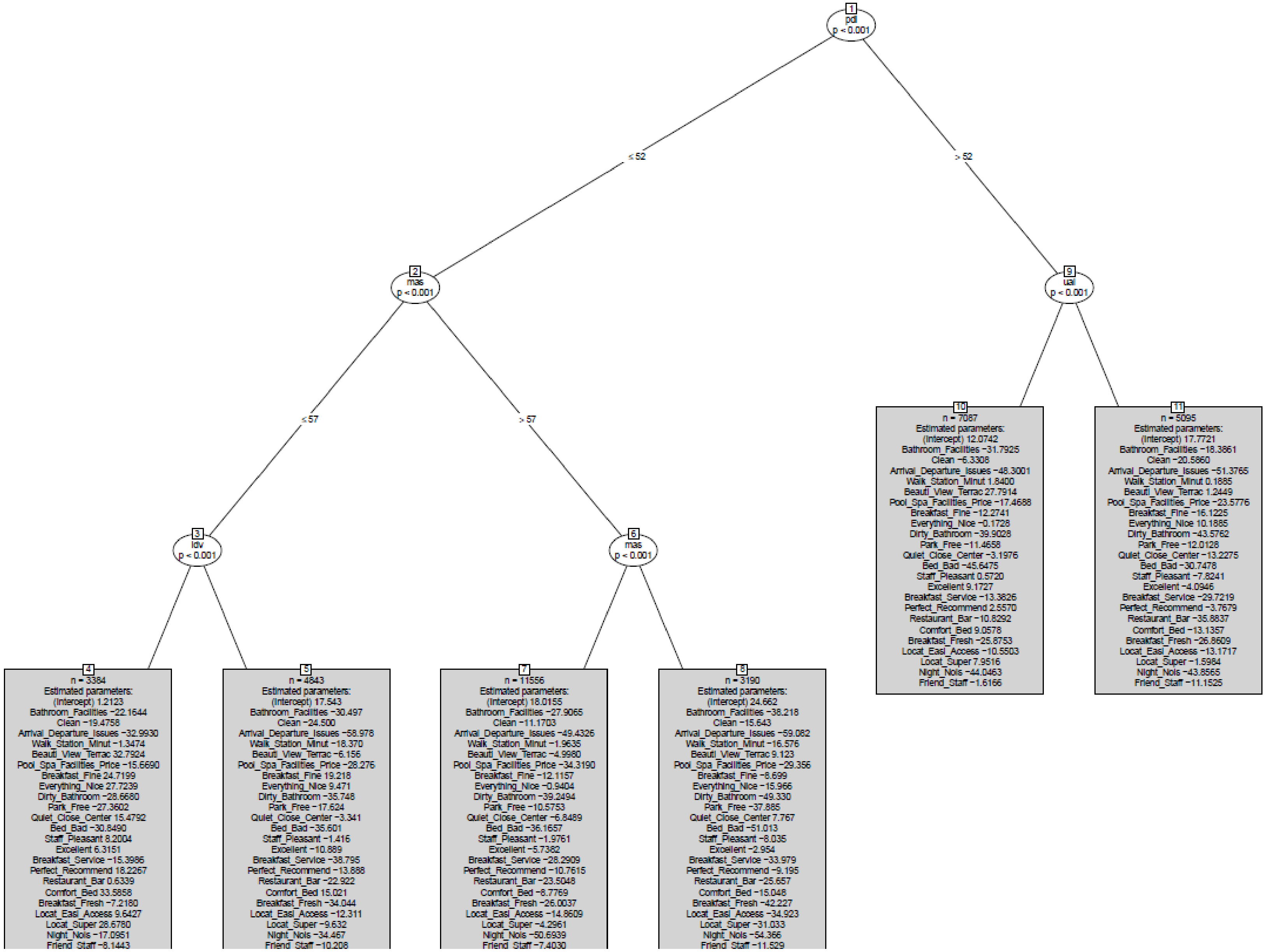
**Appendix 3:** Logistic regression tree for business hotels



**Appendix 4:** Logistic regression tree for residential hotels



**Appendix 5:** Logistic regression tree for city hotels



## References


Andrews, D. W. (1993). Tests for parameter instability and structural change with Unknown Change Point. *Econometrica*, *61*(4), 821.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2015). Understanding satisfied and dissatisfied hotel customers: Text mining of Online Hotel Reviews. *Journal of Hospitality Marketing & Management*, *25*(1), 1–24.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M. & Lafferty, J. D. (2009). Topic models. In *Text mining*, 101-124. Chapman and Hall/CRC.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*.

Booking.com. (2022). *About Us*. Booking.com. Retrieved April 2022, from https://www.booking.com/content/about.en-gb.html?aid=356980&label=gog235jc-1DCBQoggJCBWFib3V0SDNYA2ipAYgBAZgBCbgBF8gBDNgBA-gBAYgCAagCA7gCgJ7rkgbAAgHSAiQzNTA5YTRkYS1mZTllLTRiNjItOThhMi0xY2M3N2NjNjY0NWLYAgTgAgE&sid=784c8d670b610be08bd57b02fa868d6b&keep_landing=1&

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, *35*(6), 953–975.

Casillas, A., Ezeiza, N., Goenaga, I., Pérez, A., & Soto, X. (2019). Measuring the effect of different types of unsupervised word representations on medical named entity recognition. *International Journal of Medical Informatics*, *129*, 100–106.

Chang, V., Liu, L., Xu, Q., Li, T., & Hsu, C. H. (2020). An improved model for sentiment analysis on Luxury Hotel Review. *Expert Systems*.

Chatterjee, S. (2020). Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach. *International Journal of Hospitality Management*, *85*, 102356.

Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting more out of biomedical documents with Gate's full lifecycle open source text analytics. *PLoS Computational Biology*, *9*(2).

Day, G. S. (1971). Attitude change, media and word of mouth. *Journal of Advertising Research*, *11*(6).

Engel, J. F., Blackwell, R. D., & Miniard, P. W. (1995). *Consumer behavior*. Fort Worth: Dryden Press.

Finch, B. J. (1999). Internet discussions as a source for consumer product customer involvement and Quality Information: An exploratory study. *Journal of Operations Management*, *17*(5), 535–556.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent Dirichlet allocation. *Tourism Management*, *59*, 467–483.

Harrison-Walker, L. J. (2001). The Measurement of Word-of-Mouth Communication and an Investigation of Service Quality and Customer Commitment As Potential Antecedents. *Journal of Service Research, 4*(1), 60–75.

Haubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, *19*(1), 4-21.

Hofstede, G. (2013). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Sage.

Hofstede, G. (2016). *Dimension data matrix*. Geert Hofstede. Retrieved April 2022, from https://geerthofstede.com/research-and-vsm/dimension-data-matrix/

Hothorn T., & Zeileis A. (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research*, *16*, 3905-3909.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04*.

Huber, J., & McCann, J. (1982). The impact of inferential beliefs on product evaluations. *JMR, Journal of Marketing Research (Pre-1986)*, *19*(000003), 324.

Jalilvand, M. R., Esfahani, S. S., & Samiei, N. (2011). Electronic word-of-mouth: Challenges and opportunities. *Procedia Computer Science*, *3*, 42–46.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.).

Jin, W., Ho, H. H., & Srihari, R. K. (2009). OpinionMiner: a Novel Machine Learning System for Web Opinion Mining and Extraction. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '09*.

Kacen, J. J., & Lee, J. A. (2002). The influence of culture on consumer impulsive buying behavior. *Journal of Consumer Psychology*, *12*(2), 163–176.

Katz, E., & Lazarsfeld, P. F. (1955). *Personal influence: the part played by people in the flow of mass communications*. Free Press.

Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., & Merugu, S. (2011). Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. *Proceedings of the 2011 SIAM International Conference on Data Mining*.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. *Machine Learning*, *59*(1-2), 161–205. https://doi.org/10.1007/s10994-005-0466-3

Li, F., Han, C., Huang, M., Zhu, X., Xia, Y. J., Zhang, S., & Yu, H. (2010). Structure-Aware Review Mining and Summarization. *Proceedings of the 23rd International Conference on Computational Linguistics*.

Liu, B. (2010). Sentiment Analysis and Subjectivity. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing*. essay, Chapman & Hall/CRC.

Lu, B., Ott, M., Cardie, C., & Tsou, B. K. (2011). Multi-aspect sentiment analysis with topic models. *2011 IEEE 11th International Conference on Data Mining Workshops*.

McCarthy, L., Stock, D., & Verma, R. (2010). How Travelers Use Online and Social Media Channels to Make Hotel-choice Decisions. *Cornell Hospitality Report*, *10*(18).

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*.

Moghaddam, S., & Ester, M. (2013). The FLDA model for aspect-based opinion mining. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*.

O'Brien, T.,V. (1971). Tracking consumer decision making. *Journal of Marketing (Pre-1986), 35*(000001), 34.

Petersen, J. A., Kushwaha, T., & Kumar, V. (2015). Marketing communication strategies and consumer financial decision making: The role of national culture. *Journal of Marketing*, *79*(1), 44–63.

Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research*, *44*(4), 333–348.

Ray, A., Bala, P. K., & Rana, N. P. (2021). Exploring the drivers of customers' brand attitudes of online travel agency services: A text-mining based approach. *Journal of Business Research*, *128*, 391–404.

Reisenbichler, M., & Reutterer, T. (2018). Topic modeling in marketing: Recent advances and research opportunities. *Journal of Business Economics*, *89*(3), 327–356.

Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234–243.

Smith, A. (2013). (rep.). *Civic Engagement in the Digital Age*. Pew Research Center.

Statista. (2022, February 4). *Top online travel companies by market Cap 2021*. Statista. Retrieved April 2022, from https://www.statista.com/statistics/1039616/leading-online-travel-companies-by-market-cap/

Sun, T., Youn, S., Wu, G., & Kuntaraporn, M. (2006). Online word-of-mouth (or mouse): An exploration of its antecedents and consequences. *Journal of Computer-Mediated Communication*, *11*(4), 1104–1127.

Tiedemann, J., & Thottingal, S. (2020). Building open translation services for the World. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Tiedemann, J. (2016). OPUS – Parallel Corpora for Everyone. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*.

Tsytsarau, M., & Palpanas, T. (2011). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, *24*(3), 478–514.

UNWTO. (2019). (rep.). *International Tourism Highlights*.

van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15*(2), 121–140.

Woodside, A. G., & King, R. I. (2001). An updated model of travel and Tourism Purchase-Consumption Systems. *Journal of Travel & Tourism Marketing*, *10*(1), 3–27.

World Bank (a). (2022). *GDP per capita, PPP (current international $)*. Retrieved from https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

World Bank (b). (2022). *Individuals using the internet (% of population)*. Retrieved from https://data.worldbank.org/indicator/IT.NET.USER.ZS

World Bank (c). (2022). *Unemployment, total (% of total labor force) (national estimate)*. Retrieved from https://data.worldbank.org/indicator/SL.UEM.TOTL.NE.ZS

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, *44*, 120–130.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2).

Zhang, J. (2019). What's yours is mine: Exploring customer voice on airbnb using text-mining approaches. *Journal of Consumer Marketing*, *36*(5), 655–665.

Zim, C., Niepert, M., Stuckenschmidth, H., & Strube, M. (2011). Fine-Grained Sentiment Analysis with Structural Features. *Proceedings of the 5th International Joint Conference on Natural Language Processing*