ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

What makes for a memorable holiday?

Analyzing the impact of online travel review attributes and cultural variations on consumer travel preferences and satisfaction

Master's thesis Business Economics

Data science and marketing analytics

Barbu Andresoiu | 508868

Supervisor: Andreas Alfons

Second assessor: Bas Donkers

Date final version: 30th of July 2024

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This thesis explores the drivers behind customer satisfaction in the package travel industry, an industry niche that has not yet been thoroughly explored in the literature. The paper's focus is on Sunweb, a Dutch tour operator active in multiple international markets. The study utilizes text analytics through Latent Dirichlet Allocation to identify topics present in user generated reviews scraped off TrustPilot. It then uses these topics to try to explain the ratings given by Sunweb's customers on Trustpilot. The reviewers' nationality and Hofstede's cultural dimensions are also included, to assess whether there are clear differences in in the importance of those topics depending on the origin and culture of the user. Ordinal logistic regression and random forests are used to assess how these attributes affect the user star rating from 1 to 5. While the cultural aspects do not provide much insight, the influence of the topics on the ratings can give valuable insights into how tour operators and specifically Sunweb can improve their services. The findings mostly confirm hypotheses based on previous research on how intangible aspects tend to influence reviews positively. The main managerial takeaways are to stimulate repeat customers to give reviews as those are generally positive, and to create an accessible online environment for customers to find their preferred vacation types and attributes.

Contents

Introduction

A vacation abroad is a thing many people look forward to, of any age, gender or culture. While the trip itself is the main goal, the whole process is exciting and contributes to the experience, from choosing the purpose and destination of the trip, booking it, the actual trip, and of course reflecting on it on the way back and the years that follow.

In the past, travel plans would usually happen through word-of-mouth recommendations or through expert advice, such as from travel agencies. You would ask travel agencies to plan a trip for you, or to offer you a package deal of transport, accommodation and other wants depending on the customer. Since the advent of the internet, more and more travel arrangements are made online, with the process being very consumer friendly and putting the decisions in the hands of users who can themselves decide on and book their travel needs. And while most players in the travel market have their own websites, aggregators are increasingly popular. Websites such as booking.com have put the accommodation selection one click away, and airline aggregators also give the customer the option to choose the flights best suited for them. Same holds for car rental agencies and other excursions that travelers would like to partake in. Most of this can be found online.

Some customers, however, still prefer getting a package holiday, organized by someone else. While hotels make up most of the revenue earned in the travel industry, package holidays are in second place and are projected to keep growing over the next few years[1]. Travel agencies and tour operators are parties offering these trips and they have also been switching to the online medium in the form of ecommerce stores that sell planned trips.

The accessibility of the internet has given consumers the power to broadcast their experiences with different companies in the form of online reviews, and the travel industry is no stranger to this phenomenon. Gretzel & Yoo (2008) found that more than half of travelers read reviews before planning their trip at least once, and more recently, more than 80% of Trip Advisor users said the reviews on the website help them feel confident in their travel decisions, according to an independent study on how essential reviews are in the booking process[2].

[1] Statista. (n.d.). *Travel & Tourism - Worldwide | Statista Market Forecast*. Retrieved from https://www.statista.com/outlook/mmo/travel-tourism/worldwide

[2] TripAdvisor. (2014, February 11). *Eighty percent of TripAdvisor users read at least six to 12 reviews before choosing a hotel*. TripAdvisor Media Center. https://tripadvisor.mediaroom.com/2014-02-11-Eighty-percent-of-TripAdvisor-users-read-at-least-six-to-12-reviews-before-choosing-a-hotel

What the internet has also further facilitated is the spread of online services around the world. This also holds true for travel companies being accessible from a multitude of countries. Cultural differences play a pivotal role in the field of consumer behavior, affecting preferences and attitudes towards products and services, with advertising practices also differing across countries and cultures (Solomon et al., 2016). Therefore, understanding cultural nuances is critical for businesses looking to expand or cater to diverse consumer segments across multiple markets.

While customer surveys, focus groups, and other traditional research methods can shine light on these cultural differences, they present some disadvantages such as duration of data collection and biases caused by interviewer influence, respondent bias and survey learning effects, or just sample selection. The ample availability of online reviews and advancements in natural language processing and translation algorithms present a great opportunity to gain extra insights into consumer behavior from readily available data, written in a natural setting. In this case, the posting feedback for a vacation that had just passed. The constant flow of online reviews also presents an opportunity to investigate how reviews differ across time, and whether they reflect improvement or regression of the company's services, as well as possible societal changes in preferences.

Most literature that covers the topic focuses on hotel reviews (Schuckert et al., 2015; Yang et al., 2018). In these cases, the reviews refer to the hotel room or apartment booked on platforms such as booking.com or AirBnB, not to the platform itself, which is an intermediary. Tour operators and travel agencies plan the whole trip, taking care of transport, accommodation, and possible auxiliaries such as car rental or special equipment. Reviews given to travel agencies should tend to tell us something about the whole experience of booking with the company at hand.

The aim of this paper is to answer the following research questions:

What attributes influence the overall experience and satisfaction levels of customers in the travel industry when booking package holidays?

How do national and cultural factors influence the satisfaction levels of customers in the travel industry when booking package holidays?

How does the influence of those attributes vary across national and cultural segments?

This will be done by conducting an exploratory analysis into reviews for Sunweb, found on the popular review website TrustPilot and available for all 7 countries Sunweb operates in: The Netherlands, Belgium, France, Germany, The United Kingdom, Denmark and Sweden. Using text analytics and classification methods, the aim of this research is to gain insight into what topics, and therefore details of the purchased trip, affect customer satisfaction, how these details and their importance vary across the different countries, and across the different cultural nuances of those countries. The attributes will be extracted from the reviews, the cultural factors will be based on cultural dimension theory and the overall star rating given in each review will be used as a measure for customer satisfaction.

Relevance

The insights from this analysis should reflect the positives and negatives of how customers experience their whole trip, and what details tend to be of importance. This adds to the existing research in customer review analytics as well as to consumer behavior in the tourism industry.

As mentioned previously, the bulk of the literature at the time of writing mainly covers reviews pertaining hotels. This paper aims to add to the literature by focusing on travel agencies/tour operators which oversee the whole trip, from the booking process to the transport and accommodation, and customer support. This is a different segment of travel industry suppliers that would be interesting to look at.

From a management perspective, this paper should give tourism companies insight into what tends to be important to their potential customers, and therefore strengths they should promote and weaknesses they should improve on. Xie et al. (2020) also suggests that travel managers should focus on a customer centric approach, online and offline, and really take note of customer feedback. The fact that reviews have an increasing social influence on consumers booking their trips (Book et al., 2015) also makes considering reviews in managerial decisions of vital importance.

The geographical and cultural aspect of the research should also add to the extent to which things are appreciated and criticized in different countries and depending on which cultural aspects. This adds to existing cross-cultural research into consumer behavior and from a managerial perspective the findings can be used by companies to see how they should adapt depending on the markets they operate in. Vermeulen & Seegers (2009) found that online hotel reviews had a stronger effect on customer consideration in the case of lesser-known hotels. In the case of a travel agency wanting to expand to a new market, the cultural aspect of this research can give insights into what is appreciated in that culture, which should lead to strong reviews to help the initial low recognition in the market.

As for the company whose data is used in this analysis, the results can be used as recommendations for actions soon. The framework can also further be used to gain updated insights into what affects their customer satisfaction.

The paper will be structured as follows. First, a literature review will be conducted, of previous works related to cross cultural consumer behavior, online consumer reviews and the influence of cultural factors on those consumer reviews, focusing on the travel industry to stay relevant to the question at hand. A description of the data will follow, including the collection and initial processing. Next, the topic modeling method will be described as well as the classification methods to be used ulteriorly, and why they are suitable for this research problem. The results of the models will then be presented, followed by the interpretation and answers to the research questions and formulated hypothesis. The paper will conclude with a discussion section which serves to look into the limitations of this particular paper and what possible future research ideas would be of valuable contribution to this topic of interest.

Purpose

The travel industry is in constant flux, continually evolving due to dynamic consumer preferences, technological advancements and world events. To keep up in this landscape, it is of utmost importance that travel companies stay in touch with emerging trends and patterns, at the industry level, but also at the individual level. Therefore, this research is primarily exploratory in nature, aiming to generate insights and understanding of how customers judge package travel deals.

Exploratory research is characterized by an open ended and flexible approach, seeking to understand the topic at hand and generate ideas. This method is particularly useful, in the case of trying to understand consumer reviews for a particular company such as Sunweb. Its exploratory nature makes it so the research is not bound by previous studies in this area and can be conducted in such a way to help Sunweb gain their own insights into their own data. The main objective of this research is to understand consumer preferences for package holidays, while also controlling for the reviewers’ countries and cultural elements.

While the primary focus of the paper is exploratory, there is a significant body of literature into customer reviews and reviews in the travel industry (not package holidays specifically) that cannot be ignored. To enrich the exploratory findings, the research will include a hypothesis testing component, based on findings from existing literature. The existing literature will be explored in the literature review, as well as the resulting hypotheses.

Literature review

This section aims to gain some background information in order to understand the travel industry and online reviews, as well as how cultural factors can influence those. Consumer behavior and cultural dimensions, topic modeling in the travel industry and consumer reviews across cultures will be explored. Finally, even though this is mainly an exploratory study, some hypotheses will be formed based on the literature in order to anchor the exploratory findings within the established body of literature.

Consumer behavior and cultural dimensions

Consumer behavior explores how individuals make decisions about what products or services to consume, and how both internal and external influences affect those decisions. As society changes, so does this field, which examines factors like personal, psychological, and situational elements affecting consumers (Stávková et al., 2008) as well as the decision-making process itself, an example being the popular five-step model consisting of problem recognition, information search, alternative evaluation, purchase, and post-purchase behavior (Kotler & Keller, 2012). Research in the field largely revolves around these decision-making steps and the factors that affect them.

Culture is one of those influences of the consumer decision-making process. Culture is a complex concept that is difficult to define. An example of a definition is "Culture is a system of inherited conceptions expressed in symbolic forms by means of which men communicate, perpetuate, and develop their knowledge about and attitudes toward life." (Geertz, 1977). Definitions, however, usually emphasize how cultural elements shape what is considered normal within a society. Hofstede (1980) further proposed that culture is the mental framework distinguishing one group from another, highlighting differences and similarities across cultures. Cultural studies can take an etic approach, comparing variables common to all cultures, or an emic approach, focusing on understanding cultural meanings from within. The etic approach offers an outsider's perspective, while the emic approach provides an insider's view of culture.

Dutch psychologist Gert Hofstede developed a method to quantify cultural characteristics through cultural dimensions. Originally, four dimensions were defined: "Power distance," "Individualism vs Collectivism," "Masculinity vs Femininity," and "Uncertainty avoidance" (Hofstede & Bond, 1984). Later, two more dimensions were added: "Long term vs Short term orientation" and "Indulgence vs restraint" (Hofstede, 2011). Power distance measures acceptance of unequal power distribution, with high scores indicating hierarchical societies and low scores suggesting equality and questioning of authority. Individualism vs collectivism assesses whether priority is given to personal goals as opposed to societal interests, with individualistic cultures valuing personal freedom and collectivist cultures valuing harmony. Masculinity vs femininity evaluates which traditionally masculine or feminine traits are valued, with masculine traits including assertiveness, competition and success and feminine traits including nurturance, cooperation and quality of life. Uncertainty avoidance shows comfort with ambiguity and change, with some societies creating rules to avoid uncertainty while others are more tolerant. Long term vs short term orientation reflects the focus on long-term planning or short-term gratification, with long-term cultures prioritizing future rewards and adaptability. Indulgence vs restraint measures the extent to which societies allow for gratification of desires versus controlling them.

Topic modeling for reviews in the travel industry

Analyzing text reviews in the tourism industry is a known phenomenon. Online consumer reviews have been employed to study tourist opinions on hotels, airlines, restaurants and attractions. However, the majority of the existing research has mainly focused on the hotel sector (Schuckert et al., 2015). The analysis of online ratings is applied in the tourism industry across multiple steps of the consumer decision making process. Gavilan et al (2018) shows how in the consideration phase, web users tend to trust low ratings more than high ratings. The number of reviews does moderate this. Having many positive ratings does increase trustworthiness.

Studying the topics found in online hotel reviews, Berezina et al (2015) discovered that the most important attributes found in online reviews were value for money, contact with hotel staff, and room furnishing attributes. An interesting finding is that tangible components, such as the room quality and money issues, play a significant role in negative reviews, while intangible aspects such as the staff competence and overall experience were the main aspect of positive reviews.

In a topic modeling study of reviews for Disneyland, Luo et al (2020) has similar findings, with topics such as value for money, food price, and cleanliness, all tangible aspects, being more present in negative reviews, while positive reviews often talked about a happy experience or the fantasy feeling, an intangible element. This research also included a geographical component since it included multiple Disney theme park locations around the world. Some interesting insights are that food and dining options tend to appear in reviews for Disneyland

Paris, shopping options are popular in reviews for Hong Kong Disneyland. Interestingly the topic non-human characters, referring to the Disney characters is very popular in reviews left by users from Philippines, India, Malaysia, and Singapore. These findings suggest that marketing efforts can be targeted based on which topics are important for each destination but also on the origin of the visitors. While the focus was on Airbnb reviews, Ding et al (2020) also included a cultural component, comparing Malaysian users to international users and also suggested targeting customers based on their expectations. Interestingly, it was also found that host attributes are a common topic in reviews, highlighting the importance of this intangible element.

Consumer reviews across cultures in the travel industry

While more limited, there are also instances of cultural elements being included in travel review text analysis. In an application of Hofstede's cultural dimensions on hotel reviews, Leon (2019) found that these dimensions can explain 9.90 per cent of reviews variation and 4.50 per cent of rating variation.t

Certain articles, however, only focus on one of the dimensions. Focusing on uncertainty avoidance, Litvin (2019) found that customers from low uncertainty avoidance countries tend to give higher star ratings, which coincides with Litvin et al.'s (2004) finding that low uncertainty avoidance travelers tend to invest greater effort into travel planning.

Looking at flights, Chatterjee & Mandal (2020) suggests that consumers from countries with lower individualism are more prone to positive ratings and recommendations than customers from countries with high individualism. High individualist consumers also show more importance to process attributes such as in-flight entertainment than outcome attributes such as value for money. Low uncertainty avoidance shows a lower importance of tangible attributes such as seat comfort or inflight entertainment in comparison to higher uncertainty avoidance. Lastly, higher long-term orientation seems to correlate with less importance of food and beverages, an outcome attribute. Furthermore, Stamolampros et al. (2018) found that customers from more power distance societies are more critical to staff, and more prone to complain about baggage fees or delays, with the opposite being true to the other end, with low power distance passengers praising staff and in-flight services. A very similar effect is seen in the case of individualism. Long term orientation also has a strong effect. Passengers from long term-oriented cultures are more sensitive to check in price or staff assistance, while short term-oriented cultures are more sensitive to extra fees and the general baggage policy of the carriers. The other three dimensions show little variation in the prevalence of the topics.

Looking at the satisfaction of American and Chinese tourists in restaurants, it was found that Chinese tourists, coming from a country that features a larger power distance, collectivism, and is less indulgent, tend to give higher ratings than American tourists, for whom the highly individualistic culture shows itself in being more open to giving low ratings and expressing dissatisfaction (Jia, 2020). The individual aspect is in line with the findings of Stamolampros et al. (2018), while the power distance element is not. Chinese tourists also focused on what to eat, while American tourists focused on why and how, showing a difference in the tangible vs intangible aspects of a dining experience. An interesting counterintuitive finding was that Chinese tourists tend to be put off by large crowds, which doesn't coincide with the collectivist nature of their society.

Hypotheses based on literature review

The literature pertaining to this field of online review analysis with cultural aspects taken into consideration is quite messy, with many articles focusing on different parts of the tourist experience, such as flights, hotels or attractions, while also focusing on different variables and topics that appear in the datasets at hand. Therefore, this paper will contribute to this field by looking at the correlation of topics with nationality and cultural aspects, while focusing on the whole travel experience brought by a package holiday deal. While the findings will depend on the topics found in the current data, some hypotheses can be formed based on previous literature.

H1: Tangible topics have a negative effect on review rating.

H2: Intangible topics have a positive effect on rating.

H1 and H2 are based on the findings of Berezina et al (2015) and Luo et al (2020). In this paper, this distinction between tangible and intangible will be made based on the topics found in the data used in this research.

H3: Higher individualism has an overall negative effect on reviews.

H4: Higher uncertainty avoidance has an overall negative effect on reviews.

Regarding the cultural aspects, uncertainty avoidance and individualism have the clearest link to travel reviews. The other dimensions will therefore be explored in further detail in this study.

As in the case of Luo et al (2020), the expectation is that reviews will show connections between countries and certain topics. However, hypothesis formulation is tough since the topics for travel agencies and the nationalities used will be different from those present in the review of theme park attributes.

Data

The first three paragraphs in this section will describe how the data was collected and processed before applying topic modeling. The final paragraph will describe what was done after topic modeling was applied, to prepare the data for explanatory modeling.

Scraping and cleaning reviews

The data used was extracted from TrustPilot.com, an online review website, scraping reviews of the various iterations of the Sunweb brand. These iterations are Sunweb Netherlands, Belgium, Denmark, France, Germany, Great Britain and Sweden.

The reviews section contains the following aspects. The review title, the review text itself, the name of the reviewer and the number of reviews they have posted on the website and the reviewer's location, the date of the experience (the vacation), the date the review was posted, and a star rating from 1 to 5. Some of those reviews also contain a reply from the Sunweb customer service team.

Reviews were selected from the period 6-08-2009 until 10-05-2024. For each iteration, the native language was selected. The only exception was Sunweb Belgium for which reviews in both French and Dutch were selected.

The reviews were scraped using the Google Chrome extension Instant Data Scraper, a tool which, as the name suggests, analyzes the contents of the webpage and scrapes the details in a spreadsheet format. The tool gives you the option to select the aspects you are interested in, so in this case the only ones that were scraped were the review title and text, the star rating and the dates of the experience and of the review. This was arranged in a tabular format, with a column corresponding to each of those attributes.

For each iteration of Sunweb, the reviews were downloaded in separate documents (except Belgium due to the two common languages spoken there, for which two documents were created). Each document received an extra variable representing the Sunweb publication it belonged to. Following that, the documents were concatenated into one large document containing the scraped attributes and the publications. The final reviews document contained the following columns: publication, review title, review text, star rating, date of experience, date of review, and reviewer country. In total, there were 9895 reviews in the document. The distribution of the reviews per country was as follows: 777 from Belgium, 387 from Germany, 2644 from Denmark, 1794 from France, 229 from Sweden, 653 from Great Britain and 3411 from The Netherlands.

Further processing of the text data

A dependent variable was created using the star rating. On Trustpilot, the star rating is shown in the form of an image that corresponds to the rating. There are five possible images, with one up to five of the stars filled in. When scraping the website, the rating wasn't saved as a number from one to 5, but as the source and name of the image file corresponding to the star rating. The number of stars per image is indicated in its title, for example with "stars-1.svg" up to "stars-5.svg". Using these titles, the variable was first converted to numeric from 1 to 5 and then to an ordered factor, with 1 being the lowest and 5 the highest.

Before starting the translation, an R function was run to detect the language of the review, namely detect_language from the gcld2 package (Google's compact language detector 2). As mentioned before, six languages were present in the dataset, meaning that they had to be translated to English. This was done using the *deeplr* package in R, using the DeepL translation algorithm. A filter was set to only translate the reviews that were not already in English and after translation, the main dataset now contained all the reviews in English. This is done to standardize the data, making it uniform to ensure consistency in analysis, and increase the dataset used for modeling ratings, as opposed to analyzing each language individually. The topic modeling technique to be discussed in the next section (LDA) is also suitable for translated reviews as it focuses on individual words and is therefore less susceptible to slight language nuances that are possibly lost in translation. While there are certain limitations and no translation is perfect, the combination of an AI based translation that is designed to keep the nuances of the language that is being translated, and a model that focuses on individual words and how there are linked, is a good compromise tradeoff for consistency in the review language. One note is that translation was not successful for each review. 431 reviews were not translated by the DeepL algorithm and therefore removed, which left 9464 reviews at that point.

After translation, the first step in processing the review text data was to convert all reviews to lowercase. This is because computers can't distinguish between "hotel" and "Hotel" for example. All numerical characters were eliminated, and all punctuation was removed. Next, common stop words were removed. These are words that don't add much value to the analysis, such as common articles, prepositions, conjunctions, and pronouns. Stemming was also applied, which consists of bringing words to their root form, meaning that words like "walker" or "walking" were turned into "walk". Finally, extra whitespaces were removed.

After the first preprocessing steps, tokenization followed for the review column, which means the text was split into each individual (stemmed) word. By doing this, all unique stemmed words in the reviews were extracted. Pruning the vocabulary consists of removing words that are too rare or too common to improve the quality of the analysis. Pruning was applied by removing words that appeared in less than 100 documents or more than 5000 documents. This is done to remove possibly misspelled words and to remove words that don't appear often since they aren't useful for analysis, but also words that appear in too many documents as those will not help differentiate the reviews from one another. Following the first pruning stage, a document-term matrix (DTM) was created. This is a matrix that describes the frequency of terms occurring in a collection of documents. After looking at the frequency of the words in the DTM, a noticeable thing was that many words were left that do not convey any information. Therefore, an extra list of words to be removed was created, those words being the following. *"also", "alway", "anoth", "bad", "book", "can", "differ","even", "fine", "get", "good", "great", "hotel", "just", "like", "look", "take", "well", "without", "anyth", "everyth", "mani", "much", "never", "nice", "perfect", "realli", "recommend", "still", "super", "thank", "thing", "think", "alreadi", "although", "either", "probabl", "quit", "seem", "thing", "veri", "vacat", "without"*.

Many of these words convey clearly positive or negative sentiments and would therefore be associated with positive reviews. While they would be useful for a sentiment analysis, that is not the goal of this research. The goal is to find out what attributes present in reviews affect the star rating, and positive or negative words would detract from actual vacation related words. Other words included are words that are to be expected in many travel reviews, such as “book”, “vacat” from “vacation”, or “hotel”.

Cultural Dimensions

The dataset used for cultural dimensions was retrieved from https://geerthofstede.com/ and contains data on the six cultural dimensions for 111 countries, with 65 countries containing values for each dimension. All countries that Sunweb operates in coincide with Hofstede’s research, therefore they had values for each dimension and no missing values were present. As previously mentioned, the cultural dimensions are power distance (PDI), individualism (IDV), masculinity (MAS), uncertainty avoidance (UAI), long-term vs short-term orientation (LTOWVS) and indulgence vs restraint (IVR). The dimension scores are reported from a minimum of 0 to a maximum of 100. As a few examples, a country with a high score for masculinity can be considered to have a masculine culture, while a low score can be associated with a feminine culture. For power distance, a high score suggests strong hierarchies and lower scores suggest a flatter organizational or hierarchical structure. High scores for individualism suggest a culture focused on individualism and personal freedom, while a lower score suggests a collectivism culture. High scores for uncertainty avoidance show that the culture isn't too fond of uncertainty, while lower scores suggest the culture is more comfortable with uncertainty. High scores for long-term vs short-term orientation suggest that long term thinking prevails in the cultural landscape, prioritizing future rewards, while lower scores would suggest cultures where short term gratification is more common. Lastly, high vs low scores for indulgence show whether societies allow and promote the gratification of desires vs controlling those.

Only the seven countries used in the research were kept from all the countries, so the Netherlands, Belgium, Denmark, Germany, Great Britain, France and Sweden. The final change that was made was in the country codes so they would correspond to those in the dataset containing reviews. They were changed from three letter country codes to two letter country codes. NET to NL, BEL to BE, DEN to DK, FRA to FR, GER to DE, GBR to GB, and SWE to SW, respectively. The final change was to divide all the scores by 100 and bring them on a scale from 0 to 1. This was to use the same scale as for the topic probabilities.

Post topic modeling adjustments

Some final adjustments were made to the dataset when the topics were discovered. The topic modeling technique used, and the resulting topics will be discussed in the methodology and results sections. This section explains how the dataset was affected.

After the 12 topics were found, each review received topic probabilities for each of those 12 topics, on a scale from 0 to 1, and those were presented in a data frame containing the probabilities for the 12 topics and the review ID as variables The other variables of interest from the reviews data frame were the star rating, the reviewer’s country and the ID to merge. These three variables were then merged with the topic probability data frame using an inner join. No observations were lost.

Hofstede’s cultural dimensions for the six countries of interest were then merged with the data frame containing ID, country, star rating, and the 12 topic probabilities. First, however, only the countries of interest and their codes were selected, so 'NL', 'BE', 'DE', 'DK', 'FR', 'GB' and 'SE'. This led to another 286 observations being dropped. The topics data frame was then merged with the dimensions using a left join, by country code. The final data frame to be used for modeling contained 9178 observations.

Methodology

Topic modeling using Latent Dirichlet Allocation

To find topics in the review text the chosen method was Latent Dirichlet Allocation (LDA from now on), an unsupervised machine learning model commonly used in natural language processing for the identification of themes and topics in text-based data, reviews in this case. An advantage of this method of topic identification is that it allows for multiple topics to be found in one document (Egger & Yu, 2022). This is helpful in this research as the expectation for reviews of a travel agency is that multiple aspects of the holiday experience are to be mentioned, and not one global topic.

This method follows the following assumptions. Each document is a collection of words, a "bag of words" in text analytics jargon. The order and the grammatical role of the words is therefore not considered in the model. This fits well with the translated nature of the reviews. Subtle language specific nuances can be lost when translating reviews from their original language to English. However, individual words should be accurately translated, making the bag of words approach of LDA suitable for translated reviews.

The number of topics *k* must be decided beforehand. While there is no exact science for this selection, statistics and domain knowledge can certainly help. Two statistics that can be used to test the performances of a different number of topics are model coherence and perplexity, which show differences in model fit based on the number of topics chosen.

Another way to determine the number of topics is with the *Griffiths2004*, *CaoJuan2009*, *Arun2010* and *Deveaud2014* statistics, plotted against each other.

The *Griffiths2004* (Griffiths & Steyvers, 2004) statistic is based on the log likelihood of the model, which is a measure of how well the model explains the data. The statistic is computed by running the LDA model for different number of topics and plotting the log likelihood, the optimal number of topics can be observed at the value where the log likelihood starts to level.

The *CaoJuan2009* (Cao et al., 2009) statistic measures cosine similarity between topics, which is a measure of how similar the topics are to each other. The point of this statistic is to visualize at what number of topics the similarity is minimized, which indicates that the topics are as distinct as possible from each other.

The Arun2010 method (Arun et al., 2010) uses the LDA model and a technique called singular value decomposition to find the optimal number of topics. LDA generates two distributions, namely the probability of each topic in each document (document-topic distribution) and the probability of each word in each topic (topic-word distribution). What SVD does is it decomposes the term-document matrix (a matrix that shows the frequency of each term per document, so in each review in this case) into three matrices that capture underlying patterns in the data. What the Arun2010 method does is measure how different the LDA distributions are from the SVD distributions, by checking different distributions for different numbers of topics. The best number of topics is when the divergence is minimized.

The Deveaud2014 (Deveaud et al., 2014) method is another measure of the distinctiveness of topics. The Jensen-Shannon Divergence (JSD) – a method of measuring the similarity between two probability distributions – is calculated between each pair of topics for each number of topics. The best number of topics is that which corresponds to the average JSD being maximized. This means the topics are more distinct from each other, with less overlap between them.

As *Griffiths2004* and *Deveaud2014* need to be maximized while *CaoJuan2009* and *Arun2010* need to be minimized, plotting the scores together gives a good visual representation of the best number of topics.

Since there is no clear, specific answer to the question of the optimal number of topics in LDA, multiple trials will be run with different numbers of topics. The final number of topics will be chosen based on the interpretability and perceived relevance of the different numbers of topics.

Modeling review scores based on extracted topics and cultural factors

After the topics had been extracted and the topic probabilities had been assigned to each review, modeling of the review scores followed. The ordinal rating variable was used as the dependent variable and the topics, nationality and cultural factors will be used as the independent variables. In the case of the topics, each topic related variable will show the extent to which that topic is present in a certain review.

Three models will be constructed for each of the methods to be named hereinafter and trained on the training portion of the dataset. The dependent variable will be the user rating for all three models.

- The first model will only feature the topics as independent variables.
- The second model will feature the topics as well as the nationality of the reviewer
- The third model will feature the topics and Hofstede's cultural factor scores

The following methods will be used to get closer to answering the research question of which factors affect customer satisfaction, also when taking into account national and cultural factors. New insights should emerge, as well as answers for the hypotheses formed earlier, relating to tangible and intangible factors, and certain cultural characteristics that have had clear links in previous research.

Before modeling, the data will be split into a training sample and a testing sample, with 80% of the data being used for training, and 20% for testing. The model fit will then be tested on the training dataset, and out-of-sample predictions using the model will be run on the test portion of the dataset.

Ordinal logistic regression

Ordinal logistic regression is a statistical method used to model the relationship between one or multiple predictor variables and one ordinal outcome variable (Agresti, 2010). It is similar to the classic binary logistic regression, however instead of their only being two possible outcomes, multiple outcomes of ordered nature are possible. This makes the method suitable for our outcome variable since the star rating is of ordinal nature.

Ordinal logistic regression, the outcome variable is assumed to have a continuous underlying latent variable, and the ordinal variable valued serve as thresholds along this continuous variable. The model estimates the cumulative probability of a certain category and all those below it.

The cumulative probabilities are modeled using the logistic function, like the one used in binary regression, but modified to take multiple ordered categories into consideration. The probability of an observation belonging to the category k and those below it, is given by: $logit\big(P(Y \leq k|X)\big) = \alpha_k - \beta_1 X_1 - \cdots - \beta_p X_p$, where *P (Y ≤ k | X)* is the cumulative probability of the outcome variable falling into category k or those below it, α_k is the threshold parameter for category *k*, and β_1 up to β_p represent the regression coefficients and X_1 to X_p represent the predictor variables.

One advantage of this method is its ease of interpretability which makes it suitable for a base model. The results give insights into how each of the different attributes used as independent variables affect the user star rating, showing whether the effect is positive or negative and its strength.

The links between the predictors and the outcome are easy to interpret, which is useful for the purpose of this research, namely getting insights into which factors in a planned trip affect the overall experience, and how.

The statistics used to compare the different ordinal regression models are the AIC (Akaike information criterion) which represents model fit while penalizing for the number of parameters. A lower value indicates better fit. The BIC (Bayesian Information Criterion) holds the same principle, but with stronger penalization for more parameters. AIC is often used for model selection when prediction is of interest, so to select the best model to predict the next sample. Due to its harsher penalty, BIC is better suited for hypothesis testing, and trying to figure out the true model (Aho et al., 2014). While both criteria can be used individually, using both can prove beneficial when comparing models as agreement of both can show more robustness (Kuha, 2004). Since the main purpose of this research is not of predictive nature, but to find out which model explains the data best, the BIC is more important, however this does not mean that the AIC will not be considered. The Log Likelihood is a measure of how well the data is explained by the model. It is the natural logarithm of the likelihood function, which evaluates the probability of the observed data under the model. Higher log likelihood values indicate a better fit of the model to the data. Deviance is a goodness of fit measure, similar to the residual sum of squares in linear regression. The lower the deviance, the less the model deviates from a perfect fit.

Random Forest classification

Random forests are an ensemble learning method used primarily for classification and regression tasks. In this case the method is used for classification. To understand random forests, which are an extension of the decision trees machine learning model, decision trees need to be explained first.

Decision trees

A decision tree (Kotsiantis, 2013) is a supervised learning method that can be used for classification and regression problems. In this case, the method is used for classification. The outcome variable can either be binary or have multiple classes. The aim of the method is to split the data based on characteristics with the goal of classifying it based on the target variable. A decision tree starts with a root node at the top, which then branches out into internal nodes, which then further split the data, or leaf nodes, which represent the final classification label. The point of each internal node is to split the data into homogeneous subsets. This rarely happens, so the solution is to look for the characteristic that provides the cleanest split. In order to find the best attribute for the split at each node, the Gini index is used in this case, which evaluates how well the feature will split the data into homogeneous subsets. A limitation of this model is that in the case of complex data, overly complicated decision trees can result that tend to overfit the data. This problem can be solved by pruning the tree, which consists of replacing some of the sub-trees created by internal nodes, with leaf nodes. This leads to more bias but reduces the variance that comes with overly complex trees.

According to James et al. (2013), some advantages of decision trees are their interpretability, especially in graphic displays of how the data is split at each criterion, and it is believed that they closely mirror human decision making compared to other machine learning algorithms. Other than the overfitting issue, trees can be non-robust, meaning that a small change in data can affect the final estimated tree, and they often have lower levels of predictive accuracy.

Random Forests

Random forests (Breiman, 2001) use a technique called bagging (bootstrap aggregating) which consists of sampling different subsets of data which are then used to train different decision trees. In the case of classic bagging, trees are then trained using all the predictors in the dataset. In random forests, each decision tree is trained using a random selection of variables, meaning that each tree is trained on a different set of variables, making the trees less correlated with each other opposed to bagging where one very strong feature can influence all the trees (James et al., 2013). This also helps when there is multicollinearity present in the data since not all variables are used each time. Random forests are highly effective in predictive modeling due to their ability to handle overfitting and work well with large datasets (Breiman, 2001). For classification, the final prediction is made by taking the majority vote of all the trees. The number of trees and the number of random variables used to train each tree are parameters that can be changed.

A common rule for the number of random variables in a tree is to take $\sqrt{p}$, with *p* representing the number of features. The largest model contained 18 features, meaning that $\sqrt{18} \sim 4.24$. This was rounded up to 5 and used for all three models to maintain consistency. 500 trees were chosen for the number of trees, since the higher the number of trees, the higher the stability of the forest (Probst et al., 2019).

Looking at the effects of the variables in a random forest is not as straightforward as in classic regression techniques. Due to many trees being constructed in this method, the visual representation possible in decision trees is not possible. One measure for predictor effects is variable importance. This is a measure used to determine the significance of each feature (or variable) in predicting the target variable. In this case, importance will be judged by mean decrease in impurity. This is done by adding up the amount that the Gini index decreases at splits using a given predictor, averages over all the trees (James et al., 2013). This is done for all trees in the forest and the average decrease in impurity is calculated for each predictor. Features that lead to large decreases are considered more important.

While importance is very useful to see which model attributes contribute the most, it doesn’t show us whether the effects are positive or negative. Partial dependence plots (PDPs) can be used for this purpose. Partial dependence plots are a useful tool for interpreting machine learning models, such as random forests by illustrating the relationship between one or more predictors and the outcome variables, while averaging out the effects of other predictors (Hastie et al., 2013). A PDP for a single predictor shows how the outcome variable changes as the value of that predictor varies, holding the other predictors constant at their average. For two predictors, a PDP shows the combined affect of those predictors on the outcome variable. The two predictors are plotted on the x-axis and y-axis and the outcome is presented as three-dimensional plot or a contour plot. When one of the two predictors is categorical and each category is used in combination with the other predictor, the plot is two dimensional once again, treating the effect of the non-categorical predictor as if it were calculated in a subset of that category. The plots can then be compared to see if how each categorical value influences the non-categoric predictor.

Predictions and model comparison

While the goal of this research is to gain insights into what affects the travelers’ experience, not to create a model to predict future reviews, running predictions can still be useful. Some reasons for this are for cross model comparisons, but also to identify anomalies in the data, for example.

The following statistics will be used to judge the models' prediction power.

Accuracy: a measure of a classification model's performance, defined by the following formula: $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$, representing the ratio of correctly predicted instances to the total number of instances. A high accuracy indicates that the model correctly predicts the class labels for a large proportion of the instances.

P-Value [Acc > NIR]: This p-value tests the hypothesis that the accuracy of the model is greater than the No Information Rate. The NIR is the accuracy that would be achieved by a model that always predicts the most frequent class. It is used as a baseline measure of accuracy for comparison purposes. A low p-value (typically < 0.05) indicates that there is a statistically significant difference between the model's accuracy and the NIR, suggesting that the model has predictive power beyond random chance.

Kappa: Cohen's Kappa is a measure of interrater reliability. It compares the predicted values and the actual values of the outcome variables, while also taking chance into account. A value of 0 indicates agreement by chance, and values lower than 0 indicate agreement worse than chance. The closer to 1 the value comes, the more indication that the model's prediction accuracy is considerably better than chance.

Results

This following section discusses the results and is structured as follows. First, the main results will be discussed, starting with the LDA topic modeling results. These topic modeling results are the backbone of the analysis, since to find out what attributes found in reviews influenced the score, the attributes themselves needed to be found in the text data. This will be followed by the model results, which use those topics as well as the countries and cultural dimensions to predict the star rating given by users. The main aim of these results is exploratory; to find out how each those features influence the outcome.

The ancillary results will be discussed next. These consist of the hypotheses mentioned in the literature review section and models' prediction power and discussion thereof. Hypotheses H1 and H2 are concerned with the nature of those topics, and whether they are considered tangible or intangible and how this affects the score. These hypotheses will be discussed at the end of the results section, once the topics have been discussed and placed into either a tangible or intangible category. Hypotheses H3 and H4 which look at how two cultural dimensions affect the score, based on previous research, will also be discussed then.

Main results

LDA topics

The metrics concerning the different numbers of topics (*Griffiths2004*, *CaoJuan2009*, *Arun2010* and *Deveaud2014*) and the perplexity scores plot can be found in Appendix A. The metrics that need to be maximized and those that need to be minimized show gradual increases and decreases, respectively, the higher the number of potential topics gets. However, there were some jumps at 8 and at 12 topics. The perplexity score kept decreasing with an increasing number of topics. This is expected, however there was no clear elbow present in the curve, or a clear steep drop, so this metric proved not to be very useful. Due to these factors, three iterations were computed, with 8, 10 and 12 topics. The final choice was for 12 topics due to the increased specificity and interpretability. These topics and their interpretation based on the words in the topics can be found in table 1 below. The word concentrations per topic can be found in figure B1 in Appendix B. To give insight into the other possible amounts of topics, the concentrations of words for 8 and 10 topics can also be found in Appendix B, in figures B2 and B3. From now on, topics 1 through 12 will be referred to as T1 through T12, except during interpretation, when their actual meaning is of interest.

Table 1

Topic definitions and important words in those topics

Topic number and interpretation	Terms
Topic 1: Skiing and Accommodation	ski, apart, accommod, holiday, pass, site, use, easi, price, includ
Topic 2: Family and Place	onli, star, stay, peopl, place, noth, famili, children, littl, far
Topic 3: Pricing and Extra Costs	price, pay, offer, websit, extra, person, euro, made, cost, review
Topic 4: Room and Food	room, food, staff, clean, pool, beach, everi, bed, restaur, inclus
Topic 5: Guidance and Tours	guid, inform, help, tour, time, app, need, return, car, experi
Topic 6: General Feedback	day, one, got, onli, first, told, next, two, arriv, ask
Topic 7: Refunds and Money	money, cancel, back, now, refund, voucher, receiv, custom, due, trip
Topic 8: General Complaints	becaus, want, say, don, come, back, know, see, someth, one
Topic 9: Past Experiences	year, friend, went, top, arrang, definit, problem, organ, servic, satisfi
Topic 10: Customer Support	call, phone, contact, email, custom, servic, answer, via, respons, tri
Topic 11: Travel Logistics	bus, airport, hour, flight, arriv, transfer, check, wait, befor, departur
Topic 12: Travel and Service Quality	trip, travel, time, servic, compani, last, sever, experi, use, turkey

Ordinal logistic regression results

Since topic assignment is based on the probability of each topic appearing in each review, this means that the sum of the probabilities is always one. Due to this attribute, the issue of multicollinearity arises. To fix this, a simple approach is to exclude one topic from the analysis. The consequence of this is that the effect of each topic in the model can only be interpreted in relation to the excluded topic.

To choose the topic that would serve as a reference, the topic definition and relevance was taken into consideration. Looking at the topics and their definition, there are two topics that don't convey much information about travel attributes. Topic T6 provides some general feedback, with words such as *day, one, got, onli, first, told, next, two, arriv, ask.* These words don't provide much information about what is going on on the trip and in the review itself. The same holds for topic T8, with words such as *becaus, want, say, don, come, back, know, see, someth, one*, that don't convey much information about the trip itself. The cluster of words does however convey a more negative sentiment than topic T6.

A general topic that doesn't provide clear information is therefore best used as a reference topic. Since these general words were split into two topics, the solution was to combine them by adding up the probabilities, to create topic T13, which we just described as "general" and to use it as the reference. Topics T6 and T8 are dropped, and topic T13 is used in the analysis, but also dropped since it's the reference topic.

The results of the three models can be found in table 2. On significance, most topics are significant in all three models, the only exceptions being topic 10 which is insignificant in all instances, and topic 2 which is only significant in models 2 and 3, when adding extra variables (the countries in model 2 and Hofstede's cultural dimensions in model 3). As for the effect sizes of the topics, they do not change much between models and there are no instances of the nature of the effect (positive or negative) changing.

As for the interpretation of the effects, the largest positive effect belongs to topic 9, which focuses on past experiences and therefore refers to repeat customers. Other strong positive effects belong to topics T1, T5 and T12. T1 refers to skiing and accommodations, T5 refers to guidance, tours, and information at the location, and topic T12 refers to general service quality. Topic T4 and T11 also have a slightly smaller positive effect. These topics refer to hotel attributes such as room and food in the case of T4, and to travel logistics in the case of T11.

The only topic with a negative effect in this case is topic T7, with a very small effect, and referring to refunds and money.

A curious outcome is that all but one topic have positive effects. An important note, as mentioned previously, is that these effects are in reference to removing topic 13 which represents general feedback. What can be insinuated from this is that topic 13 had a negative effect, as removing it increases positivity so blatantly. The only case in which that was not the case was for topic T7, relating to money issues. This is to be expected as people having money issues with the company would not be expected to be very positive about the company. This also means that the effect of mentioning money issues is much more negative than represented by the model.

In the case of the countries present in model 2, it is important to note that these countries are in reference to The Netherlands, which is the reference country. The only significant effects in this case are for Belgium and Denmark (negative), and for the UK (positive). This means that reviews posted from these countries tend to be more negative in the case of Belgium and Denmark and more positive in the case of the UK. The effect sizes, however, are very small, meaning that the differences are not that extreme.

Concerning model 3 which includes Hofstede's cultural dimensions, the only insignificant effect corresponds to mas, meaning masculinity. PDI, IDV and LTOWVS have slight positive effects, of roughly the same size. Same holds for UAI and IVR, but with negative effects of similar sizes. This means that cultures defined by more power distance, individuality and long-term orientation tend to give higher reviews, and cultures defined by uncertainty avoidance and indulgence tend to give lower reviews.

Another important note for the effect sizes is that for regression analysis, the effect is tied to a one unit increase in the independent variable. In the case of both the topics and the cultural dimensions, these variables are expressed in a value between 0 and 1. A one unit increase is therefore impossible. In this case, a one hundredth increase is how they should be judged, meaning that the effect sizes need to be divided by 100. Taking T1 as an example, a 1% or 0.01 increase would correspond to a 0.22 increase int the logarithmic odds of getting into a higher rating category, a lot smaller than 22.03. And in the case of the cultural dimensions, taking PDI as an example, a 1% increase would lead to a 0.03 increase in the logarithmic odds of going to the next category. This is however more in line with the scale of the country effects, which are based on categorical - and therefore dummy - variables that have a 1 unit increase from 0 to 1.

Comparing the models based on AIC, BIC, Log-likelihood, and deviance, an improvement can be seen from model 1 to models 2 and 3. Models 2 and 3 have the same exact values for these metrics, meaning that using countries or cultural dimensions leads to the same outcome and the model used should be based on the desires of the research.

Table 2
Ordinal regression results

	Model 1	Model 2	Model 3
T1	22.03 (1.05) ***	19.65 (1.10) ***	19.65 (1.10) ***
T2	2.27 (1.26)	2.60 (1.26) *	2.60 (1.26) *
T3	3.42 (1.26) **	3.08 (1.27) *	3.08 (1.27) *
T4	10.43 (0.98) ***	10.81 (0.99) ***	10.80 (0.99) ***
T5	21.45 (1.19) ***	22.00 (1.19) ***	22.00 (1.19) ***
T7	-3.81 (1.14) ***	-4.83 (1.15) ***	-4.83 (1.15) ***
T9	51.20 (1.35) ***	50.60 (1.38) ***	50.60 (1.38) ***
T10	2.23 (1.19)	2.30 (1.20)	2.29 (1.20)
T11	8.32 (1.08) ***	8.30 (1.08) ***	8.30 (1.08) ***
T12	21.89 (1.28) ***	23.30 (1.32) ***	23.30 (1.32) ***
countryBE		-0.35 (0.08) ***	
countryDE		0.57 (0.13) ***	
countryDK		-0.22 (0.07) **	
countryFR		0.10 (0.07)	
countryGB		0.50 (0.11) ***	
countrySE		-0.22 (0.17)	
pdi			3.15 (0.88) ***
idv			3.21 (0.97) ***
mas			-0.24 (0.22)
uai			-4.36 (0.88) ***
ltowvs			1.87 (0.48) ***
ivr			-5.12 (0.75) ***
1|2	10.31 (0.68) ***	10.13 (0.69) ***	9.32 (0.83) ***
2|3	10.78 (0.68) ***	10.60 (0.69) ***	9.80 (0.83) ***
3|4	11.26 (0.68) ***	11.09 (0.69) ***	10.28 (0.83) ***
4|5	12.35 (0.69) ***	12.19 (0.69) ***	11.38 (0.84) ***
AIC	16640.67	16571.61	16571.61
BIC	16737.29	16709.64	16709.64
Log Likelihood	-8306.33	-8265.80	-8265.80
Deviance	16612.67	16531.61	16531.61
Num. obs.	7344	7344	7344

Random forest results

An advantage of random forests is that the model can deal better with multicollinearity than classic regression models. Because of that feature, no variables were excluded from the model. Looking at the random forest variable importance, there aren't any differences in topic importance between the three models (only topics, countries added, cultural dimensions added). The variable importance for model 3 can be found in figure 1. The variable importances for models 1 and 2 can be found in figures C1 and C2 in appendix C.

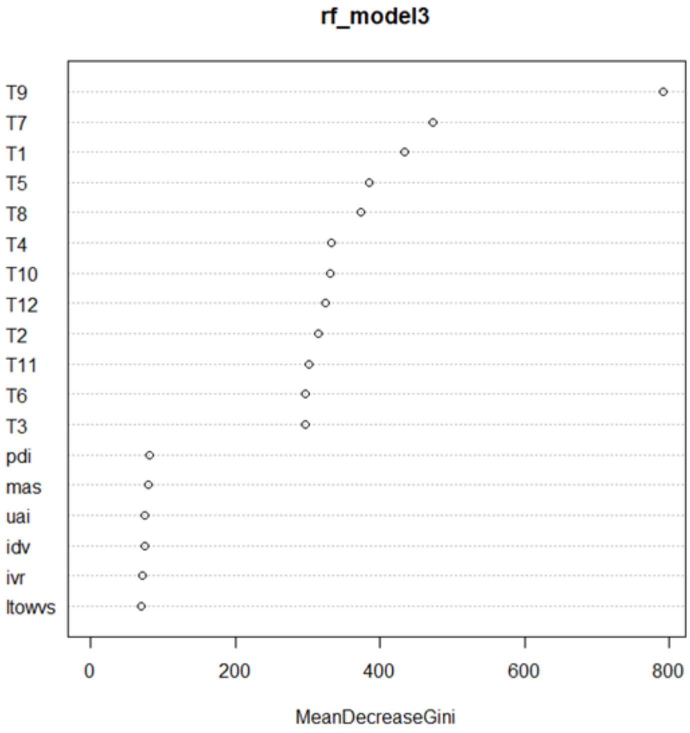

Figure 1: Variable importance for random forest model 3

The most important topic by far was T9. The other important topics were T7 and T1 and afterwards all topics had similar importances. The countries and the different cultural dimensions did not have much importance on the outcome. This corresponds to the previous models, showing that topics that mention repeat experience had a high influence on rating, as well as the topic concerning money. T1 and T5 were also quite important in this case, and they were followed by topic 8 which was excluded from the previous analysis. Topic 6, also excluded, was one of the last ones in the importance list. This explains the reference category being negative as it looks like the general topic with a more negative sentiment in T8 slightly outmatched the neutral one in T6.

Figure 2 shows the partial dependence plots for each variable, which illustrates how each independent variable affected the outcome variable. In the case of topics T9, T1 and T12, the effects are clearly positive, just like in the previous models, and topics T4 and T5 have positive effects, but lower than is the case for the previous three. The rest of the topics lean more negatively. While this is a change from the previous model which contained all but one positive topic, it can be explained by the negative effects of the reference topics T6 and T8.

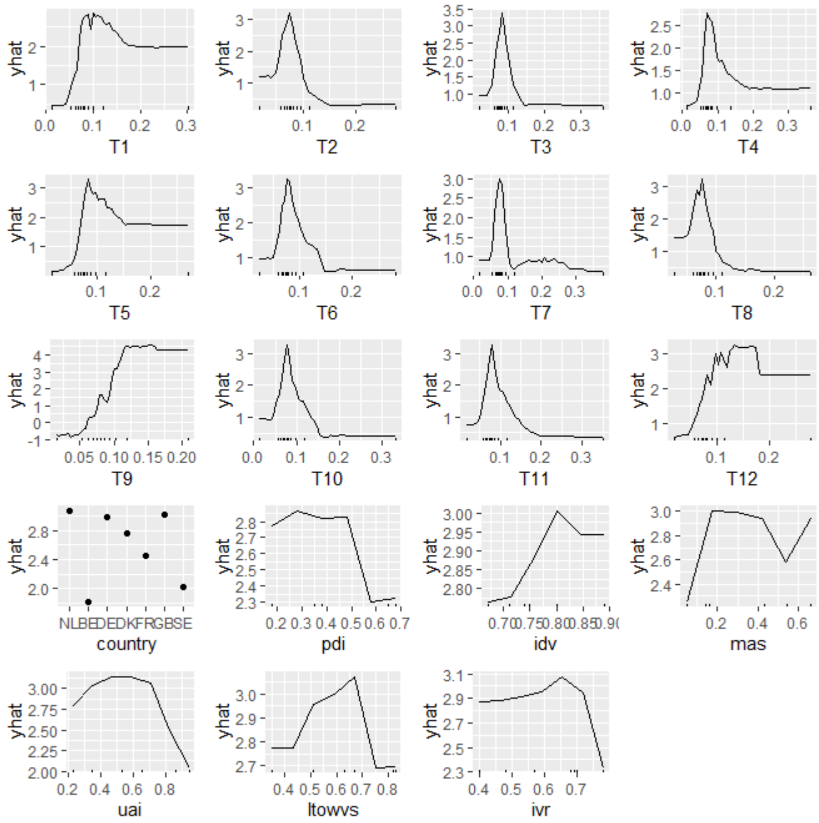

Figure 2: Partial dependence plots

For the effects of different countries, Belgium and Sweden seem to have more negative effects than the Netherlands. Great Britain and Germany are roughly on the same level, and Denmark is slightly lower.

For the cultural dimensions, PDI shows a negative drop the larger it becomes, as is the case for UAI and IVR. IDV has a mostly positive effect, as does mas. LTOWVS also has a positive effect until a sudden drop when it gets large.

Important to note is that the countries and cultural dimensions had a very low importance in comparison to other variables, meaning that these effects aren’t very important.

One other aspect that was of interest in this paper were interactions between each topic and the country or culture of the review. Due to computational limitations, and the fact that the importance of either countries or cultural dimensions was low in the models, the decision was to use the countries for interaction effects. These were modeled using partial dependence plots with two variables, each topic and the country. These plots can be found in figures D1 through D12 in Appendix D. The tables show that there aren't clear differences in topic effects per different countries, with all plots following very similar curves. There are some exceptions that deserve a mention, not in the direction of the curves but in the effect size.

Ancillary results

Hypotheses discussion

For discussing positive or negative effects, the results of the random forest models will be taken into consideration. This is because, as mentioned previously, the ordinal regression leaving one variable out led to some uncertainty about the actual effect of certain topics, since they were in reference to a mostly negative reference topic.

The hypotheses H1 and H2 focused on how tangible and intangible aspects affect review score, with the expectation that tangible aspects have a negative influence, and intangible ones a positive one. While a binary approach of looking at topics as either tangible or intangible should mean that each topic can be placed into one of those categories, not all topics are easily placeable. The clearly positive topics were T1, T9 and T12. T1. T9 and T12 are obviously intangible topics, since T9 refers to past experiences and T12 refers to the travel and service quality which can't exactly be measured. Topic T1 is also mostly intangible, although not as clearly as previous topics. T5 also had a slight positive effect. Guided tours are somewhere in between tangible and intangible. However, for this research they are considered tangible as they are a part of the holiday experience and not an object that was part of the holiday. T4 also had a positive effect, however this is a very tangible topic, referring to room and food quality.

In the case of the negative effects, T3 and T11 are tangible topics, concerning pricing and extra costs, and travel logistics. T2, T7 and T10 on the other hand, concerning family preferences, refunds and money issues, and customer support lean towards being intangible as the experience derived from these topics depends per customer.

Based on these facts, it is tough to accept or reject hypotheses 1 and 2, as there is evidence on both sides, but not enough to draw a firm conclusion. While there were some intangible topics with negative effects, most of the positive effects were in fact caused by intangibles, meaning that H1 is more strongly supported by the data than H2.

Hypothesis H3 can clearly be rejected based on this research, as both models showed a positive effect for individualism. Same can be said for hypothesis H4, as both models showed negative effects for uncertainty avoidance.

It would still be advisable to conduct further research into those cultural dimensions in this context, for the effect size and importance was not very high for either of the dimensions.

Predictions

Ordinal regression models

Based on the in-sample prediction statistics present in table E2 in Appendix E, all models have a prediction accuracy around 0.59, with slight improvements between model 1 (0.5927) and models 2 (0.5967) and 3 (0.5959). The same holds for Cohen’s Kappa of 0.3552 in model 1 and 0.365 in models 2 and 3. All models had a significant P value.

For the out of sample predictions, model 1 has a prediction accuracy of 0.5862, and a Cohen’s Kappa of 0.3552. Models 2 and 3 are completely identical in their scores across the board, from their accuracy of 0.5927 and Cohen’s Kappa of 0.365. All models had a significant P-value. This shows that there is no difference in out-of-sample prediction power between models 2 and 3 and that using either countries or cultural dimensions in this context makes no difference. These statistics can be found in table E4 in Appendix E.

The confusion matrices used for these prediction statistics can be found in Tables E1 and E3 in Appendix E. What these tables show is that the models failed to predict the non-extreme outcomes and only managed to predict in a binary manner, only predicting scores of 1 and of 5.

Random forest models

The in-sample prediction accuracy for the first random forest model is 0.9721 and sees an improvement to 0.9843 in model 2 and 0.9845 in model 3. Cohen’s Kappa goes from 0.9608 in model 1 to 0.978 and 0.9783 in models 2 and 3, respectively. The P-value is significant for all models. The in-sample confusion matrix in table F1 in Appendix F shows that as opposed to the previous ordinal regression models, the random forest models also manage to predict the middle categories between 1 and 5. The rest of the metrics can be found in table F2 of Appendix F.

The statistics for the out-of-sample predictions are much lower than in-sample. Model 1 only has an accuracy score of 0.5921 which improves to 0.5965 and 0.5927 for models 2 and 3, respectively. P-value is significant for all models. Cohen’s Kappa is 0.3925 for model 1, slightly improving to 0.3939 in model 2. For model 3, this score is lower than in model 1 at 0.388. Looking at the confusion matrix in table F3 in Appendix F, the predictions once again end up on the lowest and highest categories, though not as rigidly as in the case of the ordinal regression where there were no predictions at all for the middle categories. The rest of the metrics can be found in table F4.

Model comparison

The model results were very similar both in the cases of the independent variables used for the models and the two modeling approaches (ordinal regression and random forest). In the case of the ordinal regression, the difference between choosing the model containing countries and the model containing cultural dimensions depends on what the goals in question are, as the performance did not differ too much. In the case of the random forests, the differences weren’t very large, but once again model 2 seemed to be the best fit. Also, the fact that interactions could be plotted without having to deal with excessive computational power and time was an advantage.

Between the modeling types, the random forest gives a clearer view of how each variable affects the outcome. This is because this modeling technique can deal with multicollinearity and no coefficients needed to be dropped, which showed that there were coefficients that had negative effects which seemed positive in the first place when looking at the ordinal regression results. Another reason why the random forest model is preferable is its predictive power. While the out-of-sample predictions were similar between the two modeling techniques, the in-sample prediction was a lot better using the random forest. Not only that, but the models also demonstrated to be capable of predicting other categories than the lowest and the highest.

Conclusion

The goal of this research was to find what attributes of a vacation affected the overall experience, expressed through user generated reviews and star rating. Since the company used for this specific research was Sunweb, a company internationally active in multiple markets, the research also contained a national and cultural component that would give insights into how different nationalities, cultures, and therefore markets give weight to different attributes of the vacation experience. Since the study focused on a tour operator, it added research in a particular area of larger travel research that has been overlooked so far. Due to the mixed nature of the study, containing hypotheses based on previous research, but also an exploratory part to find out what tour operators, specifically Sunweb in this case, should focus on. This section will look back at the research questions of the paper as well as on the managerial implications that resulted from the exploration.

Research questions

What attributes influence the overall experience and satisfaction levels of customers in the travel industry when booking package holidays?

According to this research, using LDA topic modeling and Sunweb reviews, the attributes found in the reviews were: skiing and accommodation, family and place, pricing and extra costs, room and food, guidance and tours, general feedback, refunds and money, general complaints, past experiences, customer support, travel logistics, travel and service quality. General feedback and complaints can be ignored in the case of this question, as they are of a very general nature. how these attributes affect the ratings will be discussed in the next section on managerial implications.

How do national and cultural factors influence the satisfaction levels of customers in the travel industry when booking package holidays?

According to the model results, the importance of the national and cultural factors in determining the user ratings were very small. While there was some variation which is discussed in the results section, the size and importance of the effects were too small to make any general claims about it.

How does the influence of those attributes vary across national and cultural segments?

Unfortunately, this question could not be answered to satisfactory extent. For one, due to computational issues, interactions between cultural dimensions and the attributes mentioned above were not computed. As for national factors, or the interactions between countries and topics in the random forest PDPs, there was no clear evidence of any serious variation, with very similar patterns in each country. To add to that, the variable effects in the case of logistic regression, and importance in the case of the random forest, were very small for those factors.

Managerial implications

This part will focus on how travel agencies and tour operators can use the findings of this paper, regardless of how they compare to previous studies. It should be noted that the findings strongly apply to Sunweb since their data was the one used in the research.

The strongest importance was given to the topic mentioning past experiences. This means that repeat customers are very satisfied with the service. This makes sense intuitively, otherwise they wouldn't return. However, this still provides some valuable insight to the company. For one, it means that return customers are not often disappointed to return, which shows consistency of service. For the company to present better reviews, repeat customers should be targeted and stimulated to leave reviews as this would bump up the score and improve brand image. This ties together with the topic focusing on ski and accommodation, showing that ski holidays are a very strong product offered by Sunweb, and therefore repeat customers of ski holidays should be a very strong focus. The guidance and information during vacation also showed a positive effect. This should strongly be promoted in Sunweb's marketing, showing that they care for customers while on location, not just before they have completed a transaction. In the case of room and food attributes, the effect was positive. This shows that customers are satisfied with these important attributes of their experience. To keep this satisfaction, an effort should be made to match customers with their desired vacation type. This can be done through content making clear what they offer, but also other possible marketing strategies such as interactive booking procedure based on the customer's preferences.

Looking at the negative aspects, the first one that jumps out is family experience. This does not necessarily mean that Sunweb is not fit for families, especially with the wide selection of vacations offered. An action to improve this aspect would be to make sure family-oriented customers actually book family-oriented accommodations. This can be done through better highlighting family pages, or it could be a part of the previously mentioned marketing idea of an interactive procedure, where families with children can show their preferences. Some other negative topics such as travel logistics and customer support are tough to judge. There can be problems with travel logistics that are out of the company’s control that can lead to frustration either way. The same can be said for negative opinions on customer support, as appealing to support is most likely already caused by a problem, not a positive experience. Therefore, a recommendation would be to investigate negative reviews with high scores on those topics and look into what the specific problems were. For companies that have enough data, this could also be done with another topic modeling analysis. When it comes to negative sentiments caused by topics concerning pricing and extra costs, these should be a top priority. Transparency into extra costs is very important, early in the booking process. As is the case for refund policies, they should be communicated very clearly. Although, once again, a look into the negative reviews high in those topics is recommended, as it is also possible that some customers do not pay attention to the information.

Lastly, the fact that the general topics are mostly negative can be explained intuitively. An unhappy customer that doesn’t want to write a fleshed-out review is more likely to drop a quick negative general review, than a happy customer with no complaints.

Discussion

Limitations & future research

While this study added valuable insights into the field of consumer behavior, focusing on cross cultural differences in the travel industry, it is nevertheless subject to certain limitations. Acknowledging these constraints, the implication of the research results as well as possibilities for future research can be better understood.

An obvious limitation is the use of data pertaining to one specific tour operator, namely Sunweb. The company is active in multiple countries and therefore has customers of multiple nationalities and cultures leaving reviews. However, these reviews still reflect the overall sentiment towards this company, and while the cultural factors are used as a moderator, a different company with differences in offers, price range, or brand identity, could attract different topics of focus in reviews, all other things equal. A simple idea for future research would be reapplying the techniques used in this paper to another travel company with activities in multiple countries. This framework could also be expanded to multiple companies in order to gain market research insights for tour operators, on their own operations as well as on their competitors'.

While Hofstede's dimensions are an established framework to analyze culture, it is still an arbitrary method of conceptualizing an intangible concept and therefore influences the findings in this research. Different methods of defining culture, such as Trompenaars cultural dimensions, or the GLOBE method might have led to different results and would thus be interesting to apply in a redo of this research. Another limitation of the cultural component is the assumption of cultural homogeneity. Cultures can differ between different regions of the same country. However, most countries used for this research are not of very large scale. Using the same scores for a country like the United States would be a bigger limitation, rather than for individual European countries. Also, cultures evolve over time, and since the Hofstede method has been established for quite some decades, the method is sure to have missed some cultural changes.

For ease of research and cross inter country analysis, all reviews were translated to English. While the algorithm used considers cultural nuances, there are differences between languages that cannot be captured by an algorithmic translation. Another thing to consider is the data cleaning process, consisting of stemming, removing stop words, and the specific list of words mentioned previously. These affect the LDA analysis, and other choices in the data cleaning process could have led to different results.

Another aspect of the data is the specific countries used in this research. While there certainly are cultural differences between The Netherlands, Belgium, France, Germany, The UK, Sweden and Denmark, these are all Western European countries, or at least considered Western countries, which certainly share some characteristics in terms of expectations. An interesting follow-up would be to research this topic while including countries from different cultural regions, such as the Middle East, Latin American or South-East Asia.

The data was also unbalanced, with certain countries having many more reviews than others, as described in the data section. This reflects Sunweb's presence in those countries but does also bring a limitation. Since the number of reviews in the scope used wasn't large enough to sample equal amounts, a larger number of reviews, possibly for a different travel company or with a broader time scope could be used to balance out the presence of different countries in the overall dataset. Another interesting application would be to add a time element to the analysis, to reflect how reviews have changed over time. This would reflect what the company had improved on or what aspects have fallen off or still need improvement.

The topic modeling method also poses some limitations as this was a research choice, and other methods could highlight different aspects. An example of a limitation of this method is the presence of multiple topics and therefore the topic probability distribution. This is not very realistic, not all topics appear in each review, even if the measure in which they appear is very small. However, topic modeling methods that only assign one topic per review would also not be very desirable as a lot of information would be lost. A solution for a further research paper that still uses LDA is hard assignment. A way to do this would be to decide on a threshold, and if the probability is higher than said threshold, the topic is assigned to the review. This would lead to some topics being assigned while others are not, and each topic would turn into a dummy variable.

The nature of the data also poses a limitation. Data based on probabilities is known as compositional data and what helps analysis are certain transformations that can make the proportions easier to interpret. This would help in regression analysis, as opposed to the dropped variable approach used in this paper.

On that note, the models used are also just two out of many options, and other methods could also be interesting to use in research. In the predictions, a very clear limitation of the ordinal regression was that it only predicted the lowest and the highest category. The model is very easy to implement and interpret, so a way to deal with this limitation could be to transform the outcome variable into a binary positive/negative variable. This can be done multiple ways, clustering the lowest two and highest two and removing the middle category, including the highest category in the positive or negative category, or even going for a very extreme approach of saying that the only positive review is that of 5 stars, and everything else is negative. While each of these methods have upsides and downsides, it is up to the management and their purpose and goals.

Another factor was the absence of validation techniques when running the ordinal regression and random forest models. The reason this was not included was mostly due to time constraints, and the bulk of the time being spent on the creation of the topics, such as the extent to which the vocabulary was pruned and the decision on the number of topics, with multiple tests having been run.

The last limitation lies in the fact that this paper only used quantitative analysis. A qualitative based research paper with interviews, asking travelers to expand on their experiences would generate a lot of insights into this topic. Going back to the limitation of language translation, conducting interviews with multiple interviewers who speak multiple languages could generate interesting insights that could be translated afterwards, only to communicate the results.

In conclusion, as in any research, many limitations are present in this paper, meaning that many different follow-ups could be conducted, with changes from the data used, the cleaning procedure, models and interpretation. From a business perspective, a competitor analysis could be very useful, using different approaches or even copying the methods in this paper. For example, this paper used roughly 9000 reviews, and this for all Sunweb publications. Tui, a much larger travel company that also specializes in package trips has more than 94000 reviews on Trustpilot, and that is only for the United Kingdom. Analyzing this data would yield many more insights into package trips reviews. However, it would also require more time, resources and processing power, and it is better suited for a longer-term project.

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636. https://www.jstor.org/stable/43495189

Agresti, A. (2010). Analysis of Ordinal Categorical Data. In *Wiley series in probability and statistics*. https://doi.org/10.1002/9780470594001

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Lecture notes in computer science* (pp. 391–402). https://doi.org/10.1007/978-3-642-13657-3_43

Berezina, K., Bilgihan, A., Çobanoğlu, C., & Okumuş, F. (2015). Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews. *Journal Of Hospitality Marketing & Management*, *25*(1), 1–24. https://doi.org/10.1080/19368623.2015.983631

Book, L. A., Tanford, S., Montgomery, R. J. V., & Love, C. (2015). Online Traveler Reviews as Social Influence: Price Is No Longer King. *Journal Of Hospitality & Tourism Research*, *42*(3), 445–475. https://doi.org/10.1177/1096348015597029

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011 Chatterjee, S., & Mandal, P. (2020). Traveler preferences from online reviews: Role of travel goals, class and culture. *Tourism Management*, *80*, 104108. https://doi.org/10.1016/j.tourman.2020.104108

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Ding, K., Choo, W. C., Ng, K. Y., & Ng, S. I. (2020). Employing structural topic modelling to explore perceived service quality attributes in Airbnb accommodation. *International Journal Of Hospitality Management*, *91*, 102676. https://doi.org/10.1016/j.ijhm.2020.102676

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*. https://doi.org/10.3389/fsoc.2022.886498

Eighty percent of TripAdvisor users read at least six to 12 reviews before choosing a hotel. (2014, 11 februari). MediaRoom. https://tripadvisor.mediaroom.com/2014-02-11-Eighty-percent-of-TripAdvisor-users-read-at-least-six-to-12-reviews-before-choosing-a-hotel

Gavilán, D., Avello, M., & Martínez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, *66*, 53–61. https://doi.org/10.1016/j.tourman.2017.10.018

Geertz, C. (1977). *The interpretation of cultures*. Basic Books.

Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. In Information and communication technologies in tourism 2008 (pp. 35-46). Springer, Vienna.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hastie, T., Tibshirani, R. J., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. http://catalog.lib.kyushu-u.ac.jp/ja/recordID/1416361

Hofstede, G. (1980). *Culture's consequences: International Differences in Work-Related Values*. SAGE Publications, Incorporated.

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in context. *Online Readings in Psychology And Culture*, *2*(1). https://doi.org/10.9707/2307-0919.1014

Hofstede, G., & Bond, M. H. (1984). Hofstede's culture dimensions. *Journal Of Cross-cultural Psychology*, *15*(4), 417–433. https://doi.org/10.1177/0022002184015004003

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. In *Springer texts in statistics*. https://doi.org/10.1007/978-1-4614-7138-7

Jia, S. (2020). Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews. *Tourism Management*, *78*, 104071. https://doi.org/10.1016/j.tourman.2019.104071

Kotler, P., & Keller, K. L. (2012). *Marketing Management 14th ed*. https://slims.bakrie.ac.id/textbook/index.php?p=show_detail&id=156

Kotsiantis, S. (2013). Decision trees: a recent overview. Artificial Intelligence Review, 39(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kwon, H., Ban, H., Jun, J., & Kim, H. (2021). Topic Modeling and Sentiment Analysis of Online Review for Airlines. *Information*, *12*(2), 78. https://doi.org/10.3390/info12020078

Leon, R. (2019). Hotel's online reviews and ratings: a cross-cultural approach. *International Journal Of Contemporary Hospitality Management*, *31*(5), 2054–2073. https://doi.org/10.1108/ijchm-05-2018-0413

Litvin, S. W. (2019). Hofstede, cultural differences, and TripAdvisor hotel reviews. *International Journal Of Tourism Research/The International Journal Of Tourism Research*, *21*(5), 712–717. https://doi.org/10.1002/jtr.2298

Litvin, S. W., Crotts, J. C., & Hefner, F. (2004). Cross-cultural tourist behaviour: a replication and extension involving Hofstede's uncertainty avoidance dimension. *International Journal Of Tourism Research/The International Journal Of Tourism Research*, *6*(1), 29–37. https://doi.org/10.1002/jtr.468

Luo, J. M., Vu, H. Q., Li, G., & Law, R. (2020). Topic modelling for theme park online reviews: analysis of Disneyland. *Journal Of Travel & Tourism Marketing*, *37*(2), 272–285. https://doi.org/10.1080/10548408.2020.1740138

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews. Data Mining And Knowledge Discovery/Wiley Interdisciplinary Reviews. Data Mining And Knowledge Discovery*, *9*(3). https://doi.org/10.1002/widm.1301

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and Tourism Online Reviews: Recent Trends and Future Directions. *Journal Of Travel & Tourism Marketing*, *32*(5), 608–621. https://doi.org/10.1080/10548408.2014.933154

Solomon, M. R., Bamossy, G., Askegaard, S., & Hogg, M. K. (2016). *Consumer Behaviour PDF eBook: A European Perspective*. Pearson Higher Ed.

Stamolampros, P., Korfiatis, N., Kourouthanassis, P. E., & Symitsi, E. (2018). Flying to Quality: Cultural Influences on Online Reviews. *Journal Of Travel Research*, *58*(3), 496–511. https://doi.org/10.1177/0047287518764345

Stávková, J., Stejskal, L., & Toufarová, Z. (2008). Factors influencing consumer behaviour. *Zemědělská Ekonomika*, *54*(6), 276–284. https://doi.org/10.17221/283-agricecon

Tan, P., Steinbach, M. M., & Kumar, V. (2008). Introduction to Data Mining. In *Routledge eBooks* (pp. 151–206). https://doi.org/10.4324/9780080878096-12

Vermeulen, I., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, *30*(1), 123–127. https://doi.org/10.1016/j.tourman.2008.04.008

Xie, L., Guan, X., Cheng, Q., & Huan, T. (2020). Using customer knowledge for service innovation in travel agency industry. *Journal Of Hospitality And Tourism Management*, *45*, 113–123. https://doi.org/10.1016/j.jhtm.2020.08.001

Yang, Y., Park, S. W., & Hu, X. (2018). Electronic word of mouth and hotel performance: A meta-analysis. *Tourism Management*, *67*, 248–260. https://doi.org/10.1016/j.tourman.2018.01.015

Appendix A

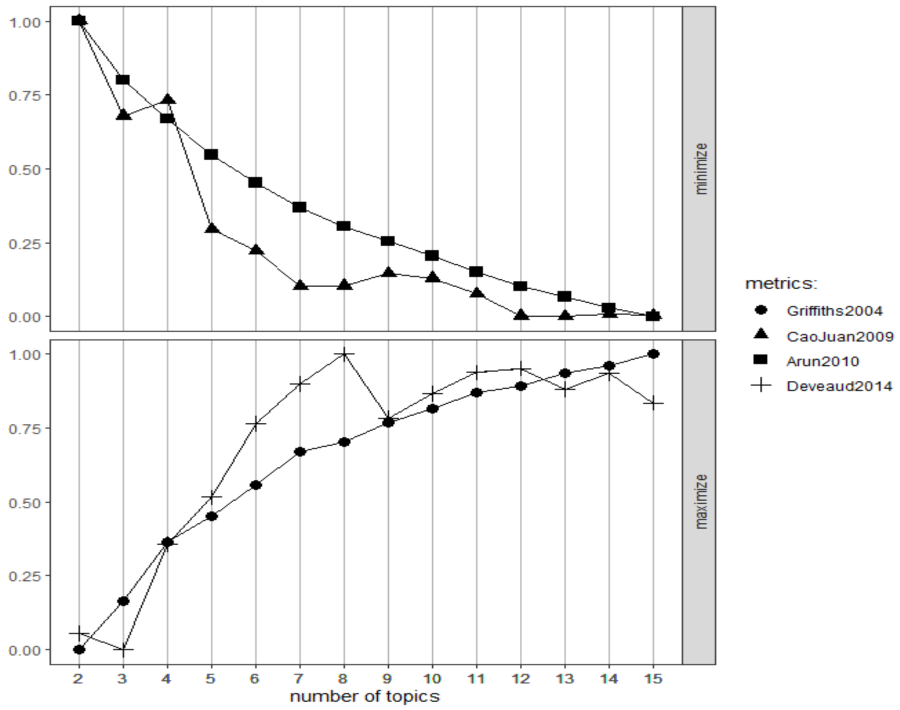

Figure A1: Metrics for different numbers of topics

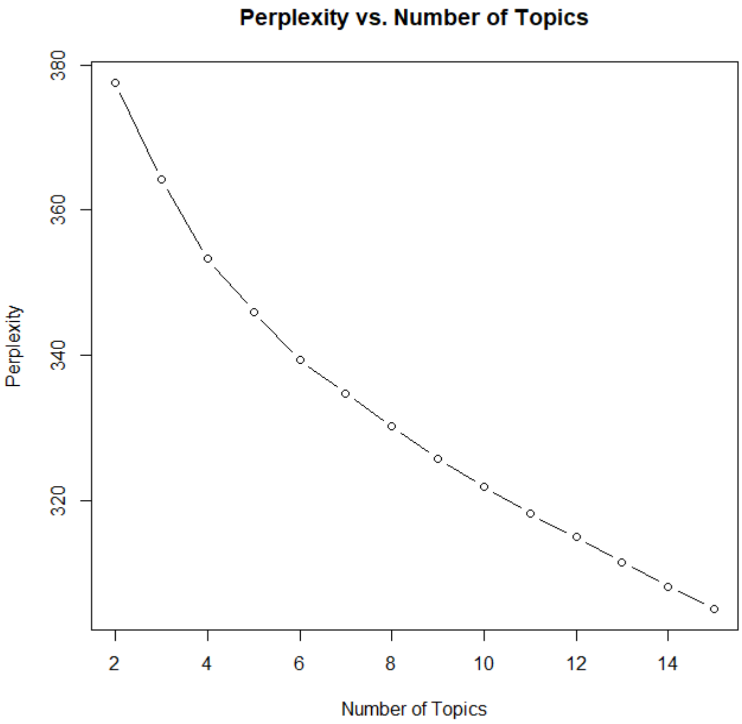

Figure A2: Perplexity for different numbers of topics

Appendix B

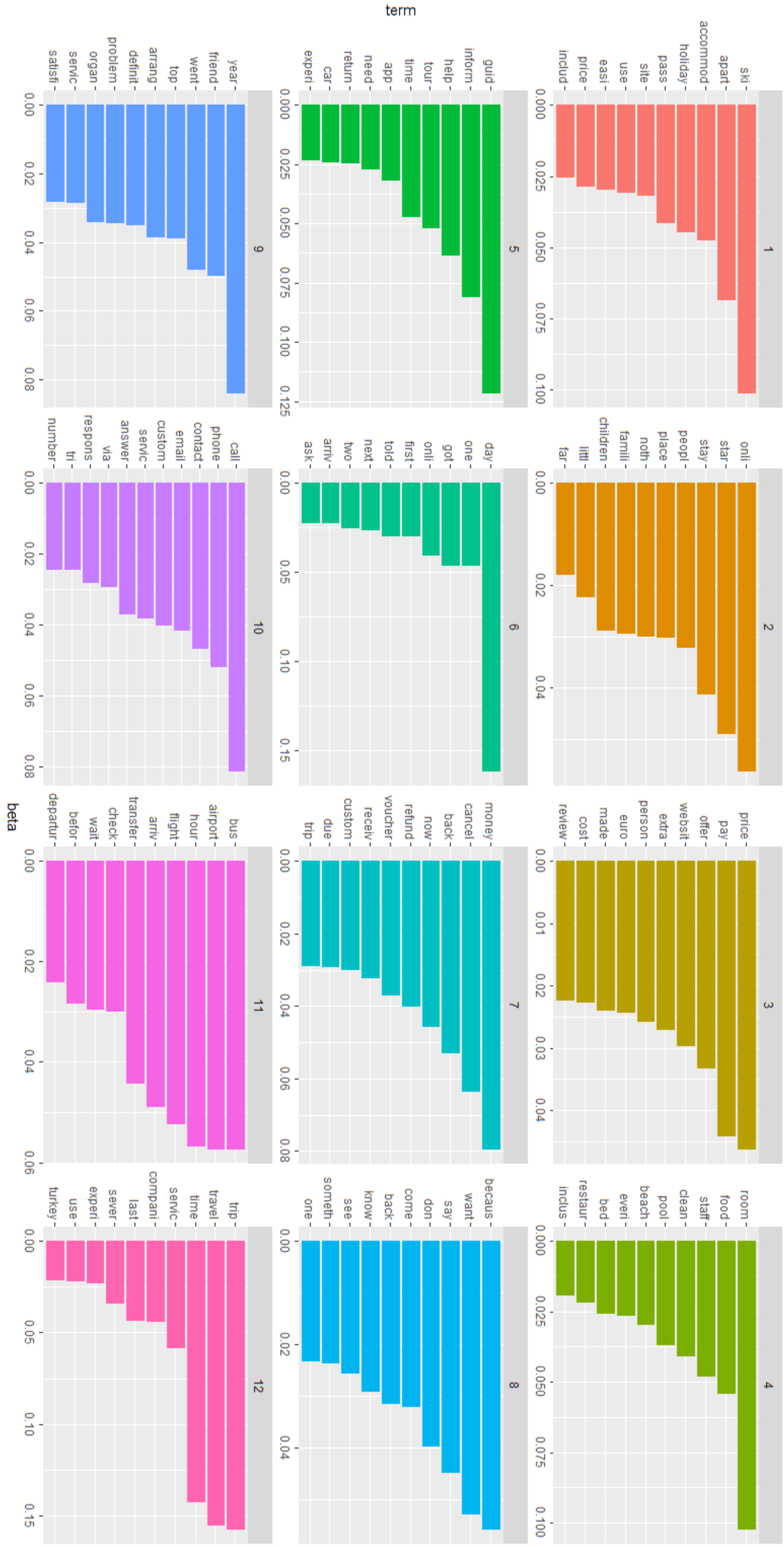

Figure B1: Word concentrations per topic for 12 topics

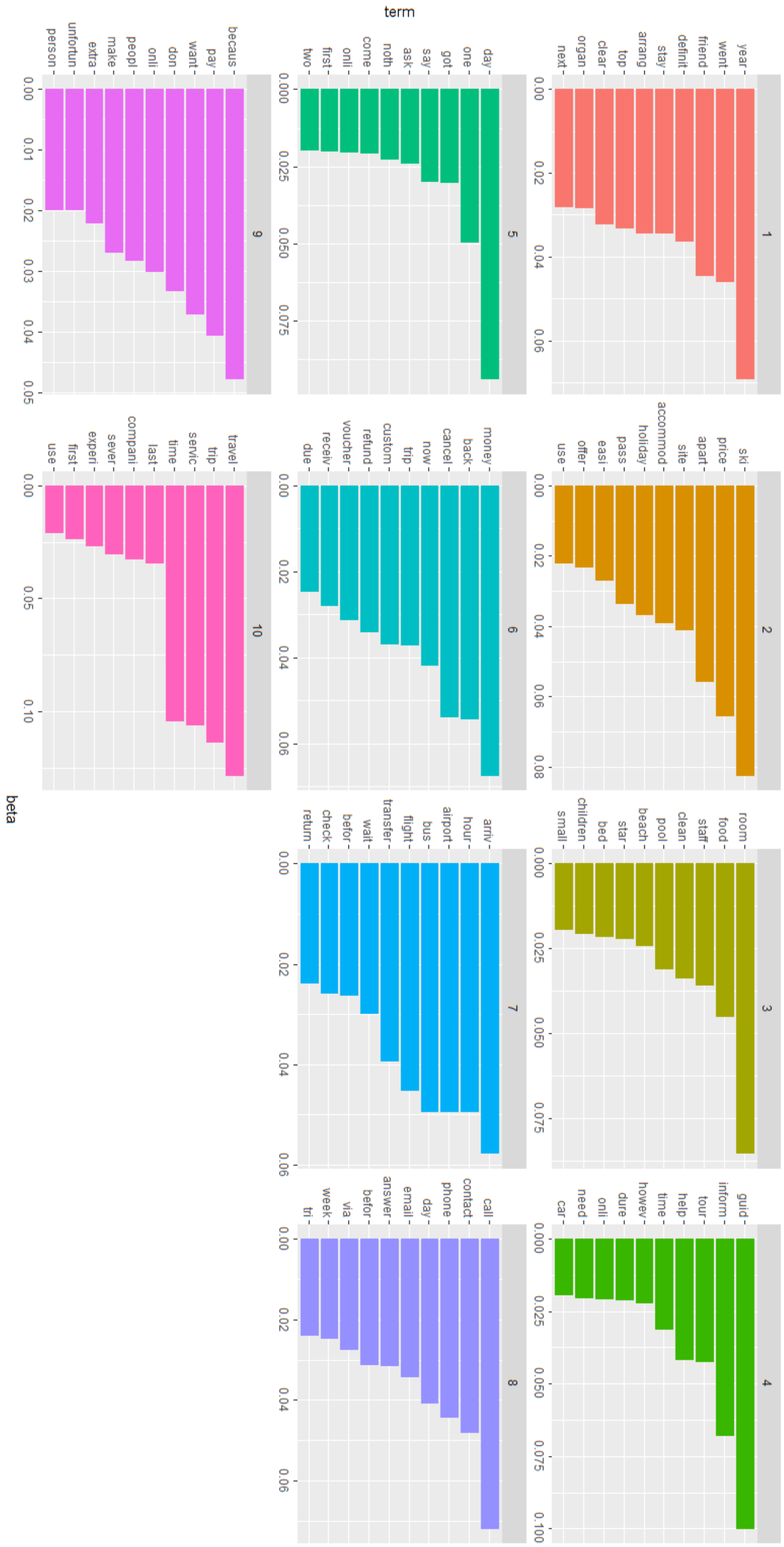

Figure B2: Word concentrations per topic for 10 topics

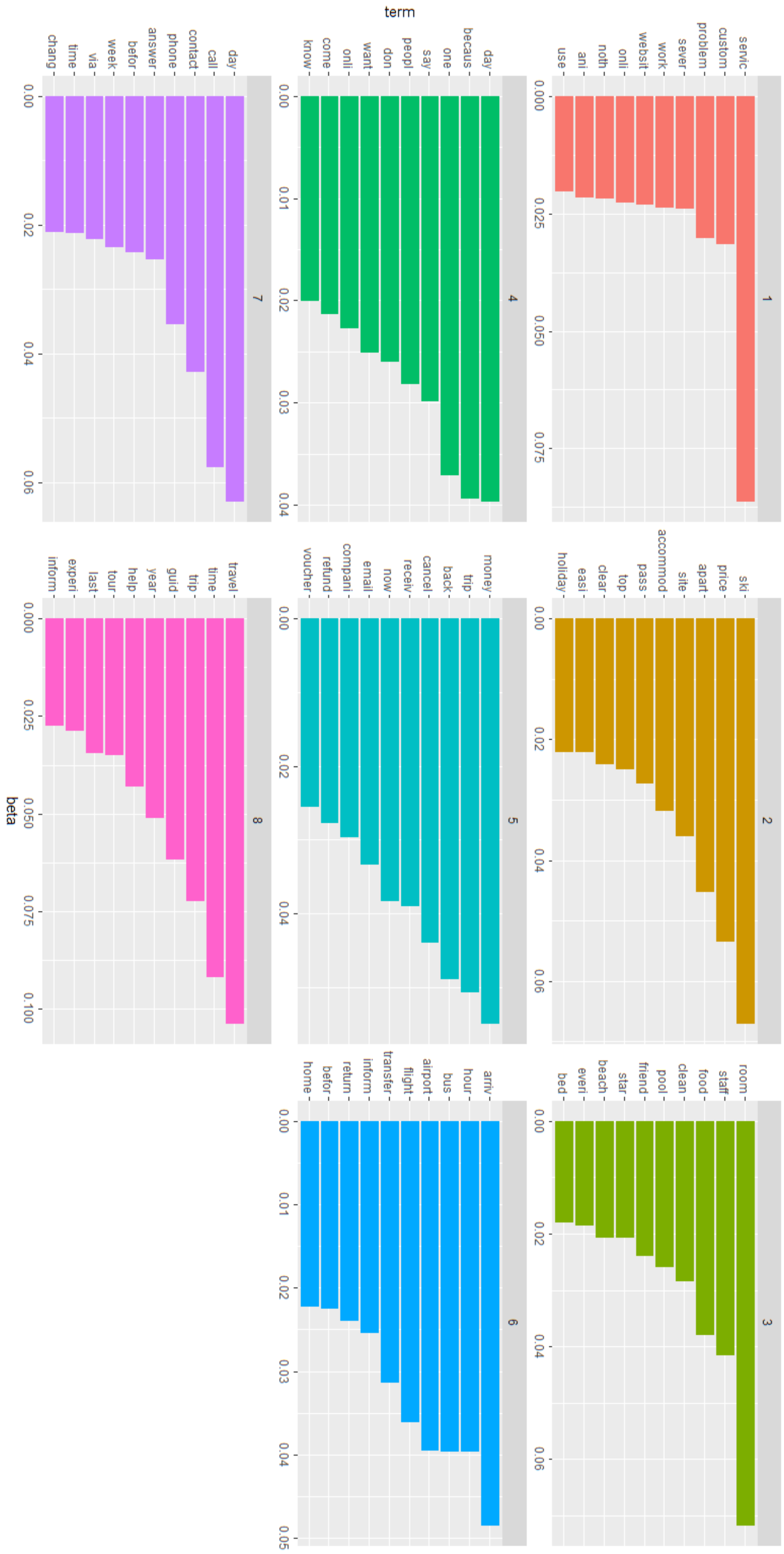

Figure B3: Word concentrations per topic for 8 topics

Appendix C

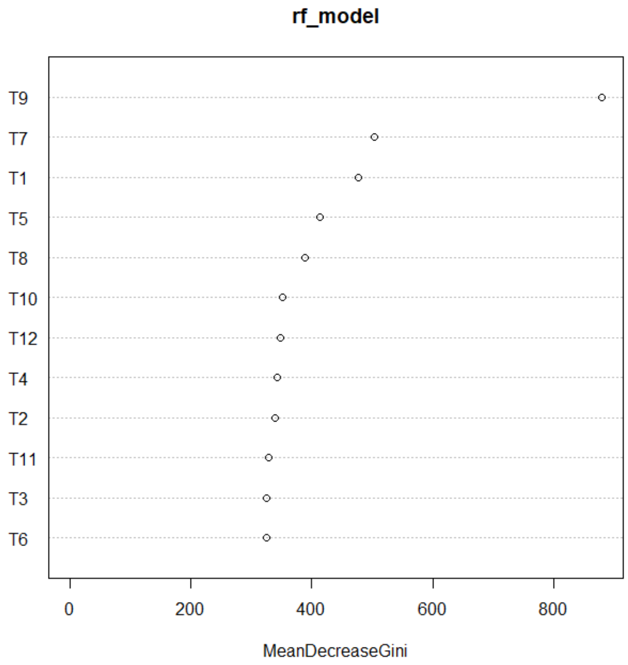

Figure C1: Variable importance for random forest model 1

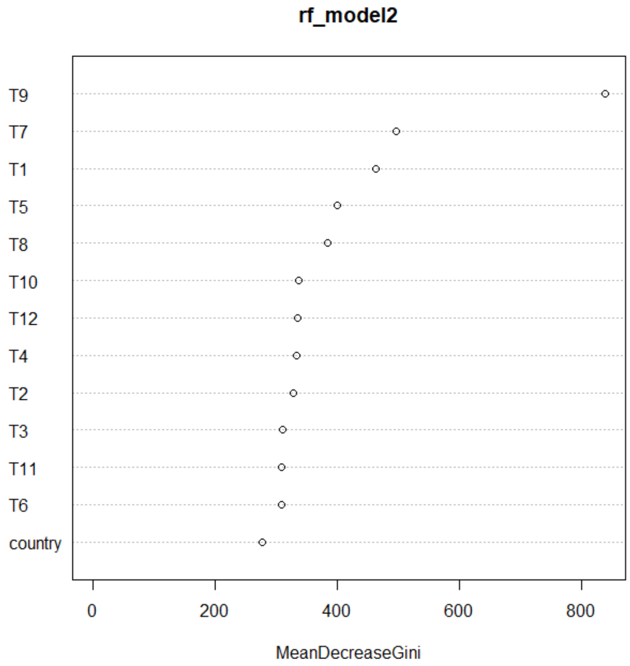

Figure C2: Variable importance for random forest model 2

Appendix D

Partial

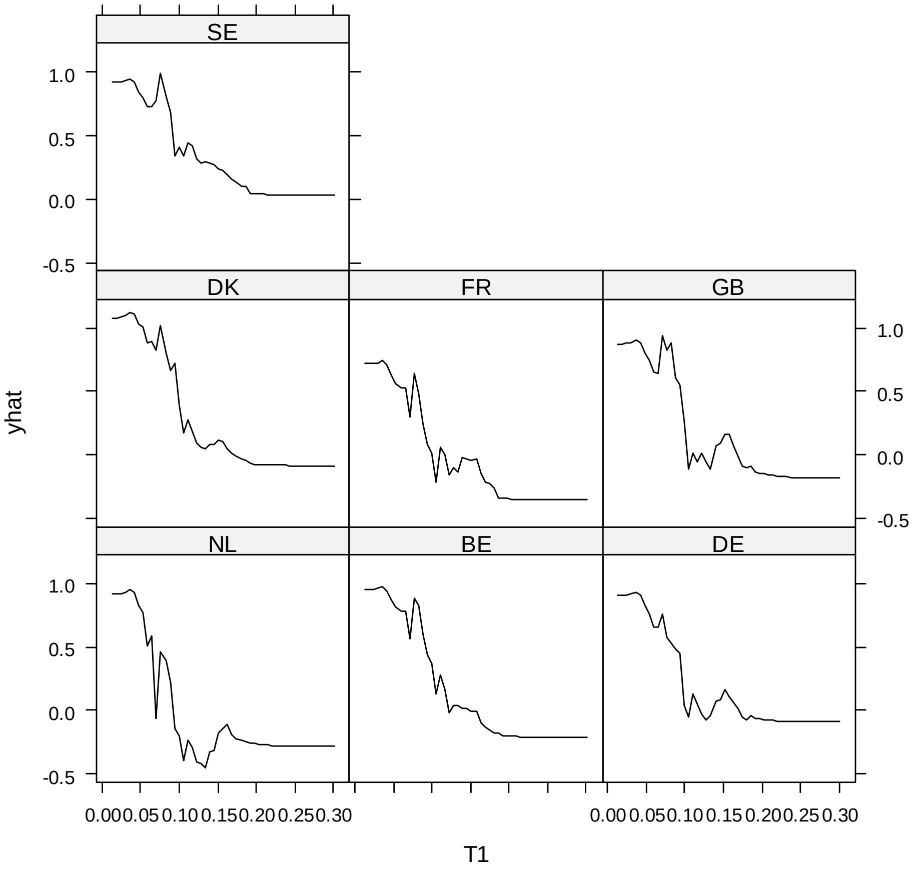

Figure D1: Partial dependence plots of Topic T1 and countries

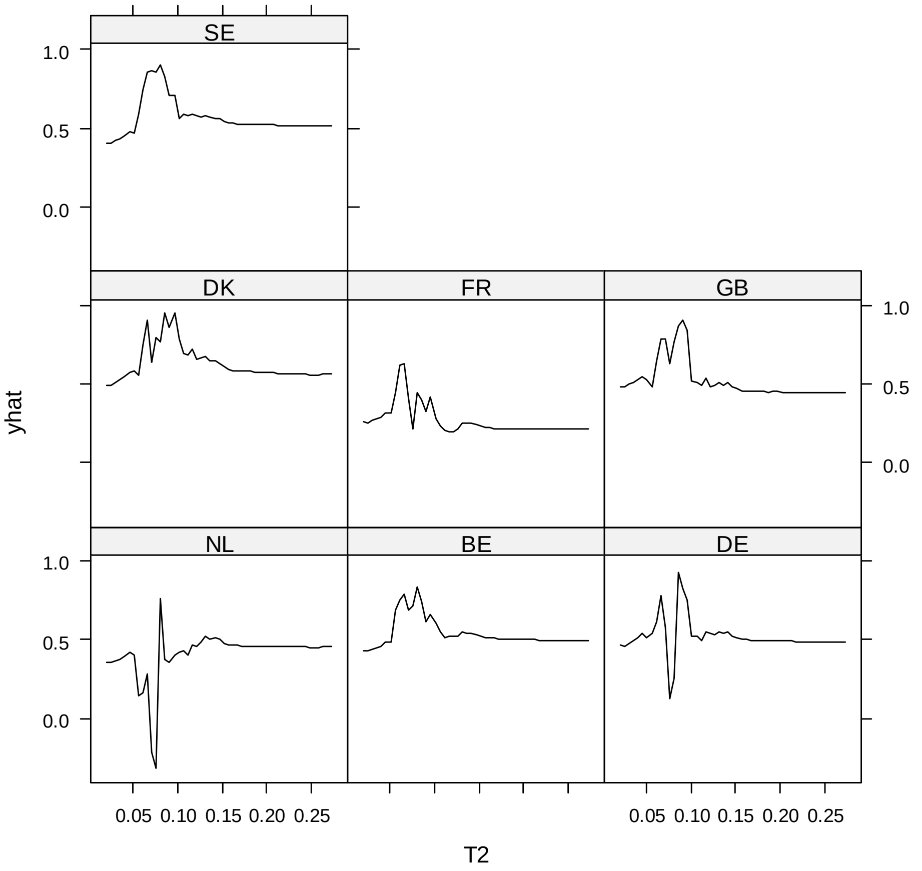

Figure D3: Partial dependence plots of Topic T2 and countries

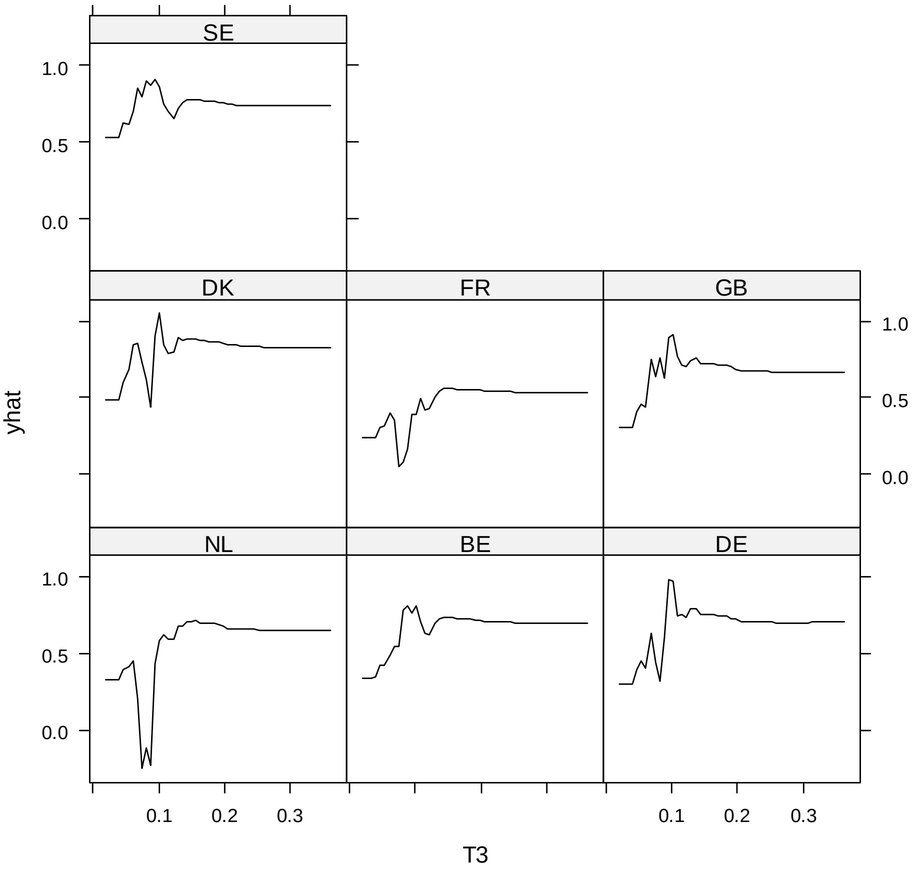

Figure D4: Partial dependence plots of Topic T3 and countries

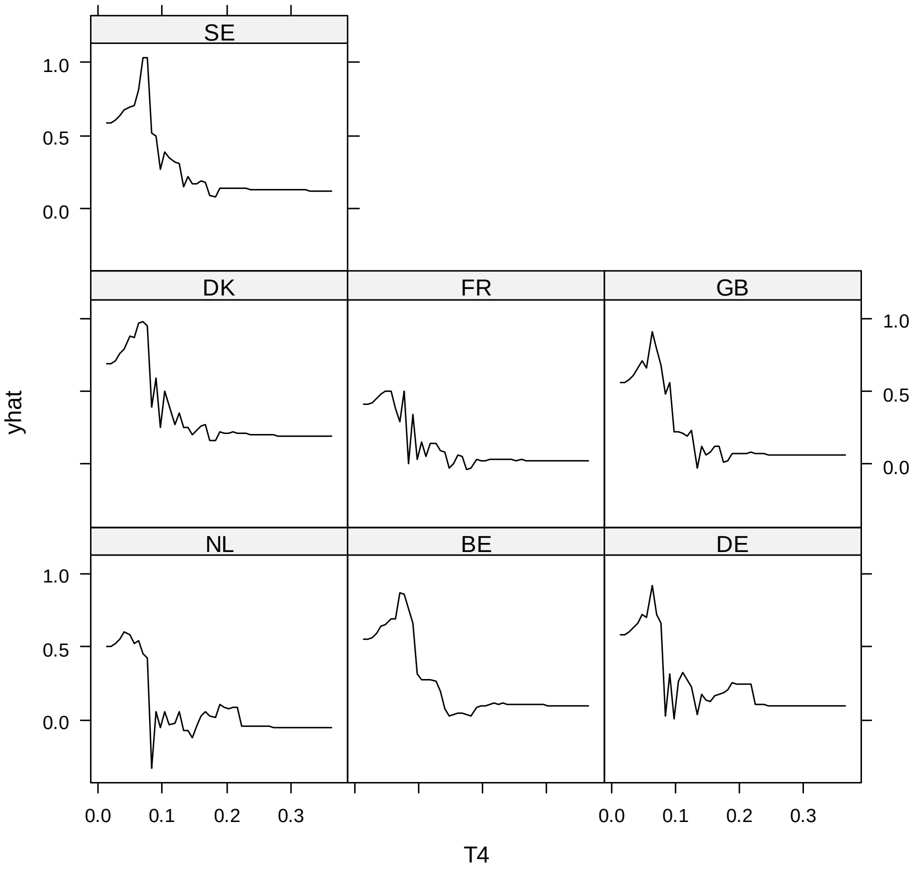

Figure D5: Partial dependence plots of Topic T4 and countries

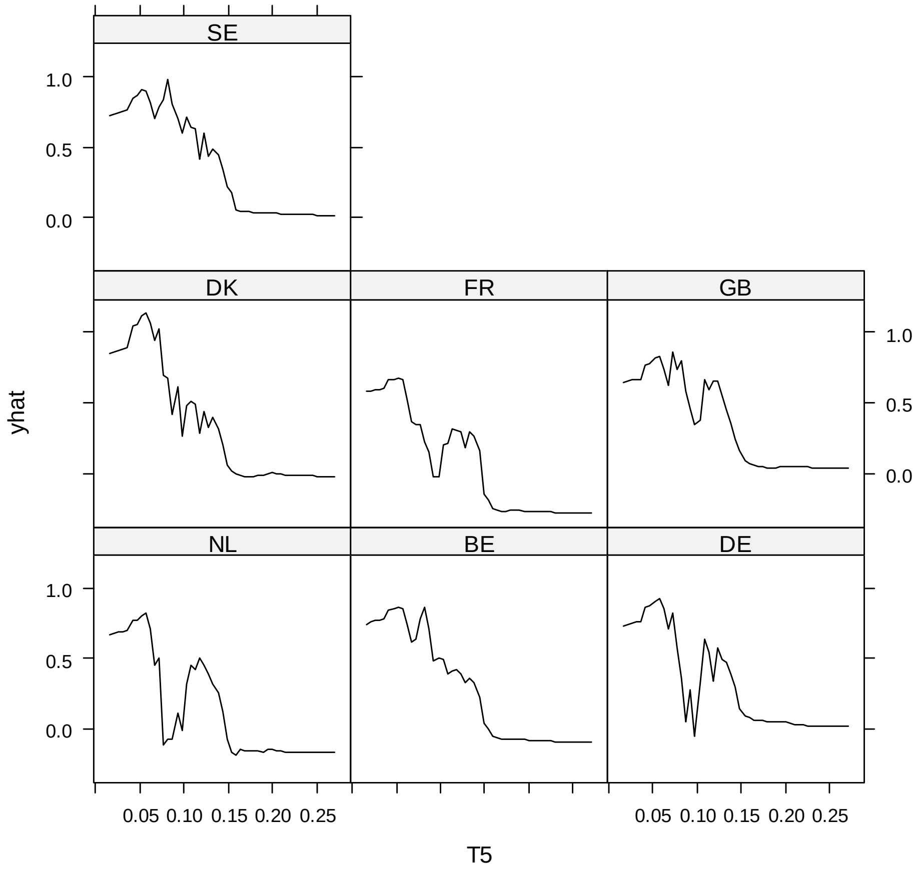

Figure D6: Partial dependence plots of Topic T5 and countries

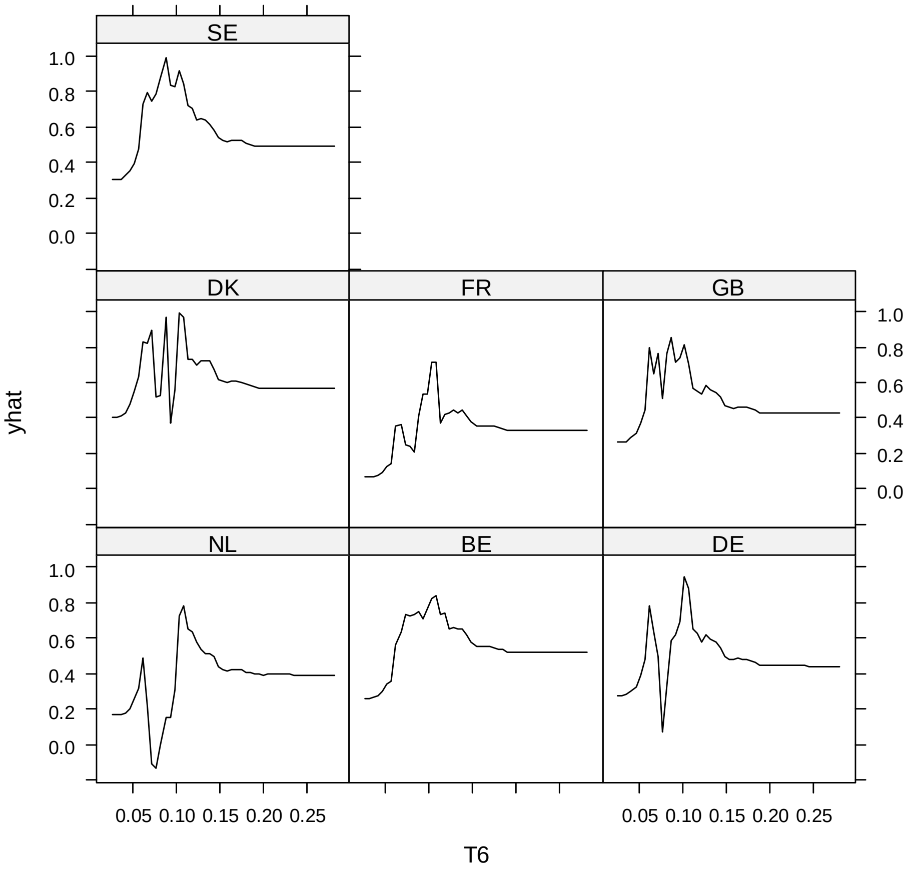

Figure D7: Partial dependence plots of Topic T6 and countries

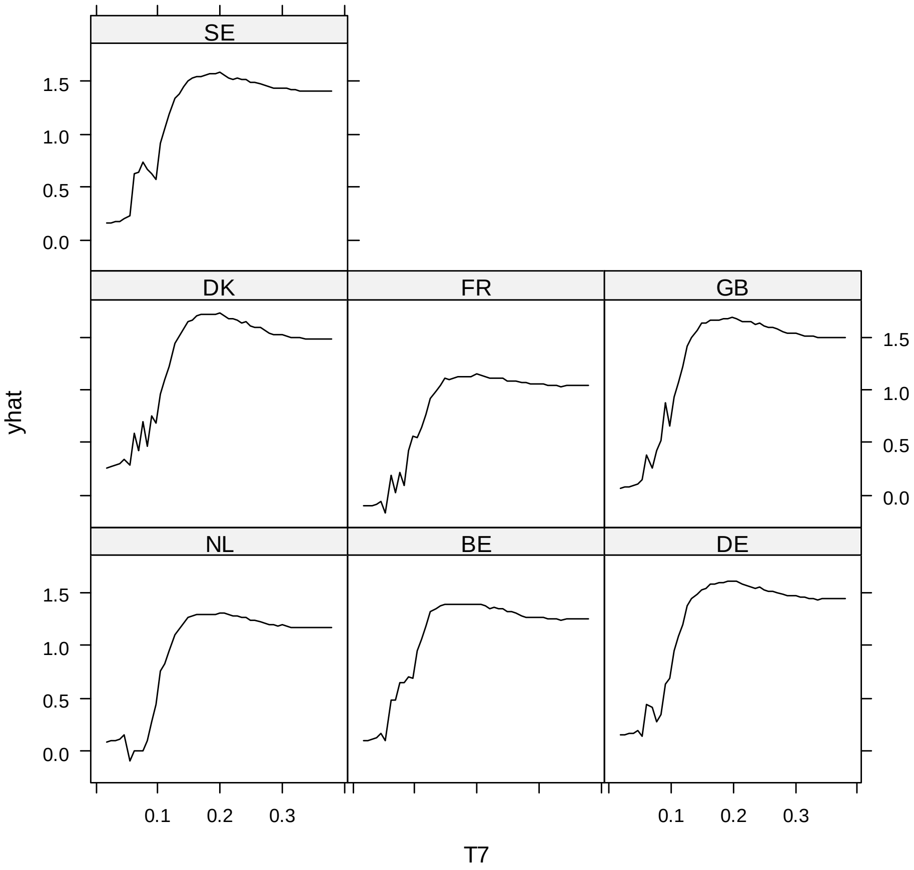

Figure D8: Partial dependence plots of Topic T an7d countries

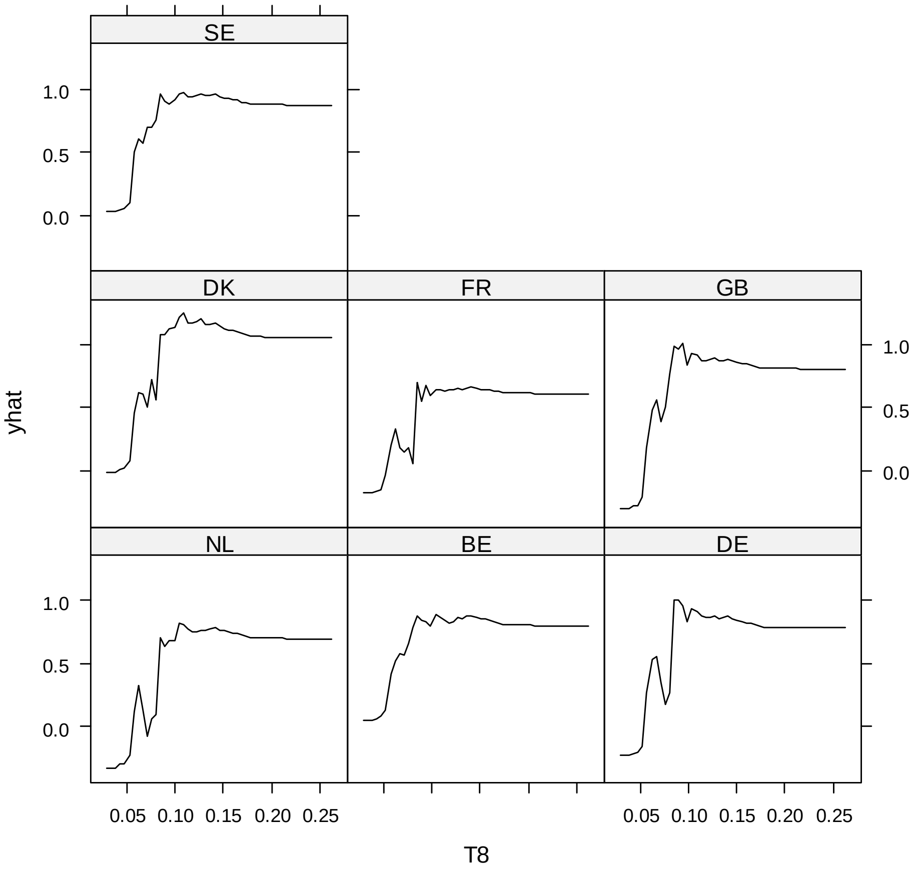

Figure D9: Partial dependence plots of Topic T8 and countries

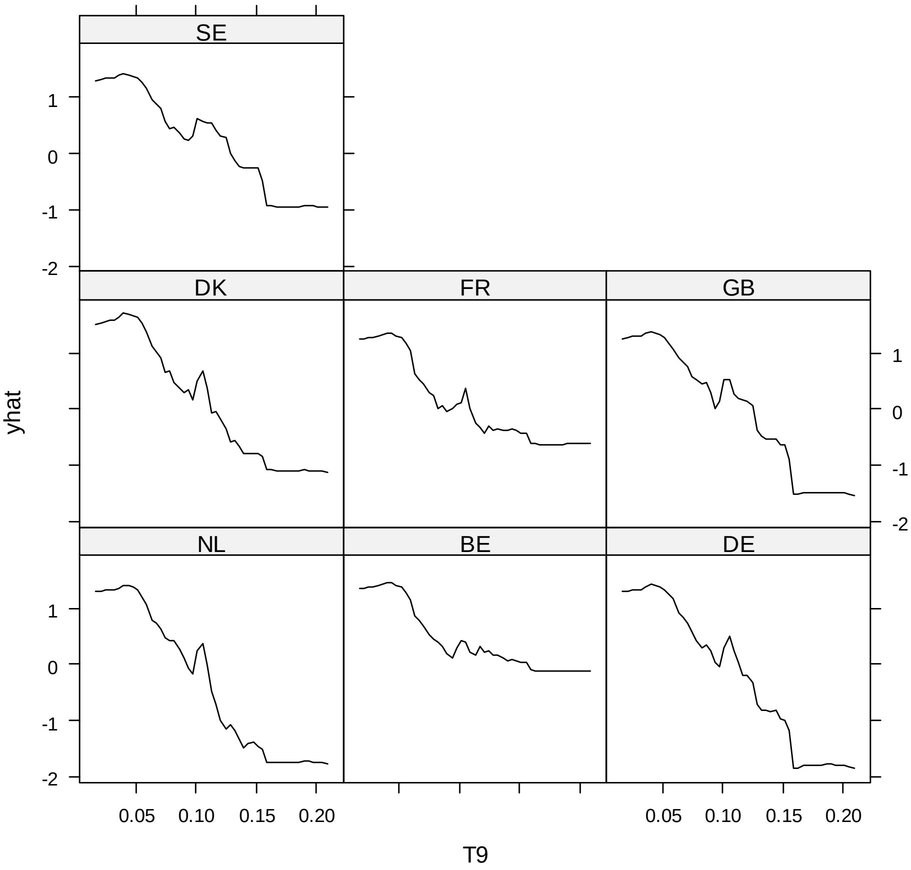

Figure D10: Partial dependence plots of Topic T9 and countries

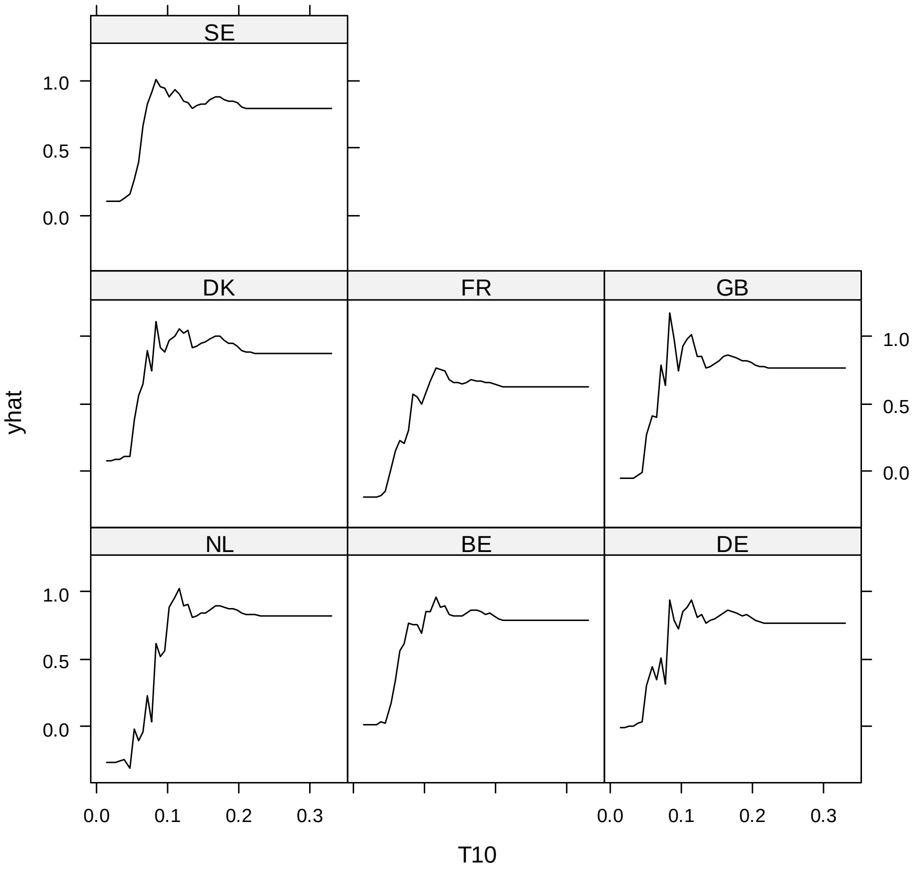

Figure D11: Partial dependence plots of Topic T10 and countries

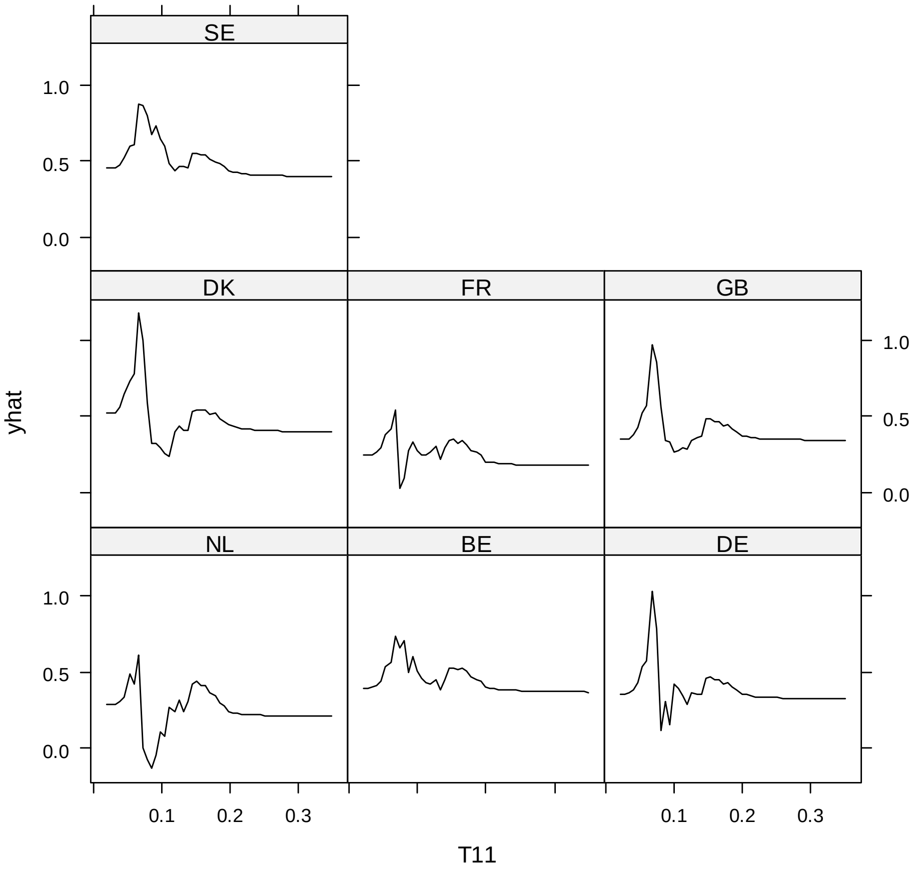

Figure D12: Partial dependence plots of Topic T11 and countries

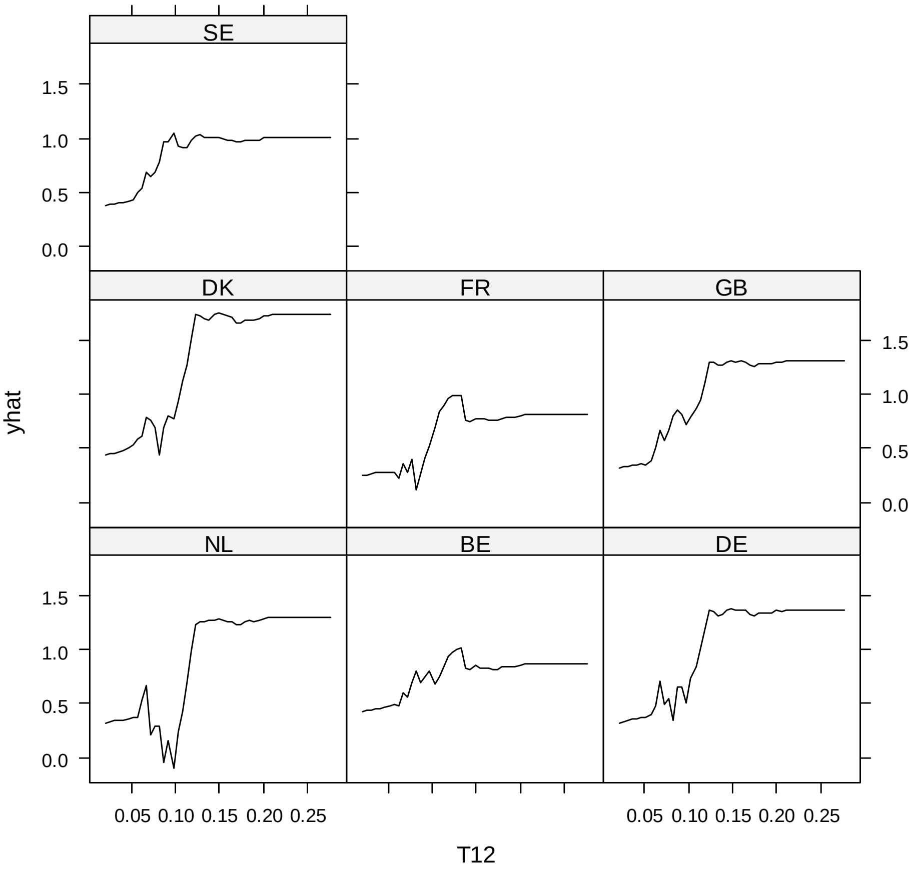

Figure D13: Partial dependence plots of Topic T12 and countries

Appendix E

Table E1

Ordinal regression models within-sample predictions confusion matrices

Model	Prediction \ Reference	1	2	3	4	5
1	1	427	69	63	63	82
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0
	5	95	48	63	264	660
2	1	1710	320	236	256	296
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0
	5	381	148	269	1056	2672
3	1	1713	326	239	255	305
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0
	5	378	142	266	1057	2663

Table E2

***Ordinal* regression models within-sample prediction metrics**

Statistic	Model 1	Model 2	Model 3
Accuracy	0.5927	0.5967	0.5959
95% CI	(0.5698, 0.6153)	(0.5854, 0.6079)	(0.5845, 0.6071)
No Information Rate	0.4046	0.4041	0.4041
P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16	< 2.2e-16
Kappa	0.365	0.3715	0.3705

Table E3

Ordinal regression models out-of-sample predictions confusion matrices

	Prediction \ Reference	1	2	3	4	5
Model 1	1	423	71	61	65	90
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0
	5	99	46	65	262	652

Model 2	1	427	69	63	63	82
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0
	5	95	48	63	264	660
Model 3	1	427	69	63	63	82
	2	0	0	0	0	0
	3	0	0	0	0	0
	4	0	0	0	0	0
	5	95	48	63	264	660

Table E4

***Ordinal* regression models out-of-sample prediction metrics**

Statistic	Model 1	Model 2	Model 3
Accuracy	0.5862	0.5927	0.5927
95% CI	(0.5632, 0.6088)	(0.5698, 0.6153)	(0.5698, 0.6153)
No Information Rate	0.4046	0.4046	0.4046
P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16	< 2.2e-16
Kappa	0.3552	0.365	0.365

Appendix F

Table F1

Random forest models within-sample predictions confusion matrices

Model	Prediction \ Reference	1	2	3	4	5
1	1	2065	2	0	5	6
	2	0	458	0	1	0
	3	0	1	489	2	4
	4	4	2	3	1216	47
	5	22	5	13	88	2911
2	1	2074	1	0	2	0
	2	0	463	0	1	0
	3	0	0	493	1	1
	4	4	0	4	1254	22
	5	13	4	8	54	2945
4	1	2079	1	0	3	4
	2	0	463	0	1	0
	3	0	0	495	1	3
	4	5	0	4	1260	28
	5	7	4	6	47	2933

Table F2

Random forest models out-of-sample predictions confusion matrices

Statistic	Model 1	Model 2	Model 3
Accuracy	0.9721	0.9843	0.9845
95% CI	(0.9681, 0.9757)	(0.9812, 0.9871)	(0.9814, 0.9872)
No Information Rate	0.4041	0.4041	0.4041
P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16	< 2.2e-16
Kappa	0.9608	0.978	0.9783

Table F3

Random forests models out-of-sample predictions confusion matrices

Model	Reference/Prediction	1	2	3	4	5
1	1	430	61	57	49	70
	2	1	5	3	7	1
	3	8	1	6	4	6
	4	22	17	20	60	80
	5	61	33	40	207	585
2	1	423	62	61	47	69
	2	3	1	3	3	2
	3	8	1	5	7	4
	4	13	19	20	55	57
	5	75	34	37	215	610
3	1	428	64	63	49	73
	2	2	1	0	2	2
	3	7	1	6	9	4
	4	14	16	20	49	60
	5	71	35	37	218	603

Table F4

Random forest models out-of-sample predictions confusion matrices

Statistic	Model 1	Model 2	Model 3
Accuracy	0.5921	0.5965	0.5927
95% CI	(0.5693, 0.6147)	(0.5736, 0.6191)	(0.5698, 0.6153)
No Information Rate	0.4046	0.4046	0.4046
P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16	< 2.2e-16
Kappa	0.3925	0.3939	0.388