Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Data Science and Marketing Analytics

From Word-of-Mouth to Click-of-Mouse

Unraveling Cultural Nuances from Multilingual Online Hotel Reviews

Ioannis Nikolaidis (657983)

Supervisor:	Dr. A. Tetereva
Second assessor:	Dr. C. Bellet
Date final version:	8th of August 2023

Abstract

The internet's advent has transformed the way consumers share their experiences. Previously driven mainly by word-of-mouth, now online reviews significantly influence consumer decisions, especially in the hospitality industry, due to the plethora of options on online travel agencies. Hoteliers are now competing not just with rivals but also against consumers' past experiences, necessitating continuous adaptation to meet the evolving demands. This thesis presents a methodological framework to extract insights about consumer satisfaction from online reviews. The framework uses novel and cutting-edge text analytics methods to extract topics and aspects from reviews, while also considering cultural heterogeneity as a distinguishing factor based on the reviewer's nationality. Data from global top travel destinations informed these findings, and the inclusion of multi-lingual reviews, translated using deep models, sets this work apart. This structure highlights multiple determinants of satisfaction with model performances ranging from 86-89% accuracy. Significant findings include the link between cultural dimensions like Individualism and review valence and distinct characteristics across cultural profiles. For instance, reviewers from Masculine societies prioritize practicalities, whereas Feminine societies value aesthetics and comfort. Managers can utilize these insights to tailor their marketing strategies, optimize costs and enhance service quality.

Keywords: text analytics, BERT, Top2Vec, SVD, cultural dimensions, consumer heterogeneity, NLP

Contents

List of Figures

List of Tables

Chapter 1

Introduction

To what extent has the internet influenced the way consumers choose a travel destination?

In the past, travelers relied on guidebooks and word-of-mouth to plan their trips. However, the rise of Web 2.0 applications has made it possible for travelers to access a wealth of information and experiences shared by other travelers. This has led to a huge increase in online user-generated content (UGC) on hotels, travel destinations, and travel services [1]. UGC can be a valuable source of information for travelers, as it provides firsthand accounts of experiences from other users. This can help travelers make informed decisions about their trips.

According to prior research, the Internet has become an increasingly important tool for travel planning [2]. A growing number of travelers are using the internet to search for travel-related information, read online reviews, and book travel arrangements [1]. Based on a study by Complete, Inc. (2007), one-third of travel purchasers visited a message board, forum, or online community before making an online travel purchase [1]. This suggests that a significant number of travelers believe that online reviews can help make travel decisions. Another study found that approximately 74% of travelers consider reviews when planning trips for pleasure [3]. Most travelers find reviews to be a valuable source of information when making travel decisions. This is especially true for experience goods, such as accommodation, where the quality is often unknown before consumption [4].

Travelers rely on word-of-mouth and online reviews to make inferences about the quality of an experience good before they purchase it. Prior studies have suggested that the influence of user reviews is particularly significant for experience goods [5]. This is because experience goods are intangible and heterogeneous, making it difficult for consumers to assess their quality before they purchase them. With the rising numbers of internet users daily, it is understandable that Electronic Word of Mouth (eWOM), especially in the hospitality industry, has become a major source of information for hospitality and tourism industry consumers [4]. Electronic WOM is also a significant medium that travel marketers should understand and harness, by extracting useful insights and adjusting their strategies accordingly. To achieve this, the latest techniques such as text analytics and natural language processing can be employed. For instance, Opinion Mining and sentiment analysis (OMSA) can offer valuable information to marketers as it is a procedure that deals with the extraction of opinions, sentiments, and emotions from the text. OMSA has been shown to be well-suited to various types of market intelligence applications [6].

In this thesis, we carry out an extensive analysis of reviews using advanced text analytics and machine learning techniques, targeting the identification of sentiments crucial to customers within the hospitality sector. We employ novel, state-of-the-art methodologies to extract valuable insights from textual reviews. By using techniques such as aspect-based sentiment analysis, we can better understand how customers feel about different aspects of a hotel, taking cultural variations and themes discussed into consideration.

1.1 Research Question

This study will look at how different hotel features, the topics discussed in reviews, and cultural diversity affect the review valence. We will do this by analyzing multilingual user-generated content from Booking.com, with a focus on top tourist destinations around the world to ensure a wide range of cultural perspectives and opinions. Hofstede's cultural dimensions will be integrated into the analysis to provide insights into cultural influences. The central research question is therefore as follows:

"What particular attributes of a hotel, review themes and cultural elements substantially affect the satisfaction of consumers?"

To answer the research question, the following methodological process will be applied:

1. Initially, the text is rid of superfluous noise, the language of the review is identified, and if not English, is translated accordingly.
2. The BERT algorithm for sentiment analysis will be leveraged to categorize reviews on a scale from 1 to 5, forming our dependent variable.
3. The Top2Vec algorithm will be employed to glean topics from reviews, without any preliminary preprocessing since the algorithm inherently possesses this capability.
4. In order to extract attributes, the text will be meticulously processed to form the TF-IDF matrix, experimenting with diverse tokenization strategies (unigrams, bigrams, 1-skip-1 : ngrams).
5. Singular Value Decomposition (SVD) will be employed to manage the dimensionality of the large, sparse TF-IDF matrix.

6. A simple, interpretable classification method (Logistic Regression) will be used to construct a sentiment analysis model that classifies reviews based on derived topics and attributes. Concurrently, the cultural dimension is accounted for by integrating Hofstede's dimensions, with the goal of quantifying statistically significant associations.

7. A penalized logistic regression will be utilized to assess the relevance of each independent variable toward the prediction of the review sentiment, providing an additional criterion besides statistical significance.

8. Logistic Regression Trees, that allow for the selection of regressors and partitioning variables will be employed. This would facilitate a comprehensive examination of the intricate synergies between cultural dimensions and other variables, and how these affect consumer satisfaction.

After preprocessing, topic and feature extraction, and classification stages, the components of significance will undergo detailed analysis and interpretation.

1.2 Relevance and Contribution

Consumer-generated reviews are an important part of the decision-making process for many consumers. These reviews can provide valuable insights into the products and services that are offered by a business, as well as the experiences of other consumers. As a result, it is important for businesses to understand what is meaningful to consumers and to adjust their products and services accordingly. This thesis will provide a novel approach for managers to efficiently examine user-generated content. It will also explore the features that consumers deem important to them during their stay. This information can be used by businesses to improve their products and services, to better meet the needs of their customers, and to possibly develop effective marketing strategies.

This research contributes to the existing literature by investigating reviews from top travel destinations using the latest methods (Transformers and self-attention mechanisms) in textual analysis. It also emphasizes the importance of cultural diversity and different opinions by not only considering English reviews but also using deep neural translation models to preserve review information from over 15 different languages. This allows for a better and more precise assessment of cultural heterogeneity.

Chapter 2

Literature Review

2.1 Consumer Behaviour and Cross-Cultural Research

Generally, consumer behavior can be conceptualized as a detailed examination of the components and attributes that an individual evaluates when choosing to consume a particular product or service. Stávková et al. [7], defined consumer behavior as the examination of how individuals reach decisions pertaining to their purchases, desires, needs, or actions, in relation to a specific product, service, or company. An important perspective when aiming to understand consumption behavior is viewing it as a social [8], and psychological process [9].

The decision-making involved in consumption is not merely transactional; rather, it is an intricate process in which individuals invest their own resources such as time, money, and effort to obtain products [10] [11]. This approach, encapsulates the complete consumption cycle, starting with the decision of where, when, and why to acquire a product, continuing with how frequently the product is used after the purchase, and concluding with how the product is evaluated [8]. Understanding these behavioral patterns is considered crucial for businesses, as it can contribute to a better understanding of how potential customers might interact with a certain product and assist in the identification of untapped market opportunities [9].

Several exploratory models have been proposed to unravel the rationale of the decision-making of consumers. Among others, the 'black-box' model, hypothesizes that consumer decisions are driven by external stimuli. These stimuli encompass a wide array of factors ranging from marketing messages, product sampling and availability, and promotional campaigns to pricing strategies [12]. It is perceived that while external factors can influence consumers, it does not necessarily mean these influencing elements solely originate from the marketing strategies employed by businesses. An emphasis on the 'self' as a profound influencer of social behavior can be given [13]. The 'self' as an aggregate of individuals' overt or covert declarations using first-person pronouns, underpins all aspects of social motivation. This broad concept of 'self' passes through all social impulses, impacting on how individuals gather, comprehend, and assess information [13]. Empirical examples underscore 'self' influences on behavior, including links to self-definition [14], identity prominence [15], self-observation [16], memory access [17], and self-esteem [18] related independence from group norms. Furthermore, Snyder's research [19] pinpoints that individual distinctions in terms of monitoring social situations(high self-monitors) against self-introspection(low self-monitors) have meaningful implications. These differences are associated with how people observe their emotions, develop beliefs, and how they relate their attitudes to behaviors. The aforementioned statement can be generalized for every aspect of social motivation and its components can be referred to as elements of a cultural group's subjective culture [20].

While certain facets of the 'self' might be perceived as universally common, such as basic human needs declared by expressions like 'I want to sleep' or ' I am hungry', there are also characteristics that are culturally specific. For instance, conceptions of the 'afterlife' can vary greatly, and are often influenced by an individual's religious beliefs, cultural mythology, worldview, or the language prevalent in their culture [13]. Moreover, the self can be thought of as a complicated structure consisting of universal aspects, culture-specific elements, and personal, public, and collective perceptions [21]. The weight of these aspects is differentiated across different cultures, influencing social behavior and identity formation [13]. Although, at the time there was evidence of variations of the self across cultures [22], the specification of the way the self establishes certain features of social behavior in different cultures was undeveloped.

According to the above-mentioned studies, it is deduced that consumer behavior and culture are two concepts that are significantly interrelated. Elaborating further on the subject matter, major shifts in political boundaries and consumer market structures, underpinned by robust socio-cultural dynamics, have led to a transformation of consumer behavior patterns. The integration of regions has eliminated many, if not all, barriers between markets while simultaneously expanding and unifying market entities [23]. The reshaping of consumer behavior became particularly noticeable through cultural interaction following the initial wave of consumers migrating from emerging market economies to industrialized ones. Enhanced mobility, whether through migration or travel, led to consumers being widely exposed to a diverse array of products and lifestyles from different countries with varying cultures and idiosyncrasies [23].

Concurrently, advances in communication technology have eradicated the difficulty of immediate interaction, which resulted in linking markets globally and subjecting consumers to a plethora of external influences beyond their national restrain [23]. Furthermore, changes in intrapersonal and socio-cultural communication patterns, caused by affordable traveling, satellite communication links, the Internet, and other components, resulted in a worsened patterning of consumer behavior after an extensive conflation of traits. Links between geographically scattered groups are becoming rapidly established [24]. Subsequently, with consumer behavior evolving into a phenomenon of increased complexity, academic attention has been directed towards the development and comprehensive analysis of frameworks for cross-cultural consumer behavior across previous centuries and extending into the present day. Culture, considered among the most abstract factors affecting human actions, is developed by numerous interconnected influences. These include language, education, ecological factors, and socialization processes within a society's economic, political, religious, social, and technological frameworks [25].

Culture has been described in various ways in an attempt to capture its meaning. One broad interpretation sees culture as the human-created portion of our environment [26]. Another definition is that "Culture is that complex whole which includes knowledge, belief, art, morals, custom and any other capabilities and habits acquired by man as a member of society" [27]. A different perspective, as presented by Geert Hofstede [28], postulates culture as the collective mental framework that distinguishes one group or category of people from others. Hofstede's concept represents cross-cultural comparisons, and these research methodologies are referred to be 'etic,' a word taken from linguistics. The approach would be 'emic' if the emphasis was mainly on the traits of one culture [29]. Cross-cultural research aims to identify a range of customary behavioral characteristics that can assist in investigating similarities or differences among multiple cultures. This exploration is justified by the suggestion that individuals who speak different languages (for example English and Chinese), reside in geographically distant locations (like England and Australia), or have experienced different historical eras may possess distinct subjective cultures [13].

2.2 Cultural Dimensions

Several cultural dimension systems have been proposed. For instance, Triandis [13] studied dimensions of variation of cultural contexts that have direct relevance to the way the self is defined and the link between culture and the self. The dimensions proposed were that of "Cultural Complexity, "Individualism versus Collectivism", and "Tight versus Loose" cultures. Regarding cultural complexity, it was accessed that the more complex the culture, the more confused likely to be the individual's identity [30]. Regarding the second dimension, it was mentioned that individuals that are part of collectivistic cultures, pay more attention to ingroups and outgroups and regulate their behavior than is the case in individualistic cultures [20].

This study will primarily emphasize a more extensive approach that was presented by Geert Hofstede [31], who has defined four dimensions related to anthropological or societal issues. These dimensions were labeled as: "Power Distance", "Uncertainty Avoidance", "Individualism versus Collectivism" and "Masculinity versus Femininity".

Thoroughly, Power Distance refers to the acceptance of unequal power allocation in institutions and organizations by less powerful individuals. In other words, it is primarily an anthropological or societal issue intertwined with social inequality and an individual's dominance over others [31]. Uncertainty Avoidance is associated with the way a society deals with conflicts and aggression while aiming to answer the question of "how much threatened do people feel when they are in uncertain situations?" [31].

Individualism versus Collectivism can be perceived as a spectrum with two opposing ends. On the one end, individualism is characterized as a state where individuals are expected to care for themselves and their immediate family alone. On the other hand, collectivism, at the opposite end of the spectrum, represents a scenario where individuals are part of larger in-groups or collectivities that offer them support in return for their loyalty. This dynamic essentially pertains to an individual's degree of reliance on a group and their self-perception as an 'I' or a 'we' [31]. Regarding the fourth dimension, Masculinity is associated with societies where their dominant values are success, money, and things whereas, on the opposite side, Femininity refers to societies in which the dominant values are compassion, empathy, and quality of life.

Following additional research, two more dimensions were added: "Long Term versus Short Term orientation," which is related to the choice of focus for people's efforts in terms of the future or the present and past, and "Indulgence versus Restraint," which is related to the gratification versus control of basic human desires related to enjoying life [32], concluding in the final six dimensions. Regarding the relevance of cultural dimensions and the extent of how much they affect customer reviews, research by Ramona Diana Leon [33] demonstrated that the variation of Hofstede's cultural dimensions can explain 9.9 % of reviews variation and 4.5 % of rating variation.

In this research, Hofstede's cultural dimensions are assumed to be definitive. They are employed to examine their potential influence on a reviewer's assessment of accommodation within the hospitality sector, while encompassed by other variables such as topics.

2.3 Sentiment Analysis and Consumer Reviews

Nowadays, the volume of text-based information on the internet is growing day by day [34]. Increasing volumes of customer feedback obtained from online reviews have become an increasingly important subject for research in the area of customer Sentiment Analysis (CSA) as the field of IT has evolved [35]. While it's evident that there's an abundance of information available, the task of accurately analyzing and identifying key aspects or topics within consumer reviews presents a substantial challenge. Efforts to manually process this information exceed human capabilities due to the extensive time required [34], therefore it is not considered an option. To solve this problem, Sentiment analysis, also referred to as opinion mining, is a procedure that involves discerning and categorizing the sentiments or emotions (whether positive, negative, or neutral) expressed within consumer reviews. This process employs methods from text analysis, machine learning, and natural language processing [36].

Textual information is mainly divided into two categories: facts and opinions [34]. Facts are associated with subjective information such as "the location of the hotel is the A area". Conversely, opinions correspond to the actual sentiment of their presenters. In other words, they signify how the reviewers felt about a product or a service and ideally also elaborate on the reasons for their approval or disapproval. Opinions can take various shapes and scales., hence platforms are tailored to extract the necessary information depending on the product or service in question. For instance in Amazon, there is a designated review section that allows customers to share their thoughts on the products they've purchased and rate them on a set scale. Another case could be the movie review platforms such as IMDb or Rotten Tomatoes, where reviews tend to delve deeper into the reasons behind an individual's liking or disliking of a movie and its specific elements. Relevant to this study's context is booking.com, a platform focused on the accommodation industry. On this site, although not mandatory, reviewers are encouraged to provide a title for their review, rate the accommodation overall as well as its specific features, and distinguish between the positive and negative aspects of their stay during the review. Consequently, given the complexity of opinions in terms of variety and size, sentiment analysis becomes an indispensable tool in determining the factors that guide review valence.

There are several categories of sentiment analysis. Specifically, prior studies defined four different techniques - Word Level, Sentence Level, Feature level, and Document-level sentiment analysis [34]. Word-level sentiment analysis is considered a common and effective technique as it links sentiment words to a class (e.g. "brilliant", "perfect", and "very good" as positive sentiment). It mainly relies on databases that pair adjectives with respective classes and its reliability depends on the comprehensiveness of the sentiment word set, including synonyms and antonyms. Thus, it is understandable that a significant constraint of this approach is its inability to take into account negation words. A statement such as "the breakfast was not very good" might be mistakenly classified as positive. This occurs because the phrase "very good" matches with a lexicon, but the negation term "not" is ignored. The sentence-level version analyzes sentiment at a more granular level. It goes the idea of the word level sentiment analysis by including negation rules such as "no", "not" and "never" among others. Negative verbs such as "stop" and "problem" signify a negative sentiment and overall it includes an analysis of different verb forms and combinations. This method has its own limitations. Even though it does take into account negation words, it may inaccurately classify a sentence as positive if it contains more words related to positive sentiment than negative, without fully examining the significance of each feature. The feature-level analysis, a more sophisticated sentiment analysis method, targets individual attributes within the reviews. It is capable of assigning positive or negative weights to each identified feature. The aggregation of feature weights is what determines the overall sentiment, and it generally utilizes mathematical or statistical formulas for sentiment prediction. Document-level sentiment analysis classifies a whole document, like a review, as expressing either positive or negative sentiment. This process often employs methods such as SentiWordNet to focus on the document's term profile and to draw out terms with a specific Part of Speech (POS) label [34]. In this technique, the sentiment score is compiled through a particular combination of weight assignments and aggregation procedures. Two principal factors that need to be taken into account are which POS tags ought to be pulled out, and what the optimal method is for distributing weight to the scores of various POS tags.

The objective of this analysis is to identify features that significantly determine the review sentiment. Hence, the sentiment method of this study is considered as aspect-based sentiment analysis. This type of analysis involves three main steps: identifying pairs of sentiment and their corresponding targets within the text, categorizing these pairs based on predetermined sentiment values such as positive or negative, and then aggregating the sentiment values for each aspect to develop a concise summary [37]. In essence, aspect-level sentiment analysis involves collecting sentiments related to specific aspects and summarizing the final sentiment. Furthermore, various methods exist for aspect detection, ranging from frequency-based to intricate machine learning strategies [37]. Frequency-based methods, mainly rely on frequently used single or compound nouns in the text but may introduce noise due to potential misinterpretation. More challenging to implement are syntax-based methods, which demand a thorough understanding of the language to distinguish relationships between word pairs. Notably, unsupervised machine learning is a proven and popular method for extracting key aspects from text [37]. Several unsupervised techniques have been proposed to automatically model different aspects. Methods based on Latent Dirichlet Allocation (LDA) present each document as a combination of various aspects (or topics), providing a word distribution for each aspect [38]. A significant drawback of LDA is its use of the bag-of-words (BoW) method to represent texts, which essentially overlooks the semantic relationships between words in the text. Generally, traditional topic modeling techniques like LDA demand extensive corpus preprocessing; meticulous choice of parameters, including deciding the number of topics to be generated; suitable model evaluation; and the interpretation of the produced topics based on common sense and subject-matter expertise [39]. More recently, neural models have shown their competence in extracting topics with better coherence [40], and hence, these methods could be employed.

Given the above considerations, the Top2Vec method was employed for the initial stage of feature extraction, specifically for the purpose of extracting topics and the words associated with each. One benefit of Top2Vec is its ability to automatically determine the number of topics, in contrast with other methods that require presetting this number. Each topic identified by the model is linked with a set of generated keywords, enhancing the comprehension of the projects' semantics [41]. The purpose of this algorithm is to find topic vectors based on an input corpus, and it does this by firstly creating document and word embeddings using neural network models such as Doc2vec or Word2vec, then using dimensionality reduction and clustering techniques to group similar vectors and spot dense areas in the vector space, which are interpreted as topics [42]. It's noteworthy that Top2Vec facilitates, and offers a hierarchical reduction tool, which is beneficial given the large number of extracted topics from this method that demands exhaustive analysis. Initiating the topic reduction process with a target of ten topics is a recommended strategy for Top2Vec [43]. Accordingly, in this research, the reduction process will be performed, after topic extraction, aiming for ten final topics.

Thara & Sidharth [44] proposed a structure to apply Singular Value Decomposition (SVD) on the expansive sparse TF-IDF matrix to manage its dimensionality problem, in an effort to conduct aspect-based sentiment classification. Other studies followed a similar approach as well [45]. While this technique has been commonly used in past research, it has not been integrated with an interpretable model, presenting a novel approach in this thesis. After the extraction of aspects, the next step involves their classification. This can be typically achieved through various methods, such as syntax-based analysis, supervised learning, or unsupervised learning methods [37]. In this thesis, the objective is to analyze hotel reviews, assign them to the appropriate sentiment category, and explore the factors that influence the sentiment classification of a specific review. In order to achieve that, it is preferable to use interpretable supervised techniques. These methods contribute to identifying significant correlations and assessing their link with the response variable, which is why logistic regression models were employed. Similar model approaches have been conducted numerous times in the past with decent results [46] [47] [48].

2.4 Related work in the Hospitality Industry

A study by Subroto & Christianis [49] aimed to classify peer-to-peer accommodation guest reviews in Indonesia's top 10 cities based on ratings. They employed a Classification and Regression Tree (CART) model to identify impactful attributes in reviews and found terms like 'dirty', 'bad', 'toilet', and 'never' were often correlated with lower ratings. They used predefined topics (linking words with a certain theme) rather than computational methods like Latent Dirichlet Allocation (LDA) to form topic-attribute associations. Their Random Forest model, used to predict ratings based on these topics, achieved 60.09 % accuracy. This modest result may be due to the simplistic topic extraction approach and lack of investigation into the specific impacts of individual topics and potential cultural influences.

Another study by Calheiros [50] expanded the scope of sentiment polarity analysis by integrating text mining techniques and introducing Latent Dirichlet Allocation (LDA) for topic modeling in the research. The analysis was performed on over 400 customer reviews, and identified inherent relationships using LDA modeling within two aspects of the hotel industry: sentiment classification and hospitality issues concerning a selected eco-hotel. The major contribution of this study was the application of LDA topic modeling to discover customer emotions triggered by various hotel aspects, intertwining sentiment polarity and hotel domain semantics. This research uncovered significant relationships between topics and review valence, with some notable of them being the topics related to food, location, hospitality, and romance which is related to aesthetics. Surprisingly, their findings did not highlight "people" or "decoration," elements that the hotel managers believe to be primary reasons for visitors' interest.

A study by Hu et al. [51] proposed the Structural Topic Model (STM), a novel text analysis method, to identify dominant themes in negative reviews due to the limitations of the Latent Dirichlet Allocation (LDA) method. Analyzing 27,864 TripAdvisor hotel reviews, the study identified ten major negative topics, of which two were novel categories, namely "bugs" and "room type". Interestingly, typical dissatisfaction factors such as transportation, decoration, Wi-Fi, and food service did not emerge as prominent negative aspects in the STM model. The study, while informative, had shortcomings, the most notable of which were its restrictive emphasis on data from a single city and platform up to 2013 and its lack of critical hotel and consumer qualities such as culture.

A study by Ramona Diana Leon [33] took the analysis one step further and addressed the cultural aspect in reviews by analyzing 1,821 comments for the Catalonia Sagrada Familia hotel across 77 countries. The aim of this research was to identify the role of cultural distinctiveness in a hotel's online customer behavior, particularly in review ratings. Another important aspect of their study was the use of topics, which was supported by the research of Rossetti et al (2015), who underlined the significance of topic modeling for sentiment classification. The study leveraged LDA topic modeling to uncover customer sentiments on hotel issues. The study discovered that guests who are more likely to provide comprehensive reviews generally come from cultures characterized by low power distance, collectivism, masculinity, low uncertainty avoidance, long-term orientation, or indulgence. Conversely, those who are more inclined to diverge from average previous ratings are usually from cultures with high power distance, individualism, femininity, high uncertainty avoidance, long-term orientation, or indulgence.

In reviewing prior studies, it's clear that many researchers have primarily focused on the hospitality industry for textual analysis of reviews. However, there are certain limitations in these studies that this research aims to address in depth. Firstly, the cultural aspect, which is vital in an industry involving global customers, is often disregarded.

Secondly, the vast majority of previous studies did not consider multilingual reviews, primarily focusing on English reviews or those from a single selected country. This lack of diversity poses a challenge for customer segmentation and marketing development in the accommodation industry. Lastly, the methodological framework of past sentiment analysis studies, especially in the field of topic extraction, appears somewhat repetitive, indicating a need for novel approaches and refreshed methods for topic modeling and sentiment analysis.

Chapter 3

Data

The information for this study was gathered from Booking.com, one of the world's top online platforms for booking travel and accommodation. As of December 2022, Booking.com leads the online travel agent (OTA) market, with a market capitalization of over 78,000 million dollars and a huge global revenue of 17.09 billion dollars. Because of its size and reach, Booking.com was an excellent choice for collecting and analyzing customer reviews from all over the world. The data was collected using an API Scraper, which pulled information from more than 400 hotels in around 23 top travel destinations (Forbes, 2019), ensuring a wide range of cultures and viewpoints. The research focused on top tourist attractions to ensure a broad perspective, rather than focusing on a single market. This was done to account for a variety of factors, such as different pricing structures, diverse destination offerings, and a multitude of cultural perspectives. The approach thus ensured a more comprehensive and versatile representation of global tourism trends, which is essential for generating robust insights and valid conclusions. The final data set used in the study includes 9051 reviews from reviewers in 30 different European countries. The selection of European countries was made with the condition that they are included in Hofstede's analysis, ensuring a corresponding cultural dimension for each observation.

3.1 Data Scraping

An API scraper was used to gather information from top travel destinations on Booking.com. To ensure that the data set was not biased towards non-popular hotels, each review was scraped based on a query that excluded hotels with fewer than 200 reviews. Since negative reviews are less common than positive reviews, the filtering process started with lower-rated reviews to ensure that a significant portion of negative reviews were included in the data set. In addition, the filtering process was designed to emphasize consumer data, such as the review date, the reviewer's nationality, the reviewer's travel type, the overall rating, the title of the review, and the positive and negative aspects of the review. Unfortunately, the API scraper did not provide an option to filter by reviewer country. As a result, the initial data set consisted of approximately 32,000 reviews, but only the European countries were retained after cleaning.

The review text on Booking.com is divided into three sections: the title, the positive aspects, and the negative aspects. However, some reviewers did not follow this format, and some even wrote their entire review in the title section or wrote "nothing positive" or "nothing negative" in the positive or negative sections. This can seriously cause data quality problems as it might affect the accuracy and consistency of the data analyzed. To address this issue, the data was cleaned to remove observations with 5 or fewer words in the positive or negative sections. The review text column was then constructed by appending the review title, then the positive aspects (if any), and then the negative aspects (if any).

Another task that required further action was the existence of reviews that were not in English. As this analysis aims to perform cross-cultural research, it is important to maintain information from several nationalities. For that reason, it was decided that reviews that were not in English should be translated using a robust method to minimize information loss of accuracy and content/context. The language identification in the dataset was performed using the Python package *langdetect*, which is a tool that leverages Google's language detection library. The *langdetect* package can identify the language of text with a high degree of accuracy (Danilák, 2014;2019). The reviews were translated using the *Opus-MT* deep learning translation model, which was developed by the Language Technology Research Group at the University of Helsinki. This model was chosen because it is able to maintain the context of the original text while translating it into another language. This is important for sentiment analysis, as it allows the model to understand the overall sentiment of the review, even if it is translated from another language. (Tiedemann & Thottingal, 2020). The *Opus-MT* model was trained on a massive text collection called OPUS, which contains more than 3.2 billion sentences. This gives the model a large corpus of data to learn from, which helps it to translate text more accurately (Tiedemann, 2016). The reviews have been translated from a variety of languages, ranging from languages such as French, Spanish, Dutch, and Italian, to Russian, Polish, Latvian, Serbian, Greek, and Finnish among others. After the translation of reviews, a function was employed to calculate the number of words in each review and construct a new column named *num_words*. This variable will be used to investigate the relationship between the length of a review and its valence.

3.2 Preprocessing

The review text is a collection of all the raw reviews that were scraped from the website. This thesis performs a thorough analysis of textual data, so each review (document) must be processed in a way that makes the analysis efficient. This signifies that any information that might be redundant for the models should be left out. As mentioned in the methods section, the reviews were classified based on sentiment using BERT, and topics were extracted with Top2Vec. These techniques did not necessitate extensive preprocessing since BERT is designed to function with noise and Top2Vec automatically performs the required preprocessing steps. The only intervention that was made in the first stage was to remove leftover elements from HTML web scraping, such as the Unicode \r that was identified in several cases. Thus, the processing of the text in this analysis begins with the aim generate a Term Frequency - Inverse Term Frequency (TF-IDF) matrix and apply a dimensionality reduction technique so as to include the resulting dimensions in the model.

The initial step involved converting all documents to lowercase, which is necessary because computers cannot inherently distinguish that "Terrible" and "terrible" are the same word. Subsequently, all numerical characters were eliminated as they are often irrelevant when it comes to textual analysis. All punctuation was also eliminated due to its potential to compromise the further preprocessing steps. Then, stopwords, which are common English words such as "and", "the", and "of", which usually lack substantial meaning, were removed. Furthermore, the process of stemming was executed to reduce words to their root form, helping in the consolidation of various forms of the same word. For instance, words like "swimming" and "swimmer" would both be reduced to "swim". Prior to tokenizing the review column, extra whitespaces were eliminated.

After these preprocessing steps, the review column was tokenized and a vocabulary consisting of all the unique words in the reviews was extracted. Since some words, such as slang terms or misspelled words, did not correspond to anything specific, the vocabulary was substantially pruned. This pruning was achieved by removing words that appear in fewer than 50 documents and words that appear in more than half of all documents, with the usage of an appropriate function. Such words, being either extremely uncommon or excessively common, are typically not useful for analysis.

Following this, the pruned vocabulary was vectorized to generate a document-term matrix (DTM). In some cases, after the data was pruned, certain words were not found in any review, resulting in rows of zeros in the DTM. These rows, lacking any beneficial information, were subsequently identified and removed from the matrix.

Finally, the DTM was transformed into Term Frequency-Inverse Document Frequency (TF-IDF) format and Singular Value Decomposition (SVD) was applied to compress this large sparse matrix into a more manageable 10-dimensional format, to carry on with the analysis.

3.3 Cultural Dimensions

Hofstede's cultural dimensions were obtained from the Hofstede Institute's website. The countries that were included in this analysis were selected based on their match with Hofstede's research, in order to ensure that there would be no missing values in the data. The cultural dimension scores were then appended to the review dataset by matching the reviewer's nationality to the country from Hofstede's data.

3.4 The Happiness Index

When looking for additional variables that could influence customer satisfaction, there was decided to incorporate a measurement that corresponds to the general satisfaction of a person within a society. Instead of using raw metrics like GDP per capita, inflation, and other macroeconomic factors, we decided to use an index that represents the people's perspective of social welfare. To serve that purpose, the Happiness Index was utilized as it takes into account various aspects of the economy and societal functioning, providing a more comprehensive view.

Thoroughly, the Happiness Index is a metric incorporated by the World Happiness Report. It is based on the Cantril ladder, which is a single-item question that asks respondents to rate their lives on a scale of 0 to 10, with 0 being the worst possible life and 10 being the best possible life. The report also includes data on six other factors that are considered to contribute to happiness: GDP per capita, social support, healthy life expectancy, freedom, generosity, and perceptions of corruption. It is important to mention that the metric incorporated, is not based on any index of these six factors. Instead, the scores are based on individuals' own assessments of their lives. This means that the report is a measure of relative happiness and not absolute happiness. As a result, the Happiness Index scores were gathered for every country present in the data, and these scores were then aligned with the reviewer's nationality and the corresponding year, given that the reviews were from 2022 and 2023.

3.5 Descriptive Statistics

Upon a detailed examination of the extracted data, as illustrated in Figure 3.1, it is evident that the review rating variable ranges on a 1 to 10 scale. A significant portion of the observations lies within the 7-10 bracket, suggesting a prominent imbalance in the data set and a skew toward positive reviews. Transforming such a broad scale into a binary response variable presents a challenging task since it involves a degree of subjectivity from the researcher, in determining which ratings qualify as high, low, or neutral. As a resolution to this issue, the decision was made to utilize BERT for classifying sentiments on a 1 to 5 scale.

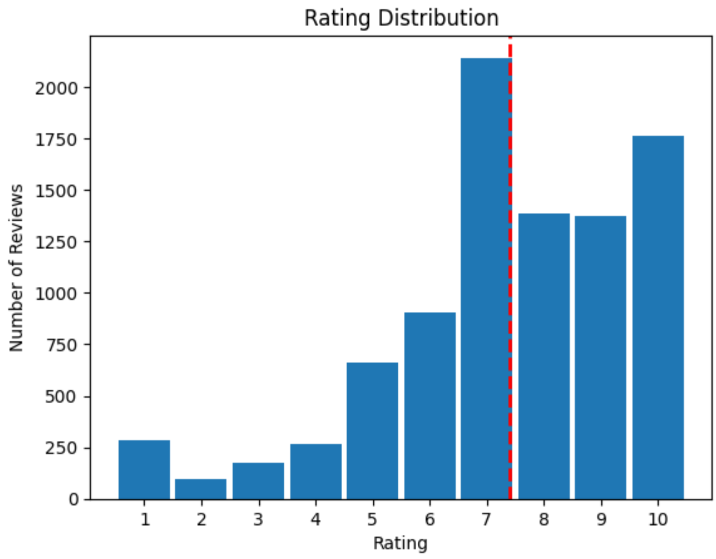

Figure 3.1: Rating Distribution[1]

Figure 3.2 shows that the majority of reviews fall within the 4-5 interval, after the application of BERT sentiment classification in the reviews. This suggests that the dataset is imbalanced, with the majority of reviews being positive. To address this issue, we upsampled the observations of the minority class (low sentiment) to match the majority class.

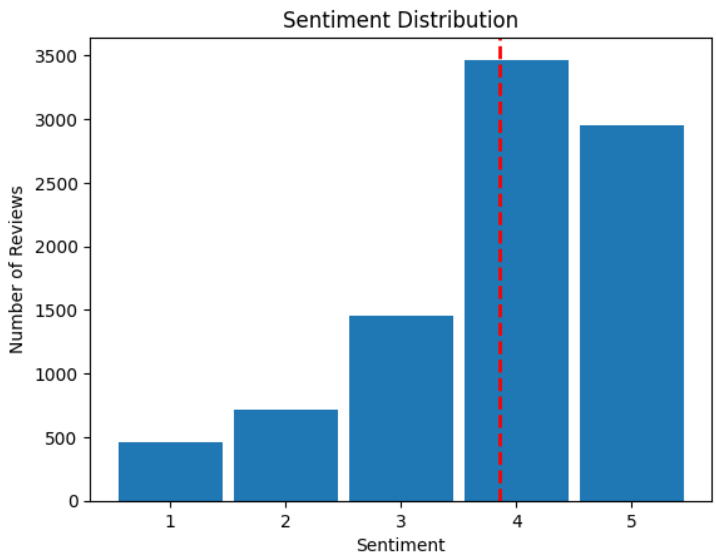

Figure 3.2: Sentiment Distribution[2]

[1]Note: The red line represents the average rating which is approximately 7.5
[2]Note: The red line represents the average sentiment score which is approximately 3.8

3.6 Feature Engineering

To proceed with sentiment analysis, it is important to define a suitable response variable. The goal is to transform the outcome variable into a binary format, which will make it easier to interpret and classify. Using BERT, it was possible to categorize sentiments as low (1,2), high (4,5), and exclude observations of neutral sentiment (3). Specifically, the response variable was set to 1 for high-sentiment reviews and 0 for low-sentiment ones. The purpose of this separation is to gain insights from extremely negative or positive experiences and identify the themes or aspects that influence a certain sentiment. The neutral category was removed because it could pose classification challenges and potentially impair the models' performance due to its redundancy. As shown in Figure 3.2, there was a significant imbalance and skew toward positive sentiments. To minimize the potential negative consequences of this imbalance, the minority class was upscaled. This prevents the model from being overly trained to classify the majority class at the expense of the negative cases, which are critical to our analysis.

Chapter 4

Methodology

This section outlines the diverse methodologies employed in this study. It begins by presenting the research framework, followed by an exploration of top2vec for topic extraction. Subsequently, the applied classification methods, namely logistic regression and logistic regression trees, will be explained. The section concludes with an explanation of the models' evaluation process.

4.1 Research Framework

This research comprises five sequential steps. Initially, reviews from 23 top-rated destinations are harvested using an API scraper on Booking.com. The collected data undergoes transformation and cleaning processes in preparation for translation. After that, non-English reviews are translated into English utilizing a deep translator. Following translation, the reviews are piped into the Top2Vec algorithm for automated processing based on algorithmic requirements, leading to the extraction of high-quality vectors for topics through n-gram tokenization. Ten key topics are identified and hot-encoded to designate each review's association with the topics. Before proceeding, BERT (Bidirectional Encoder Representations from Transformers) is employed to carry out sentiment analysis. It categorizes sentiments into five different, which will later be used to formulate the binary response variable corresponding to either high or low review valence. The reviews then undergo additional processing and tokenization to establish a document term matrix. The purpose of this matrix is to apply Singular Value Decomposition (SVD) dimensions, summarizing the general essence of the n-grams, instead of conducting a selection of n-grams manually.

The final dataset is a combination of the hot-encoded topics, SVD dimensions, cultural dimensions, the number of words per review, and the happiness index for each document. These variables are the ones used for predicting the response variable. Lastly, three predictive models are deployed, specifically logistic regression, LASSO logistic regression, and Logistic Regression trees. The outcomes of these models are then investigated to answer the research question. This framework is illustrated in Figure 4.1.

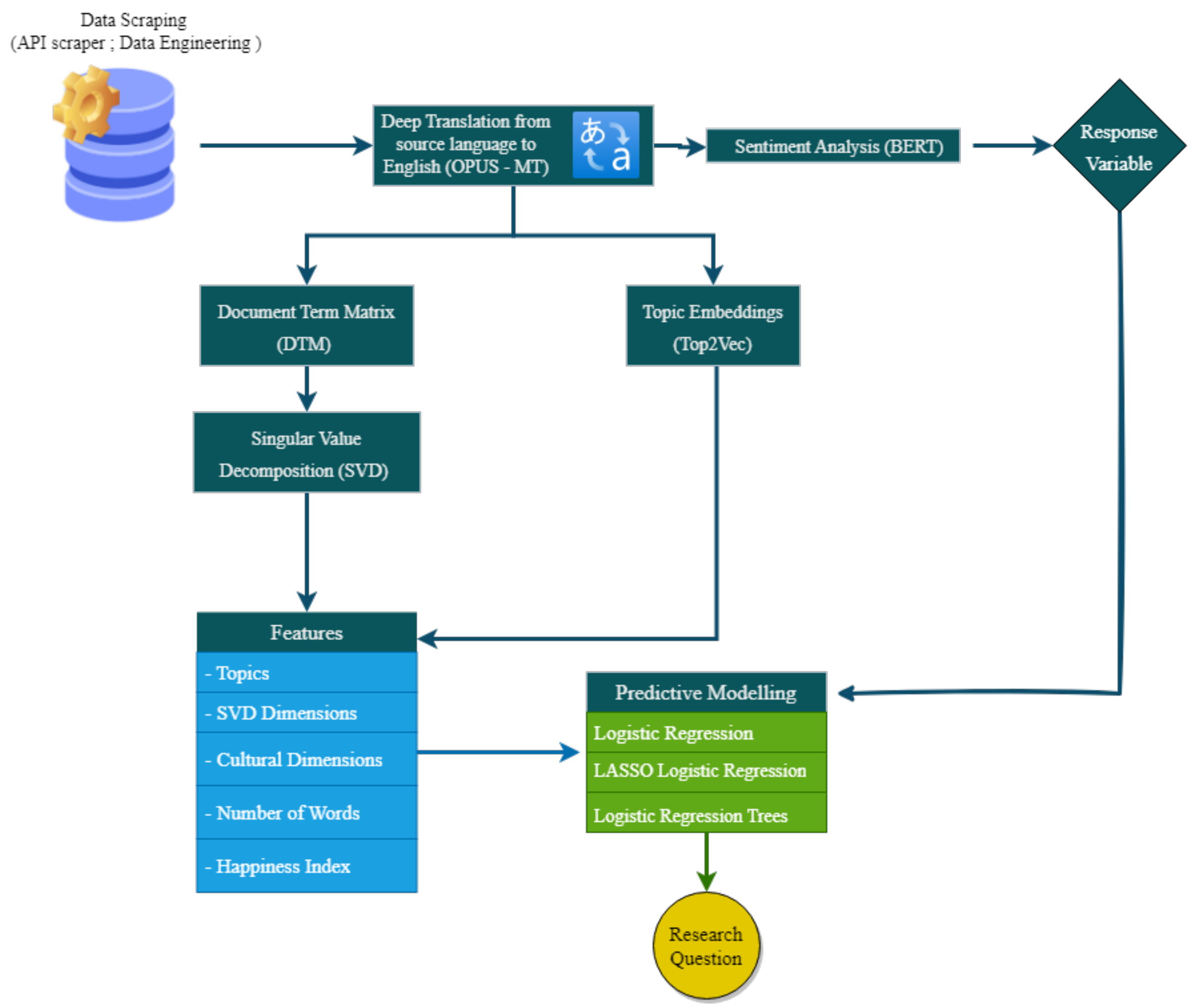

Figure 4.1: Summary of Research Framework

4.2 BERT Sentiment Classification

In this analysis, we leverage the power of the BERT (Bidirectional Encoder Representations from Transformers) model. BERT is a neural language model that was developed by Google. It is pre-trained on a large corpus of text, which allows it to understand language semantics deeply.

BERT belongs to transformer neural network architecture that was initially created to solve the problem of language translation. Compared to the older architectures that used to serve similar purposes such as Long Short Term Memory (LSTM), BERT is faster to train as words can be processed simultaneously and the context can be learned better as BERT can learn context from both directions at the same time and not learning it separately to concatenate it in the end like LSTM's did. Generally, Transformer architectures consist of encoders and decoders, but to be specific BERT consists of multiple stacked encoders thus the explanation will be emphasized on those.

Thoroughly, what these encoders do is get input tokens and generate embeddings (vector representations of words) simultaneously for each word. Similar words have closer numbers in their vectors. Before that, a certain order of words should be specified. The model employs "positional encodings" to maintain sequence order since it lacks both recurrence and convolution. These encodings are added to the input embeddings and have the same dimensionality d_{model}, permitting summation. In the study of Vaswani et al. [52], sine and cosine functions of varying frequencies were defined for the positional encodings:

$$PE_{(\text{pos},2i)} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right)$$

$$PE_{(\text{pos},2i+1)} = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right)$$

where pos is the position and i is the dimension. Each positional encoding dimension corresponds to a sinusoid with wavelengths forming a geometric progression from 2π to $10000 \cdot 2\pi$. This choice was made as it was hypothesized that it would enable the model to learn to attend by relative positions.

The architecture of the encoder entails an assembly of six ($N = 6$) identical layers. Each stratum can be dissected into two distinct components. The initial component utilizes a mechanism known as multi-head self-attention, whereas the latter constitutes a straightforward, position-wise, fully connected feed-forward network. A technique known as the residual connection is implemented circumventing these two components, followed by layer normalization [52] . This means that the output emanating from each sub-component is encapsulated by the formula $\text{LayerNorm}(x+\text{Sublayer}(x))$, where $\text{Sublayer}(x)$ represents the function executed by the sub-layer itself. In order to accommodate these residual connections, every sub-layer in the model, in addition to the embedding layers, produce outputs with a dimension of $d_{\text{model}} = 512$ [52].

A noteworthy mechanism of the transformer model is indeed the multi-head attention. It is utilized to determine the significance of each word in a sentence context. This mechanism generates query, key, and value vectors per word, utilizing multiple attention heads to calculate attention scores independently. By applying the attention function in parallel across these heads, the Transformer allows for diverse perspectives on the same input sentence. The output vectors from each head are merged and projected to yield final values. This approach enables the model to capture various types of relationships and dependencies among the words in a sentence, akin to viewing the sentence through multiple lenses simultaneously, each emphasizing a different facet [52].

The formula given in the text is the formal definition of how multi-head attention is calculated:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{4.1}$$

Where each head is computed as:

$$\text{head}_i = \text{Attention}(QW_{Qi}, KW_{Ki}, VW_{Vi}) \tag{4.2}$$

Where the projections are parameter matrices:

$$W_{Qi} \in \mathbb{R}^{d_{\text{model}} \times d_k} \tag{4.3}$$

$$W_{Ki} \in \mathbb{R}^{d_{\text{model}} \times d_k} \tag{4.4}$$

$$W_{Vi} \in \mathbb{R}^{d_{\text{model}} \times d_v} \tag{4.5}$$

$$W_O \in \mathbb{R}^{hd_v \times d_{\text{model}}} \tag{4.6}$$

BERT is capable of solving a variety of problems such as Neural Machine Translation, Question Answering, Sentiment Analysis, and Text Summarization among others. In order to accomplish these tasks, BERT requires to undergo some steps in order to gain an understanding of Language and how it works. Therefore this model is pre-trained in order to understand language and context and then fine-tune the model to learn a specific task, based on the problem needs solving which in this study is Sentiment Analysis.

The goal of pre-training BERT is to make it learn what is language and what is context. BERT learns language by training on two unsupervised tasks simultaneously. They are Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In Masked Language Modeling, BERT takes in a sentence with random words filled with masks. The goal is to output the masked tokens, which helps the model to understand the context of the sentence. For example, if the sentence is "The strong brown horse came first in the race", BERT might mask the word "brown". The model would then have to learn to predict the correct word, which in this case is "brown". This helps the model to understand the relationship between the words in the sentence and the context of the sentence.

The strong {MASK1} horse {MASK2} first in the race

In the case of next sentence prediction, BERT takes two sentences and it determines whether the second sentence actually follows the first sentence like a binary classification problem. Two example sentences can be seen below:

A: John is happy — B: He got a good grade in his thesis

For this study, the sentiment classification was executed utilizing the bert-base-multilingual-uncased-sentiment model. Incorporating 12 layers within its encoder stack, this model underwent training on a diverse set of reviews in multiple languages. It has exhibited superior predictive accuracy specifically for English reviews. As such, to optimize its performance, all reviews within our dataset were translated into English. The model's architecture is illustrated below:

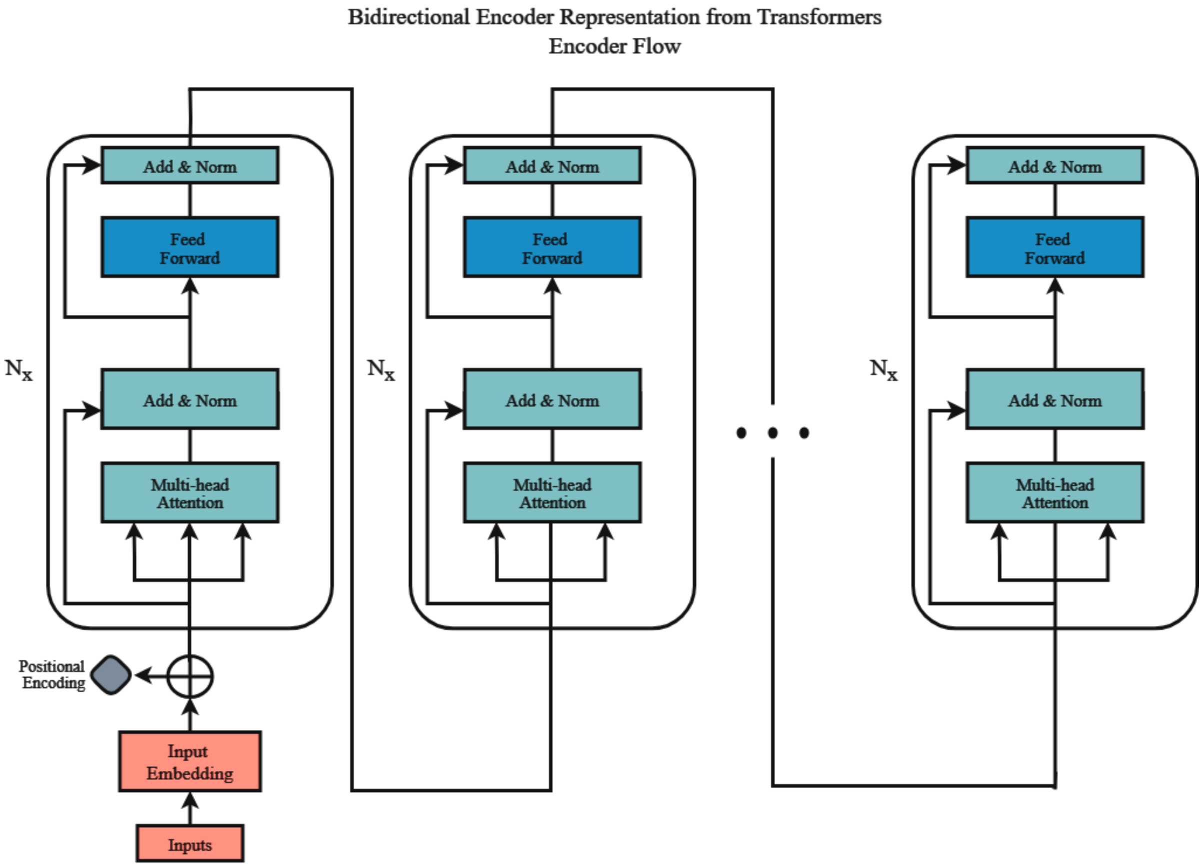

Figure 4.2: Summary of Research Framework

4.3 Top2Vec Distributed Representations of Topics

Top2Vec is a topic modeling algorithm that automatically identifies, extracts, and generates topics from text-based data. It uses pre-trained word vectors to create meaningful and embedded topics, documents, and word vectors. This gives an intuitive and comprehensive overview of the topics discussed in a dataset. The method can be summarized into 5 essential steps.

The first step is to create joint documents and word vectors using semantic embedding. The idea is that documents that are similar should be close together in the embedding space, while those that are dissimilar should be far apart [42]. This is typically done using deep learning models that have been trained on large corpora of text. These models learn to represent words and documents as vectors that capture their semantic meaning. The embedded document vectors are high-dimensional, so it is necessary to reduce their dimensionality while preserving as much variability as possible. This helps to identify dense areas, where each data point represents a document vector. Uniform Manifold Approximation and Projection (UMAP) is a commonly used technique for this purpose [42]. UMAP is effective because it can preserve the global and local structure of high-dimensional data, allowing us to effectively identify clusters in our data.

The next step is to apply Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), a density-based clustering algorithm that identifies dense areas of similar documents. Each document is assigned to a specific cluster or categorized as noise if it doesn't belong to any dense area. The algorithm looks for areas in the reduced-dimensional space with a higher density of document vectors, which symbolize semantically similar content. Top2Vec calculates centroids in the original, high-dimensional space, rather than the reduced embedding space. This is done by taking the arithmetic mean of all document vectors within a dense cluster, as identified in the previous step [42]. The result is a topic vector for each cluster which represents the core of that topic. The final step in the topic modeling process is to identify the words that are most closely related to each topic vector. This is done by calculating the cosine similarity between each topic vector and all of the word vectors in the vocabulary. The words with the highest similarity to a topic vector are considered to be the most descriptive of that topic. This process is often referred to as "topic-word mapping" [42].

Consequently, Top2Vec is a robust unsupervised method for identifying and presenting topics in textual data. It combines the richness of semantic embeddings, the efficiency of dimensionality reduction techniques, and the clarity of clustering algorithms. Top2Vec seamlessly integrates these elements into a coherent model, generating interpretable topics that provide significant insights into the underlying textual data.

4.4 Singular Value Decomposition

The concept of Singular Value Decomposition (SVD) originates from a principle in linear algebra. According to this principle, it's possible to express any given matrix A as the product of three distinct matrices, namely U, S, and the transpose of V, where both U and V are orthogonal matrices [44]. This principle is typically denoted as follows:

$$A_{m\times n} = U_{m\times m} S_{m\times n} V_{n\times n}^T$$

Subject to the conditions $UU^T = I$ and $VV^T = I$, the matrix U's columns comprise the orthonormal eigenvectors of AA^T, whereas the orthonormal eigenvectors of $A^T A$ are encompassed within the columns of matrix V. The diagonal components of the matrix S correspond to the square roots of the Eigenvalues of either the U or V matrix, arranged in decreasing order. These diagonal components found in S are recognized as Singular Values [44].

SVD is often used to identify the low-rank approximation of a specific matrix. In a situation where there are k singular values, it's feasible to recreate the matrix using only r values (with $r < n$). This method of determining the low-rank approximation essentially filters out noisy data [44]. Furthermore, the low-rank approximation achieved through SVD captures the most influential characteristics of the original data matrix. It facilitates abstracting and focusing on the most important semantic elements while excluding noise and irrelevant features. This allows for a more robust and accurate identification of sentiments in reviews [44]. Moreover, SVD assists in identifying latent or hidden structures within the data, which are crucial for sentiment analysis. For example, synonyms or words that are often used together in positive or negative reviews can form semantic structures that SVD can identify and leverage. This ability can lead to more complex and accurate sentiment analysis, detecting nuances that other methods may overlook.

In this study, Singular Value Decomposition (SVD) was utilized aiming to address the problem of high dimensionality in the sparse TF-DF matrix and feature extraction. The method itself does not provide a direct solution for associating words with specific dimensions or components. When applying Singular Value Decomposition (SVD) for text analysis, it becomes possible to link specific words to individual dimensions by integrating the SVD approach with a Term Frequency - Inverse Document Frequency (TF-IDF) representation of the text data. In detail, while performing SVD on the TF-IDF matrix, each resulting dimension corresponds to a linear combination of the original words in the constructed vocabulary.

This linear combination's weights show how much each word adds to a given dimension. By examining these weights, it is possible to identify which words are mostly associated with each dimension, therefore when it comes to SVD dimensions in the results section, the interpretation will be based on the 30 words with the highest absolute weights for each of the 10 dimensions derived. The task was performed by Algorithm 1 below:

Algorithm 1: SVD Feature Extraction for Interpretation

Result: Prints top 30 terms for each SVD dimension

Initialization:

Define the term names: $term\_names \leftarrow pruned\_vocab.term$

Define the number of dimensions: $num\_dimensions \leftarrow$ number of dimensions in $svd\_out.u$

for $dim \leftarrow 1$ **to** $num\_dimensions$ **do**

  $term\_weights \leftarrow svd\_out.v[, dim]$

  $ordered\_indices \leftarrow$ order $term\_weights$ in decreasing order

  $top\_tokens \leftarrow term\_names[ordered\_indices[1:30]]$

  print("Dimension", dim, ":")

  print(top_tokens)

end

4.5 Logistic Regression & LASSO

Logistic Regression is a statistical model that predicts the likelihood of a binary outcome. The outcome can be either 0 or 1, such as whether a customer clicks on an ad or not, or in this analysis, whether a review is positive or negative

The logistic regression model is a linear model, but the output is not a linear function of the input. Instead, the output is a sigmoid function, which is a nonlinear function that takes a real number as input and outputs a number between 0 and 1.

The equation for the logistic regression model is:

$$p(y=1|x) = \frac{1}{1+\exp(-wx)}$$

where y is the binary outcome, x is the input vector, w is the weight vector, and $p(y=1|x)$ is the probability of y being equal to 1 given x.

The weight vector w is learned from the data using maximum likelihood estimation. This means that the weight vector is chosen to maximize the probability of the observed data.

Least Absolute Shrinkage and Selection Operator (LASSO) Logistic Regression, is a variation of logistic regression that penalizes the size of the weight vector. This can be useful for reducing overfitting, which is a problem that can happen when the model learns too much from the training data and does not generalize well to new data.

The equation for LASSO logistic regression is:

$$\min_w \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j |w_j|$$

where λ is a hyperparameter that controls the amount of regularization.

4.6 Logistic Regression Trees

Logistic regression trees, as explored in this research, fuse the concepts of simplistic decision trees with logistic regression, creating a scenario where the data is initially segmented based on defined parameters, followed by the application of a logistic regression model at each node [53]. This unification of two prevalent methods yields a singular tree that can be analyzed through the lens of the logistic regression outcomes discovered in each leaf. These regression results offer quantifiable effects of predictor variables, enabling the establishment of relationships among predictor, response, and split variables.

The tree employs a procedure that recursively partitions the data based on chosen splitting variables, resulting in binary splits. Every node contains a logistic regression model designed for the given partition of data. One of the prime advantages of such a model is the easy identification of disparities in predictors based on the data partitions they operate on.

For this thesis, the logistic regression tree model will be employed to discern significant data divisions founded on cultural aspects. At every node, logistic regression models will be applied, featuring Top2Vec extracted topics as predictors and review rating as the response. This methodology facilitates the exploration of how topics vary in importance contingent on different cultural and economic indicators, paving the way for clearer connections between satisfaction factors and individual background.

Logistic regression trees is a machine learning model that combines the concepts of decision trees and logistic regression. Decision trees recursively partition the data into smaller and smaller subsets, based on the values of the predictor variables. Logistic regression, on the other hand, is a statistical model that predicts the probability of a binary outcome, such as whether a customer is satisfied or not. Logistic regression trees work by first partitioning the data into different segments based on the values of the predictor variables. Then, a logistic regression model is fit to the data in each segment. The results of the logistic regression model are used to determine which variable to use for the next split. This process is repeated until the tree reaches a stopping criterion [53]. The advantage of logistic regression trees is that they can provide both qualitative and quantitative insights into the data. The qualitative insights come from the decision tree itself, which can be used to understand the relationships between the predictor variables and the outcome variable. The quantitative insights come from the logistic regression models, which can be used to estimate the effects of the predictor variables on the outcome variable.

In this thesis, logistic regression trees will be used to explore the relationship between customer satisfaction and cultural factors. The data will be partitioned into different segments based on cultural indicators, and logistic regression models will be fit to the data at each node. The results of the models will be used to identify the topics that are most important for customer satisfaction in different cultural groups.

Logistic regression is a powerful tool for exploring key satisfaction factors. However, it does not provide a simple way to understand the relationships between these factors. Interaction effects can be modeled, but this can be computationally challenging and difficult to interpret when there are many predictor variables. Additionally, interaction effects may not accurately represent the true differences in consumer preferences between cohorts, as the model would be run on the entire dataset and not on its segments.

The data is recursively partitioned by executing a parameter instability test. If any splitting variables indicate instability, the node will be bifurcated [54]. Zeileis et al. [54] presented the recursive partitioning algorithm consisting of the following steps:

1. The model is fit using all observations in the node, and the parameter set β is estimated by minimizing the respective objective function, which in this case is the negative log-likelihood.

$$\min_{\beta} \sum_{i=1}^{n} \log(1 + \exp(-y_i \beta^T x_i))$$

2. Parameter estimates are evaluated by running the fluctuation test to detect any instability.

$$\lambda_{supLM}(W_j) = \max_{i=1,\dots,n} \left(\frac{i}{n} \cdot \frac{n-i}{n} \right)^{-1} ||W_j \left(\frac{i}{n} \right) ||^2$$

where i stands for a given observation, n represents the number of observations in the node, and W_j is the partial sum process of the scores derived from the model being run.

3. If instability is identified, the parameter with the highest instability is selected as a splitting variable.

4. The split point is then calculated to ensure that the objective function pertaining to the model is optimized.

The logistic regression tree model is implemented in the R programming language using the 'partykit' package [55] [54]. The parameters of the model can be tuned to ensure that the model is not overfitting the data. The minsplit parameter corresponds to the minimum number of observations in a node to be considered for a split. The maxdepth parameter regulates how many layers the tree may grow. The results of the model can be plotted, and the splits along with the consequent logit models can be analyzed per node to detect any differences amongst cohorts.

4.7 Model Evaluation

To evaluate the logistic regression and tree models, we will use four measures to select the best model: accuracy, sensitivity, specificity, and AIC. Accuracy is a measure of how well our model classifies predictions. It is defined as the number of correct predictions divided by the total number of predictions.

Accuracy is a measure of how well a model predicts the correct class for a given observation. It is calculated as the number of correct predictions divided by the total number of predictions.

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

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Since our data is unbalanced, relying solely on Accuracy may be misleading. Sensitivity and specificity offer insights into individual class predictions. When data skews heavily towards one class, a high-accuracy model might still lack sensitivity, thus examining both sensitivity and specificity gives a holistic view of model performance on such datasets.

Sensitivity is a measure of how well a model predicts the positive class for a given observation. It is calculated as the number of true positives divided by the total number of positive observations.

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

Specificity is a measure of how well a model predicts the negative class for a given observation. It is calculated as the number of true negatives divided by the total number of negative observations.

$$Specificity = \frac{TN}{TN + FP}$$

Akaike Information Criterion (AIC) is a measure of model fit that penalizes models for their complexity. A lower AIC value indicates a better-fitting model. AIC is calculated as follows:

$$AIC = 2k - 2ln(L)$$

where k is the number of parameters in the model and L is the likelihood of the model.

The likelihood of the model is a measure of how well the model fits the data and it is calculated as the probability of the observed data given the model. The AIC penalizes models for their complexity by adding a term to the likelihood that is proportional to the number of parameters in the model. Simply put, this means that models with fewer parameters will have lower AIC values compared to models with more parameters. In general, models with lower AIC values are preferred because they are more likely to have a good fit for the data.

Chapter 5

Results

This chapter presents the results derived from applying various models. We begin by examining the topic extraction from the hotel reviews. After that, we provide a rationale for the selection of the final models. Lastly, the interpretation of the models takes place. The objective is to understand the relationships between topics, features, and cultural dimensions with the response variable. This analysis will contribute to gaining insights into consumer behavior and the significance of these associations.

5.1 Topic Modelling

This study employed the Top2Vec algorithm with the goal of extracting topics from hotel reviews. The choice of this specific algorithm was motivated by several advantages it offers. Notably, its ability to automate necessary pre-processing steps distinguishes it from traditional topic modeling methods like LDA, which require extensive preprocessing, including lowercasing, tokenization, removal of stopwords, punctuation, numbers, and rare or infrequent terms. The Top2Vec algorithm not only generates high-quality vectors known for their accuracy but also incorporates an in-depth exploration of the relationship between documents and their corresponding topics. Moreover, it provides the functionality to search documents and topics based on a keyword list or a query, and can even identify similar words and documents. The Top2Vec model operates with a set of parameters. The first, labeled as "documents", determines the input corpus, which is expected to be a list of strings.

The "speed" parameter controls the model's training pace. For our analysis, we chose the "deep-learn" option, which, despite requiring a significant amount of training time, produces the highest-quality vectors. Given that the number of reviews in this study was relatively small, the extended training time was not considered an issue. The parameter "workers", determines the number of threads used in training the model. Larger numbers typically result in faster training. However, in our case, four workers were chosen.

The execution of the model yielded a total of 175 topics. As mentioned before, Top2Vec allows hierarchical topic reduction. With its application, the 175 were reduced to 10. The words clouds from the extracted topics are illustrated in Figure 5.1 :

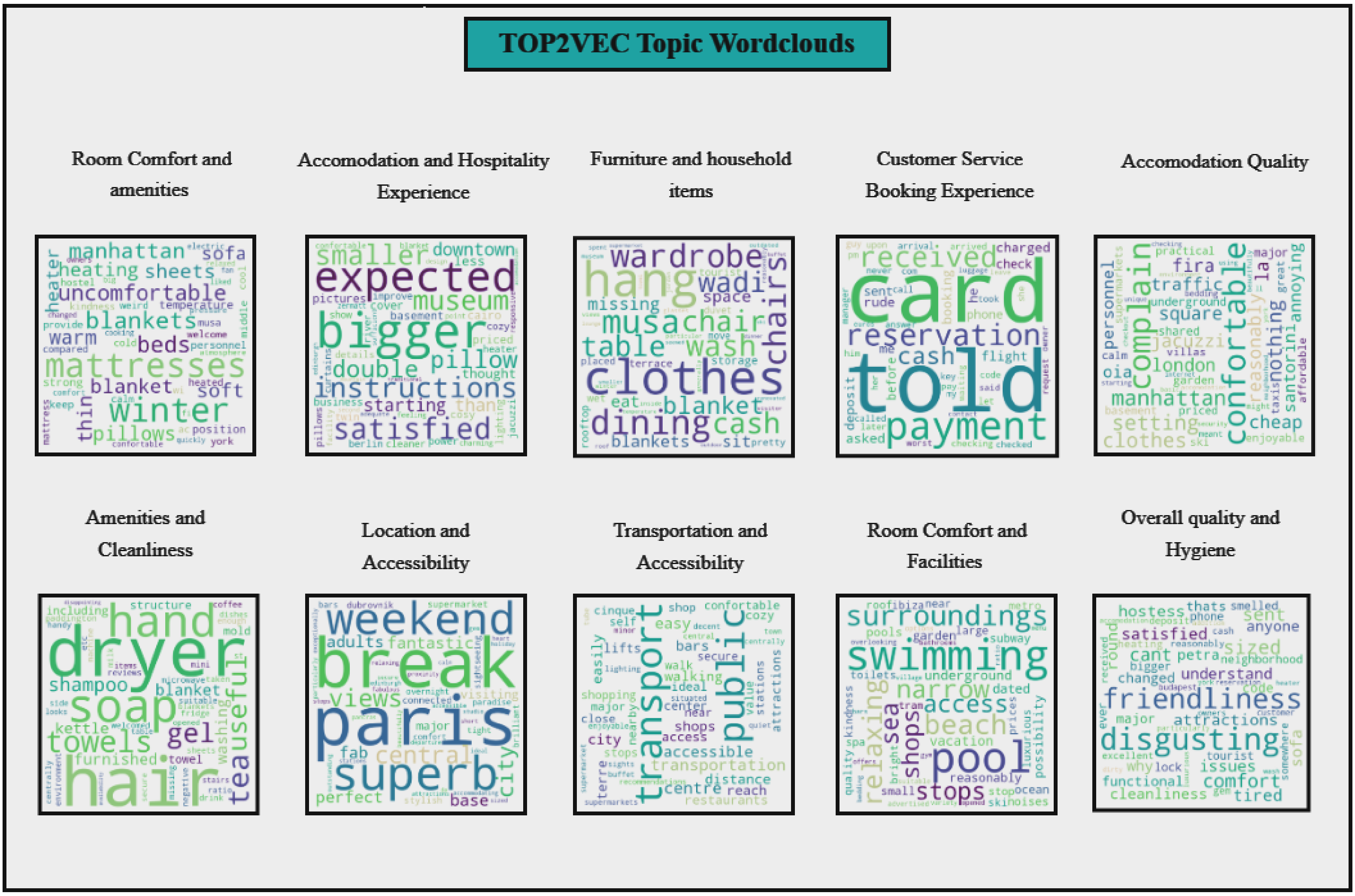

Figure 5.1: Top2Vec Topic Wordclouds

5.2 Predictive Modelling

After extracting and associating topics with each document, we employed a logistic regression model to classify reviews based on their rating (high or low). This process allowed us to gain an overview of the associations between independent variables and the response variable. Next, we incorporated Hofstede's cultural dimensions by combining these with the data frame so that the corresponding scores were matched with the nationality of each reviewer. This step was performed with the objective of evaluating the influence of culture on consumer satisfaction in the hospitality industry. After that, in order to integrate the content of the reviews and assess their impact on review valence, we created a TF-IDF matrix. This was done using three different tokenization methods: unigrams, bigrams, 1-skip-1 unigrams, and 1-skip-1 bigrams. Singular Value Decomposition (SVD) was then applied to reduce the dimensionality of the resulting large sparse matrix. The final model selection was based on overall accuracy and the Akaike Information Criterion (AIC) score. Importantly, we also addressed the issue of imbalanced classes in the review valence, where low valence was the minority class, by applying upsampling to the dataset. This assists the model in learning both classes equally well. The table below provides a comparison of the various tokenization methods utilized.

Metric	Unigrams	Bigrams	1-skip-1:Unigram	1-skip-1:Bigram
Accuracy	88.65%	70.58%	85.35%	68.67%
Sensitivity	89.15%	62.06%	83.24%	65.08%
Specificity	88.14%	78.41%	87.45%	72.25 %
AIC Score	6155.61	9038.095	7129.40	11988.04

Table 5.1: Comparison of different tokenization methods for the Logistic model.

Note: The values indicate the performance metrics obtained using different tokenization methods. Unigrams and Bigrams refer to the type of n-gram used, while 1-skip-1:Unigram and 1-skip-1:Bigram refer to the n-gram model with a skip factor of 1. Accuracy, Sensitivity, and Specificity are measured in percentages, while the AIC Score is a relative quantity.

The comparison of the performance of the simplistic model using various tokenization methods signified that the unigram method proved to be the most efficient, providing the highest accuracy and the lowest AIC score. Including bigrams in both forms (ordinary and skip-grams) seemed to considerably decrease the overall performance and substantially increase the AIC Score, hence for the final models in this analysis unigram tokenization was employed. The results presented are derived from the application of repeated 10-fold cross-validation. This process enhances the stability and dependability of model performance estimates by taking the average of the results over numerous different data splits.

5.3 Interpretation of Logistic Regression

Starting with the simple logistic regression, the variable IDV which corresponds to the Individualism dimension appears with a coefficient of -0.2182 meaning that as the level of Individualism is increased by one standard deviation the log odds of observing the high valence rating class decreases by approximately 0.21, a result at a 0.1% significance level, ceteris paribus. As mentioned before, individualism represents a societal structure in which individuals focus primarily on their personal needs and those of their immediate family. This is in contrast to collectivism, where individuals advocate for strong group cohesion and offer their loyalty in return. It can thus be inferred that individualists, with potentially higher standards for products or services, might also have heightened quality expectations within the hospitality industry. Consequently, a negative correlation is anticipated between individualism and review valence. As the degree of individualism escalates, the likelihood of encountering highly-rated reviews diminishes. This is attributed to the increased challenge of satisfying the expectations of reviewers coming from highly individualistic cultures. Specifically, the United Kingdom stands out among the countries examined with the highest Individualism score of 89. Although the United Kingdom is no longer a part of Europe, its departure from the EU exemplifies its strong individualistic culture, as it is the only sovereign country to have left the EU. Some other examples of countries with high Individualism scores include The Netherlands and Hungary, both scoring 80, making them the leading individualistic societies in Europe. This finding confirms previous studies in the cross-cultural research field regarding review rating, which supported that people from individualistic societies are tending to give lower ratings compared to those reviewers who are from collectivistic societies [56].

Moreover, another cultural dimension that appeared significant in predicting the review valence was that of the Long Term Orientation (LTO). In detail, LTO with a coefficient of approximately -0.09 indicates that as LTO is increased by one standard deviation the log odds of observing a high review valence is decreased by 0.09,ceteris paribus, at a 1% significance level. Interestingly, this finding contradicts earlier research which proposed that individuals from cultures oriented towards long-term relationships tend to avoid giving negative feedback, as they aim to preserve the established relationship with the product or service provider [57]. Nevertheless, there were also studies such as those of Stamolampros et al. [58] which showcased negative relationships between LTO and a review's valence. Undeniably, long-term-oriented cultures are observed as pragmatic because individuals are seeking to promote or pinpoint possible improvements for a better future. In the context of this analysis, a more negative review valence might actually signal a variety of issues that a reviewer hopes to see addressed in the future when it comes to long-term-oriented cultures. Some examples of this study are Ukraine and Germany scoring 86 and 83 on this dimension, respectively.

When examining topics, the one pertaining to furniture and household items was found to be statistically significant at a 1% level. Accounting for the category of reference (which corresponds to overall quality and hygiene as the most neutral topic) the coefficient of 0.34 suggests that if a review falls under the topic of furniture and household items, the log odds of observing a high review sentiment increases by an additional 0.34 points, beyond what it would if the review belonged to the baseline topic of overall quality and hygiene. As a result, it is understood that in the evaluations studied, reviewers who mentioned furniture and home objects were more satisfied, meaning that these items were present and functional and lived up to the guest's standards. Additionally, it signifies that in the context of living conditions and surroundings, reviews were associated with more positive sentiments.

Another topic that appeared to be important was that of the customer service and booking experience. Indicatively, its coefficient of -1.17 indicates that if a review belongs to this topic, the log odds of observing a high sentiment is 1.17 less when compared to the baseline category, ceteris paribus. This topic, generally covers the overall customer journey from the booking process, arrival, stay, and departure in a hotel. It mainly covers negative interactions and therefore it is considered reasonable that this topic influences negatively a review's valence. It is comprehended that the customer service and the booking experience referring to the procedures prior to the actual accommodation are considered crucial determinants of a review's valence. The booking and reservation process is a critical touchpoint in the overall customer journey in the hospitality industry. It corresponds to the initial interaction that customers have with a hotel, setting an expectation for the entire stay. Negative experiences in these areas can significantly undermine a hotel's reputation, as reviewers are likely to highlight these issues in their feedback. In this examination, they were primarily stated in a negative context.

In addition, the topics of Accessibility and Location, and Accessibility and Transportation were the ones that portrayed significant relationships with the response variable. Thoroughly, the coefficients of 0.94 and 0.50 are indicating that reviews in these categories have 0.94 and 0.50 higher odds of being associated with a high review valence compared to the baseline topic, ceteris paribus. These results hold statistical significance with a probability of 0% and 0.1% performing a type I error, respectively. Overall, location and transportation, both play vital roles in yielding a positive review in various ways. One of them could be in the context of convenience and accessibility, as in general, the easier it is for guests to reach the accommodation and navigate through the surrounding area, the more positive their experience will probably be. This is particularly important for visitors interested in discovering local sights, as they would greatly appreciate a hotel located in close proximity to attractions. Drawing comparisons with past research, Calheiros [50] through textual analysis of reviews from an eco-hotel, revealed that the topic related to location was associated with very positive sentiments, confirming the results derived in this study.

Lastly, an interesting observation from the data analysis reveals that all the Singular Value Decomposition (SVD) dimensions derived from the document term frequency matrix were statistically significant, apart from the ninth dimension. For clarity and conciseness, the interpretation will focus on the dimensions with the greatest, positive or negative, influence on certain sentiment classes.

Upon examining aspects that appear to decrease the likelihood of predicting a high review valence, the third and fifth dimensions seem to have the strongest negative correlation. Focusing specifically on the third dimension, it incorporates words such as "fair", "bad", "everything", "nothing", "bathroom", "terrible", "slept", "dirty", "poor", "size", "towels", "conditions", "smoke", "never", "noise", "toilet", "cleanliness", and "awful" among others. The large coefficient of this variable signifies that a review containing these words is likely to be categorized as a sentiment of 1 or 2 out of 5. The third dimension indicates that shortcomings in meeting customer expectations concerning overall hygiene and cleanliness can lead to negative reviews. It features words like "fair", "bad", "terrible", "dirty", "poor", and "awful", which mainly represent a negative sentiment, while "fair" suggests a somewhat neutral sentiment. Additionally, terms such as "slept", "smoke", and "noise" may imply issues with sleep quality, a factor of high importance to guests. Notably, the presence of absolute terms like "nothing" and "everything" seems to encapsulate the overall guest experience and could reflect extreme sentiment states such as 'nothing was good' or 'everything was terrible'.

An examination of the 5th dimension, which showed the most significant negative impact on review sentiment, unraveled an association with words such as "scam", "pictures", "broken", "booking", "info", "lack", "money", "exchange", "nothing", "work", and "freezing", among others. These words suggest a prevalent sense of disappointment linked to the booking experience and the quality of service provided at reception. Furthermore, it appears that expectations created by the pictures displayed on the booking website did not align with the actual condition of the rooms, leading to further disappointment. The words "freezing" and "broken" may hint at issues with room heating, indicating potential concerns with the physical comfort of the guests. The presence of these words, collectively, contributes to the negative sentiment in the reviews, suggesting areas in need of improvement to enhance guest satisfaction. Subroto and Christianis' [49] findings, which showed a negative correlation between review ratings and words like "dirty," "toilet,", "bad", "never", and "broken," are verified by these results. This study, therefore, broadens the spectrum of potential words used in negative reviews.

In analyzing factors that enhance the probability of a higher sentiment, it was found that the first and sixth dimensions demonstrated the strongest positive correlations. Words such as "good," "exceptional," "lovely," "location," "staff," "clean," "friendly," "central," "breakfast," "quiet," "price," "value," and "view" were all associated with the second dimension. Similar terms were found in the fourth dimension, along with others like "host," "kind," "personal," "loved," "beach," and "studio."

Upon comparing these observations with those linked to negative reviews, a clear pattern emerges. Indeed, customer service, hygiene, and staff demeanor were found to be vital determinants of review ratings. The frequency of words such as "location" and "central" indicates that guests favor accommodations with convenient access to sights and transport, which facilitates better maneuverability during their stay. The mention of "breakfast" is also noteworthy, as it appears to be heavily discussed among guests with higher sentiment scores. Finally, the mention of "beach" could suggest a preference for hotels located near the seaside, probably for those seeking summer accommodations. This observation further strengthens the results of our study. Table 5.1 shows the summary of the model.

Coefficient	Estimate	Std. Error	z value	P-value
(Intercept)	-0.4614	0.2063	-2.237	0.0253*
Number of Words	0.0404	0.0349	1.157	0.2471
Power Distance	0.0063	0.0604	0.105	0.9165
Indulgence	-0.1296	0.0687	-1.885	0.1593
Masculinity	-0.0032	0.0392	-0.082	0.93453
Uncertainty Avoidance	0.0195	0.0665	0.293	0.76916
Long-Term Orientation	-0.0954	0.0379	-2.782	0.0241*
Individualism	-0.2182	0.0687	-2.782	0.0054**
Happiness Index	0.0631	0.0533	1.184	0.2363
Room Comfort & Amenities	0.1677	0.1652	1.016	0.3097
Accommodation & Hospitality_Exp	0.0398	0.1494	0.267	0.7895
Furniture & Household Items	0.3482	0.1496	2.195	0.0281*
Customer Service & Booking Experience	-1.1772	0.1976	-5.370	7.86e-08***
Accommodation & Quality	0.1039	0.1391	0.747	0.45512
Cleanliness & Amenities	-0.1709	0.1452	-1.177	0.2392
Accessibility & Location	0.9495	0.1798	5.280	1.29e-07***
Accessibility & Transportation	0.5024	0.1858	2.703	0.0068 **
Room Comfort & Facilities	0.0102	0.0102	0.064	0.9491
Overall Quality & Hygiene	-	-	-	-
Dim1	241.6	37.66	-6.414	1.42e-10***
Dim2	-610.5	1112	5.489	4.05e-08***
Dim3	-484.1	20.70	-23.387	<2e-16***
Dim4	-356.7	11.49	31.041	<2e-16***
Dim5	-878.7	35.87	-24.494	<2e-16***
Dim6	123.0	25.49	4.825	1.40e-06***
Dim7	29.15	15.95	1.828	<2e-16***
Dim8	-118.9	7.54	-15.754	<2e-16***
Dim9	17.15	11.85	1.732	0.0676
Dim10	-172.4	8.63	-19.963	<2e-16***

Table 5.2: Logistic Regression Model Summary

*Note: Statistical Significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.*
Due to the appearance of different scales, the data was standardized prior to the analysis. Associations between the independent variables and the response are interpreted as a one-standard deviation change. 'Dim' variables are referring to the 10 Singular Value Decomposition dimensions extracted from the TF-IDF matrix.

5.4 Interpretation of Logistic Regression with LASSO penalty

After getting an initial understanding of the application of a simple logistic regression, a LASSO Logistic model was employed. Under the constraints of the regularization parameter, the objective was to identify those factors that contribute the most to estimating the response variable. The simpler model made it easy to investigate for effects based on statistical significance. In LASSO, the variables that have not been shrunk to zero are the ones that contribute the most in terms of prediction, under the constraints of the regularization parameter, and will thus be the ones presented. Variables

In terms of cultural dimensions, the only coefficients that shrunk to zero were those of the Masculinity and Uncertainty Avoidance dimensions. Individualism and Long Term Orientation showed similar relationships with the response variable as in the simpler model, though their coefficients were smaller due to the penalty effect. Power Distance and Indulgence emerged as crucial dimensions for the model's predictive capabilities, both exhibiting a negative relationship with the response variable. In detail, the result indicated that as Power Distance increases by one standard deviation, the log odds of observing a high sentiment are decreasing by approximately 0.029, ceteris paribus, taking into account the regularization effect. Similarly, an increase of one standard deviation in the dimension of Indulgence decreases the log odds of observing a high sentiment by approximately 0.028.

Contrasting with prior research, Furrer et al. [59] found that customers from high power distance cultures are more tolerant of failures from powerful service providers, a discovery that diverges from the outcomes of this study. That differentiation may be due to the disparity in services examined. Furrer et al. [59] probed the links between culture and service quality in retail banking—a field vastly different and considered of higher status than the hospitality industry—potentially explaining the discrepancy in findings. A similar study examining the hospitality industry by Ramona Diana Leon [33] verified the findings of this study by pointing out that customers from cultures with high power distance and those from indulgent cultures tend to diverge from previous average ratings, suggesting a negative link between the rating and the power distance dimension. Another study conducted aiming to examine the influence of power distance in hotel reviews, also found that reviewers from high power distance cultures, typically assign lower ratings [60].

Considering topics and SVD dimensions, the associations paralleled those in the simple logistic model, and hence, will not be elaborated further.

Coefficient	**Value**
(Intercept)	-0.4827
Number of Words	.
Power Distance	- 0.0295
Indulgence	- 0.0284
Masculinity	.
Uncertainty Avoidance	.
Long-Term Orientation	-0.0353
Individualism	-0.0118
Happiness Index	.
Room Comfort & Amenities	0.0058
Accommodation & Hospitality Experience	-0.074
Furniture & Household Items	0.2075
Customer Service & Booking Experience	-1.4195
Accommodation & Quality	.
Cleanliness & Amenities	-0.4878
Accessibility & Location	-0.7286
Accessibility & Transportation	-0.2236
Room Comfort & Facilities	.
Overall & Quality & Hygiene	-0.1382
Dim1	249.35
Dim2	-17.997
Dim3	-286.82
Dim4	-133.01
Dim5	-479.21
Dim6	50.42
Dim7	117.92
Dim8	228.08
Dim9	.
Dim10	82.61

Table 5.3: Lasso Regression Results

Note: The symbol '.' represents coefficients that were reduced to zero by the Lasso regression.

5.5 Interpretation of Logistic Regression Trees

Aiming to expand the understanding of how cultural dimensions are influencing customer satisfaction in hotel reviews, a logistic regression tree was employed after performing the analysis with two more straightforward methods. The Logistic Regression tree informs about the interactions between the splitting variables and the regressors toward the prediction of the response variable. In detail, topics, SVD dimensions, the count of words in a review, and the Happiness Index were used as regressors, while the cultural dimensions served as the partitioning variables. The application of this model was considered necessary as more straightforward methods, despite suggesting potential causal links between cultural dimensions and customer satisfaction, did not clearly indicate how these dimensions interact with specific topics or words within a review.

Upon plotting the logistic regression tree, it is perceived that the cultural dimension of Masculinity vs Femininity performs the main split. This specifically refers to a split based on whether the Masculinity dimension has a value greater than 43 or less than 43. This means that the primary division relies on cases leaning towards a moderately masculine culture or moving towards a more feminine culture. Prior studies showed that the perceptions of service quality seem to differ significantly between masculine and feminine-oriented cultures as individuals from highly masculine cultures are keener to provide feedback compared to those from feminine cultures [61]. Moreover, it is more probable for highly masculine individuals to provide negative complaints regarding poor service quality, as they have higher expectations and are less tolerant than those of feminine cultures [62]. Additionally, regarding the overall evaluation, masculine cultures are rating lower [63]. Thus, given that masculine cultures are seen as more critical, this specific distinction was anticipated and confirmed by previous analyses.

Following the split of the root node, the first end node that yielded statistically significant results represents a profile of reviewers who come from cultures that are perceived as feminine ($MAS <= 43$), exhibit a moderate to strong tendency to avoid uncertainty ($UNAV <= 63$), lean moderately to strongly towards long-term orientation ($LTO > 38$), and exhibit characteristics ranging from restraint to moderate indulgence ($IND <= 68$). The results indicated that the first (Room Comfort and amenities) and fourth (Customer Service Booking Experience) topics, were negatively associated with the response variable, decreasing the log odds of observing a high review valence. These results can be interpreted by relating a certain country of this analysis with the aforementioned profile.

Intriguingly, The Netherlands aligns well with this particular profile, recording a score of 14 for Masculinity, 53 for Uncertainty Avoidance, 67 for Long Term Orientation, and 68 for Indulgence. This demonstrates that The Netherlands has a modest inclination towards avoiding uncertainty. These cultures have a relatively low tolerance for unconventional matters and an inherent desire to abide by the rules and maintain order. Moreover, it is observed that the Dutch culture prioritizes indulgence over restraint, indicating the significance they place on leisure time. Connecting these aspects with the analysis, it is understood that the comfort of the room and amenities are things that conduce to comfort, convenience, or enjoyment, hence they are considered important for cultures that value indulgence.

Furthermore, the significance of customer service and booking experience reflects the urge from such cultures to have a decent amount of control and certainty. A plausible explanation could be that individuals from such cultures desire a hassle-free booking experience devoid of misunderstandings or unpleasant surprises, coupled with swift and effective customer service. Consequently, the data examined demonstrated a tendency towards negative feedback from such types of cultures, indicating that their expectations about accommodation comfort, facilities, booking experience, and customer service, were not reached.

In addition, the analysis revealed that with each additional word in a review for the specific profile examined, the likelihood of a high sentiment rating decreases by 0.008, assuming all other variables remain constant. This suggests that cultures falling into this category tend to be more expressive about unsatisfactory experiences. The culture profile in question tends to exhibit moderate to low uncertainty avoidance, partially aligning with Litvin's [64] findings that individuals from low uncertainty avoidance cultures write longer reviews than those from high UAI cultures, by examining TripAdvisor hotel reviews. Hofstede [31] further supports this, supporting that low-UAI cultures view "time as free," while high-UAI cultures see "time as money." However, while the Netherlands fits the profile identified by the tree model, it doesn't imply the results specifically pertain to this country. Therefore, while Litvin's [64] study indicated a relationship based only on low uncertainty avoidance cultures, our results show a slightly broader range, including cultures that exhibit moderate to low levels of uncertainty avoidance, thus differing from prior research findings.

Another interesting situation was the categorization of cultures that are slight to very feminine and score between 64 and 88 on the Uncertainty Avoidance scale. One country that resonates with these qualities is France which scores 43 on Masculinity and 86 on Uncertainty Avoidance. In a comprehensive analysis, it can be observed that the French lean towards situations that offer certainty and are discomforted by ambiguity or the unknown. This cultural predilection towards certainty is predominantly driven by high levels of anxiety, therefore society has developed coping mechanisms to manage such feelings. Uncertainty is more pronounced in the professional environment, and this is where the cultural inclination towards femininity serves as a balancing factor, driving their society towards prioritizing quality of life. This can be evidenced by the significantly fewer weekly working hours and more annual vacation days when compared to other European nations.

Individuals with a combination of slight femininity and high uncertainty avoidance might prefer well-designed rooms with a warm, relaxing, and inviting ambiance, which relates highly to aesthetics and contributes to the overall quality of their stay. They may also value well-equipped facilities that support their interests and demands throughout their stay, such as a fully supplied kitchen for making meals or an adequate fitness facility for keeping up with their workout habits. Aligning with these cultural traits, the results of the analysis show that there's an inverse relationship between the 10th topic, which concerns room comfort and amenities, and the review sentiment, highlighting the importance of these aspects.

Next, a distinct pattern emerged for cultures that are characterized as being somewhat feminine to moderately masculine (with Masculinity scores between 44 and 64) and having a predisposition towards Indulgence ($IND > 50$). This particular cultural profile is placing a high value on the overall hospitality experience provided by an accommodation, as the second and the fifth topic related to these qualities appeared statistically significant. Once again, these constituents were negatively associated with the review valence.

One country matching this description is Belgium. It appears that people from such cultures value the evaluate their hospitality experiences holistically. For Belgians, the accommodation experience includes more than simply the physical amenities. It also includes the quality of service, the mood, and the general friendly attitude of the hotel. Additionally, considering the SVD dimensions, there was exhibited a strong inverse association between the 6th dimension and the response variable. The 6th dimension corresponds to a range of words that cover the spectrum of accommodation holistically, including words such as "bedsheets", "breakfast", "value", "friendly", and "noisy" to name a few. This particular dimension incorporates a variety of characteristics that relate to the comprehensive quality of hospitality and accommodation, thereby confirming that every small detail is crucial for this distinct group of customers. Consequently, in order to satisfy this group, hospitality service providers must emphasize not just physical comfort and conveniences, but also intangible aspects of hospitality such as warm, friendly service and a pleasant environment as this certain profile appears to be the most demanding one.

Also noteworthy was the scenario involving Masculine cultures that score 65 to 66 on Masculinity. Cultures representative of these scores are typically viewed as masculine, characterized by assertiveness and decisiveness, and often derive their status from material possessions. One example is Germany and the United Kingdom which score 66 on Masculinity. Within these particular cultures, it seems that practical matters hold a significant weight in the perception of service and product quality. The analysis revealed that topics such as the customer service and booking experience, as well as the ease of accessibility and location of the accommodation, played a crucial role in forming the sentiments reflected in the reviews. One might argue that the nature of these aspects aligns well with the pragmatic orientation of these masculine cultures.

Indicatively, customer service and booking experience could potentially mirror the assertive and decisive nature of these cultures. A smooth and efficient booking process, in synergy with responsive and competent customer service, are elements that directly cater to the preference for decisiveness and efficiency. Moreover, the location and accessibility of accommodations could relate to their Utilitarian viewpoint. Proximity to important sites, ease of transportation, and practicality of location can significantly contribute to an overall positive experience, matching their desire for efficiency and utility. This observation is also linked with the perception that individuals from more masculine cultures have a stronger motivation to provide feedback, particularly in relation to practical elements of their experience. Therefore, it is not surprising that these cultures appear to place a higher emphasis on these areas when it comes to the hospitality industry. This suggests that providers catering to such cultures should prioritize the efficiency and practicality of their offerings to meet the expectations of this demographic.

Similarly, there was another distinctive cultural profile for societies that fall within the 67 to 79 range on the Masculinity scale. Switzerland and Italy, countries falling in the 67-79 bracket on the Masculinity dimension, illuminate another intriguing dimension of cultural influences on review tendencies. The review behavior of these cultures emphasizes the importance of customer service, the booking process, and accessibility & transportation. This may be due to the high premium these masculine societies place on efficiency and pragmatism. In such societies, a successful hospitality experience extends beyond luxury and comfort to include practical convenience and seamless transactions. Their perception of the booking process's efficiency can be a determinant of overall service quality, with mishaps potentially leading to negative reviews. Moreover, the location's accessibility and transportation convenience can significantly influence service perception, particularly for goal-driven visitors who value their time.

It's worth noting that for the final two profiles associated with high masculinity, a strong negative correlation was observed with the 2nd and 4th SVD dimensions. The second dimension is unique in that it's related to highly negative and critical reviews. Meanwhile, the fourth dimension incorporates words about value and money and is the only dimension connected to the term "work". Figure 5.2 below provides examples of words related to the two dimensions mentioned earlier:

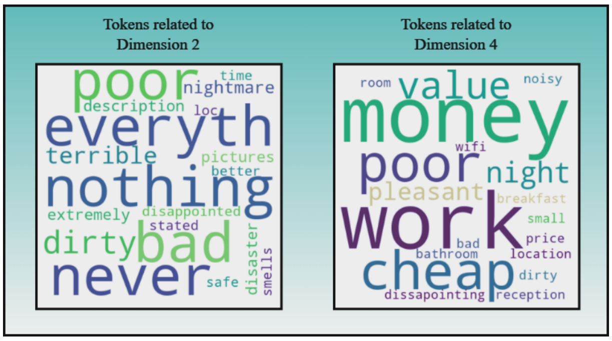

Figure 5.2: Wordclouds for Dimensions 2 and 4.

Note: The words that appeared in the word clouds were the most frequently used words for the two dimensions. The words were ranked by their weight, and the top 20 words were used in the word clouds.

As can be seen, the word clouds predominantly present terms related to a more pragmatic culture (Masculinity), emphasizing material value over interpersonal connections. The second dimension concentrates on negative experiences articulated in a very critical manner, while the fourth dimension revolves largely around value, money, and work.

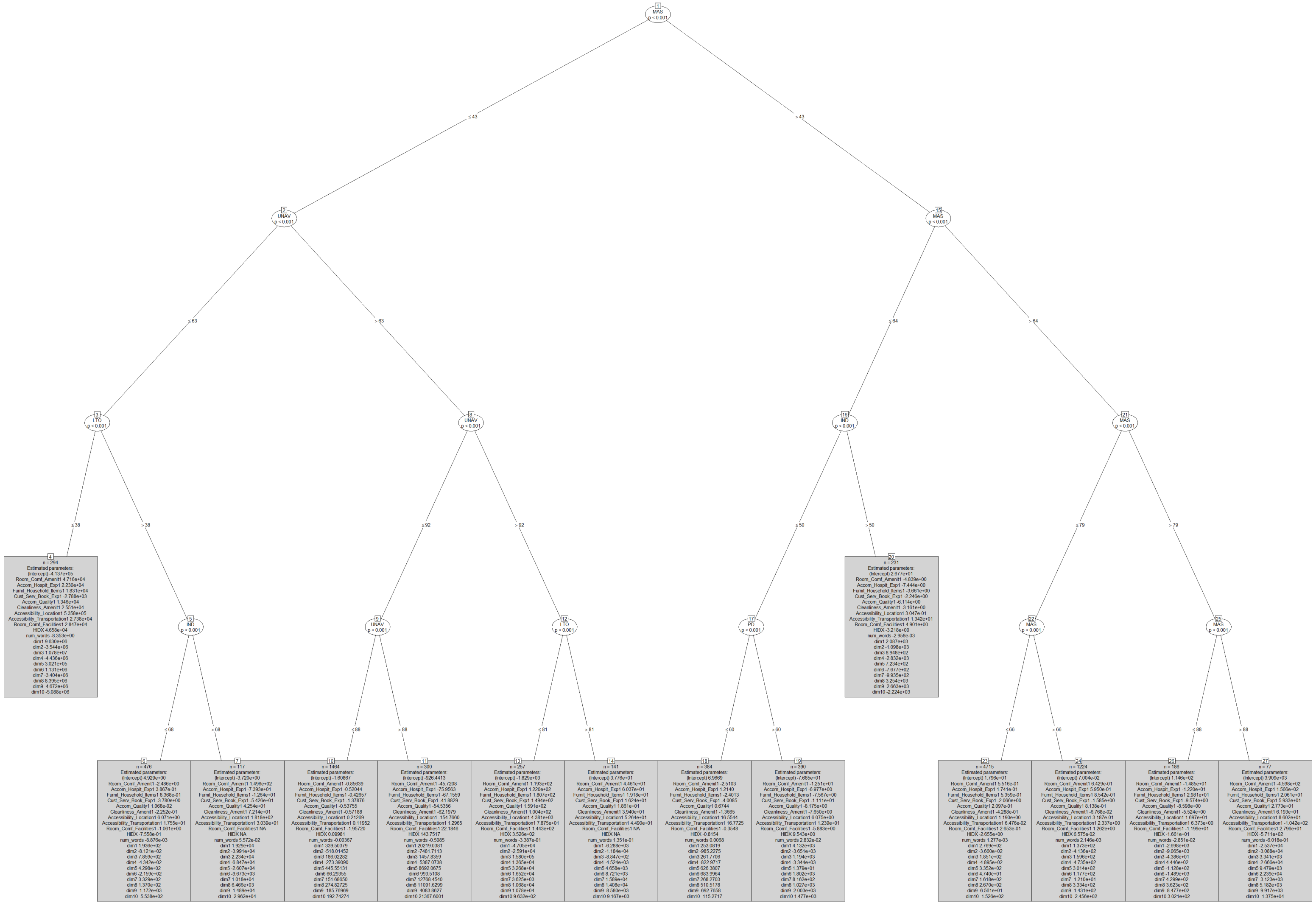

Figure 5.3: Logistic Regression Tree

5.6 Comparison of Model Performances

Regarding LASSO regression's hyperparameters, a repeated 10-fold cross-validation approach determined the optimal lambda value as 0.001, under the condition that the alpha parameter remained static at 1. This selection was guided by the objective of achieving maximum accuracy. For the Logistic Tree model, a thorough search over a hyperparameter grid was conducted. This grid explored a range from 10 to 100 (in steps of 10) for the minsize parameter, 1 to 5 (in steps of 1) for maxdepth, and 0.01 to 1 (in steps of 0.01) for alpha. The highest accuracy was obtained from a model configured with minsize equal to 70, maxdepth at 5, and alpha set to 0.9.

Metric	Logistic	LASSO Logistic	Logistic Tree
Accuracy	88.65%	86.66%	89.00%
Sensitivity	89.15%	88.60%	89.85%
Specificity	88.14%	84.71%	88.14%
Kappa	77.29%	73.31%	77.99%

Note: This table compares the performance of three models: Logistic, LASSO Logistic, and Logistic Tree. Accuracy, Sensitivity, Specificity, and Kappa are the metrics evaluated. The percentages represent the scores obtained by each model on these metrics.

Table 5.4: Performance of Logistic, LASSO Logistic, and Logistic Tree Models

The logistic tree model, which was the most intricate method utilized, yielded the highest accuracy among the approaches tested. However, McNemar's Chi-square test revealed that the differences in accuracies between the logistic tree and the simpler logistic model were not statistically significant. Following that, a different approach was used for the evaluation. The F1 scores, using different data subsets, were compared using both the simple logistic model and the logistic tree, as relying solely on the overall accuracy is not a particularly robust choice. The F1 score is a measure of the accuracy of a binary classifier and it is calculated by taking the harmonic mean of the precision and recall scores. The precision score is the proportion of positive predictions that were actually positive, and the recall score is the proportion of actual positives that were correctly predicted as positive. The subsets considered were constructed based on the results derived from the logistic regression tree and the cultural profiles constructed.

To examine whether the regression tree performs better at predicting certain cultures, a statistical significance test was conducted. The Wilcoxon signed-rank test was used on all five different subsets. This non-parametric statistical test compares two paired groups by calculating the difference between the pairs of numbers and ranking these differences. The test then evaluates whether these ranks are randomly distributed or whether one group has consistently higher or lower ranks than the other. Additionally, because it considers both the direction and magnitude of differences, it is a signed-rank test. The null hypothesis for this test is that the median difference between the paired observations is zero.

Thoroughly, for the first cultural profile, the test's p-value of 0.3132 indicated that the null hypothesis cannot be rejected, keeping a 0.05 threshold for statistical significance. The F1 scores for the Simple Logistic and the Logistic Tree were 0.8888 and 0.8897 respectively, so for that subset, the Logistic Tree performed slightly better without a statistically significant difference. However, it is important to note that failing to reject the null hypothesis does not mean that the null hypothesis is true. It simply means that the test did not provide enough evidence to conclude that there is a difference between the scores. The test could have been underpowered, or the difference between the scores could be very small. For cultural profiles 2, 4, and 5, the logistic tree model performed better in terms of F1 score. The difference in F1 score between the two models was statistically significant in these cases. The only cultural profile where the simple model performed significantly better than the tree model was profile 2. The results of the statistical test are shown in Table 5.5 below:

Profile	**Simple Logistic F1**	**Logistic Tree F1**	**Wilcoxon test p-value**
Cultural Profile 1	0.8888	0.8897	0.3132
Cultural Profile 2	0.8853	0.9132	$< 2.2e - 16$
Cultural Profile 3	0.9534	0.9313	$< 2.2e - 16$
Cultural Profile 4	0.8926	0.8945	$< 2.2e - 16$
Cultural Profile 5	0.8581	0.9038	$< 2.2e - 16$

Note: This table compares the F1 scores from the Simple Logistic and Logistic Tree model, according to different cultural profiles. The Wilcoxon p-value column indicates the p-value result of the Wilcoxon test between the F1 scores of the two models for each profile. The scientific notation $< 2.2\text{e} - 16$ means that the p-value is a very small number around zero.

Table 5.5: F1 score comparison using the Wilcoxon Test for Different Cultural Profiles

Where:

- Cultural Profile 1: $MAS \leq 43 \wedge UNAV \leq 63 \wedge LTO > 38 \wedge IND \leq 68$
- Cultural Profile 2: $MAS \leq 43 \wedge 64 \leq UNAV \leq 88$
- Cultural Profile 3: $44 \leq MAS \leq 64 \wedge IND > 50$
- Cultural Profile 4: $65 \leq MAS \leq 66$
- Cultural Profile 5: $67 \leq MAS \leq 79$

Each model had something unique to offer to investigate the relationships between the independent variables and the sentiment of the review. The logistic model helped quantify relationships, while LASSO pinpointed the variables that are pivotal towards a certain prediction. The tree model, through its unique ability to partition data using cultural dimensions, allowed us to transform extensive data into more manageable and interpretable subsets based on important variables. Individual cultural profiles and associations with particular attributes and topics emerged as a result of this. Consequently, the extracted various cultural profiles, offer nuanced insights into how cultural score combinations impact review sentiment, proving valuable for strategic marketing approaches.

Chapter 6

Conclusion and Limitations

Conducting thorough research into what constitutes customer satisfaction through textual data is crucial for any business operation. During the early days of the Internet, it was relatively simple to monitor customer sentiments and adjust according to their expectations. Even so, in the contemporary era with an ever-growing number of internet users, consumers have found it simpler to share feedback on products and services. Hence, with the rapid increase in data availability, the task of identifying suitable methodologies to extract meaningful insights is becoming increasingly challenging. Previously, consumers primarily depended on word-of-mouth for purchasing decisions, while currently, the vast majority turn to websites (forums) and reviews for guidance. Given the importance of positive reviews in decision-making, businesses must understand and meet customer needs to boost their market reputation. This understanding contributes to the generation of further positive feedback, ultimately uplifting the business's position in the market of interest. Innovative textual analysis and machine learning techniques can help businesses understand customer viewpoints and what affects their satisfaction. Customers and markets have dynamic relationships, so it is difficult to predict individual choices. Nonetheless, customers can be grouped by their behaviors and preferences. Culture significantly influences decision-making thus businesses can tailor their marketing strategies to specific consumer profiles, helping to retain or attract customers.

This thesis aimed to answer the question *"What particular attributes of a hotel, review themes and cultural elements substantially affect the satisfaction of consumers?"* by utilizing a novel methodology and utilizing recent machine learning and NLP techniques. The study underscored significant relationships and differences between cultural variety and review sentiment, offering key insights for hospitality industry marketers. It revealed a substantial negative link between societal individualism and hotel review sentiment, implying that markets with high individualism might pose increased challenges for positive customer reviews. High-quality product and service expectations characterize such societies, prompting hoteliers to closely meet their needs and respond to their feedback. Likewise, a trend for future-orientated customers to leave lower sentiment reviews was identified. Contrary to previous studies suggesting long-term oriented cultures are less likely to leave negative feedback, this research suggests these customers now provide more negative feedback, seeking future issue resolution. In other words, their desire to give useful feedback for better future experiences surpasses their concern about negative relationships with hoteliers. Analysis of hotel reviews highlighted key themes such as furniture and household items, customer service and booking experience, and accessibility in terms of location and transportation, as crucial to review sentiment. While location and transportation factors can't be altered for existing businesses, their significance should inform decisions on the placement of new hospitality ventures. Optimal locations are easily accessible and close to public transport, as it improves the guest experience, leads to better reviews, and enhances the establishment's reputation. Additionally, cleanliness and a friendly staff emerged as influential factors in review sentiment. Negative experiences often involved unclean accommodations or unfriendly hosts, while positive ones highlighted clean facilities and kind hosts. Therefore, these basic elements of hotel services should be consistently upheld to high standards.

The logistic tree model was used to thoroughly examine cultural heterogeneity. The model revealed that Masculinity is the primary cultural factor that influences review valence. Overall, five cultural profiles were identified. The first profile encompasses cultures leaning towards femininity, with a moderate to high tendency to avoid uncertainty and a balance between long-term orientation and indulgence. These customers value room comfort, aesthetics, amenities, and a straightforward booking experience, with lengthier reviews often signaling dissatisfaction. The second profile refers to cultures with moderate to high femininity and uncertainty avoidance, emphasizing on overall room comfort and design. Customers from moderately feminine cultures prioritize aesthetics over practicality. For such guests, offering rooms with pleasing views, decor, and new furniture, tailored to their cultural preferences, could enhance satisfaction. In contrast, profiles four and five, which are associated with strong masculine traits, value efficient booking, location, and good customer service more than room aesthetics. These pragmatic cultures are outspoken in their negative reviews, often using words like "terrible" or "nightmare". This highlights the importance of delivering based on feedback, as negative reviews can hurt a property's reputation. For these guests, a quick booking process and clean rooms with practical amenities are essential. On the other hand, the third profile, a mix of feminine and masculine traits leaning towards indulgence, observes and evaluates the hospitality experience holistically. They weigh the service quality, ambiance, and room amenities equally. This group has high standards and seeks an overall quality experience, requiring friendly and attentive service for full satisfaction.

From a methodological standpoint, this thesis offers a new approach to studying consumer opinions and their cultural nuances. It uses a combination of text analytics techniques, including Top2Vec, SVD, logistic regression, and logistic regression tree. These techniques are used to extract common themes and topics from reviews, reduce the dimensionality of the TF-IDF matrix, determine the relationships between the extracted aspects and customer satisfaction, and understand how cultural dimensions interact with other variables respectively. The study also incorporates multilingual reviews, uses a transformer model for labeling, and employs a novel topic extraction method. This allows for a more comprehensive exploration of cultural heterogeneity and achieves an accuracy of 86-89%.

This thesis introduced a novel method to analyze consumer preferences, highlighting significant cultural differences, yet there is room for further exploration in this field. Expanding the study to non-European countries could offer insights into global cultural variances. Further, examining evolving cultural profiles by categorizing travelers (e.g., families, couples, solo travelers) can enrich understanding and aid in creating effective recommendation systems. This deep dive could uncover specific preferences, like seasonal destination choices, among different cultural and traveler groups. Another limitation of this analysis is the limited availability of negative reviews. This is a common problem when analyzing consumer-generated text from online travel agencies (OTAs), as some hoteliers may delete negative reviews or even provide fake positive reviews that could inflate the overall sentiment for an accommodation property. In addition, neutral reviews should also be examined because they represent a substrata of customers who may need minor interventions to turn them into positive sentiments. This is important for businesses that want to improve their reputation quickly in the short term, as it is more difficult to satisfy highly unsatisfied customers. Finally, more sophisticated classification methods - which are more robust in terms of accuracy - could be applied. These methods could be combined with black box interpretability techniques, such as LIME (local interpretable model-agnostic explanations) or Shapley Values, to provide explanations for the predictions made by the models.

Bibliography

[1] Q. Ye, R. Law, B. Gu, and W. Chen, “The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings,” *Computers in Human Behavior*, vol. 27, no. 2, pp. 634–639, 2011, web 2.0 in Travel and Tourism: Empowering and Changing the Role of Travelers. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563210000907

[2] M. Sigala, A. Lockwood, and P. Jones, “Strategic implementation and it: gaining competitive advantage from the hotel reservations process,” *International Journal of Contemporary Hospitality Management*, vol. 13, no. 7, pp. 364–371, 2001.

[3] U. Gretzel and K. H. Yoo, “Use and impact of online travel reviews,” in *Information and communication technologies in tourism 2008*. Springer, 2008, pp. 35–46.

[4] S. W. Litvin, R. E. Goldsmith, and B. Pan, “Electronic word-of-mouth in hospitality and tourism management,” *Tourism Management*, vol. 29, no. 3, pp. 458–468, 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0261517707001343

[5] L. R. Klein, “Evaluating the potential of interactive media through a new lens: Search versus experience goods,” *Journal of Business Research*, vol. 41, no. 3, pp. 195–203, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296397000623

[6] B. Pang, L. Lee *et al.*, “Opinion mining and sentiment analysis,” *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[7] J. Stavkova, L. Stejskal, Z. Toufarova *et al.*, “Factors influencing consumer behaviour,” *ZEMEDELSKA EKONOMIKA-PRAHA-*, vol. 54, no. 6, p. 276, 2008.

[8] R. Benda-Prokeinová, K. Dobeš, L. Mura, and J. Buleca, “Engel’s approach as a tool for estimating consumer behavior,” 2017.

[9] A. Krishna and N. Schwarz, “Sensory marketing, embodiment, and grounded cognition: A review and introduction,” *Journal of Consumer Psychology*,

vol. 24, no. 2, pp. 159–168, 2014, sensory perception, embodiment, and grounded cognition: Implications for consumer behavior. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1057740813001162

[10] J. Novotny and P. Duspiva, "Factors influencing consumers' buying behavior and their importance for enterprises/faktory ovlivnujici kupni chovani spotrebitelu a jejich vyznam pro podniky," *E+ M Ekonomie a Management*, vol. 17, no. 1, pp. 152–167, 2014.

[11] M. Horáková, "Consumer behavior of college students in the czech republic," *Journal of Competitiveness*, 2015.

[12] M. Grzegorz, "Value-based inventory management," *Romanian Journal of Economic Forecasting*, vol. 9, no. 1, pp. 82–90, 2008.

[13] H. C. Triandis, "The self and social behavior in differing cultural contexts." *Psychological review*, vol. 96, no. 3, p. 506, 1989.

[14] P. M. Gollwitzer, R. A. Wicklund, and J. L. Hilton, "Admission of failure and symbolic self-completion: Extending lewinian theory." *Journal of personality and social psychology*, vol. 43, no. 2, p. 358, 1982.

[15] A. G. Greenwald, C. G. Carnot, R. Beach, and B. Young, "Increasing voting behavior by asking people if they expect to vote." *Journal of Applied Psychology*, vol. 72, no. 2, p. 315, 1987.

[16] M. Snyder, "Self-monitoring of expressive behavior." *Journal of personality and social psychology*, vol. 30, no. 4, p. 526, 1974.

[17] R. H. Fazio and C. J. Williams, "Attitude accessibility as a moderator of the attitude–perception and attitude–behavior relations: An investigation of the 1984 presidential election." *Journal of personality and social psychology*, vol. 51, no. 3, p. 505, 1986.

[18] R. C. Ziller, "Self-other orientations and quality of life," *Social Indicators Research*, vol. 1, pp. 301–327, 1974.

[19] M. Snyder, *Public appearances, Private realities: The psychology of self-monitoring.* WH Freeman/Times Books/Henry Holt & Co, 1987.

[20] H. C. Triandis and V. Vassiliou, "Interpersonal influence and employee selection in two cultures." *Journal of Applied Psychology*, vol. 56, no. 2, p. 140, 1972.

[21] H. E. Tajfel, *Differentiation between social groups: Studies in the social psychology of intergroup relations.* Academic Press, 1978.

[22] R. A. Shweder and R. A. LeVine, *Culture theory: Essays on mind, self and emotion.* Cambridge University Press, 1984.

[23] S. P. Douglas and C. Samuel Craig, "The changing dynamic of consumer behavior: implications for cross-cultural research," *International Journal of Research in Marketing*, vol. 14, no. 4, pp. 379–395, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167811697000268

[24] M. Featherstone, "Global culture: An introduction," *Theory, Culture & Society*, vol. 7, no. 2-3, pp. 1–14, 1990.

[25] R. H. Munroe, R. L. Munroe, and H. S. Shimmin, "Children's work in four cultures: Determinants and consequences," *American Anthropologist*, vol. 86, no. 2, pp. 369–379, 1984. [Online]. Available: http://www.jstor.org/stable/678969

[26] M. J. Herskovits, "The contribution of afroamerican studies to africanist research," *American Anthropologist*, vol. 50, no. 1, pp. 1–10, 1948. [Online]. Available: http://www.jstor.org/stable/663945

[27] E. B. Tylor, *Primitive culture: Researches into the development of mythology, philosophy, religion, art and custom.* J. Murray, 1871, vol. 2.

[28] G. J. Hofstede and P. Pedersen, "Synthetic cultures: Intercultural learning through simulation games," *Simulation & Gaming*, vol. 30, no. 4, pp. 415–440, 1999.

[29] J. W. Berry, "Imposed etics—emics—derived etics: The operationalization of a compelling idea," *International journal of psychology*, vol. 24, no. 6, pp. 721–735, 1989.

[30] T. G. Dragonas, "The self-concept of predolescents in the hellenic context," Ph.D. dissertation, University of Aston in Birmingham, 1983.

[31] G. Hofstede and M. H. Bond, "Hofstede's culture dimensions: An independent validation using rokeach's value survey," *Journal of cross-cultural psychology*, vol. 15, no. 4, pp. 417–433, 1984.

[32] G. Hofstede, "Dimensionalizing cultures: The hofstede model in context," *Online readings in psychology and culture*, vol. 2, no. 1, p. 8, 2011.

[33] R. D. Leon, "Hotel's online reviews and ratings: a cross-cultural approach," *International Journal of Contemporary Hospitality Management*, vol. 31, no. 5, pp. 2054–2073, 2019.

[34] N. Raghuvanshi and J. M. Patil, "A brief review on sentiment analysis," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).* IEEE, 2016, pp. 2827–2831.

[35] P. K. Jain, R. Pamula, and G. Srivastava, “A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews,” *Computer Science Review*, vol. 41, p. 100413, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574013721000538

[36] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, “Deep learning for aspect-based sentiment analysis: A comparative review,” *Expert Systems with Applications*, vol. 118, pp. 272–299, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418306456

[37] K. Schouten and F. Frasincar, “Survey on aspect-level sentiment analysis,” *IEEE transactions on knowledge and data engineering*, vol. 28, no. 3, pp. 813–830, 2015.

[38] S. Brody and N. Elhadad, “An unsupervised aspect-sentiment model for online reviews,” in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 804–812.

[39] W. Chen, F. Rabhi, W. Liao, and I. Al-Qudah, “Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study,” *Electronics*, vol. 12, no. 12, p. 2605, 2023.

[40] J. Huang, Y. Meng, F. Guo, H. Ji, and J. Han, “Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding,” *arXiv preprint arXiv:2010.06705*, 2020.

[41] I. Graßl and G. Fraser, “Scratch as social network: topic modeling and sentiment analysis in scratch projects,” in *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society*, 2022, pp. 143–148.

[42] D. Angelov, “Top2vec: Distributed representations of topics,” *arXiv preprint arXiv:2008.09470*, 2020.

[43] R. Egger and J. Yu, “A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts,” *Frontiers in sociology*, vol. 7, p. 886498, 2022.

[44] S. Thara and S. Sidharth, “Aspect based sentiment classication: Svd features,” in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 2370–2374.

[45] F. Abbasi, A. Khadivar, and M. Yazdinejad, “A grouping hotel recommender system based on deep learning and sentiment analysis,” *Journal of Information Technology Management*, vol. 11, no. 2, 2019.

[46] H. R. Alhakiem, E. B. Setiawan *et al.*, "Aspect-based sentiment analysis on twitter using logistic regression with fasttext feature expansion," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 5, pp. 840–846, 2022.

[47] A. Prabhat and V. Khullar, "Sentiment classification on big data using naïve bayes and logistic regression," in *2017 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2017, pp. 1–5.

[48] T. H. Jaya Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to rinca island development using doc2vec and svm and logistic regression as classifier," *Procedia Computer Science*, vol. 197, pp. 660–667, 2022, sixth Information Systems International Conference (ISICO 2021). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705092102411X

[49] A. Subroto and M. Christianis, "Rating prediction of peer-to-peer accommodation through attributes and topics from customer review," *Journal of Big Data*, vol. 8, pp. 1–29, 2021.

[50] A. C. Calheiros, S. Moro, and P. Rita, "Sentiment classification of consumer-generated online reviews using topic modeling," *Journal of Hospitality Marketing & Management*, vol. 26, no. 7, pp. 675–693, 2017. [Online]. Available: https://doi.org/10.1080/19368623.2017.1310075

[51] N. Hu, T. Zhang, B. Gao, and I. Bose, "What do hotel customers complain about? text analysis using structural topic model," *Tourism Management*, vol. 72, pp. 417–426, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0261517719300020

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[53] H. M. . F. E. Landwehr, N., "Logistic model trees," *Journal of Machine Learning Research*, vol. 59, pp. 161–205, 2005.

[54] A. Zeileis, T. Hothorn, and K. Hornik, "Model-based recursive partitioning," *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 492–514, 2008. [Online]. Available: https://doi.org/10.1198/106186008X319331

[55] T. Hothorn and A. Zeileis, "partykit: A modular toolkit for recursive partytioning in r," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3905–3909, 2015.

[56] N. S. Koh, N. Hu, and E. K. Clemons, "Do online reviews reflect a product's true perceived quality? an investigation of online movie reviews across cultures," *Electronic Commerce Research and Applications*, vol. 9, no. 5, pp. 374–385, 2010, special Section on Strategy, Economics and Electronic Commerce. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156742231000027X

[57] S. Ryu, C. W. Moon *et al.*, "Long-term orientation as a determinant of relationship quality between channel members," *International Business & Economics Research Journal (IBER)*, vol. 8, no. 11, 2009.

[58] P. Stamolampros, N. Korfiatis, P. Kourouthanassis, and E. Symitsi, "Flying to quality: Cultural influences on online reviews," *Journal of Travel Research*, vol. 58, no. 3, pp. 496–511, 2019.

[59] O. Furrer, B. S.-C. Liu, and D. Sudharshan, "The relationships between culture and service quality perceptions: Basis for cross-cultural market segmentation and resource allocation," *Journal of service research*, vol. 2, no. 4, pp. 355–371, 2000.

[60] B. Gao, X. Li, S. Liu, and D. Fang, "How power distance affects online hotel ratings: The positive moderating roles of hotel chain and reviewers' travel experience," *Tourism Management*, vol. 65, pp. 176–186, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0261517717302273

[61] H. Fang, J. Zhang, Y. Bao, and Q. Zhu, "Towards effective online review systems in the chinese context: A cross-cultural empirical study," *Electronic Commerce Research and Applications*, vol. 12, no. 3, pp. 208–220, 2013, negotiation and E-Commerce. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156742231300015X

[62] Y. Van Vaerenbergh, C. Orsingher, I. Vermeir, and B. Larivière, "A meta-analysis of relationships linking service failure attributions to customer outcomes," *Journal of Service Research*, vol. 17, no. 4, pp. 381–398, 2014.

[63] R. Song, S. Moon, H. Chen, and M. B. Houston, "When marketing strategy meets culture: The role of culture in product evaluations," *Journal of the Academy of Marketing Science*, vol. 46, pp. 384–402, 2018.

[64] S. W. Litvin, "Hofstede, cultural differences, and tripadvisor hotel reviews," *International Journal of Tourism Research*, vol. 21, no. 5, pp. 712–717, 2019.