TU WIEN Informatics

Reading between the Lines of Travelers:

Gained Insights from User-generated Content for Destination Marketing using Text Mining and Predictive Analytics

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Business Informatics

eingereicht von

Taghrid Elashkr, BSc
Matrikelnummer 00928070

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Mitwirkung: Univ.Ass. Mag.rer.nat. Dr.techn. Julia Neidhardt

Wien, 6. Oktober 2019

Taghrid Elashkr	Hannes Werthner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.ac.at

TU WIEN Informatics

Reading between the Lines of Travelers:

Gained Insights from User-generated Content for Destination Marketing using Text Mining and Predictive Analytics

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Taghrid Elashkr, BSc
Registration Number 00928070

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Assistance: Univ.Ass. Mag.rer.nat. Dr.techn. Julia Neidhardt

Vienna, 6th October, 2019

Taghrid Elashkr	Hannes Werthner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.ac.at

Erklärung zur Verfassung der Arbeit

Taghrid Elashkr, BSc
Rennbahnweg 27/3/17, 1220 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 6. Oktober 2019

Taghrid Elashkr

Acknowledgements

It has been a long journey of studying and now it is coming to an end. Although, I don't want to call it "end", because the end of a journey means the starting of another and I am sure that I am going to continue learning and gathering experiences my whole life. I would like to mention that studying at the Vienna University of Technology (TU Wien) was an amazing and challenging time for me. With all the ups and downs, I enjoyed studying at my university. I had the chance to learn about a lot of interesting topics and areas during my studies. Furthermore, I was lucky to work on my thesis project with "Österreich Werbung" (Austria National Tourist Office) as a partner, which enriched my research with practical business knowledge. This is why I thank everyone that supported me during my whole journey.

I would like to thank my supervisor Prof. Dr. Hannes Werthner for giving me the opportunity to do research on this interesting topic and for sharing his experience and expertise. My sincere gratitude goes to my co-supervisor Dr. Julia Neidhardt for her continuous support, feedback and motivation. She was open for research ideas and was there whenever I needed her assistance. That helped me to try out different interesting things and to gain more knowledge in various areas.
Furthermore, I would like to thank Prof. Dr. Wilfried Grossmann for his insightful comments and feedback. My sincere thanks go to our project partner, in particularly Mr. Holger Sicking for the fruitful cooperation.

Special thanks goes to my parents, my sisters and my brother for providing me with support and continuous encouragement throughout my studies, my thesis and during my whole life.

Thank you.

Kurzfassung

Die Nutzung des Internets im Tourismusbereich hat sich über die Jahre entwickelt, sodass Reisende nicht mehr nur Informationen konsumieren, sondern auch viele Inhalte selbst produzieren, die in verschiedenen Kanälen wie Social Media, Forums, Blogs, usw. zu finden sind. Diese Inhalte werden als unstrukturierte Daten kategorisiert und sind in Form von Beiträgen, Nachrichten und Kommentaren verfügbar. Unternehmen können dadurch mehr über ihre Kunden und deren Interessen erfahren. Die Analyse einer solchen Menge an Informationen kann sehr zeit -und kostenaufwendig sein. Daher werden diese Inhalte durch die Verwendung von Text Mining Techniken mit Hilfe von Machine Learning Methoden, und Predictive Analytics untersucht. Die vorliegende Arbeit richtet sich auf Inhalte, die in TripAdvisor United Kingdom erstellt wurden, speziell im Österreich-Forum. Die Daten werden verwendet, um neue Erkenntnisse für Reiseveranstalter, die für das Tourismusmarketing zuständig sind, zu gewinnen. Darüber hinaus werden die Daten auch verwendet, um den Zusammenhang zu den historischen Ankunftsdaten zu verstehen und um herauszufinden, wie diese Daten zur Vorhersage der zukünftigen Ankünfte nach Österreich beitragen können. Ebenso wird der Google-Suchvolumenindex auch als zusätzliche Datenquelle für die Vorhersageanalyse verwendet. Die Ergebnisse zeigen, dass Reisende das TripAdvisor-Forum nicht nur für ihre Reiseplanung nutzen und um Antworten für ihre Fragen vor Ort zu finden, sondern auch um Inspiration und Unterstützung bei der Wahl ihres nächsten Urlaubszieles, zu erhalten. Außerdem beginnen Reisende bereits drei Monate im Voraus mit der Recherche für ihre Reise. Bei Sommerreisen beginnt die Planung früher. Die geteilten Inhalte in den Foren spiegeln die Interessen der Reisende je nach Herkunftsland wider. Das Ergebnis der Zeitreihenanalyse für die Vorhersage der Ankünfte zeigte, dass das Erweitern der Modelle mit nutzergenerierten Inhalten, einschließlich Goolge-Suchdaten und TripAdvisor-Daten, die Prognose der Touristenankünfte für einige Zeitreihen verbessert, jedoch nicht für alle. Dennoch kann dieser vom Nutzer generierte Inhalt dazu verwendet werden, zukünftige Trends und Nachfragemuster zu erkennen. Der in dieser Arbeit vorgestellte Ansatz kann für andere Foren und Länder in TripAdvisor, sowie für ähnliche Datenquellen angewendet werden.

Abstract

The usage of the Internet in the tourism field has developed over the years, so that travelers do not only consume information, but also produce a lot of content themselves, which is shared in various channels including social media, travel review sites, blogs, etc. This generated content is categorized as unstructured data, which is available in forms of posts, messages, and comments. Businesses can use this content to enrich their understanding of their customers' needs. Analyzing such an amount of information can take a lot of effort and time. Therefore, text mining techniques using machine learning, and predictive analytics are used to investigate the information. This work focuses on the content created in TripAdvisor United Kingdom, particularly in the Austria forum. The data is used to gain new insights that can support travel professionals responsible for the destinations marketing of Austria. Additionally, the data is also used to understand the relationship to the historical arrivals data in order to determine how these data can contribute to the prediction of future arrivals (number of visitors) to Austria. Likewise, the Google search volume index is used as an additional data source in the prediction analyses. The results show that travelers use the TripAdvisor forum not only to get answers for questions related to their trip planning but also to get inspiration and support in choosing their next holiday destination. Additionally, travelers begin researching for their trip three months in advance, except for summer trips, for which planning starts earlier. Moreover, the forums' content reflects the travelers' interests depending on their country of origin. The outcome of the time series analysis for the prediction of arrivals showed that extending the models with the user generated content, including Google search data and TripAdvisor data, improved the forecasting of the tourist arrivals for some of the time series but not all. Nevertheless, this user generated content can be used to detect future trends and demand patterns. Finally, the approach introduced in the thesis at hand can be applied for other forums and countries in TripAdvisor, as well as for other similar data sources.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Description

In terms of the volume of online transactions, tourism has been ranked as the foremost industry [Ake09]. For both, private and public tourism organizations, Internet has become one of the most essential marketing communication channels [WF06]. Tourism organizations, especially travel agencies, hotels and destination marketing organizations, have been challenged by the development of the Internet and a lot of opportunities have arisen. The Internet offers new channels for communications and distribution of information [Ake09], where users not only consume information, but also produce a lot of content. This user generated content (UGC) is gaining massive traction as part of the purchase decision making process [OCo08]. Furthermore, it is also being used for different purposes, such as Google Trends (user search queries on Google), which offer significant benefits to forecasters, particularly in tourism [BSS15]. Therefore, the motivation has increased to investigate such UGC in the field of tourism destination marketing.

In order to promote a tourist destination, it is crucial to target potential travelers with the right message at the right time and in the right place. Travel professionals, such as destination marketers, tourism agencies, etc. aim to inspire people for a specific destination in an early phase of their decision making process. Understanding information about the channels used by travelers in their inspiration phase as well as how and when they start planning their trip are important inputs for them. One possible method to gather such information is by doing surveys and asking defined questions directed at people with specific characteristics. This method is time and cost consuming. Another method is to exploit available online content created by travelers. Information and communication technologies (ICT) have changed the tourism landscape over the past decades [WK+99, NW18]. There are multiple websites used by travelers. In literature, many studies are focusing on the user generated content (UGC) created online as travel reviews for hotels, restaurants or accommodations [MRMFFR19, NMC19]. In this study,

we want to investigate another source of information that consists of unstructured text and comprises new insights regarding traveler behavior. There are multiple sites used by travelers. For the purpose of this thesis, the widely recognized site for travel and tourism, TripAdvisor United Kingdom [1] [UK] is used in order to analyze UGC. TripAdvisor was founded in 2000 and is constantly growing. It has 490 million monthly average unique visitors and 730 million reviews and opinions [Tri18b]. It aims at enabling social interactions in the tourism industry. There is a lot of content generated by users on this site, for example, in the form of reviews, comments or ratings. Additional content comes from forums, which play an important role in sharing travel experiences and information among travelers.

This thesis examines the UGC for destination marketers with focus on the Austria forum on TripAdvisor UK. The gained results of this work are expected to be used for a better targeted marketing to inspire people to come to Austria. Furthermore, it investigates the potential use of this UGC in forecasting the arrivals.

1.2 Research Questions

Millions of travelers start planning their holidays by exploring different destinations on the Internet, on YouTube, Google or online platforms, engaging with their social networks. Travel professionals use these channels to understand the travelers' behavior and to animate potential customer as well as to plan future marketing strategies [Rhe12].

The aim of this work is to analyze different data sources (historical data and user generated content) with the goal to understand the behavior of the travelers and to create a model for the prediction of their arrivals, i.e. the number of people, who visit a specific destination in Austria.

The research questions are:

- *Which insights can be gained through social media/travel review sites to support destination marketing campaigns for Austria i.e. are these data sources used by travelers in their inspiration phase?*
- *To what extent can travel platforms, social media channels or blogs contribute to the prediction of arrivals to Austria?*

The results of this work will be gained through descriptive analyses of different data sources, new observations into social media usage in the travel domain and prediction models of the arrivals on the basis of user generated content. Due to the limitation of the data and as agreed with the project partner (Österreich Werbung/Austria National Tourism Office), this work's focus lies on travelers coming from the United States and United Kingdom. In future work, the approach can be extended to travelers coming from other countries.

[1] https://www.tripadvisor.co.uk/ShowForum-g190410-i146-Austria.html

1.3 Methodology

The methodological approach of the thesis comprises the following steps shown in figure 1.1:

Figure 1.1: Methodological approach of the thesis

1.3.1 Literature Review and Preliminary Analysis

The first step comprised of investigating of the current state of related work and methods used with regard to the tourism field. This involved understanding the planning process of travelers, getting an overview of the currently used social channels and travel review sites. Furthermore, different discussions with our project partner were conducted to identify useful information, which can be extracted from social media to support them in their marketing strategies.

Preliminary descriptive analysis of the historical data (arrivals of people from United States of America (USA) and United Kingdom (UK)) was carried out in order to understand its relationship to the user-generated content. Furthermore, tools and techniques for acquiring the data from the web and executing the preprocessing were investigated.

1.3.2 Data Acquisition

Different social media and travel review sites such as Twitter, Lonelyplanet, Reddit, Travellerspoint, etc. were analyzed to choose the most appropriate data sources. TripAdvisor UK [2] was selected as the main source for the thesis because of its rich user generated content and public user profiles. TripAdvisor offers different possibilities for the users to share their travel experiences. An essential task in this step was to identify the content relevant for the analysis. More details regarding the data will be presented in the next sections. Scripts used to acquire the data from the web were written in the programming language Python.

1.3.3 Data Preprocessing

Natural language processing was applied to prepare the content to be automatically classified using machine learning. Accordingly, different preprocessing methods were used such as stemming, deletion of stop words, lowercasing and tokenization.

[2] https://tripadvisor.co.uk

One of the faced challenges was that TripAdvisor gives its users the freedom to enter their home country as text which, in some cases, leads to an indication of invalid locations. Examples of home countries given by users include: the world, my home, ny, etc. Therefore, it was necessary to map the users' locations to a valid location (using an API) to be able to filter specific user groups from the USA and the UK and to use this information for the analysis. Different tools and programming languages were used for this step, such as Tableau software, R and Python.

1.3.4 Data Analysis and Model Development

Text mining techniques were used to explore and analyze the UGC. In order to classify the topics of the text written in TripAdvisor forums, a corpus of the data was labeled by experts. A list of classes was defined to cover the various topics and aspects discussed in the forums. It was later used by the classification models. Machine learning algorithms were evaluated and the best model was selected to automatically label the rest of the data. Additionally, entity recognition was used to filter location and date entities.

The relation between the arrivals' time series, Google index and extracted index from TripAdvisor data were investigated. Time series models were used and, based on specific evaluation measures, the models performance were compared.

1.3.5 Evaluation

On the one hand, measures were used for the evaluation of the classification models, in particularly accuracy, precision, confusion matrix, recall and F1-value. Prediction models were evaluated based on Root Mean Square Error (RMSE) and R^2. On the other hand, outputs were iteratively evaluated by an expert (the project partner) from "Österreich Werbung" (Austria National Tourist Office).

1.4 Structure of the Work

This master thesis is structured as follows:

Chapter 2 - State of the Art
This chapter presents an overview of the related work discussing the process of how travelers plan their trip and search for information about their target destinations through different channels. In addition, it provides insights about destination marketing and visitor arrival prediction.

Chapter 3 - Classifiers and Performance Measures
In this chapter, the used classifiers and performance measures for the classification of the text data are explained.

Chapter 4 - Historical Data Analysis
The historical data is described in this chapter. Descriptive analysis was used to

understand the behavior of tourists from the UK and the USA. Furthermore, cluster analyses were executed to group the regions using the arrival data.

Chapter 5 - Data Acquisition and Preprocessing
This chapter describes the chosen data source (user generated content from TripAdvisor), how the data was acquired and preprocessed for the analysis.

Chapter 6 - Travel Review Sites
Machine learning and text mining techniques were used to analyze the data, which delivered interesting insights regarding the behavior and the usage of the TripAdvisor travel forum by users from the UK and the USA.

Chapter 7 - Time Series Analysis
In this chapter time series analysis are conducted on the historical data, user generated content and Google search volume index.

Chapter 8 - Conclusion and Future Work
Finally, in this chapter, a brief summary of the thesis is provided and future work is presented.

CHAPTER 2

State of the Art

2.1 Travel Planning Process

Understanding how tourists plan their trip, why they choose a specific destination and when they start thinking about their next trip are crucial questions for destination marketing organizations in order to be able to develop their marketing plans. The travel decision making process of travelers has been one of the most researched topics in tourism and consumer behavior studies. Although the models for explaining the travel process vary among different sources, most of the literature tends to follow the classical five-stage consumer buying process [EBM90, ALA+12] illustrated in figure 2.1.

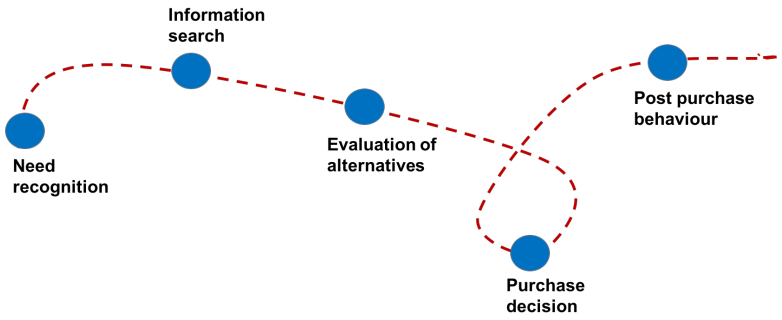


Figure 2.1: Travel planning process adapted from Engel, Blackwell & Miniard [EBM90]

The process can be divided into three main stages: pretrip, during trip and posttrip. The first and the most important stage for destination marketers is the stage, in which customers recognize a need (looking for a new destination), because that is the right time for them to promote their destinations. People use different sources to get inspiration for a specific destination. Social circle (friends, family) and online sources, for example the Internet, are critical for travel inspiration. Moreover, magazines/newspapers,

informational brochures, books, radio and travel agents are still used as a source of inspiration. According to a study conducted in 2014 by Ipsos MediaCT et al. [M+14], search engines and YouTube are top online sources of inspiration. Travel review sites and destination-specific sites are the second most used sources by travelers.

Having the need recognized, travelers start to search for information offline and online to gather more details in order to evaluate the alternatives, take the decision of selecting a destination and purchase the booking. The search may continue during the trip when travelers need to decide on what places to visit or where to eat [T+17]. This process ends with the post purchase behavior, which is about evaluating the trip and sharing experiences with others on social media or travel review sites.

Another framework for holiday planning is the decision network, which shows a generalized model for the planning process. This network represents a hierarchical structure of goals and sub-goals [Jen00]. Figure 2.2 illustrates two networks. Network A with different nodes that represent the sub-goals. The different levels of shadow of the node explain the rigidity and centrality levels e.g.: darker nodes stand for main sub-decisions as by "Travel Partners". Each node is a cluster of related concepts that defines the semantic mental model as shown in network B for destinations and activities [PF06].

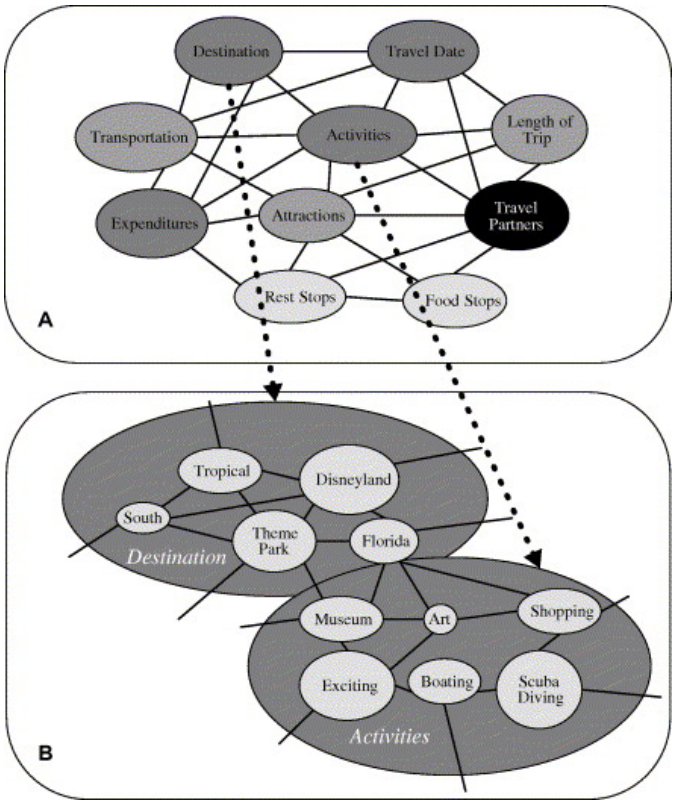


Figure 2.2: Semantic model of a vacation planner [PF06]

2.2 Travel Information Search

As shown in figure 2.1, travelers search for information in order to be able to make a decision and to evaluate alternatives. Information search can be defined as "the motivated activation of knowledge stored in memory or acquisition of information from the environment" [EBM95]. Based on this definition, information search can either be internal or external. Internal search involves identifying and retrieving knowledge from memory, while external search consists of collecting information from the marketplace [EBM95]. The undertaken amount of search of external information sources vary between individual tourists according to specific characteristics of the vacation. Depending on the amount of information searched, the extent of tourist's vacation planning and bookings can be predicted [Hyd08]. Several studies have shown that travelers tend to use four broad external information sources: family and friends, destination-specific literature, media and travel consultants [WP11].

Another channel used during planning is acquiring personal information and advice from travel agencies. Some studies predict that there will be no need for travel agents in the future as consumers turn to the Internet to make their travel planning. In 2000, Lang [Lan00] investigated the travel purchasing behavior of consumers and the future of travel agencies. The outcome was that while travelers are using the Internet to obtain information, many are still hesitant to book online, because of lack of secure payment methods (in past years). Furthermore, many people prefer to book their travel via traditional distribution channels and believe there will be a need for travel agencies [Lan00]. Nowadays, payment technologies have developed and it has become common to order and pay online. Travel agencies are supposed to develop their services according to the changes happening in the online world to maintain their competitiveness.

With the increase of the number of people using the Internet, the World Wide Web (WWW) has become prevalent for acquiring tourism information. About 95% of Internet users use the Internet to gather travel-related information. This shows that the Internet has become one of the most important sources for tourist information [PF06]. It offers travelers several online platforms to search on a given topic, share their opinions and compare alternatives. All these advantages of online tourism information search include the relatively low costs, customized information and easy comparison of different services and products. Some people prefer to start their search with online search engines (e.g., Google and Yahoo). It is easy to search using some keywords describing one's needs [HLC12]. According to a travel-tracking study conducted in 2014 with a total sample of 5000 consumers (3500 private and 1500 business), travelers indicated that they turn to the web early on in the travel process. Moreover, the study shows that 65% of leisure travelers tend to use online sources (social sites, YouTube and search) in their inspiration phase [M+14].

The increase of online social networks has not only influenced the way people communicate and share information but also the way tourism destinations are being promoted. In June 2009, Technorati, a blog search engine, indexed more than 100 million blogs. Its

2009 “State of Blogosphere” stated that 20% of the blogs were tagged as travel blogs. Travel blogs offer a channel for travelers to express their tourist experiences for a given destination. In this way they generate a lot of content, which is considered as rich content for researchers to extract marketing information. This includes monitoring destination images and products as well as understanding tourists’ needs and expectations. [BG12]

Furthermore, people tend to rely on the word of mouth (WOM), which is no longer happening only face to face, but also online. WOM is described as the communication about products and services between people, who are perceived as independent of the company offering the product or service [Wan15]. Online WOM is expected to continue growing in influencing the travel decision making processes, as the usage of the Internet during travel planning is increasing. There are variant forms for UGC on online platforms. One form in which content is created is consumer reviews and ratings. This is perceived by consumers as an important source, because they are written from a consumer’s perspective. A survey was conducted to study when and how other travelers’ reviews are used in the trip planning process. The findings showed that travel reviews are not only used in the decisive stages of trip planning but also for idea generation [GY08].

In recent years, a lot of research has been done on travel reviews. Different product and service reviews were analyzed to extract useful information for different research fields, for example, recommender systems, natural language processing and tourism marketing. Another form of generated content is in forums on travel review sites, which is rarely investigated by researchers. Therefore, one of the main aims of this work is to analyze such forums. In this thesis the focus will be on TripAdvisor UK [1], which is one of the most used travel review sites. It provides various services for people and one of these services is sharing questions and experiences in forums. People use them in different phases during their travel planning to ask questions and get feedback from other travelers or experts.

Nowadays social media is a main channel for travelers. They use them in different stages during their journey, for example at the beginning of their trip to get more details about a destination, during and after their trip to share their experiences. More details about social media will be provided in the next section.

2.3 Social Media

The Internet, and particularly the Web 2.0, has influenced the way tourism information is shared and the way people plan their trip [XG10]. Kaplan and Haenlein [KH10] describe Web 2.0 as the way in which software developers and end-users started to utilize the World Wide Web as a platform where content and applications are not only created and published by individuals, but are modified by all users in collaborative fashion. While Web 2.0 represents the technological foundation, UGC can be considered as the sum

[1] http://tripadvisor.co.uk

of all ways, in which people make use of social media. This term is usually applied to describe the various forms of media content that are generated by end-users [KH10].

Social media exist in various forums and for different purposes [XG10]. They are gaining more attention from many business executives, due to the presence and the active participation of customers on them. There are various definitions of social media in the literature. Kaplan and Haenlein [KH10] define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content". They proposed a taxonomy schema for the classification of social media according to their level of social presence richness and level of self-presentation/self-disclosure, representing six types of social media: Blogs, Social networking sites (e.g., Facebook, LinkedIn), collaborative projects (e.g., Wikipedia), content communities (e.g., YouTube, Slideshare), virtual social worlds (e.g., Second Life) and virtual game worlds (e.g., World of Warcraft). Social Networking is explained by Miguens et al. [MBC08] as online communities of people who share common interests and activities. They provide the user with various interaction, tools such as chat, video-sharing, exchange of plain messages, blogs, etc.

There are several social network applications available on the market. Figure 2.3 shows the most popular social networks worldwide as of October 2018 ranked by number of active users (in millions). Market leader Facebook was ranked as the most used social network worldwide with 2.23 billion monthly active users. A high number of social networks are available in multiple languages and enable users to connect with family, friends and other people across geographical, political and economic borders. The use of such social networks is highly diverse. Platforms, such as Facebook or Instagram mainly focus on providing users with the possibility to share photos or update their status. Other platforms rather focus on community or presenting UGC. [Sta18]

Due to the complexity of taking a decision to purchase travel products, travelers tend to do an extensive information search and rely on other travelers' experiences as means to decrease uncertainty [FBR12]. Gretzel and Yoo [GY08] conducted a web-based survey of the users of TripAdvisor and found that online travel reviews help travelers to reduce risk by making it easier to imagine what a place will be like, increase their confidence during decision making and assist them in selecting accommodation. It was found that travel reviews are not only used in the decisive stages during the trip planning to limit choices, but they are also used for idea generation. Moreover, it was shown that online travel review readers are highly educated, travel rather frequently, use the Internet extensively for their trip planning and have high incomes. Therefore, these travel review readers seem to be an important target market for travel marketers.

A study by Fotis et al. [FBR12] analyzed the perceived level of trust between holiday related sources: social media, friends' publicity and relatives, tourism websites, advertorials in mass media (e.g., TV or radio shows, newspapers and magazines), travel agents and information from other travelers in various websites. Out of these sources, friends and relatives were rated as the most trusted sources, followed by information from other

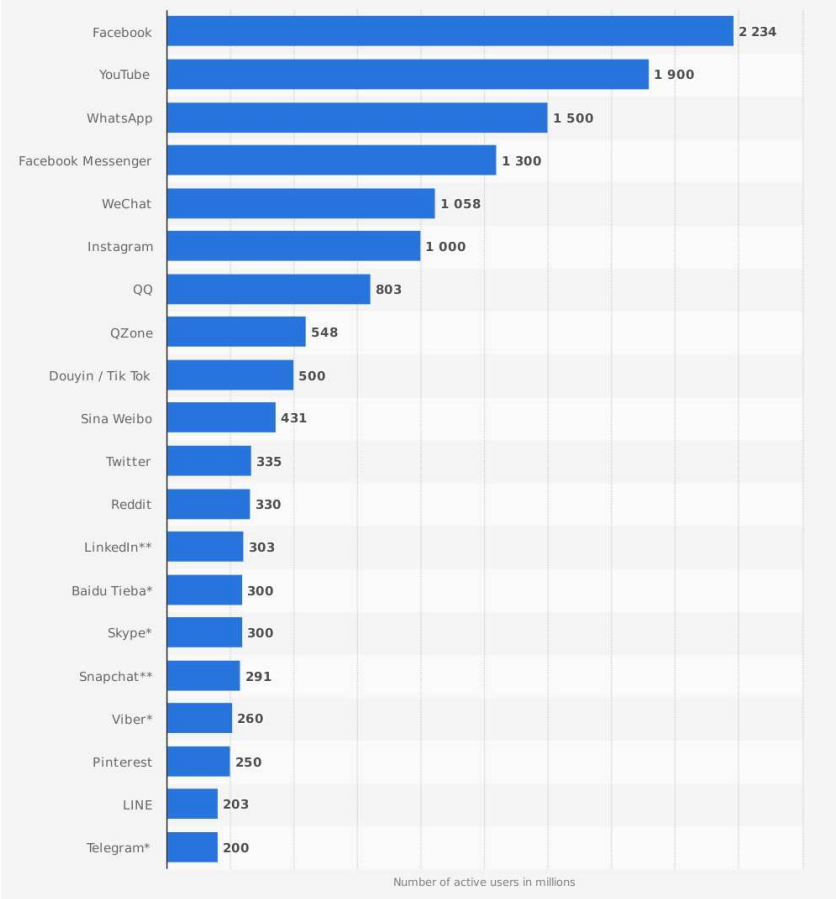


Figure 2.3: Popular social networks ranked by the number of active users (in millions) [Sta18]

travelers on various websites. Furthermore, the study showed that the usage of social media differs among national markets (e.g., comparing Australians and Russians).

2.4 Destination Marketing

How do visitors choose their destinations? When do the inspiration and information search phase for visitors start and what are the factors that influence visitors to choose a specific destination? In addition to the questions mentioned earlier, these are some of the questions that are essential for destination marketers when planning their marketing strategies. Tourism destination marketing is defined as "the management process through which the National Tourist Organizations and/or tourist enterprises identify their selected tourists, actual and potential, communicate with them to ascertain and influence their wishes, needs, motivations, likes and dislikes, on local, regional, national and international levels, and to formulate and adapt their tourist products accordingly in view of achieving

optimal tourist satisfaction thereby fulfilling their objectives." [Ana08].

Travelers have many selection of travel destinations to choose from. In terms of supply, destination marketing organizations are competing at different levels to direct travelers' attention towards their destination. Destination marketing organizations, are government agencies, travel associations and other stakeholders that are responsible for long-term marketing strategies and promotion of their respective destinations areas [T+17].

The decision to choose a destination visit does not stand alone, it consists of smaller decisions, such as: when and how to get there, and where and how to live there [Eko10]. Various aspects can influence the decision of a visitor and understanding that can help destination marketers to better fulfill their customer needs. Buhalis [Buh00] developed a framework representing the core components of destinations as illustrated in figure 2.4 where a destination can be comprised as a combination of different products and services, as well as experiences provided locally.

- Attractions (e.g., special events, heritage, natural)
- Accessibility (e.g., transportation systems and routes)
- Amenities (e.g., accommodation and retailing, catering and other tourist services)
- Available packages (e.g., packages by intermediaries)
- Activities (e.g., skiing, swimming, diving and other activities that visitors can do during their visit)
- Ancillary services (e.g., banks, posts, hospitals and telecommunications)

Figure 2.4: Six "A"s framework for the analysis of tourism destinations

On of the main goals in marketing consists of reaching consumers at the moments that can highly impact their decisions, when they are open influence, so-called "moments that matter" or a "touch point" [S+17]. Therefore, different methods are used to promote destinations in order to increase people's awareness. Designing a cost effective marketing mix is not easy because of the different needs and diversity of consumers around the world and the several cultures and linguistic backgrounds. A wide range of techniques [Buh00] are used, such as:

- above the line: includes advertising on television, press and radio as well as poster campaigns. Even though, this kind of promotion can be expensive, it helps to build a destination brand and to influence a large number of visitors to travel to the destination.
- below the line: is about participating in major tourism travel fairs in different countries to meet intermediaries and members from the public to promote their destination.

Additionally, the Internet has influenced the way in which destinations are promoted. Considering the increasing number of Internet users and their presence on different social media channels, such as TripAdvisor, Facebook, Instagram, etc., where they tend to not only consume content, but also share it, it is important for destination marketers to look into such communication channels. These channels can offer different ways for them to understand the consumers' behavior and their needs.

2.5 Visitors Arrival Prediction

Tourist arrivals are growing each year. According to the UNWTO [UNW16], tourism represents 7% of total exports. During the last 60 years, tourist arrivals increased by over 4000% and reached 1.2 billion arrivals in 2015, and the number continue to grow. Tourism destinations and suppliers strongly depend on precise predictions of future demand in order to adapt their strategies. Therefore, they benefit out of the long-term trends and seasonality of the past arrivals in predicting future arrivals using autoregressive approaches, which typically lead to quite satisfactory results [HEF+17]. The real arrivals are periodically released by government agencies. However, these releases are available with a lag of several weeks [CV12].

Tourism demand modeling and forecasting methods are generally divided into two categories: quantitative and qualitative methods. Quantitative methods are mostly used in forecasting tourism demand and are dominated by two sub-categories: non-causal time series models and the causal econometric approaches [SL08]. Qualitative techniques are used to cover changes of a large and unprecedented events that are likely to occur. These methods depend on the experience of forecasting experts [LSM08].

In the review research of Song and Li [SL08] about tourism demand modeling and forecasting, time series are defined as a model that explains a variable with regard to its own past data. Furthermore, the prediction is basically done based on historical trends and patterns (e.g., seasonality). It has been used in the last four decades, specially the integrated autoregressive moving average models (ARIMA) proposed by Box and Jenkins. Researchers have tried to improve the forecasting performance of ARIMA/SARIMA models by using other approaches. One of the approaches was about enhancing time series models to multivariate dimensions and to verify if additional information involved in time series can contribute to improve the accuracy of the forecast. This additional information can be, for example, the data generated in Internet platforms. Information technology gives people the opportunity to obtain various information on the Internet and share their opinions, which leads them to produce a big amount of data including search queries and social media [LPLH17]. This user generated content gained the attention of researchers in the last years, who investigated the usage of it to improve the forecasting models. An essential source that provides several interesting aspects is Google Trends, which is a real-time daily and weekly index of the volume of search queries that users enter into Google. This query index can be filtered by country of origin [CV12].

Figure 2.5 shows an example of Google Trends for the search query Vienna as the capital

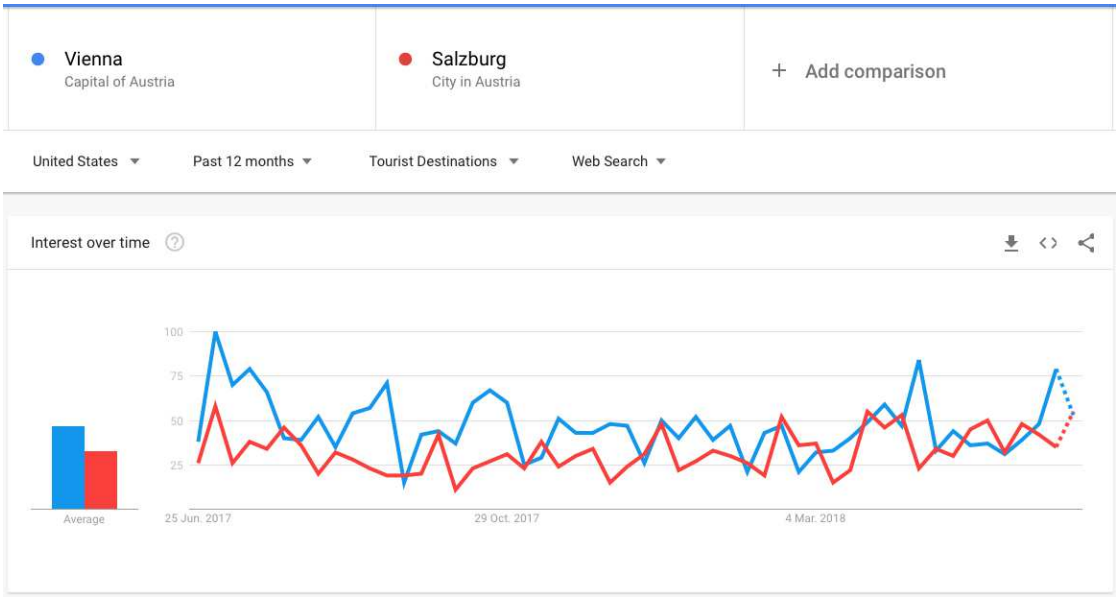


Figure 2.5: Example of Google trends comparing Vienna and Salzburg

of Austria compared with Salzburg as another Austrian city. The interface facilities filtering based on different attributes:

- Country: United States, United Kingdom, Austria, Egypt, etc.
- Duration: past hour, past 4 hours, past day, past 90 days, past 12 months, etc.
- Topic/Category: arts and entertainment, vehicles, travel, etc.
- Search type: web search, image search, YouTube search, etc.

The data can be downloaded as csv file that contains a time series where numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means that there was not enough data for this term [Goo18]. Google Trends search query were introduced by Bangwayo-Skeete and Skeete [BSS15] as a new indicator for tourism demand forecasting. Their indicator is based on an aggregated search for "hotels and flights". They investigated three approaches using their indicator: Autoregressive Mixed Data Sampling (AR-MIDAS), Seasonal Autoregressive Integrated Moving Average (SARIMA) and autoregressive (AR) and found that AR-MIDAS outperformed the other approaches in most of the forecasting experiments. The conclusion was that Google Trends information offers significantly benefit in forecasting. Hoepken et al. [HEF+17] compared both approaches and found that the usage of big data as an additional input can significantly increase the arrivals' prediction performance compared to using past arrivals alone. In their study, they used an autoregressive model

using only past arrivals as input attributes, and another model using big data information (i.e. destination price level and web search traffic (Google Trends)) as an additional input. Online information was used in the research of Pereira et al. [PRBA15] to improve the quality of transport prediction under special event scenarios.

Besides the time series and econometric models, AI techniques enhanced the forecasting in the field of tourism with new methods. They have the advantage that they do not require any preliminary or additional information about the data. Methods such as the artificial neural network (ANN) are used to imitate the learning process of a human brain. Empirical studies show that ANNs outperformed the classical multiple regression and time series models [SL08].

2.6 Conclusion

In this chapter, topics related to the thesis were investigated in order to get an overview of the state of the art in the field. Through the illustration of the planning process different stages that travelers go through during their trip planning should be made demonstrated. Furthermore, the channels used to find travel-related information were analyzed. Travelers make use of a number of sources to find information, for example, family and friends, Internet, social networks, literature, travel agencies and WOM. With regard to travel reviews, the online survey by Greztel and Yoo [GY08] showed that they are used in different stages including idea generation and trip planning. Topic destination marketing was examined in order to grasp the functions and goals of techniques used by destination marketers. This was important for investigating which information can be extracted from social media to support marketers in improving their marketing plans. Finally, the traditional and current methods applied for tourism demand modeling and forecasting were studied in order to be considered in the arrival prediction analysis.

CHAPTER 3

Classifiers and Performance Measures

Due to the increased availability of digital documents and content that need to be organized and analyzed, the automated classification or categorization of text into predefined categories has been gaining interest by researches [Seb02]. The focus of research lies on how to make machines learn to understand and interpret the data. This is often used in text mining. Text mining, also referred to as intelligent text, is defined as the discovery by computer of new, unknown information by automatically extracting information from different written resources. The main element is the linking of extracted information to new facts or hypotheses, so that new perspectives can be gained. Knowledge can be discovered through several information sources [GL+09].

In this chapter, the used classifiers and performance measures for the work will be thoroughly described.

3.1 Overview of the Classifiers

There are two types of learning, supervised learning and unsupervised learning. If the instances are given with known labels, then the learning is called supervised. This is typically used in the context of prediction or classification. The idea behind it is to map an input to output category, label or value. Otherwise, when instances are unlabeled then the learning is unsupervised. This is useful to find patterns in the data. For example, clustering algorithms are used to discover unknown but useful classes of items [KZP07].

Supervised learning is used to classify the content of TripAdvisor. The input is the title and content of the posts, and the output is the category labeled by experts. Given that the following algorithms are used when classifiying text [APA+17, IKT05]. They will be

applied for the content at hand. The results will be compared and the most suitable algorithm will be used to classify the unlabeled data.

- Naive Bayes (Multinomial)
- Logistic Regression
- Support Vector Machines (SVMs)

Naive Bayes (Multinomial)

Naive Bayes has been one of the popular machine learning classifiers for many years. Its simplicity makes it usable for various tasks. It is an algorithm that is frequently used for text classification. There are two models that are commonly used: Multivariate Bernoulli model, which represents a document as a binary feature vector representing whether each word is present or absent, and Multinomial Naive Bayes (MNB) model, which considers the number of appearance of each word in the document. Measures such as term frequency or tf-idf can be used for this classifier [KHRM06, KFPH04]. In the MNB each document (e.g., post) is represented as a collection of words and the order of words is irrelevant. The probability of a class value c given a document d is computed as follows:

$$P(c|d) = \frac{P(c)\prod_{wed} P(w|c)^{n_{wd}}}{P(d)} \tag{3.1}$$

n_{wd} is the number of times word w occurs in document d and $P(w|c)$ is the probability of appearance of word w given class c. The prior probability of class c is $P(c)$, and $P(d)$ is a constant that makes the probabilities for the different classes sum to one [FB06].

Logistic Regression

Logistic regression is one of the baseline supervised machine learning algorithms for classifications. Logistic regression is also known as maximum entropy. It can be used to classify a document into one of two classes or into many classes, so-called multinomial logistic regression. The difference between naive Bayes and logistic regression is that logistic regression is a discriminative classifier while naive Bayes is a generative classifier. A generative model like naive Bayes uses the likelihood term, which expresses how to generate the features of a document if we knew it was of class c. It tries to model the detailed properties of each class. Discriminative classifier means that it only tries to learn to distinguish the classes. It learns to assign high weight to the features that directly improve its ability to discriminate between possible classes even if it cannot generate an example of one of the classes. Generative classifier models would have the goal to Logistic regression classifier is like naive bayes a probabilistic classifier. It extracts real-valued features from the input, multiplies each by a weight, sums them and passes them through a sigmoid function to create a probability. A threshold is used to make a decision. In case

of multinomial logistic regression, the softmax function is used to compute probabilities. This description is based on the source [Lea18b] .

Support Vector Machines (SVMs)

As described by Basu et al. [BWS03] support vector machines are supervised learning models that find the best decision boundary between items (e.g.: documents) that belong to a given category or class. Depending on the defined decision boundary, unlabeled data can be categorized to a specific class. The SVMs can operate on data with large feature sets as the aim is to measure the margin of separation of the data rather than matching the features. The training of the SVMs is based on the pre-classified document. Research has shown that SVMs scale well and has good performance on large data sets. Other advantages of support vector machines are memory efficient because they use a subset of training points in the decision function (support vectors) and offer different kernels. An example for linear kernel is shown in graphic 3.1 [Lea18c]. By maximizing the margin between the two classes, the optimal separating hyperplane can be found. The middle of the margin is the optimal separating hyperplane. The points on the boundary lines are called support vectors. The Dataiku tool used for the analysis offers the possibility to run the same algorithm with different parameters to find the best performing model. The kernels that were tried during the execution are the RBF kernel, Polynomial kernel and Sigmoid kernel.

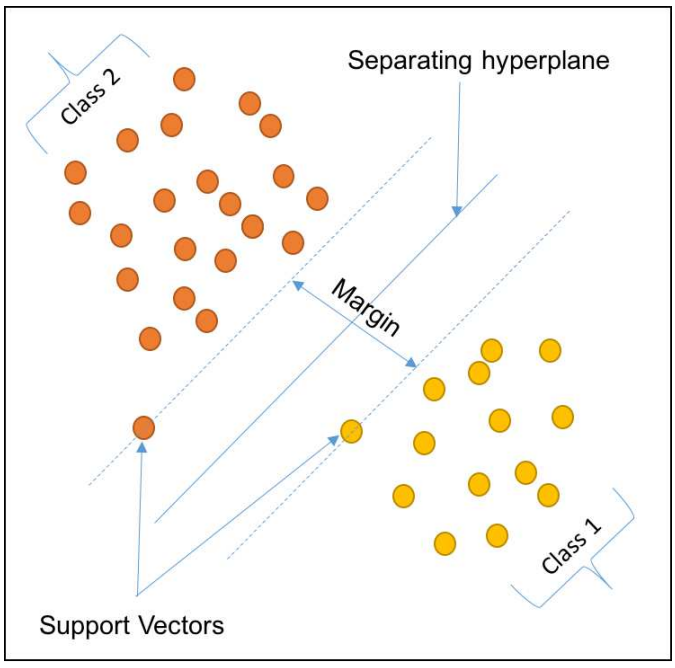


Figure 3.1: Example of a Support Vector Machine classifier

3.2 Validation and Performance Scores

3.2.1 K-Fold Cross-Validation

The following description is based on the source Sickit Learn [Lea18a]. In supervised learning training and testing a model with the same data can make the test results biased. A model that tries to predict the results that it has just trained would have a perfect score but will not be able to have good predictions on yet-unseen data, which is called overfitting. A solution for this problem is to use a part of the data as a test set. It is popular to have a split of 80% - 20% with 80% for the training set and 20% for the test set. However, when evaluating different settings for estimators, there is a risk of overfitting on the test set because the parameters can be adapted until the model performs optimally. To solve this problem, another part of the data will be used as a "validation set", which means the training is executed on the training set, after it an evaluation is done on the validation set and the final evaluation will be done on the test set. This kind of splitting is possible, but the disadvantage of it is that we reduce the number of samples that can be used to train the model. A solution for this issue is to use the cross validation. In this case the validation set is not needed anymore but the test set is still needed. In the approach called k-fold cross validation, the training set is divided into k sets and a model is trained using k-1 of the folds as a training data set. The rest of the data is used as a test set to validate the model. The performance measures are calculated based on the average of the values calculated in the loop.
"The choice of k is usually 5 or 10, but there is no formal rule. As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller" [KJ13]. A 5-fold cross validation is used for the evaluation of the used classifiers in this work.

3.2.2 Validation and Performance Scores

To choose one of the trained classifiers to label the unclassified data, different performance measures are considered. The common measures for supervised learning are precision, accuracy and recall. However, the accuracy of a classifier may be misleading because the prediction for one class may be very good and not good for the other class/es. Therefore, the confusion matrix is used to get an overview about the prediction of each class as described in the following paragraphs based on the book *Data Mining and Knowledge Discovery Handbook* [Cha09]. Table 3.1 shows the confusion matrix used for measuring the performance of the classifiers. Based on the values of the confusion matrix, precision, accuracy, recall and F1-value can be calculated as follows:

The **Precision** refers to the number of correctly predicted positive samples divided by the total predicted positive samples. This can be calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

	Predicted Class		
Actual Class		Class = Planning	Class = Not Planning
	Class = Planning	True Positive (TP)	False Negative (FN)
	Class = Not Planning	False Positive (FP)	True Negative (TN)

Table 3.1: Confusion matrix

Recall represents the number of correctly predicted positive samples divided by the number of all samples in the actual class.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

Accuracy refers to a measure based on the ratio of correctly predicted samples to the total number of samples. It can be misleading if the class distribution is uneven.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

The **F1-value** metric is a measure that combines a trade-off of precision and recall and returns a single number reflecting the "goodness" of the classifier. It is useful when the number of class distribution is imbalanced.

$$F1 - value = \frac{2 * (recall * precision)}{(recall + precision)} \tag{3.5}$$

Using the cross validation strategy, the average of the measures is taken from the overall k-folds runs.

3.3 Conclusion

Text classification is about classifying text into a specific class based on its content. In the context of the thesis at hand, classification is used to categorize the TripAdvisor posts shared in the Austria forum with the aim of identifying posts written by travelers in their inspiration phase (looking for a destination) and other phases. The labeled posts provide destination marketers with a better understanding of their customers' behavior in the forum (e.g. frequently asked questions, how they decide on a destination, their interests, etc.). The content of the posts is prepared using pre-processing text mining techniques, so that they can be processed by classification models. The classifiers Multinomial Naive Bayes, Logistic Regression and Support Vector Machines are used to build the classification models. The data is divided into a training and a test set and the results are validated through a 5-fold cross validation. The measures applied for the

evaluation are based on the confusion matrix taking into consideration: precision, recall, accuracy and F1-value. The results of the three classifiers are compared and the best performing classifier is utilized to classify the unlabeled data.

CHAPTER 4

Historical Tourism Data

In order to understand the historical arrivals data for visitors from United States and United Kingdom, descriptive analysis was conducted. Furthermore, the arrivals data was used for creating clusters of the regions of Austria.

4.1 Description of the Data Source

Each year, the World Economic Forum selects the best tourism destination and evaluates the attractiveness and development potential of about 140 countries. Austria is rated as an outstanding number twelve globally speaking and in terms of its tourist service infrastructure is even ranked as number one worldwide [iA18]. According to the Statistics Austria [Aus18], for the first time, more than 19 million nights were spent in tourism pre-summer (May and June) season 2018, which represents an increase of 4.3% compared to the previous year. Visitor historical data is used as a tool for travel tourism planning. The historical available dataset [1] was provided by our project partner. It contains arrival data for 100 regions of Austria on a monthly level in the period from January 2014 until February 2017 (38 months), specified for visitors coming from the United States of America and the United Kingdom. The regions were aggregated to facilitate analysis on the provinces' level. No details about the region of visitors are available (e.g.: in case of the USA, it is not known from which state people visit Austria). The regions of Austria were aggregated on provinces' level to enable a better overview. Austria has nine provinces: Vienna, Salzburg, Tyrol, Lower Austria, Upper Austria, Carinthia, Burgenland, Styria and Vorarlberg. The nine provinces with their corresponding regions are illustrated in figures 4.1, 4.2 and 4.3. "Other" contains arrival numbers that are not assigned to a known region. The following sections show some descriptive analysis of the data and a clustering of the 100 regions of Austria which was done based on the number of arrivals for both countries.

[1] Data is also available by statistics Austria, statistik.at

Province	Region
Burgenland	Mittelburgenland
Burgenland	Nordburgenland
Burgenland	Südburgenland
Carinthia	Bad Kleinkirchheim
Carinthia	Carnica Region Rosental
Carinthia	Hohe Tauern - die Nationalparkregion in Kärnten
Carinthia	Kärnten-Mitte
Carinthia	Katschberg-Rennweg
Carinthia	Klagenfurt
Carinthia	Klopeiner See - Südkärnten
Carinthia	Lavanttal
Carinthia	Lieser- und Maltatal
Carinthia	Millstätter See
Carinthia	Nassfeld-Pressegger See/Lesachtal/Weissensee
Carinthia	Nockberge
Carinthia	Region Villach
Carinthia	Wörthersee
Lower Austria	Donau NÖ
Lower Austria	Mostviertel
Lower Austria	Waldviertel
Lower Austria	Weinviertel
Lower Austria	Wiener Alpen in Niederösterreich
Lower Austria	Wienerwald
Upper Austria	Donau Oberösterreich
Upper Austria	Innviertel-Hausruckwald
Upper Austria	Linz
Upper Austria	Mühlviertel
Upper Austria	Pyhrn-Priel
Upper Austria	Zentralraum Oberösterreich
Salzburg	Fuschlsee
Salzburg	Gasteinertal
Salzburg	Großarltal
Salzburg	Hochkönig
Salzburg	Lungau
Salzburg	Obertauern
Salzburg	Pongau allgemein
Salzburg	Saalbach-Hinterglemm
Salzburg	Saalfelden-Leogang

Table 4.1: Regions and provinces of Austria - 1

Province	Region
Salzburg	Salzburger Saalachtal
Salzburg	Salzburger Seenland
Salzburg	Salzburger Sonnenterrasse
Salzburg	Salzburger Sportwelt
Salzburg	Salzkammergut
Salzburg	Stadt Salzburg
Salzburg	Tennengau - Dachstein West
Salzburg	Tennengebirge
Salzburg	Umgebungsorte Salzburg Stadt
Salzburg	Wolfgangsee
Salzburg	Zell am See - Kaprun
Other	Other
Styria	Schladming-Dachstein
Styria	Ausseerland-Salzkammergut
Styria	Hochsteiermark
Styria	Nationalpark Kalkalpen Region
Styria	Region Graz
Styria	Süd-Weststeiermark
Styria	Thermenland Steiermark - Oststeiermark
Styria	Urlaubsregion Murtal
Tyrol	Zell-Gerlos
Tyrol	Achensee
Tyrol	Alpbachtal und Tiroler Seenland
Tyrol	Erste Ferienregion im Zillertal
Tyrol	Ferienland Kufstein
Tyrol	Ferienregion Hohe Salve
Tyrol	Ferienregion Nationalpark Hohe Tauern
Tyrol	Imst Tourismus
Tyrol	Innsbruck und seine Feriendörfer
Tyrol	Kaiserwinkl
Tyrol	Kitzbühel Tourismus
Tyrol	Kitzbüheler Alpen - Brixental
Tyrol	Kitzbüheler Alpen-St.Johann i.T.
Tyrol	Lechtal
Tyrol	Mayrhofen
Tyrol	Naturparkregion Reutte
Tyrol	Osttirol
Tyrol	Ötztal Tourismus

Table 4.2: Regions and provinces of Austria - 2

Province	Region
Tyrol	Paznaun - Ischgl
Tyrol	Pillerseetal
Tyrol	Pitztal
Tyrol	Region Hall - Wattens
Tyrol	Seefeld
Tyrol	Serfaus-Fiss-Ladis
Tyrol	Silberregion Karwendel
Tyrol	St.Anton am Arlberg
Tyrol	Stubai Tirol
Tyrol	Tannheimer Tal
Tyrol	Tirol West
Tyrol	Tiroler Oberland
Tyrol	Tiroler Zugspitz Arena
Tyrol	Tux - Finkenberg
Tyrol	Wilder Kaiser
Tyrol	Wildschönau
Tyrol	Wipptal
Vorarlberg	Arlberg
Vorarlberg	Alpenregion Bludenz
Vorarlberg	Bodensee-Vorarlberg
Vorarlberg	Bregenzerwald
Vorarlberg	Kleinwalsertal
Vorarlberg	Montafon
Vienna	Wien

Table 4.3: Regions and provinces of Austria - 3

4.2 Descriptive Analysis

This section deals with the basic analysis of visitor arrivals from the USA and the UK with the aim of understanding the behavioral patterns in the data. Figure 4.1 displays the number of arrivals of tourists from the USA and the UK to the nine provinces of Austria between January 2014 and February 2017. The x-axis represents the provinces: Burgenland, Carinthia, Lower Austria, Upper Austria, Salzburg, Styria, Tyrol, Vorarlberg and Vienna. The y-axis represents the number of arrivals to the provinces over three years, starting from 200000 to 1000000. The histogram on the left side shows the visitor arrivals from the USA and the histogram on the right shows the visitor arrivals from the UK. The favorite provinces for US visitors are Vienna, Salzburg and Tyrol, while for UK visitors the ranking changes: Tyrol, Vienna and Salzburg. A low number of arrivals was recorded for the remaining provinces.

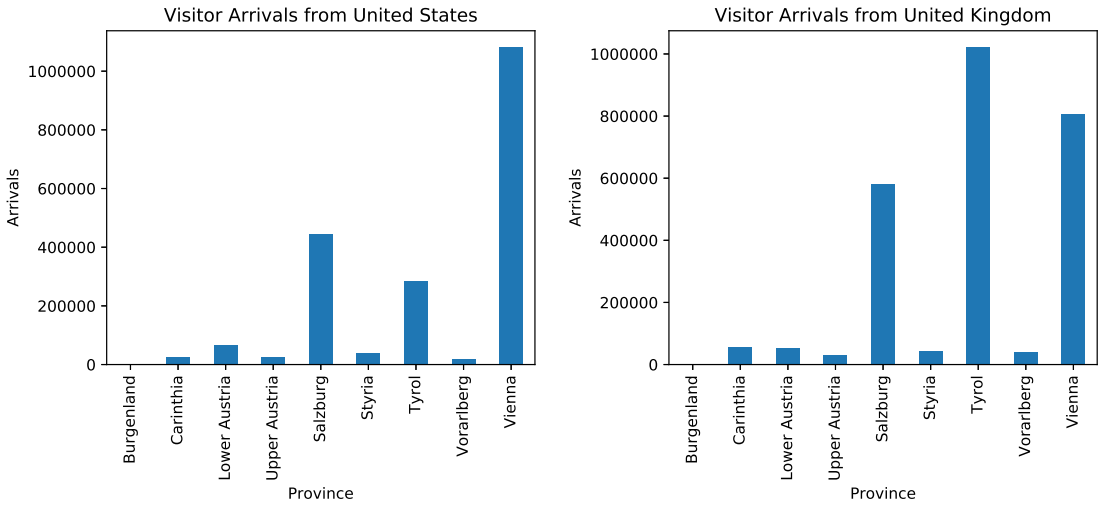


Figure 4.1: Visitor Arrivals from the USA and the UK to the nine provinces of Austria (01/2014 - 02/2017)

Figure 4.2 and 4.3 illustrate the distribution of the arrivals over all months for visitors from the USA and the UK to Austria. The x-axis shows the month at which the tourists arrived in Austria, starting from January 2014 until February 2017. The y-axis represents the number of visitors. The graphics show that the pattern of the arrivals had remained constant over the time period of three years. On the one hand, USA visitors tend to come more frequently during summer starting from May, June, July, August and until September. January and February do not seem to be of interest for visitors from the USA. On the other hand, UK visitors tend to visit Austria rather in the winter time: December, January and February, with lower number of visitors during June, July and August. The number of visitors from both countries is low in November.

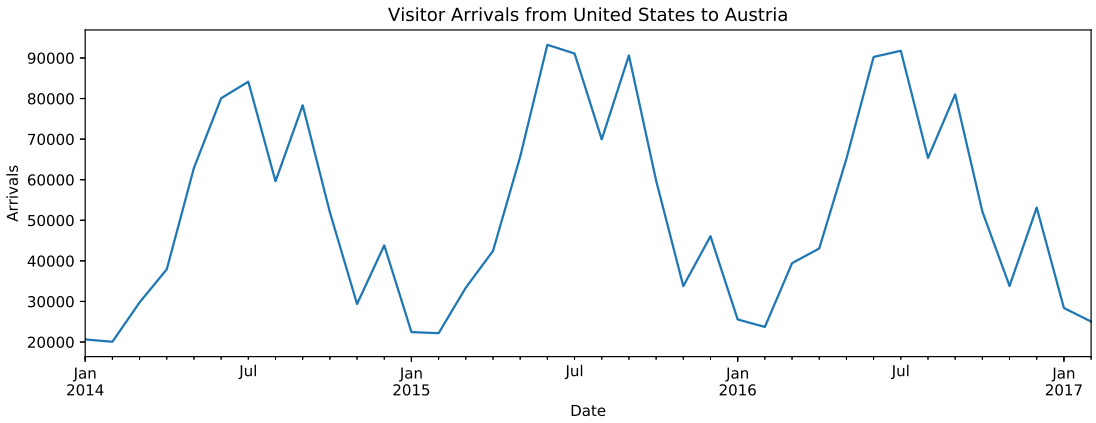


Figure 4.2: Visitor arrivals from the USA to Austria (01/2014 - 02/2017)

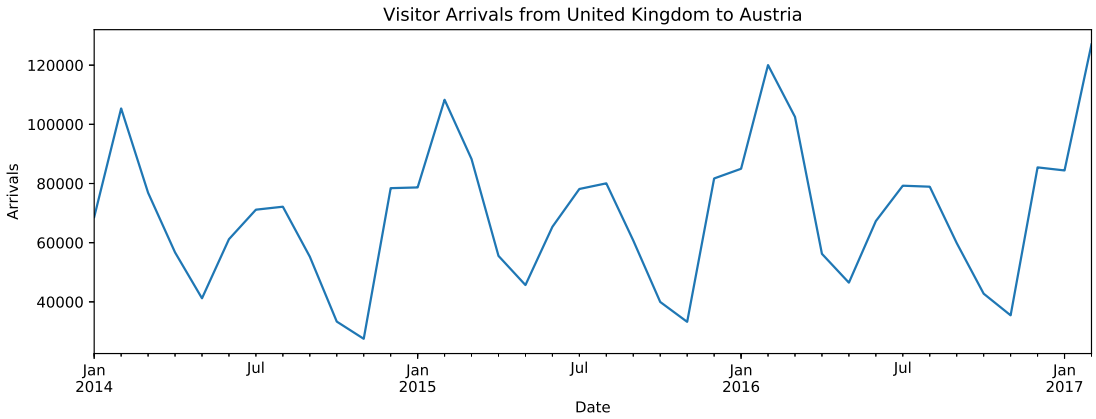


Figure 4.3: Visitor arrivals from the UK to Austria (01/2014 - 02/2017)

Figures 4.4, 4.5 and 4.6 show the tourists' arrivals behavior for the main three provinces: Vienna, Tyrol and Salzburg on a month-by-month basis. The x-axis represents the month in which the visitors arrived in the province and y-axis shows the number of arrivals over three years. Figure 4.4 displays the visitor arrivals from the USA and the UK to Vienna. During the given period of 3 years, Vienna is mainly visited during the summer months. In January, the number of visits from both countries is lower compared to the other months. This is mostly due to calender effects.

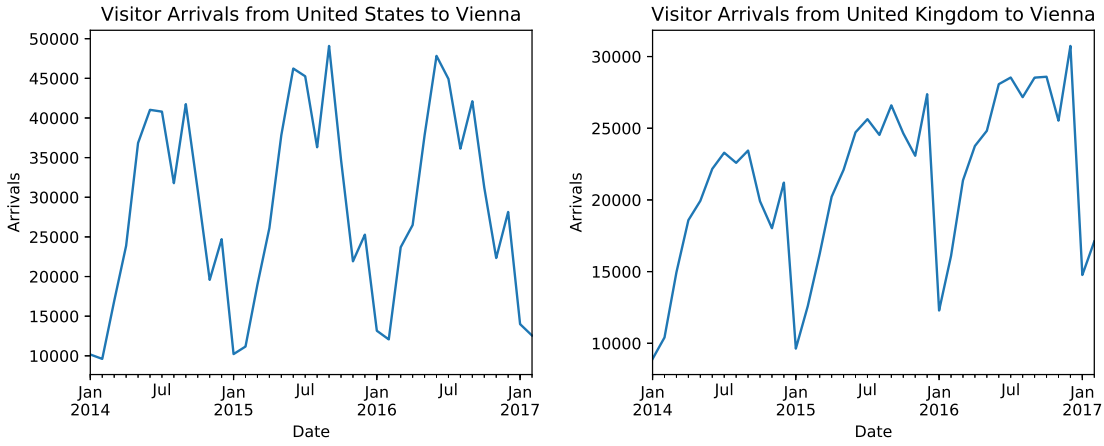


Figure 4.4: Visitor arrivals from the USA and the UK to Vienna (01/2014 - 02/2017)

Figures 4.5, 4.6 show the distribution of the arrivals to Tyrol and Salzburg. While Tyrol and Salzburg are mainly visited by travelers from USA in summer months, travelers from the UK come to Tyrol mainly in Winter. This is because of the alps and the skiing-season. Salzburg seems to receive more visitors from the UK. They visit Salzburg in summer months as well in January and February.

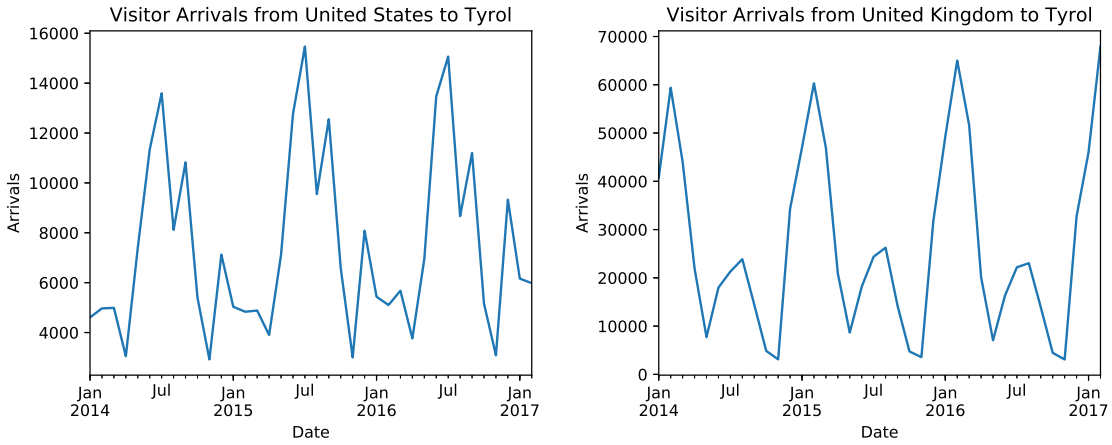


Figure 4.5: Visitor arrivals from the USA and the UK to Tyrol (01/2014 - 02/2017)

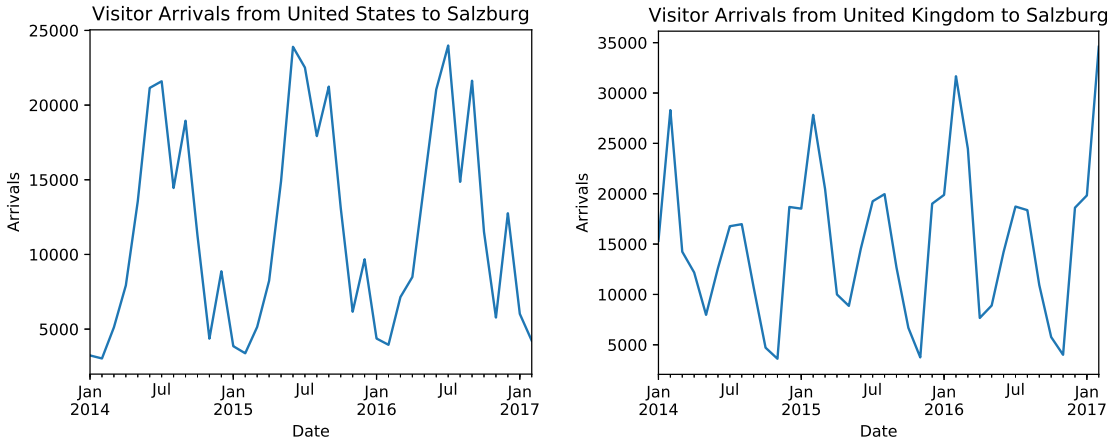


Figure 4.6: Visitor arrivals from the USA and the UK to Salzburg (01/2014 - 02/2017)

4.3 Cluster Analysis

Identifying meaningful clusters helps to understand the tourist interests, as well as to learn more about the similarity between the regions. The available data consists of 100 regions for the nine provinces of Austria. For each region, the time series, (historical arrivals) were taken as the input for the clusters. For clustering the time series, it is necessary to choose a distance measure that can identify similarities between time series. Therefore, the method Dynamic Time Warping method is used for the analysis and explained in details in the following section.

4.3.1 Dynamic Time Warping

Dynamic Time Warping (DTW) compares time series with the aim to find the optimum warping path between them [SE17]. It is commonly used for finding similarities between time series and has been widely used in the speech processing, online handwriting communities and bio informatics [RM03]. DTW is able to match various sections of a time series by warping of the time axis. The optimal alignment is determined by the shortest warping path in a distance matrix. A warping path consists of a set of contiguous matrix indices defining a mapping between two time series [EA12]. The DTW distance is considered as a robust distance measure for time series [SE17]. In other words, the basic idea behind DTW is to stretch and compress the time sequences in such a way that the distance between the two sequences is minimized [GRM15].

In the definition of the distance we assume that real numbers are used for time sequences. The distance calculation between two time series is based on the distance between the elements of the time series. Considering that we have two time series $x = x_1, x_2, ..., x_N$ and $y = y_1, y_2, ..., y_M$.

The distance for two indices can be defined by:

$$d(i, j) = |x_i - y_j| \tag{4.1}$$

The warping path is defined as a sequence $p = p_1, p_2, ..., p_L$ of index pairs $p_l = (i_l, j_l)$ fulfilling the following conditions:

- Boundary conditions: $p_1 = (1, 1)$ and $p_L = (N, M)$. That implies that the starting points and the end points of the two time series are mapped with each other.
- Monotonicity condition: $i_1 \preceq i_2 \preceq ... \preceq i_L$ and $j_1 \preceq j_2 \preceq ... \preceq j_L$. This condition guarantees that the order of the sequences is preserved.
- Step-size condition: $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$. This condition is also called symmetric step size and ensures that the value of the index is increased in each step at least from one-time series to the next one. Different step-size conditions are used in several other definitions.

The cost of the warping path is defined by:

$$D_p = \sum_{l=1}^{L} P_l = \sum_{l=1}^{L} d(i_l, j_l) \tag{4.2}$$

The optimal warping path can be defined based on the cost of the warping path as follows:

$$D(x, y) = \min D_p \tag{4.3}$$

After calculating the DTW for all pairs, the costs of the optimal warping are interpreted as a distance matrix between time series. This distance matrix is used for classification, segmentation or clustering algorithms [GRM15].

Clustering methods, in general, are divided into five categories: partitioning, hierarchical, density-based, grid-based and model based methods [SE17]. For determining the optimal number of clusters, there are different methods which can be used, such as the Elbow method, the average silhouette method and the gap static method. In the following, the methods are introduced briefly:

Elbow
The Elbow method looks at the percentage of variance explained as a function of the number of clusters. The idea is to choose a number of clusters, so that adding another cluster does not mean a significantly better modeling of the data. The first clusters will represent much information but at some point the marginal gain will drop dramatically and show a fraction in the graph. At this point, the correct number of clusters is selected [BK14].

Gap Statistic
This method can be applied to any clustering method. It compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The optimal clusters are the value that maximizes the gap statistic [sth].

Average Silhouette
This method measures the quality of a clustering which determines how well each object lies within its cluster. Each cluster is represented by the so-called silhouette, which is based on the comparison of its cohesion and separation. It shows how similar an object is to another object within the same cluster, compared to other clusters. The average silhouette width gives an evaluation of clusters and can be used to select an appropriate number of clusters. The coefficient of the silhouette lies between -1 and 1, where values close to 1 are indicating a good fit of an object to its own cluster and a bad match to other clusters, and values close to 0 or -1 indicate that the objects are either located in-between clusters or are wrongly grouped [Rou87, Ser18, sth].

Figure 4.7 and 4.8 show these methods using hierarchical clustering on the standardized data. In sum, we can see that the optimal number lies between 2 and 6.

The method used for clustering the regions is the hierarchical clustering, and the time warping distance is taken as the distance measure for building the clusters. The graphical representation used for hierarchical clusters is knows as dendogram. This is shown in figures 4.9 and 4.10. It looks like a tree that lists the regions which are clustered along the x-axis and the distance threshold or distance between the clusters(height) along the y-axis. The method *Ward.D* is used to create groups such that the variance is minimized within the clusters. At the beginning, each region is in its own cluster. Then, the closest clusters get grouped together and this procedure is repeated until all observations are in one cluster.

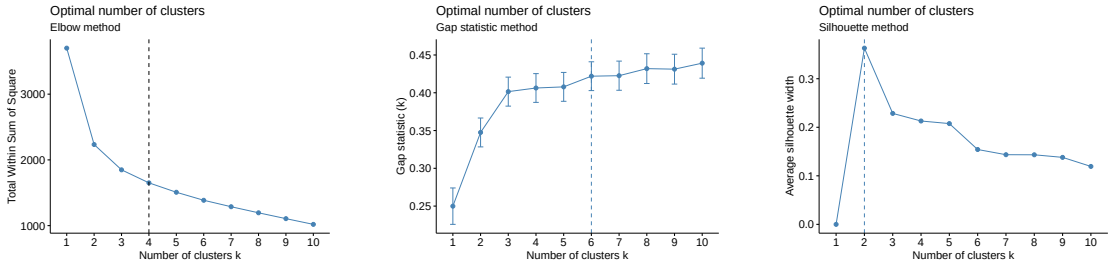


Figure 4.7: Different methods to determine an appropriate number of clusters for visitors from the USA

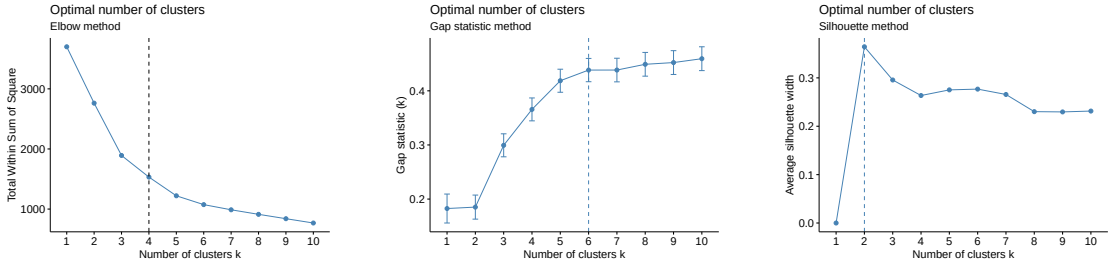


Figure 4.8: Different methods to determine an appropriate number of clusters for visitors from the UK

Figure 4.9 shows the clusters of the visitors from the United States of America to the regions of Austria. A reasonable number of clusters would be 4 clusters. The region Wien (Vienna) is in a cluster for itself, because the distance between it and other regions is high. A second cluster with “Stadt Salzburg” and “Innsbruck und seine Feriendörfer”, where the distance threshold between them and the all the other regions is low. A third cluster with mainly regions that belong to Salzburg and Tyrol as well as few regions from other provinces. “Salzkammergut”, “Arlberg”, “Donau NÖ”, “Region Graz”, etc. This shows that the favorite regions visited by tourists from the USA belong to the provinces Vienna, Salzburg and Tyrol. The fourth cluster contains the remaining regions.

The clustering of the regions visited by UK visitors are to be viewed in figure 4.10. Vienna is in a separate cluster as well given the number of visitors is clearly higher than in the other clusters. The rest of the clusters differs from the US clusters. The second cluster contains a great deal of regions that belong to Tyrol province and are visited in winter season such as “St.Anton am Arlberg”, “Paznaun-Ischgl”, “Mayrhofen”, etc. This cluster also includes three regions from the Salzburg province. The third cluster comprises cities as well as nature places that are mainly visited in summer, such as “Linz”, “Graz” and the “Salzkammergut”, etc. The fourth cluster includes remaining regions, which do not exhibit a high number of visitors.

In figure 4.11 some examples of the regions are grouped in the same cluster are illustrated, as well s the ones grouped into different clusters. The x-axis displays the month of the

year starting from 1 for January and 2 for February, 10 for October, etc. and the y-axis displays the number of arrivals per region. Graphs [a] and [b] in figure 4.11 illustrate the arrivals distribution over the 3 years for regions that were grouped in the same cluster. As we can see, the behavior and the number of arrivals over the 3 years show similar results. Graphic [c] shows the arrivals behavior for visitors from the UK to the regions of "Salzkammergut" and "Salzburger Sportwelt", which were categorized in different clusters. On the one side, the difference lies in the high number of arrivals. On the other side, there are also discrepancies in the behavior. Visitors seem to visit "Salzkammergut" more often in summer and "Salzburger Sportwelt" in winter.

In the following, further details regarding the interpretation of the clusters are provided. The number of regions in each cluster is stated in brackets.

Clustering for the Arrivals

In order to understand the characteristics of the regions in each cluster, Google Search and the online platform Hashatit[2] were used. Hashatit is a tool used to search social media content. Travelers share a lot of their travel experience in social media, such as Instagram, Facebook, Twitter, etc. The results of the search were filtered to find posts in Instagram, because the search result usually contains a picture and multiple hashtags. The search in this tool is based on hashtags (#), keywords, url, @mention shared in these channels. Looking at the hashtags for the regions mentioned in each cluster, such as "St.Anton am Arlberg", "Seefeld", "Mayerhofen", etc., a description of the regions was created. This summary is created based on manually observations. The analysis were done in *RStudio* tool.

UK Clusters

Cluster 1 (1)
Vienna is the capital city of Austria and exhibits the highest numbers of arrivals. The calculated difference between it and other regions is significant. Therefore, this cluster contains only Vienna (Wien).

Cluster 2 (13)
This cluster consists of 13 regions. The used hashtags for the regions of this cluster can be summarized in # alps, # wintersport, # nature, # ski, # snowboard and # mountainview.

Cluster 3 (14)
This cluster consists of 14 regions. The mentioned hashtags in its regions are # summer, # chill, # sun, # hiking, # alps, # flowers, # nature. The regions in this cluster show a high number of arrivals compared to cluster 4.

Cluster 4 (72)
This cluster contains the rest of the regions. It has a lot of regions that are visited during

[2]Hashatit: https://www.hashatit.com

summer and winter. The shared hashtags for these regions are similar to cluster 2 and cluster 3, but here the regions show a lower numbers of arrivals: # tree # hiking # landscape, # summer, # walk, # mountains # winter # snow.

USA Clusters

Cluster 1 (1)
Similar to by the UK clustering, Vienna also shows very high numbers of arrivals and exhibits its own cluster.

Cluster 2 (2)
This cluster contains "Stadt Salzburg" and "Innsbruck und seine Feriendörfer". These are the next regions with the highest number of arrivals. The hashtags found for these regions are # sun,# alps, # natur, # soundofmusic, # mozart.

Cluster 3 (15)
In this cluster 15 regions are presented. The regions cover winter and summer places that are visited by US visitors, but the number of visitors is much lower than in cluster 1 and cluster 2. These are a few of the hashtags found overall: # see, # wintersport, # ski, # nature,# relax.

Cluster 4 (82)
This cluster exhibits more than 80% of the regions. These regions show low numbers of visitors like in cluster 4 for the UK arrivals.

4.4 Conclusion

In this chapter, historical arrivals data from US and UK tourists that visited Austria in the period from January 2014 up until February 2017 was analyzed with the aim gaining insight into the interests and travel behavior of these tourists. Based on the analysis, the favorite regions for tourists from the USA and the UK were identified. Furthermore, the time at which travelers visit each of the provinces was determined. This output will be considered as an input while analyzing the online travel forum TripAdvisor in chapter 6 in order to investigate the correlation between the historical data and the online generated content by travelers from the USA and the UK. Additionally, hierarchical cluster analysis is applied on the regions of Austria to understand which regions show similar results over the year. The regions are grouped in 4 clusters based on the distance calculated using the dynamic time warping method to detect similarities between the different time series. The analysis of the regions' clusters mirror that travelers mainly tend to visit regions that belong to the provinces Vienna, Salzburg and Tyrol.

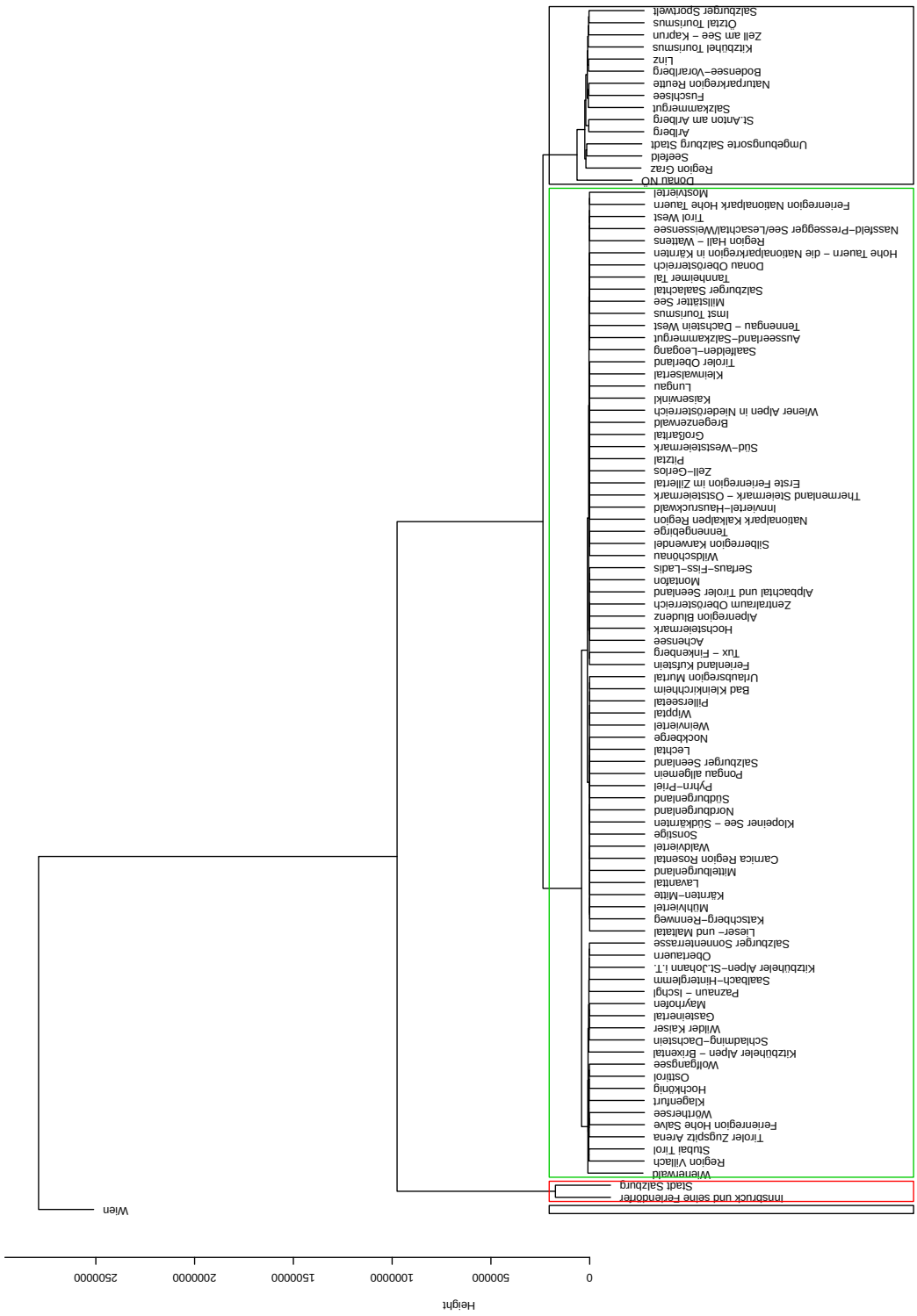


Figure 4.9: Regions clustering - Arrivals from the USA

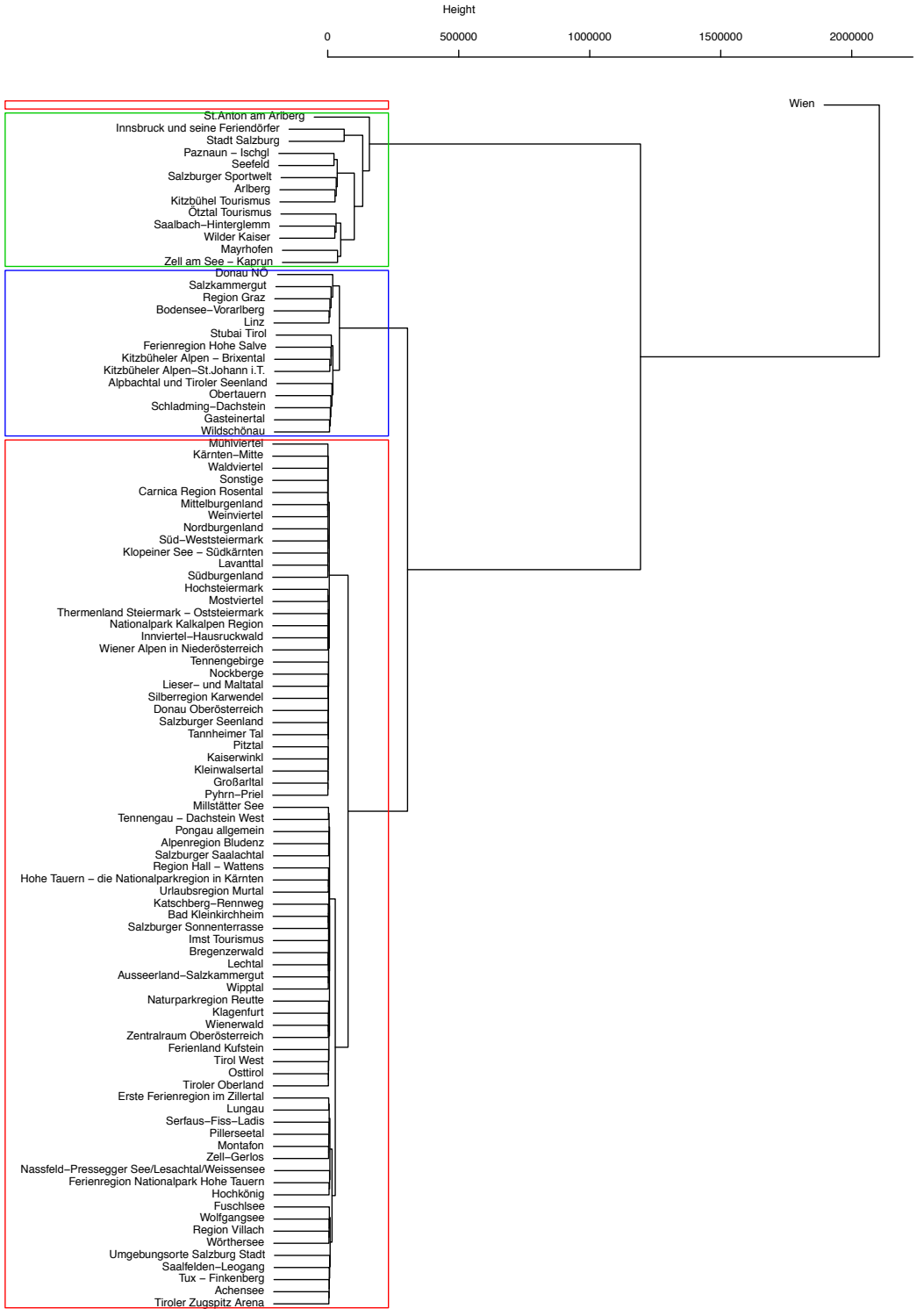
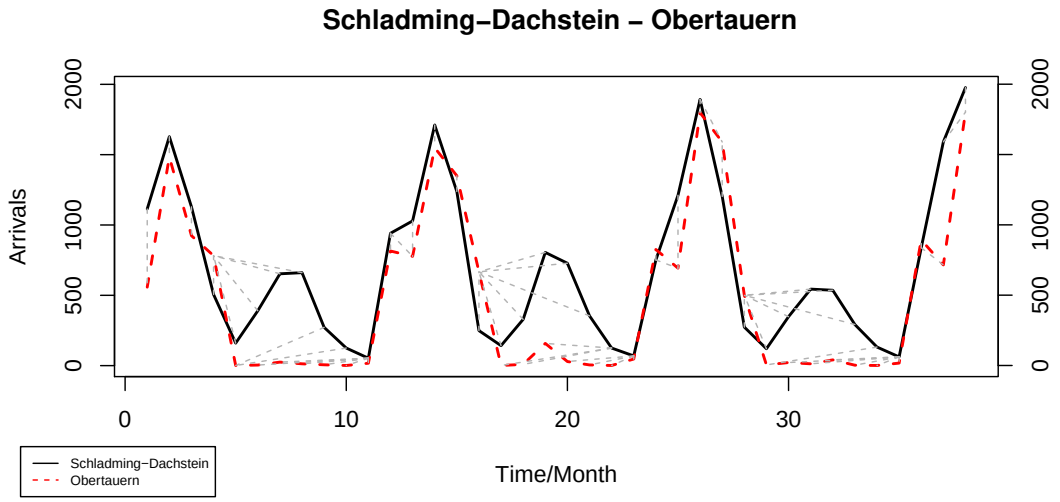
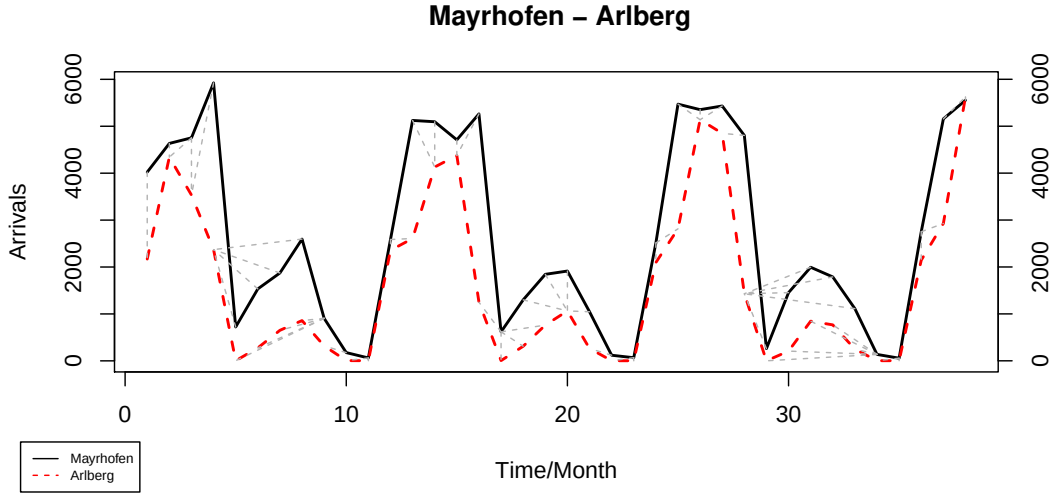


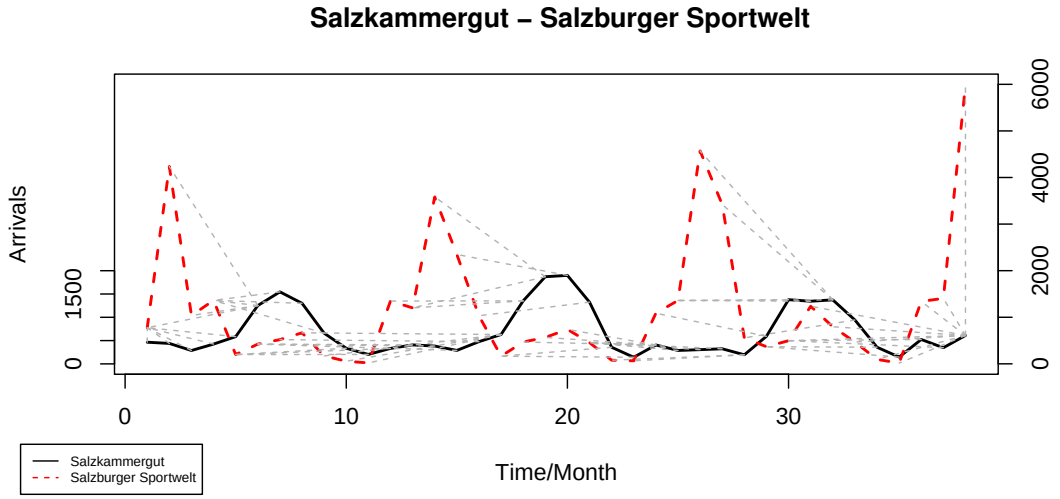
Figure 4.10: Regions clustering - Arrivals from the UK



(a) Regions in the same cluster



(b) Regions in the same cluster



(c) Regions in different clusters

Figure 4.11: Regions in different clusters - UK

CHAPTER 5

TripAdvisor Data Acquisition and Preprocessing

In this chapter, the selected data set will be thoroughly described and analyzed. In the following section different Social Media platforms will be identified in order to find the best fitting data source that provides new insights to understand the online behavior of the visitors' arrivals and their behavior in different phases of their travel journey. These insights aim to support tourism professionals such as destination marketers during the planning of their tourism marketing campaigns.

5.1 Description of the Data Source

Online social travel channels are changing the way travelers plan their trips. These websites enable users to actively interact and share reviews of hotels, accommodations and local tourist attractions [MBC08]. TripAdvisor is the world's largest travel site supporting travelers in planning their whole trip. With over 661 million reviews and opinions covering the world's largest selection of travel listings worldwide, TripAdvisor provides travelers with content that helps them decide on their travel destination, which activities to do, how to fly and where to eat. It compares prices from more than 200 hotel booking websites to find the lowest prices for its users. Furthermore, it covers around 7.7 million accommodations, airlines and restaurants. TripAdvisor is represented in 49 markets with around 490 million monthly visitors. [Tri18a, Tri18b]

TripAdvisor UK [UK] was selected as main source for our analysis because:

- The forum is public and active
- The diversity of the posts in the forum is high

- The high number of users
- The availability of the user profiles
- The language of the forum is English

Other platforms including Lonelyplanet: `https://www.lonelyplanet.com/`, Twitter: `https://twitter.com/?lang=en`, Reddit: `https://www.reddit.com/r/travel/`, Travellerspoint: `https://www.travellerspoint.com/`, etc, were analyzed but were not considered due to several reasons including missing data, e.g., the country of the user, not enough content, inactive forums and privacy issues.

TripAdvisor forums offer the users various ways to share their experiences, such as writing reviews about hotels or accommodations, and writing in the forums. There are different forums for countries all over the world. Users can start a new topic and ask questions or share their travel experiences. Since the focus of the thesis lies on Austria, the Austria Travel Forum [1] as shown in figure 5.1 is taken as the main data source. The forum has the following sub forums: Austrian Alps, Burgenland, Lower Austria, Styria, Upper Austria and Vienna Region. Each sub forum has its own sub forums.

Users can ask any questions in the forum. These can be about transport, food, sightseeing, accommodation or anything else. They can start the topic in the Austria forum or any of the sub forums. For each topic shown in figure 5.1, the following details are available: forum name, topic title, writer of the topic, number of the replies, date of the last post and the replies. As we are interested in those users that started a topic and are from the USA or the UK, it is necessary to know the home country of the users. This information is gained from each user profile, which is in some cases not available or misleading because of short abbreviations provided by users. To be able to participate in the forum, users have to create a profile on TripAdvisor. They can join TripAdvisor using their Facebook account, Google account or any email address.

5.2 Data Acquisition

The Scrapy Framework was used for acquiring the data from the web. It is an open source and a collaborative framework for extracting data from websites in a fast and simple way [Smoc17]. The framework was used to extract the posts of the forum, as well as the user profiles for users that started a so-called “topic”.

5.2.1 Austria Forum Posts

The posts of the Austria forum with all sub forums (figure 5.1) were downloaded. Each post consists of these fields: forum, topic, number of replies, last reply date, post date, post content, user name and user link, and some more. The fields forum, content, post date, user link were used for the analysis. They are described in table 5.1

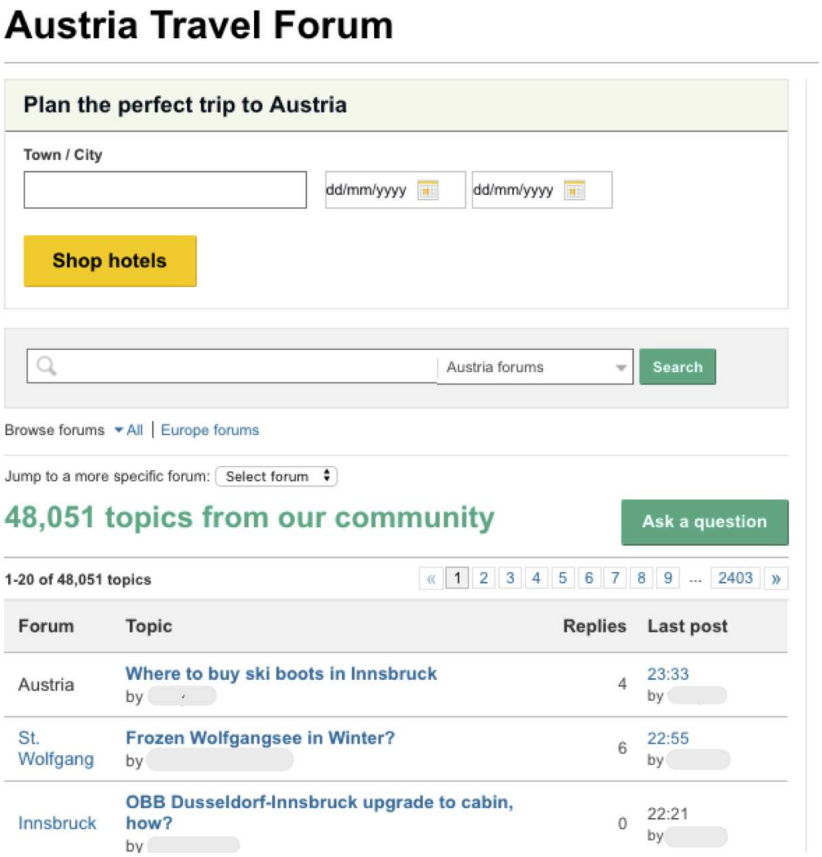


Figure 5.1: TripAdvisor UK - Austria forum

Field	Description
Forum	The name of the forum, where the post is written (e.g.: Austria forum, Tyrol forum, etc.)
Topic	The title of the post written by the user.
Content	The whole content of the post.
Post date	The date/time at which the post was created.
User name	The name of the writer of the post.
User link	The link to the profile of the writer of the post. This information is extracted in order to find out the user' country of origin.

Table 5.1: Used data for the forum posts for the analysis

The number of downloaded posts is about 43.000. The post title and content as well as the meta-data of all the posts were downloaded. The content provides detailed information about the post, as well as other information useful for the analysis. An example of a post is shown in figure 5.2. In this figure, the date of the post is mentioned which is an essential information that shows when the user started planning for his/her trip.

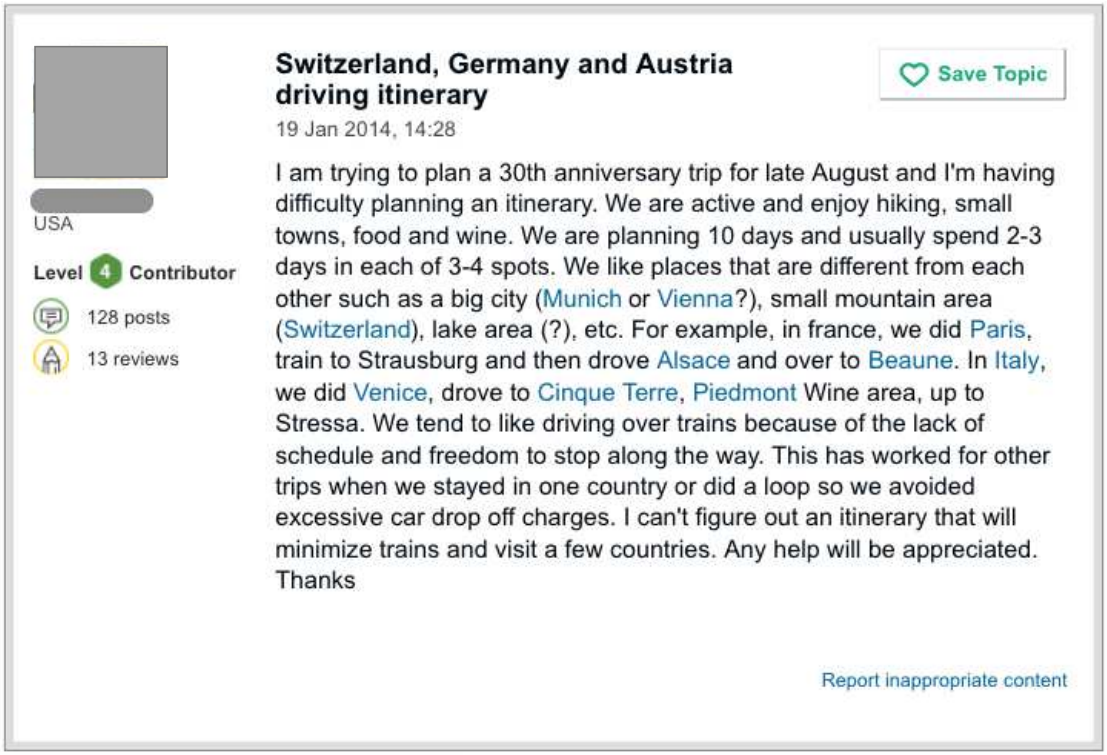


Figure 5.2: TripAdvisor UK - Example of a post written in the Austria forum

[1] https://www.tripadvisor.co.uk/ShowForum-g190410-i146-Austria.html

5.2.2 User Profiles

Since the focus of the thesis lies on users that are from the UK and the USA, all profiles of users that started a conversation in any of the Austria forums were downloaded. The number of unique users lies at around 23.000 users. These profiles are publicly available. The used fields for the analysis are: hometown, age and gender. Beside of these fields, the following fields described in table 5.2 are available in each profile.

Field	Description	Type
User name	The name of the user that started a post. This information is hidden in figure 5.3 for privacy issues.	free text
Since	The date on which the user joined TripAdvisor.	date in format month/year
Age and Gender	Age and gender of the user. This is an optional information and is not provided by all users.	categories: (13 - 17), (18 - 24), (25 - 34), (35 -49), (50 - 64), (65+), empty
Hometown	The hometown of the user. This is optional.	free text
Number of reviews	The number of reviews for hotels, restaurants, attractions or airlines.	number
Number of ratings	The number of ratings for locations.	number
Number of posts	The number of posts created by the user.	number
Number of photos	The number of photos published by the user.	number
Number of helpful votes	The number of helpful votes given by other members to these users.	number
Travel style	The travel style of the user, of which there are 6 levels. Each level is based on the contributions of the user to the TripAdvisor community. The more points the user earns, the higher his/her level.	categories
Total points	The total points the user gets from his/her contributions.	number
Badges	The number of points the user has gained from badges, such as readership, helpful reviews, top photographer, top contributor, etc.	number

Table 5.2: User profile details

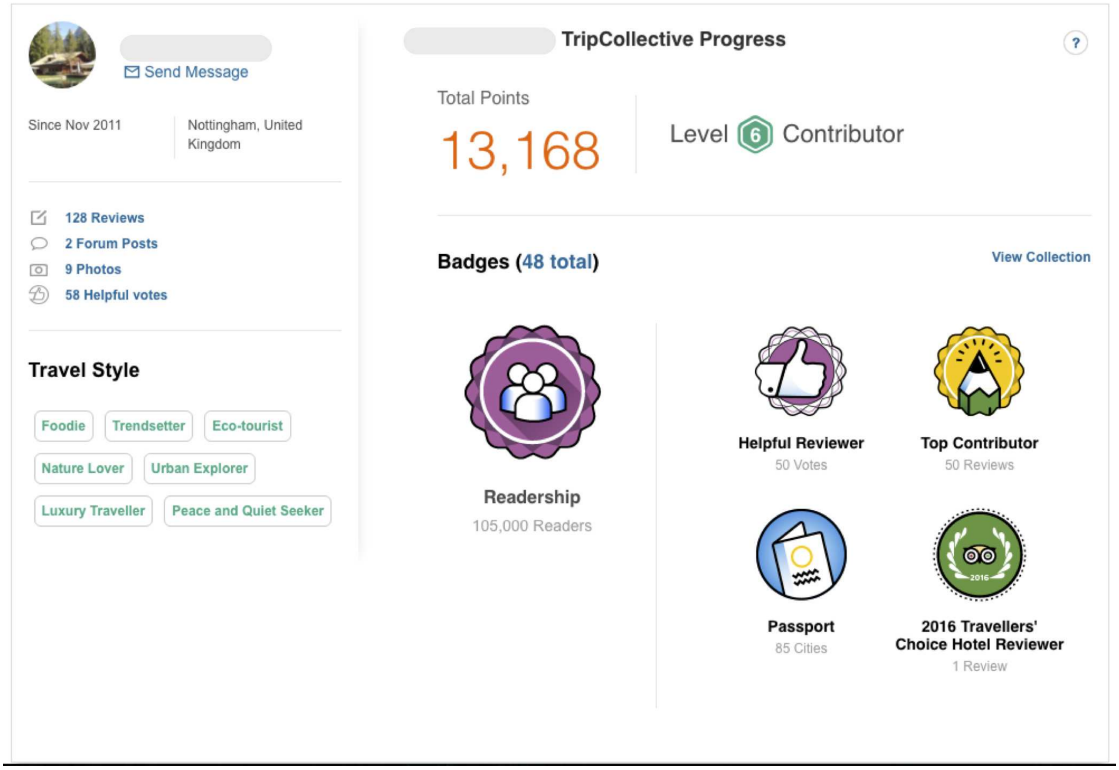


Figure 5.3: TripAdvisor UK - Example of a user profile

In figure 5.3 we can see that the user comes from Nottingham, United Kingdom and has been active in the forum since November 2011. This user did not provide his/her age or gender. He/she is a contributor, which means, an active user with a high readership.

Note: the user information is used only for the purpose of research in an aggregated form.

5.3 Data Preparation

In order to analyze and understand unstructured data created by travelers on TripAdvisor Austria forums in the form of post, machine learning algorithms are used to classify this content. The aim is to automatically structure and extract relevant information from the data to help destination marketers gain insights into topics discussed online by travelers from the USA and the UK while they plan their trip to Austria. This will reflect the interests and motivations of the travelers. The advantage here is that the analyses are based on the content directly generated by the users without asking them any direct questions. To prepare the data for the analysis, a set of the posts is labeled and then cleaned. Further details will be explained in the following sections.

The illustrated process in figure 5.4 [Zah18] is used to classify the text. The first step is about having a data set labeled to use this knowledge for the automatically text

classification. To prepare the data to be used by the classifiers, several pre-processing steps were executed such as removing stop word, stemming, etc. With a cleaned data set, different classifiers were executed. For finding the best classifier, the following measures were considered: accuracy, confusion matrix and recall. In the last step of the process the best model was selected and used to classify the unlabeled data.

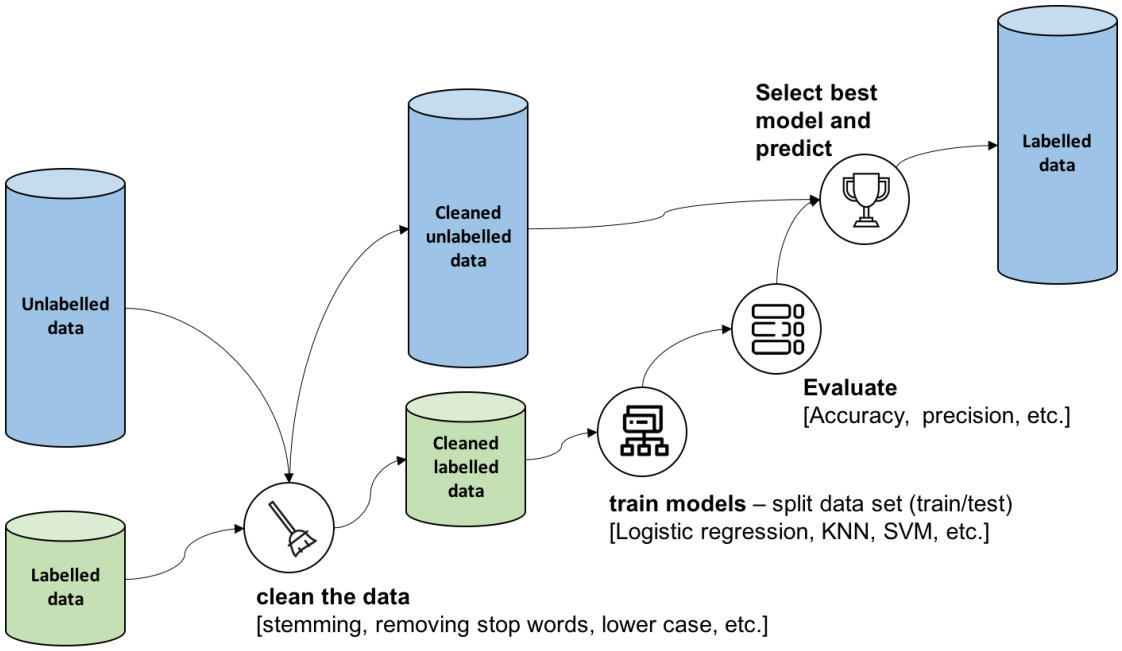


Figure 5.4: Classification process of the posts

5.3.1 Data Labeling

The process started with labeling a set of the data by the expert. Our project partner has many years of experience in the tourism field and was responsible for labeling part of the data set. Different aspects were identified in the data. Therefore, the topics were categorized into three classifications. One aspect included identifying users' behavior in the forums during their inspiration or planning phase. The idea behind it was to understand more about travelers in their inspiration phase and their use of social media channels in this phase. Another aspect comprised detecting the commonly discussed topics by travelers.

The first classification is content-based, which means each post is categorized into a topic based on the content of the post. It consists of the following ten categories:

1. Transportation: Posts about transport and how to come or go to a specific place.
2. Sightseeing: Posts about sightseeing places in Austria (e.g.: Schoenbrunn)

3. Accommodation: Posts about accommodation such as where to stay, which hotel is better, etc.
4. Culture: Posts about culture.
5. Food: Posts about restaurants, bars, or anything related to food.
6. Destination choice: Posts about choosing between two destinations, or motivation to choose a specific destination, etc.
7. Things to do: Posts about ideas and recommendations for things to do during vacation.
8. Sports: Posts about anything related to sports, such as skiing, horse-riding, etc.
9. General: Posts about general things, such as weather, safety, etc.
10. Other: Posts that do not fit any of the above mentioned categories, including questions about phone sim cards.

The second classification is phase-based and describes the phase in which the tourist is currently in. It contains the following three categories:

1. Inspiration: Getting inspiration about one's next vacation destination or when trying to choose between two destinations.
2. Planning: Asking questions about details for planning the next trip.
3. Other: Asking general questions (e.g.: weather, safety, etc.).

The third classification explains in detail the planning and inspiration phase. It consists of the following five categories:

1. Planning Local: Someone is already in Austria and is asking specific questions during his/her vacation.
2. Planning Austria: Someone has already taken the decision to come to Austria and has questions about his/her vacation (planning questions).
3. Planning International: Someone is traveling from or to Austria and other countries (e.g.: round-trip) and has questions, for example about train-connections, etc.
4. Choice Austria: Someone has decided to come to Austria but is looking for new ideas/help regarding the destinations/specific places, he/she is going to visit.
5. Choice International: Someone does not know yet where to spend their next vacation and is planning multiple destinations, e.g.: someone wants to do skiing but is looking for various places in the alps.

In the beginning, only the title of the post was used to label the data. Some examples are shown in table 5.3. In some cases, the title of the post did not provide enough information enabling a classification of the content. For example, “Vienna”, “Urgent Info”, etc. Therefore, for each post title the whole content was downloaded and used for classification. After having discussions with our expert, the focus of the classification was set to be on the 10 categories and 6 categories.

Topic	**Label 1 (Content-based)**	**Label 2 (Phase-based)**	**Label 3 (Detailed Phase-based)**
Train Munich-Kufstein-Salzburg	Transportation	Planning	Planning International
Opera Tickets	Culture	Planning	Planning local
Paris, Vienna Trip advice	Destination choice	Inspiration	Choice International
Lech or Wengen?	Destination choice	Inspiration	Choice Austria
Danube boat ride, Schonbrunn palace – which tour is best?	Sightseeing	Planning	Planning local
Hotel recommendation	Accommodation	Planning	Planning Austria
Ice Creame and Cake Café recommendations	Food	Planning	Planning local
Heat wave next week	General	Other	Other
First Skiing Holiday	Sport	Planning	Planning Austria
Salzburg and near places Planning.	Things to do	Planning	Planning Austria
Austria Vignette	Other	Other	Planning

Table 5.3: Examples of the classification of the posts in the Austria forum

5.3.2 Data Pre-processing

A challenging task in the pre-processing of the data was about finding out the country of the user. It is not mandatory for users to indicate their country of origin, but however, if the users wants to do so, a text field as shown in figure 5.5 is provided. Into this text field, anything can be entered. Some of the users enter abbreviations for their, as well as other terms including: home, the world, etc., into the location field. Therefore, it was necessary to clean the data to filter out user profiles that do not provide any information about their country or information that is not meaningful. Furthermore, it was necessary to map the abbreviations of the countries to the full location name. This was achieved using the LocationIQ API [Lab18]. Examples of mapping are: mapping the abbreviation NY to New York, United States; California to California, United States. This step was necessary in order to find users from the USA and UK.

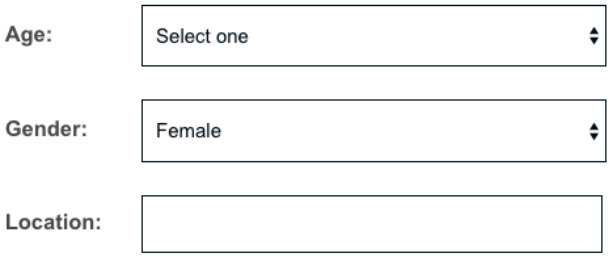


Figure 5.5: TripAdvisor UK - Form to enter age, gender and location

Table 5.4 shows an example of a post content of the forum before and after the pre-processing. The pre-processing and classification models were done using Python and the data science studio of Dataiku [Dat18] (Version 4.1.0).

Content	**Cleaned Content**
We are planning to visit Austria in mid of October for 3 days. What could be the best possible to places to visit to have a nice view of Alps (maybe snow?) and to enjoy scenic nature ? How would be the weather in Mid-October generally and is it recommended to visit during this time ?,After our 3 days visit in Austria we are thinking to continue to Black forest.,Thank you very much in advance.	plan visit austria mid octob day best possibl place visit nice view alp mayb snow enjoy scenic natur weather mid octob general recommend visit time day visit austria think continu black forest thank veri much advanc

Table 5.4: Example of a post content before and after preprocessing

- Text Normalization
  Normalizing the text is about transforming it to lower case and removing accents as well as normalizing unicode (e.g.: Café -> cafe)
- Numbers Removal
  Numbers mentioned in the text do not add value to the classification and are, therefore, removed.
- Stop Words Removal
  A vast number of words reocurrs frequently in the texts, however, such words

are essentially not meaningful because they serve as connectors in a sentence or a paragraph. Examples of these stops words: *the, I, a of, etc.*. Stop words do not contribute to the context or content of textual documents. Moreover, they do not add value to the classification of documents [KG14] or text. Therefore, it is recommended to remove them in order to enable a more efficient evaluation in terms of word frequency.

- Words Stemming
  Stemming is applied to reduce different grammatical forms of a word. It is used to identify the root or the stem of the word. For example, the words connected, connect, connecting, connection will be stemmed to the word connect. The aim of this method is to remove the prefix and suffix of the word to have accurately matching stems, which saves time and memory space [VIN15].

- Tokenization and Term Weighting
  Tokenization is used to split each document (e.g. a post from the forum) into individual words. The split is normally done based on white spaces in the text. Term weighting is about having a representative vector with which the machine learning algorithms can work. A sample method is based on counting the occurrences of each token/word. There are also other ways that affect the importance of a term in a document, such as Term Frequency (TF), Inverse Document Frequency (IDF) and Document length normalization. Term frequency of each word in a document expresses the importance of this word in the document. It is calculated based on the number of occurrences of each word in a document divided by the total number of words in the whole document. Inverse Document frequency of each word (IDF) is a weight that depends on the the distribution of each word in all documents. This means, if a word occurs frequently in all documents, then it is not as important as a word that appears only in specific documents and makes those documents recognizable. TF and IDF (TF/IDF) can be used together to determine a weight term for each word. This explanation is based on the paper written by Srividhya et al. [SA10].

- Handling Imbalanced Data
  The problem of imbalanced data set occurs, when the number of instances of one class is much lower than the instances of other classes. Machine learning classifiers tend to focus on the major classes and ignore the tiny ones, which impact the classification of the minority classes. There are different techniques for dealing with this problem [RM14]. One popular technique for balancing the class distribution is sampling, which is about changing the data by undersampling majority classes, or oversampling minority classes, or both. However, the random under-sampling of data can potentially remove certain important information, and the random over-sampling of data can lead to overfitting [Cha09]. In the last years, several techniques were introduced to handle issues, such as the Synthetic Minority Over-sampling technique (SMOTE), Adaptive Synthetic Sampling Approach for Imbalanced data sets (ADASYN), etc. [RM14].

In the context of this thesis, the classes were imbalanced, as shown in figure 5.6. This figure demonstrates the following for both classifications: the ten categories and the six categories as well as the number of manually labeled posts in each class. During the classification of the ten categories the focus lied on labeling posts written by travelers in their inspiration phase when looking for a destination for their next trip. Therefore, the class *destination choice* is the first class with 134 posts. Many of the posts conversations are related to transportation. This is why, the second most represented class is *transportation* with 117 posts. The third class is about posts with questions about activities that can be done during the trip. The other classes: *culture, sport, general, sightseeing, accommodation, food and other* are represented with an average number of around 40 posts.

Figure 5.6 additionally illustrates the number of posts in each of the six classes. The majority of the labeled posts are written in the planning phase shown in the classes *planning local, planning international and planning austria* and a reasonable number of posts is written during the destination choice in the classes *choice austria and choice international*.

The undersampling technique was used and the number of data samples was adapted until the classes were balanced, so that an optimal confusion matrix could be achieved. Undersampling was selected because many of the majority classes such as *Transportation, things to do and culture* featured a number of observations that were redundant. By removing them the data distribution did not show any significant change [DPCB15].

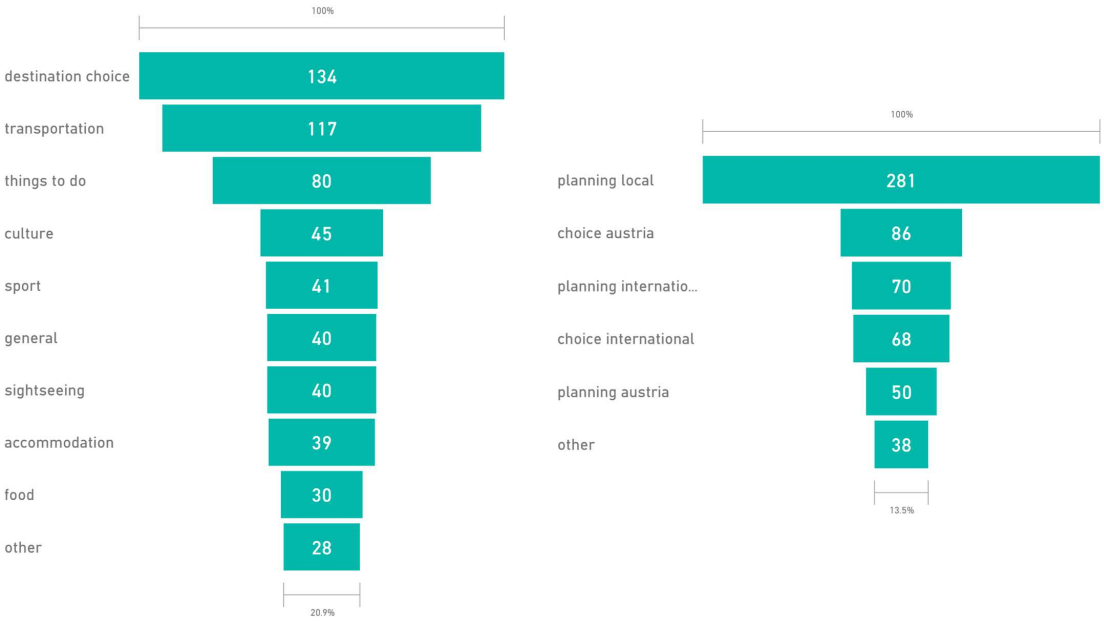


Figure 5.6: Manually labeled data (imbalanced)

5.4 Conclusion

TripAdvisor was selected as the data source to be used for the analysis. The data extracted from TripAdvisor contained the posts and users data that started a conversation in the Austria travel forum. It was essential to ensure that the chosen data source contains all the information necessary for the analysis. In the context of this thesis, it was important to be able to extract travelers' country of origin in order to analyze a specific group of users. After manually labeling the data, the data set was imbalanced, because some of the classes had more classified posts then others. This process can be improved. For example, topic analysis can be used in future work to reduce the manual work and to automatically extract meaning from texts in order to identify specific topics, which can be reviewed by experts.

CHAPTER 6

Travel Review Sites Analysis

In this section, the acquired data from TripAdvisor is analyzed in details. The data is rich and contains several information useful for the destination marketing of Austria. In particularly, for understanding the traveler behaviors of the forum, their interests and needs. The analysis of the data is mainly covering answers to following questions:

- *What are travelers writing about in the Austria forum?*
- *When do travelers from USA and UK start planning their trip to Austria?*
- *Who are the users of the forum?*

Based on observations, it was recognized that people using TripAdvisor have different needs and questions. Some of them join the forum as experts for a specific destination and share their experiences with others. Those kind of users have a high number of points and a profile filled with information about themselves. Other users participate in the forum to ask questions and get inspiration from others. These people have different behaviors. Some of them ask few questions before or during their trip. Another part of them share their exact itinerary and ask more frequently questions. Some people are there to get some inspiration and ask questions about different destinations to take a decision regarding the destination for their next trip. There are different kind of users, but those are the popular ones found during the analyses.

Travelers discuss different topics in the forum such as transport, accommodation, sightseeing, culture, etc. To gain some insights of the interesting topics for travelers, the posts were classified. Since the number of posts in the Austria forum is more than 43.000 posts written by users from overall the world, it is not efficient to label them manually. Therefore, machine learning techniques are used to train a model based on experts' knowledge and automatically classify the content of the posts. This saves a lot

of time and is easy to maintain in the future. This section analyses the content of the posts of the Austria forum to find out the topics that travelers are talking about.

6.1 Classification of The Travel Forum Posts

After the classes were defined by the expert, around 600 posts were manually classified. About 430 records were considered, because the data was imbalanced. This data was rebalanced by selecting randomly items/rows to rebalance all classes using undersampling technique. The data was divided in 80% training set and 20 % test set. The labeled data set was used for building the models to automatically classify the rest of the data, which was done using supervised learning. Based on literature research described in chapter 3, the algorithms: Logistic Regression, Support Vector Machines and MultinominalNB were chosen for the classification of the posts. The result of the classifiers was compared and the Logistic Regression algorithm was selected, because it was the best performing algorithm and the output of the confusion matrix has a good coverage overall the classes. The results of the classification models are described in the next section.

Models

Level	Feature Extraction	Classifier	Accuracy	Precision	Recall	F1-Value
Content-based (10 classes)	CountVectorizer	Logistic Regression	**0.5730**	**0.6024**	**0.5679**	**0.5695**
		Support Vector Machines	0.3587	0.5512	0.3827	0.3726
		MultinomialNB	0.4382	0.5522	0.4360	0.4459
	TfIdfVectorizer	Logistic Regression	0.5618	0.5756	0.5463	0.5449
		Support Vector Machines	0.5169	0.5172	0.5034	0.4991
		MultinomialNB	0.5506	0.5356	0.5287	0.5261

Table 6.1: Results of the supervised classification experiment for 10 classes. Bold numbers indicate the best results.

Tables 6.1 and 6.3 contain the results for each classifier. For the feature extraction of the text, two vectorizers were used: CountVectorizer and TfIdfVectorizer. Both vectorizers feature the "Bag of words" model, which is about assigning a weight for each word solely based on its frequency in the data set. TF-IDF stands for "term frequency-inverse document frequency". The weight assigned to each word depends not only on its frequency in a document but also on how frequent it is in the rest of the whole data set. CountVectorizer counts the frequency of a word and assigns that count to the weight of the word [CRV].

The evaluation of the classifiers was not only based on the accuracy, precision, recall and F1-value, but also on the confusion matrix. This is because it provides an accurate overview of how well all classes are predicted. For the content-based categorization with

the 10 categories, we can see that using count vectorization for feature extraction and Logistic Regression exhibits the best performance compared to other classifiers.

	Predicted										
Actual	**destination choice**	**transportation**	**things to do**	**culture**	**sport**	**general**	**sightseeing**	**accommodation**	**food**	**other**	
destination choice	62 %	8 %	23 %	0 %	0 %	8 %	0 %	0 %	0 %	0 %	100 %
transportation	0 %	78 %	11 %	0 %	0 %	11 %	0 %	0 %	0 %	0 %	100 %
things to do	8 %	8 %	58 %	0 %	8 %	0 %	8 %	0 %	0 %	8 %	100 %
culture	0 %	0 %	25 %	75 %	0 %	0 %	0 %	0 %	0 %	0 %	100 %
sport	38 %	0 %	0 %	0 %	63 %	0 %	0 %	0 %	0 %	0 %	100 %
general	30 %	10 %	0 %	0 %	0 %	30 %	20 %	10 %	0 %	0 %	100 %
sightseeing	11 %	0 %	22 %	11 %	0 %	0 %	44 %	11 %	0 %	0 %	100 %
accommodation	0 %	0 %	13 %	0 %	0 %	0 %	0 %	75 %	13 %	0 %	100 %
food	0 %	0 %	0 %	0 %	0 %	0 %	17 %	17 %	67 %	0 %	100 %
other	0 %	17 %	0 %	17 %	0 %	17 %	33 %	0 %	0 %	17 %	100 %

Table 6.2: Confusion matrix for 10 classes using Logistic Regression

The confusion matrix for the 10 classes is showed in table 6.2. The majority of the classes are more than 50% correctly classified. A reason why the accuracy value in table 6.1 is low is mainly because of the miss-classification of the categories “other” and “general” as well as other categories.

Level	**Feature Extraction**	**Classifier**	**Accuracy**	**Precision**	**Recall**	**F1-Value**
Detailed Phase-based (6 classes)	CountVectorizer	Logistic Regression	**0.5256**	**0.6065**	0.4780	0.4964
		Support Vector Machines	0.2308	0.2671	0.2000	0.1298
		MultinomialNB	0.5128	0.5860	0.4952	0.5148
	TfIdfVectorizer	Logistic Regression	0.4416	0.4288	0.4518	0.3792
		Support Vector Machines	0.4545	0.4702	0.4561	0.4332
		MultinomialNB	0.4545	0.4524	0.4517	0.4233

Table 6.3: Results of the supervised classification experiment for 6 classes. Bold numbers indicate the best results.

In table 6.3 the results of the classifiers for the detailed phase-based classes are listed. Again, here the best result overall the classifiers is by the combination of count vectorization and Logistic Regression. This is shown in the Accuracy and Precision values.

In table 6.4 the confusion matrix of the detailed phase-based classes are listed. Many of the predicted classes are miss-classified with the class “planning local”. By looking at the automatically categorized data, the classified data showed that this class is widely covering many aspects and not only posts where travelers are locally in Austria.

	Predicted						
Actual	**planning local**	**choice austria**	**planning international**	**choice international**	**planning austria**	**other**	
planning local	67 %	27 %	7 %	0 %	0 %	0 %	100 %
choice austria	10 %	65 %	10 %	10 %	5 %	0 %	100 %
planning international	0 %	14 %	50 %	14 %	21 %	0 %	100 %
choice international	38 %	8 %	8 %	38 %	8 %	0 %	100 %
planning austria	25 %	25 %	8 %	0 %	42 %	0 %	100 %
other	75 %	0 %	0 %	0 %	0 %	25 %	100 %

Table 6.4: Confusion matrix for 6 classes using Logistic Regression

Overall, the classification of the posts can get better by enhancing the training data and adapting the segments or classes to be more specific and different from each other.

6.1.1 Segments Distribution of the Posts

The class distribution of the classified TripAdvisor posts is shown in figures 6.1 and 6.2. In figure 6.1, the six classes: *planning local*, *planning austria*, *planning international*, *choice austria*, *choice international* and *other* are represented. More than 20.000 posts are in the *planning local* class. This gives an insight into the usage of TripAdvisor and the frequency with which travelers ask questions about local issues like, for example, transport, sightseeing, restaurants, etc. The classes *choice austria* and *choice international* are well represented and also show an interesting aspect regarding the usage of the forums of TripAdvisor. This aspect refers to users using the forum in their early making decision process to choose their next holiday destination. Real examples [1] of the posts from the data are listed in tables 6.5 and 6.6.

In figure 6.2, the ten classes *transportation*, *things to do*, *accommodation*, *sightseeing*, *destination choice*, *general*, *culture*, *food*, *sport*, *other* are illustrated. The graph shows which exact topics are discussed by travelers in the Austria forum. The most represented class is *transportation* where users ask about a route, a better way, public transport, etc. with more then 9000 posts. Many of the conversions in the forum are about things to do during the trip. Travelers also tend to ask about accommodation, recommendations and sightseeing. More then 2000 posts were classified in each of the categories of *culture, food, sports* and *general*. Posts that could not be classified into any one of the mentioned categories were assigned to the class *other*.

[1] Names of countries or cities mentioned in the text as a reference/link to the city on TripAdvisor were not crawled and are written in the text as “comma”

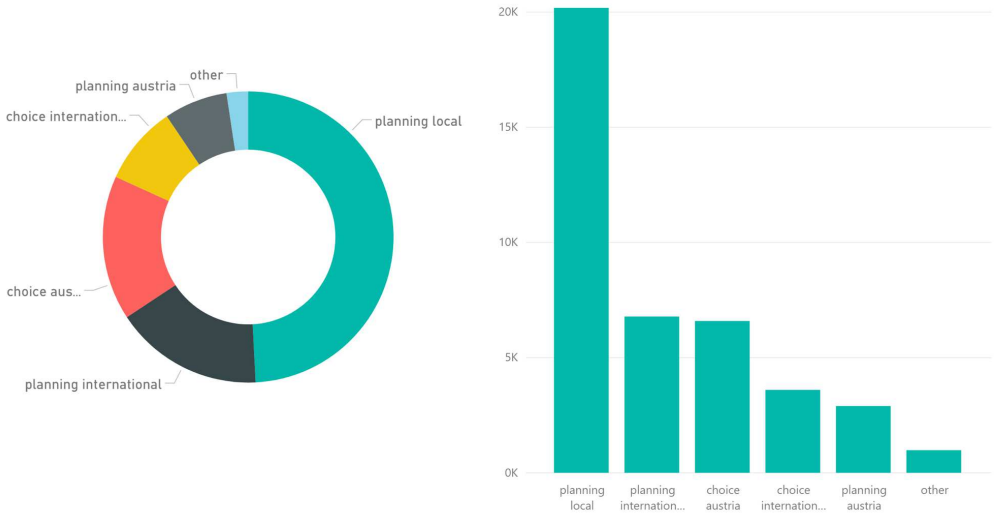


Figure 6.1: TripAdvisor - Distribution of the classified posts (6 classes)

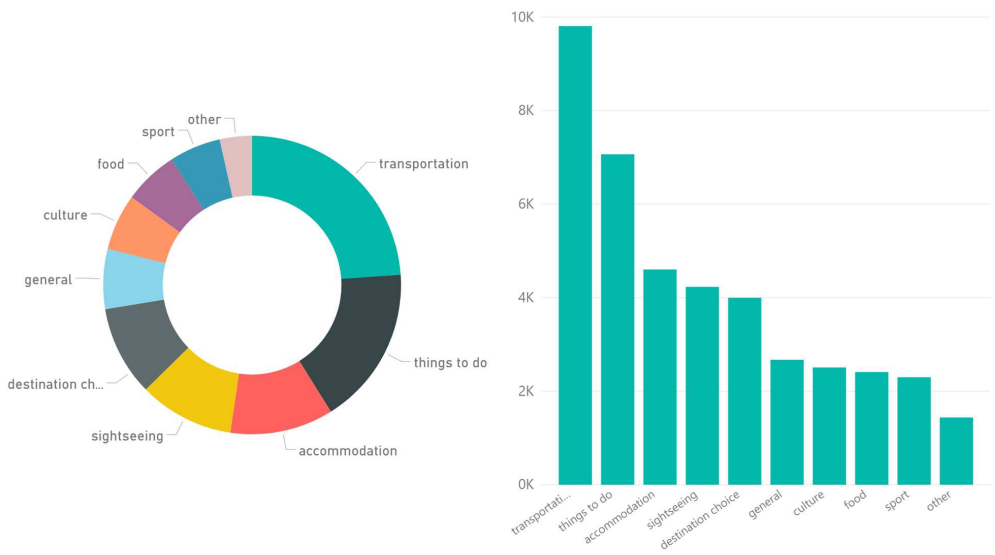


Figure 6.2: TripAdvisor - Distribution of the classified posts (10 classes)

Post	Class
"can anyone advise f the visit to zoo and hellburn palace be combined.,I wasted to visit the garden and tricks fountain of palace after visiting the zoo. can i park my car at zoo, visit it and then walk over to palace to see garden and fountain,I presume it is free to see garden and fountain unless we decide to go in a tour to see place/fountain. please advise"	planning local
"After visiting Schönbrunn, we thought it might be interesting to take a tram or bus into the city. Our hotel is on Columbusgasse, near Keplerplatz UBahn station.,I can easily figure out the UBahn route, but having trouble figuring out a bus/tram route!,Thanks in advance-"	planning local
"Hey guys, I'm trying to plan a route by car for the Wachau Valley.,We will be landing at , and taking our rental car from there, I got no idea where to start, what to see, where we can park etc.,I tried to search for trip plans but couldn't find any that traveled there by car..."	planning austria
"Hi I am traveling to austria with my husband and 6 month daughter on 12 sep. I need tips advise on places to go and also if it would be good to rent a car and see places by road n drive by road salzburg to innsbruck. All advise appreciated plz"	planning austria
"We are planning to go ,SalzburgMunich DusseldorfAms/Par and back to the ,. We plan to be 2 ntes i VieSalzMun and drive through scenic beautiful routes between ViennaSalzburg, can somebody help me to plan those beautiful scenic routes? Thank you"	planning international
"We are planning to go from Vienna to Prague & Budapest. Weighing all things up, is it better to do an organised tour of these 3 places, or individually catching trains and finding own accommodation. We don't want to waste valuable time as we can only allow about 3 days in each place. Your help appreciated. It will be summer when we go"	planning international
"Is Zell am see the best mountain region/scenic beauty spot in austria ...Which region is most popular amongst tourists ...like Bernese oberland and interlaken are hot spots in Switzerland ...similarly plz let me know abt austria .."	choice austria

Table 6.5: Examples of classified posts (6 classes)

Post	Class
"Hi, we are planning to visit , next year in the summer (2 couples and 1 child) for four days. We went here many years ago and enjoyed that there was a lot to do.,However, we would also like to visit another destination for a subsequent four days (preferably north of zell am see) and wanted to hear any suggestions? We would rather not be in big cities, but be in a region where there are lots of things to see and do to keep us all entertained. We had heard of Obertrum Am See, which is one idea but I am open to any other suggestions.,Many thanks for all your help"	choice austria
"Hi„I am a huge Christmas markets fan and have been to „ Baden-Baden, „ Aachen, „ Nurnberg, , over the years. I am researching possible destinations for next year. Somewhat constrained by direct routes from ,. I am considering twinning a trip to Munich with Innsbruck and possibly a day trip to somewhere like Regensburg. I would appreciate any thoughts - I have about 5 days to play with so a good mix of sightseeing and lively markets would be on the wish list. How would Innsbruck compare to destinations above? 1 or 2 nights there?,Thanks„K"	choice international
"I am driving from , to Lofer in late June. I am planning to stop in Cortina d'Ampezzo and Lake Misurina for lunch and sightseeing. I THINK I would like to get a little further on the first day than Lienz, but I'm not sure I want to drive all the way to Lofer in one day.,Does anyone agree that Venice to Lienz is easy one day drive - and does anyone have any suggestions about a nice place to stay between Lienz and Lofer?,Thank you!,Jeff"	choice international
"Hello, I'm currently in Vienna and my travel hot water heater just went bust. Any suggestions for a store/ travel store where I can buy a travel water heater/ travel kettle? ,Thanks!"	other

Table 6.6: Examples of classified posts (6 classes)

Post	Class
"hi..i want to drive from innsbruck to vienna , but can any one suggest me a new route avoiding salzburg on this way"	transportation
"We are in Vienna atm and have visited the Schönnbrunn Palace (which is a beauty tbh), but is it worth to visit the Beveldere Palace after visiting Schönn. Palace ? They look quit similar on pictures, so any anyput would be welcome..."	sightseeing
"I am coming to Vienna for a few days near the end of July. What is the best hostel near the center of town for a fifty year old single male traveller on a budget? Any ideas,anyone?,"	accomodation
"I will be visiting Vienna for the first time and I would like to attend a classical concert composed by a renowned full orchestra in a Vienna prestigious theatre. I will be visiting during the month of September 2017 and have a chance to visit two concerts on a Sunday 24th and a Monday 25th September. I do not like much operas but enjoy hearing an orchestra playing popular classical music. Which concert and theatre should I visit?"	culture
"Hi „Where can I eat a good quality scaloppini and schnitzel around vienna ? ,Thanks !"	food
"We are spending the month of December in Europe (our main destination is Switzerland) but our girls 16, 13, 8 want to see , and parts of ,. We plan to ski the week of 9th to the 16th in either st. Anton or grindelwald. Where should we spend Christmas and which two German cities offer the most at this time of year. I have done Europe many times but never in winter. Thanks."	destination choice
"Hi All„Me and my husband are planning for a 6 days 5 nights trip to Austria on the 2nd week of January 2016. Can anyone suggest a list of places that we can visit.We would like to include , and , in our trip.,Thanks in advance!"	things to do
"What are the chances of great skiing conditions over Thanksgiving? November 22- 28th. We are traveling to Innsbruck at that time and really want to ski the whole time. ,Thanks.."	sport
"What's the weather like in , in mid-September?"	general
"Hi. Help please. We purchased a b.free SIM card for E9.90 which should have 5 euro credit included. I have the receipt with the activation code but I can't make it work. I'm not sure which number on the receipt we need. Any advice appreciated."	other

Table 6.7: Examples of classified posts (10 classes)

6.2 Travel Planning Time

In this section, the posts created by TripAdvisor users in the Austria Q&A forums are analyzed in order to gain new insights about the time during which travelers start inquiring in TripAdvsior about their next destination.

The research focus lies on the following questions:

- *How much in advance do travelers begin researching for their trips? Can patterns be identified as in previous offline studies (as in figure 6.3) be detected based on online UGC?*
- *To what extent does the travel planning behavior differ between users from the United Kingdom and the United States of America?*

The posts of the Austria forum including all the sub forums were downloaded. The fields forum, content, post date, user link were used for the analysis. Since the focus of the study lies on users from the UK and the USA, all profiles of users that started a conversation in any of the Austria forums were downloaded. Each user profile contains information, such as user name, the date on which the user joined the forum, age, gender, hometown, number of reviews, ratings and posts, travel style, total points and badges. Out of these details the information extracted for the analysis is the hometown.

Considering the post shown in figure 5.2, we can see that the user comes from the USA and has created this post on 19 January 2014, which means the date on which the user started to actively ask in the forum about his/her trip is available. In addition, we can see that the user is planning his/her itinerary to different countries describing in details when and how he/she is going to organize the trip. Out of this text we can extract the time period, during which the trip is supposed to take place (late August). In order to automatically extract the date from the posts, Named Entity Recognition (NER) and advanced regular expression were used. NER was able to identify many types of date-related entities, for example, Monday, this week, this year, this summer, New Year's and names of the months. However, only entities that were recognized as a name of a month were considered for the analysis, for example, "January", "February", "March", etc. Posts that contained dates written in a number format such as "02.2018" or "01.04.2017" were not included. These posts were not considered given that many of the extracted numbers by NER were not related to the date. During the evaluation of the results, we recognized that some of the posts were not considered despite containing a name of a month. Therefore, advanced regular expression was used to search for date entities. Iteratively, the regex was improved until we got plausible results. Overall, the regex search was performing better than NER with an improvement of ∼9% for UK posts and ∼5% for US posts.

It was difficult to recognize the month "May" using regular expression, because it was not possible to differentiate between the month May and the word "may". Therefore, the

NER results for the month of "May" were used for the analysis, while for all the other months the extracted values from the regex search were considered.

Figures 6.4 and 6.5 are based on extracted information, including: The date of the post and the trip date from the post's content on TripAdvisor. It aims to show the time in advance during which visitors actively started to ask questions in TripAdvisor about their trip. These figures were created based on the following steps: Firstly, the duplicated posts were removed to avoid having the same post twice. This sometimes occurs a when users post the exact same content in different forms to receives various answers (e.g. in case of a round-trip). Secondly, UK posts and US posts were filtered. The number of posts without duplicates was about 40.000 of which round 20% were UK posts and 33% US posts. The next step included selecting the posts' in which users mentioned their trip date.

Figure 6.3: TripAdvisor barometer 2016 - Answer to the question: How far in advance did you begin researching for your last trip?

Each plot in figure 6.4 and 6.5 represents the time at which travelers actively write about their trip in the TripAdvisor forum, according to their trip date. The x-axis shows all the months of the year (post date) and the y-axis represents the number of posts posted in TripAdvisor by the users per month. To ensure representative results, the analyses are based on all posts in the Austria forum starting in year 2003. Different conclusions can be drawn: In figures 6.4 and 6.5 it becomes apparent that travelers tend to ask questions about their trip in the short term, which can be seen in the number of posts during the respective month itself. Another aspect is that travelers thoroughly research in advance. Overall, the research takes place within three months prior to the trip; during the fourth month, the number considerably decreases. This is confirmed by the study conducted by TripAdvisor in TripBarometer 2016 [Tri18c] (figure 6.3), which shows that more than half of the interviewees research their trip one to three months prior to traveling. Regarding the planning behavior for UK and US users, we can see that overall the search behavior

for travelers from the UK and the USA provide similar results. The difference that can be observed is the number of posts during the different seasons of the year. For example, for travelers from the UK, the number of posts in February is much higher than for US travelers and the number of posts related to summer months for travelers from the USA is higher than for UK users. In general, travelers tend to start earlier with planning for summer trip, mainly in June and July.

In the future, with the increase of users and the number of posts, the analyses can be applied on a yearly or monthly basis. Moreover, such analyses can be created for travelers from different countries.

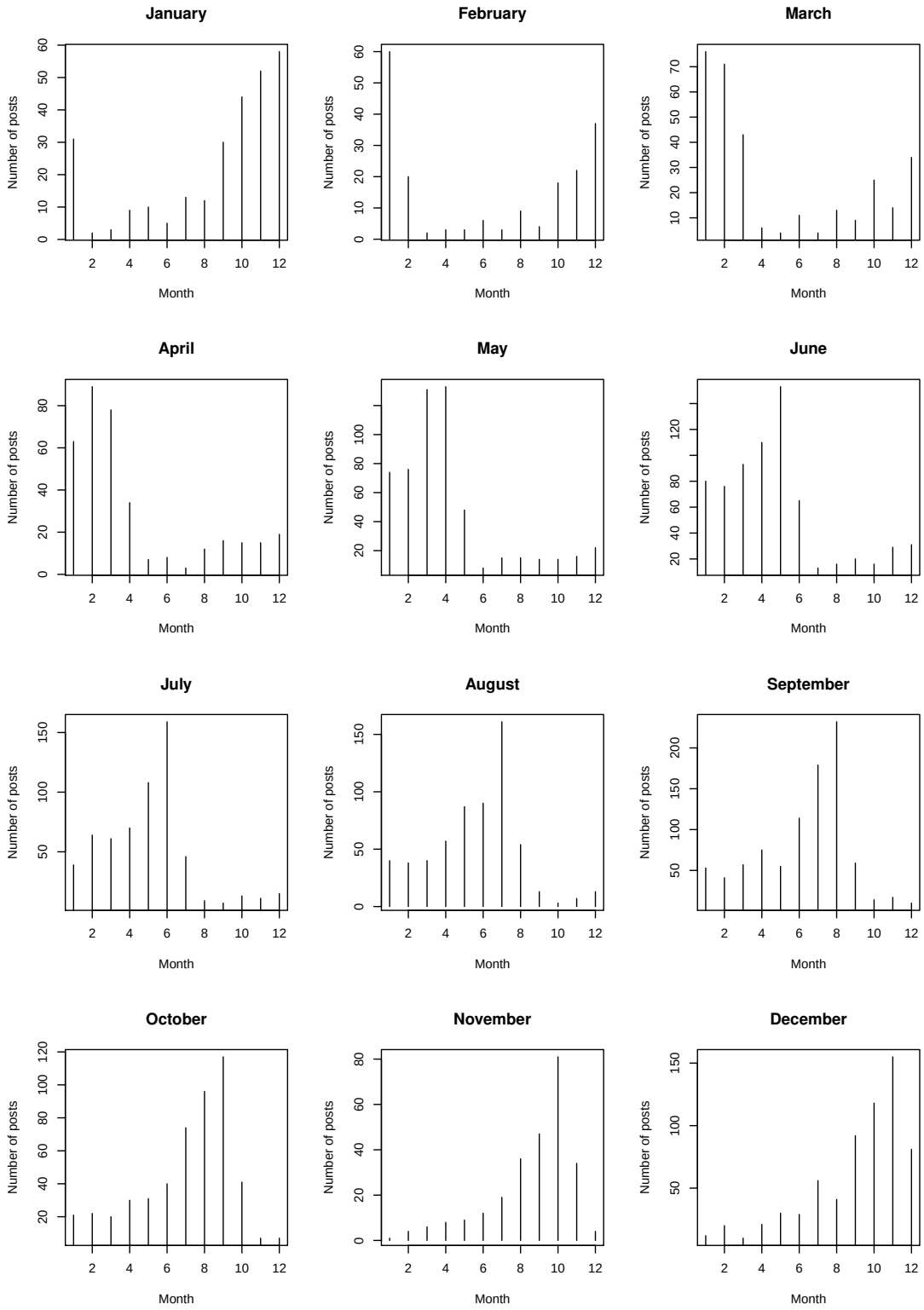


Figure 6.4: Time at which travelers from USA start inquiring about their next trip to Austria in TripAdvsior

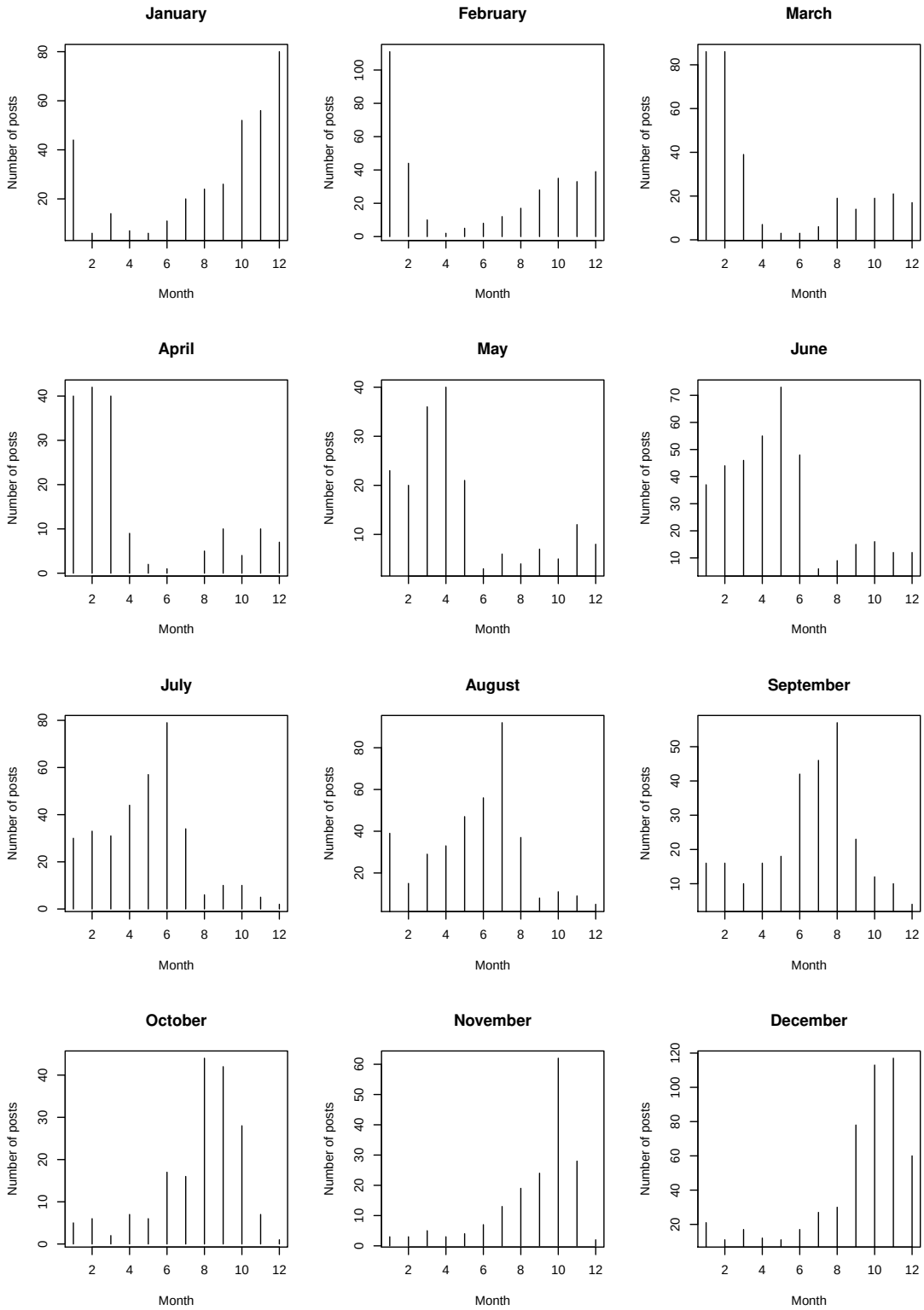


Figure 6.5: Time at which travelers from UK start inquiring about their next trip to Austria in TripAdvsior

6.3 Forums Analysis

The Travel Forum in TripAdvisor is divided into sub forums [2]. Users can browse by destination starting from a higher level to a lower level. For example, all European countries are subordinate to the Europe forums. The Austria forum has the following sub forums: Austrian Alps, Burgenland, Lower Austria, Styria, Upper Austria and Vienna Region. Each of these forums again contains other sub forums. Users can write directly in the sub forums or in the parent forums. For example, the Vienna Region forum has Vienna as a sub forum, which is the reason for having a bar for each forum in figure 6.6. In this figure, the number of posts written by users from all over the world in each forum is displayed. With about 16000 posts, Vienna has the highest number of posts, followed by the Austria and Salzburg forums. The number of posts starts to decrease in the Innsburck forum, the Austria Alps forum, the Zell am See forum, the Hallstatt forum, the Tyrol forum, etc.

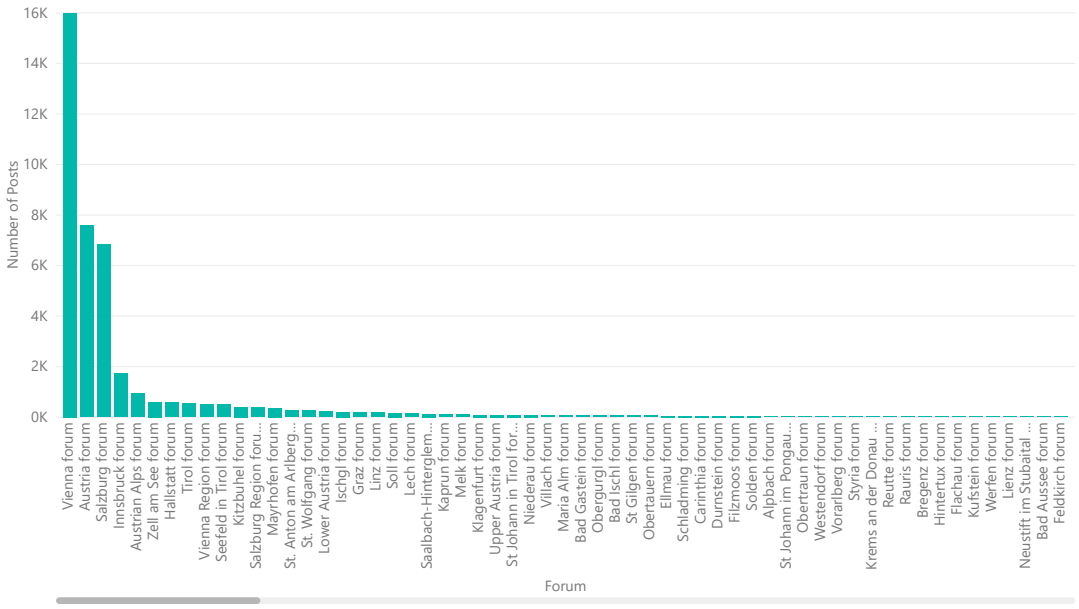


Figure 6.6: Number of posts per forum

Figure 6.7 provides more details about the topics discussed in Vienna and Tyrol forums. The x-axis represents the classes in which the posts were classified and the y-axis represents the number of posts in each class. The first diagram shows the topic distribution of the posts in the Vienna forum with a diversity of posts in different categories. As Vienna is the capital city of Austria and has a lot of attracting historical places, travelers tend to ask more about transportation, activities that can be done, accommodation, sights to

[2] https://www.tripadvisor.co.uk/ForumHome

visit, culture events to attend, etc. The number of posts written about transportation is more than 4000 posts. Culture posts are in higher demand than in the Tyrol forum. The second diagram illustrates the topic distribution in the Tyrol forum. As in the Vienna forum, the most frequently asked questions relate to transportation. Sports represent the second highest value in the Tyrol forum due to the reason that people tend to come to Tyrol for sport activities, such as skiing, snowboarding, etc., because of the alps and ski resorts. The lowest number of posts written in the Tyrol forum concerns posts about culture.

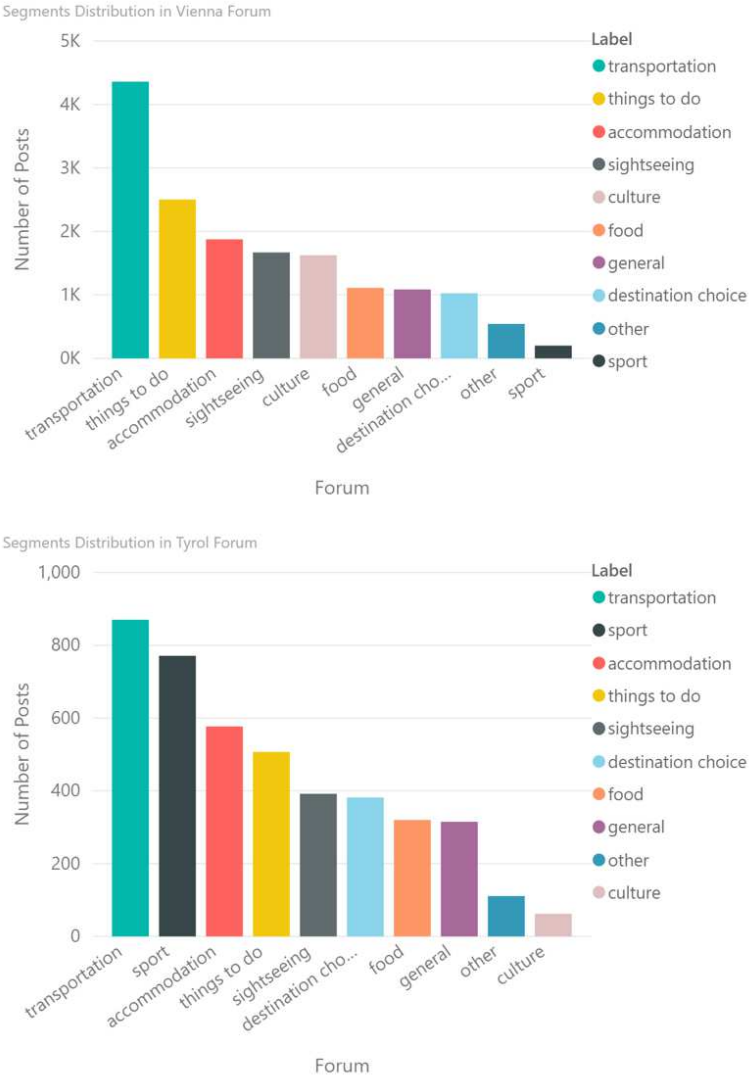


Figure 6.7: Topics discussed in the Vienna and the Tyrol forum

It is also interesting is to compare the topics written by users coming from the UK and

the USA to identify if they discuss the same topics. In this way, we can find out more about their interests. Figures 6.8 and 6.9 show the distribution of the topics in each forum for users from the UK and the USA. The x-axis represents the Austria forum and its sub forums and the y-axis represents the number of the posts as well as the category or class of the posts in different colors. The class *transportation and things to do* are the most topics represented in the most forums for group of users.

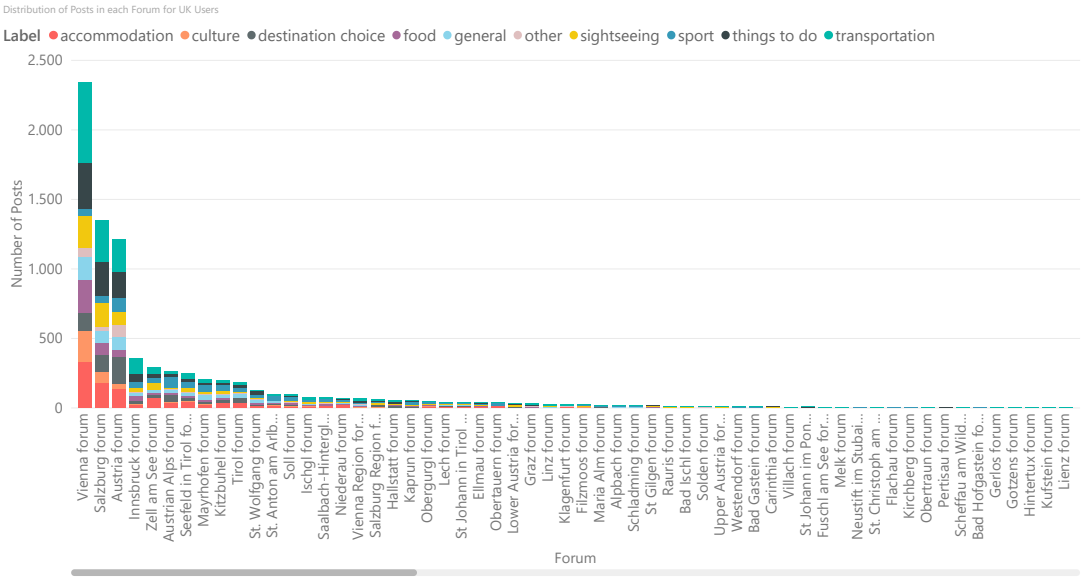


Figure 6.8: Topics discussed by UK users in Austria forums

In figure 6.8, we can clearly see that UK users write much more posts in different forums about *sport* compared to US users, especially in the Austria and Austrian Alps forum. The forums of Zell am See, Austrian Alps, Seefeld, Mayerhofen, Kitzbuhel and Tyrol are more visited by UK users than by USA users. For users from the USA, the Vienna, Salzburg and Austria forums seem to be the most frequently used forums as can be observed in figure 6.9. The posts about *destination choice* are more present in the Austria forum for both users from the UK and the USA. These analysis reflect the findings of the chapter 3, showing that the forums and sub forums of Vienna, Salzburg and Tyrol are the most active forums for travelers from the USA and UK.

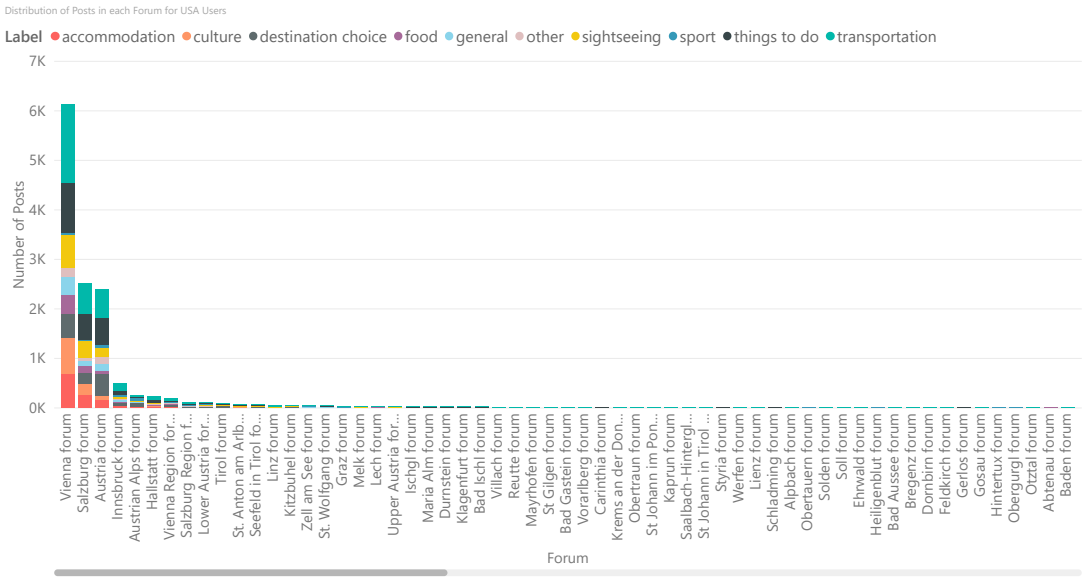


Figure 6.9: Topics discussed by US users in Austria forums

A further presentation of the posts of the forums based on the time during the year is illustrated in figure 6.10 for UK users and in figure 6.11 for US users. The x-axis shows the month of the year. The months of all years are aggregated. The y-axis shows the number of posts written in each month and the colors represent the class of the posts.

Figure 6.10 shows the discussed topics over time for UK users. At the beginning of the year, namely in January and February, a vast number of new conversations is started, specially about the topics *sport* , *accommodation*, *transportation* and *destination choice*. The month January features the highest number of posts with topics regarding *destination choice*. The topics *accommodation* and *transportation* are highly discussed over the whole year. Sports is mainly discussed in January, February, March, October, November and December. Starting from the month May and until August, the discussion about the topics *sightseeing* and *things to do* increases. The reason for the that is the summer season.

Figure 6.11 illustrates the discussed topics by users from the USA over all months. The topic *destination choice* has the heights number of posts in January, but is also discussed by the other months in the same average. Some of the topics such as, *sport*, *general*, *food*, *other* have a low of number of posts over the whole year, which reflect the travelers' disinterest regarding these topics. Posts categorized in *transportation*, *things to do* or *sightseeing* are mostly discussed by US users starting in March and go up until August.

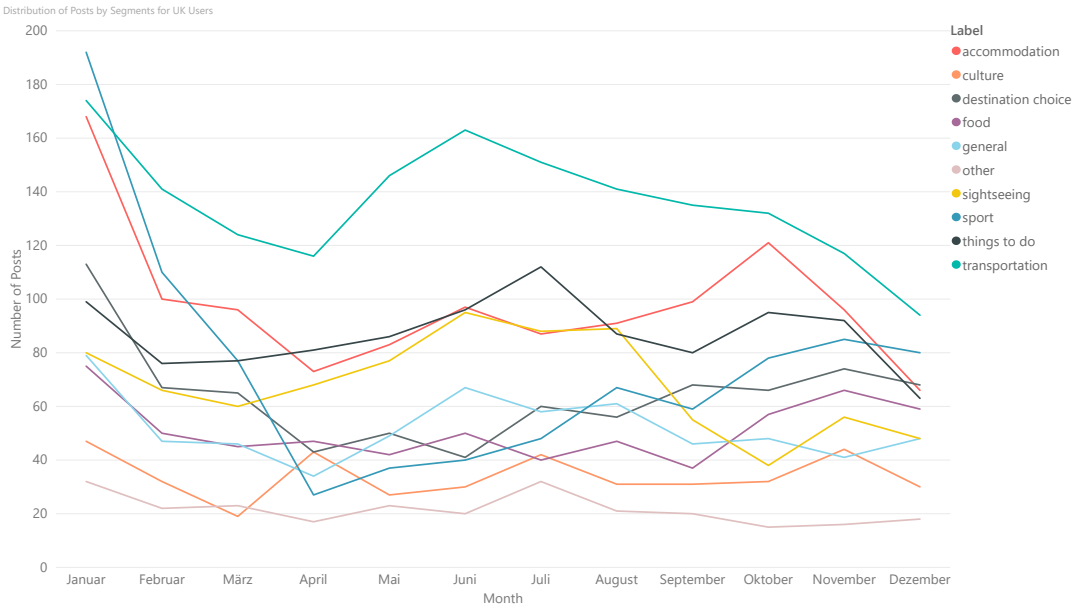


Figure 6.10: Discussed topics over time by UK users

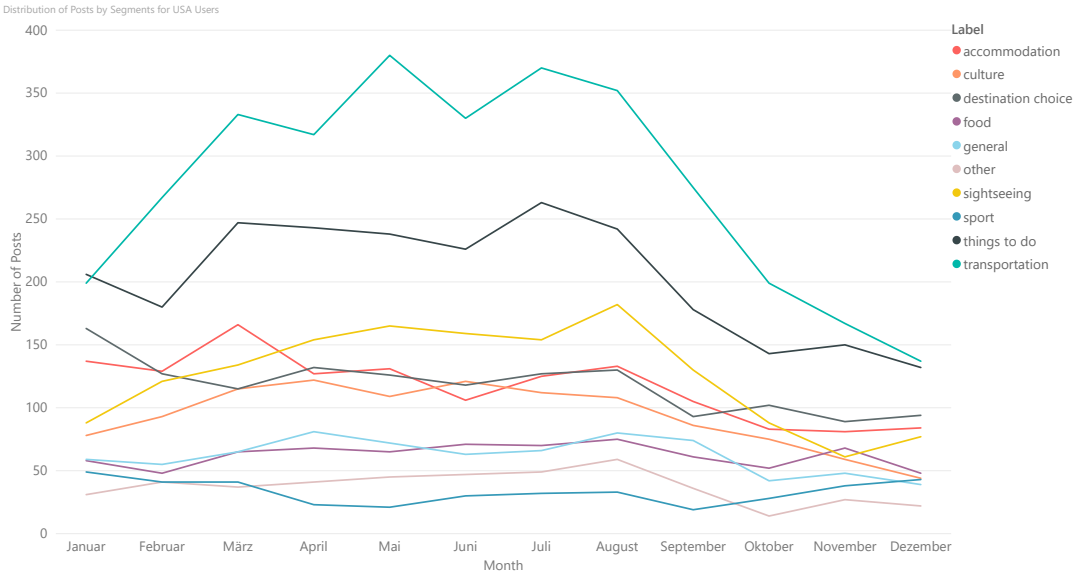


Figure 6.11: Discussed topics over time by US users

6.4 Users Analysis

6.4.1 Age, Gender and Country Distribution of the Users in the Austria Forum

To actively participate in TripAdvisor forums, users have to join TripAdvisor. They can join through their Facebook or Google accounts or by using any other email address. Each user can add as much information as he/she wishes. For the age, users can either leave it blank or select a range out of: 13-17; 18-24; 25-34; 35-49; 50-64 or 65+. For the gender, users can select between: female, male, other gender identity or they can choose to not indicate any gender. More than half of the users in the Austria forum tend to not provide their age and gender. In figure 6.12, we can see the age and gender distribution of the users. The number of users is calculated based on the distinct member link. The exact numbers of the users are represented for each group in the table.

For the users that provided information about their age and gender, we can see in figure 6.12 that the distribution of the age is generally well distributed among the different age ranges except by the range between 13-17. The biggest user group is represented by users between the age of 35 and 49. It is noteworthy that the number of young users between the age of 18 and 24 is not that high compared to the other groups. Furthermore, there is a big number of active users in the age group of 50+. Based on the gender distribution, we could say that more women are active in the forum, but this is not guaranteed, because there may be hidden information in the missing data.

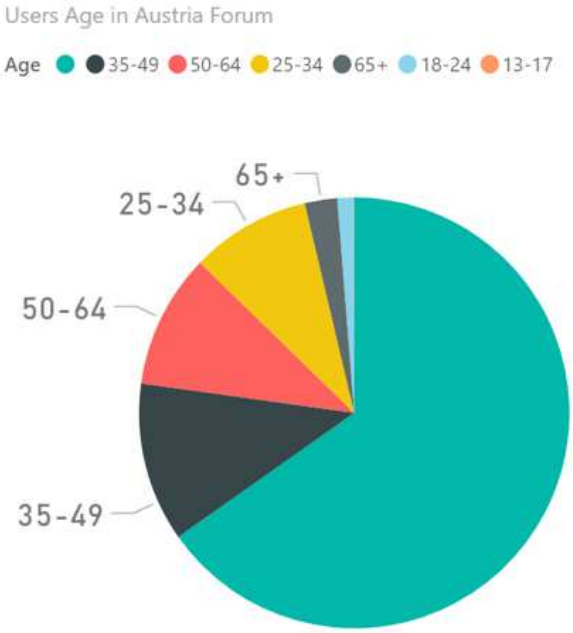
Similar to age and gender, a lot of users do not provide their home country. Some of the users provide invalid locations which get saved without any errors, because there is no validation done through TripAdvisor. Table 6.8 shows the number of users [3] coming from the USA, the UK and other countries.

Country	Number of Users
United States of America	7531
United Kingdom	5212
Other countries	10081

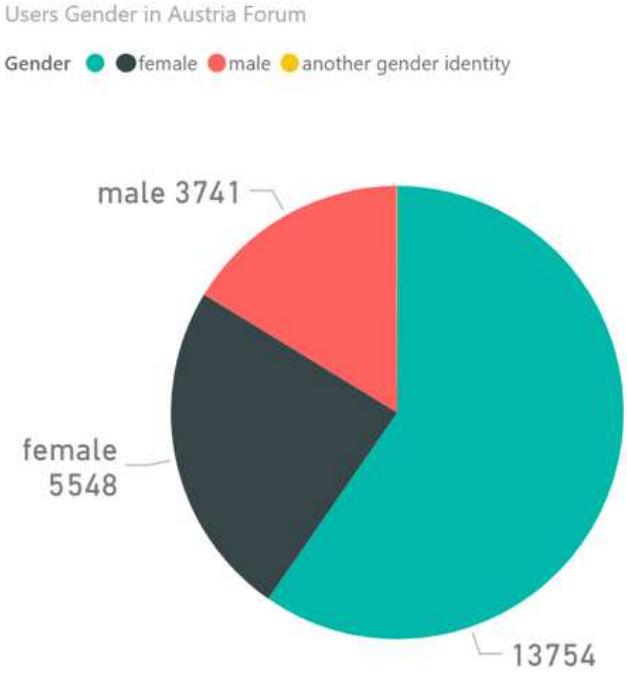
Table 6.8: Home country of TripAdvisor users that actively participated in the Austria forum

The user profiles contain other information besides age, gender and country, such as the number of posts, number of photos, total points, travel style, etc. This information can be used to further analyze the users in more detail. Based on this data, experts in the forum could be identified.

[3] Only users that started a topic conversion are considered. Users that replied to questions were not considered.



Age	Count of Users
	15053
35-49	2758
50-64	2330
25-34	2082
65+	565
18-24	282
13-17	7
Total	**22824**



Gender	Count of Users
	13754
another gender identity	7
female	5548
male	3741
Total	**22824**

Figure 6.12: TripAdvisor - Users' age and gender distribution

6.5 TripAdvisor Dashboards

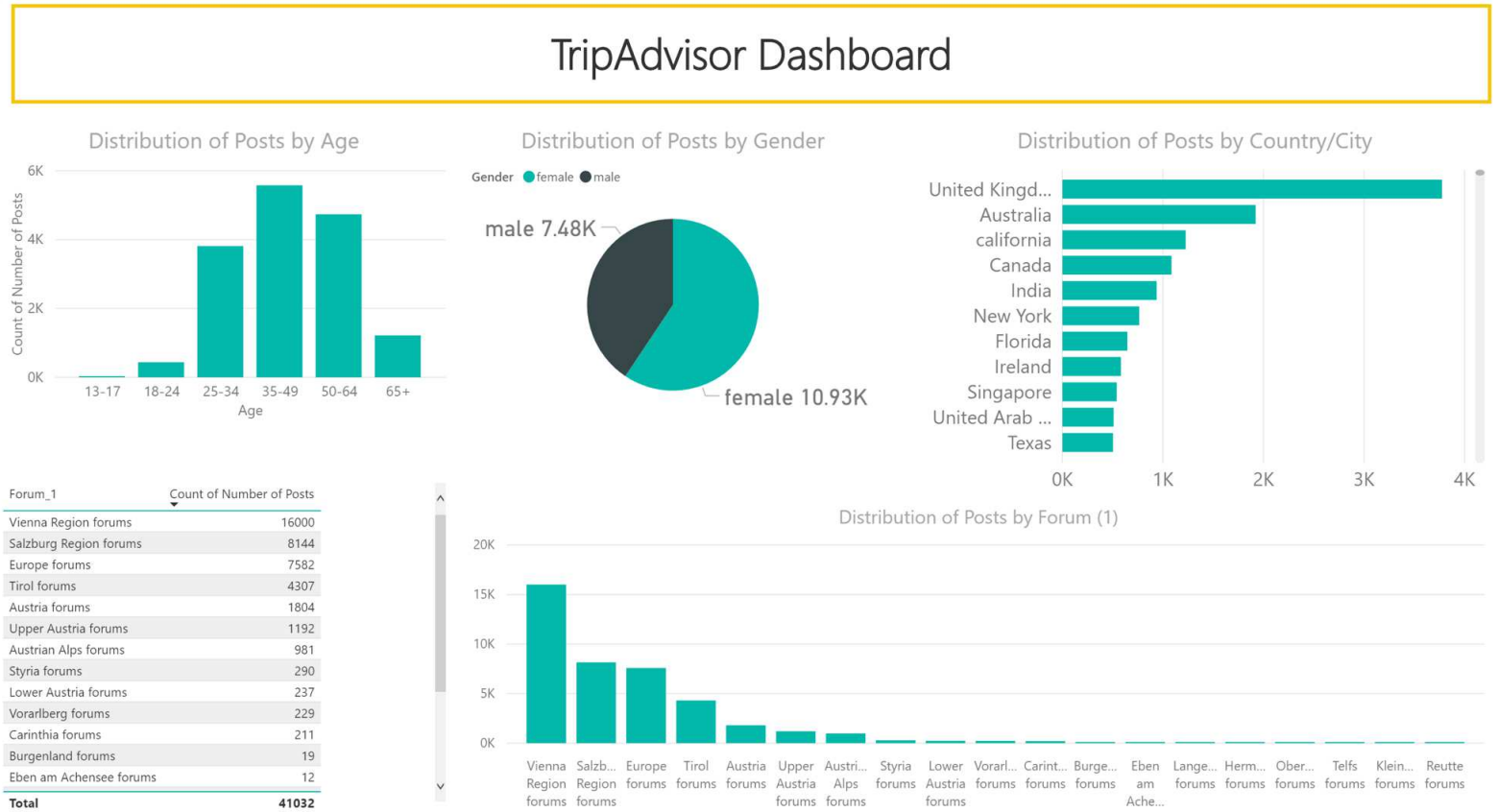
Many insights could be extracted from the data and represented in graphs. These graphs can be gathered to show a summary of all the information in one page. Figure 6.13 represents different aspects of the data in one dashboard. The dashboard shows graphics about the age, gender and country [4] of the users, as well as the number of posts in each of the sub forums of the Austria forum. This dashboard was created by the tool *Power BI* [Mic] and is interactive. This means, by clicking for example, on the *Vienna Region forum* each graphic will automatically adapt and show only the data for this forum. This makes it easy to analyze and understand the data. It is possible to see the details behind each graph by selecting the desired object.

Another dashboard is shown in figure 6.14. This dashboard displays a distribution of posts according to the age of the users and their country of origin. Furthermore, the content of the posts is listed in the table. Both post categorizations for content-based (10 classes) and detailed phase-based (6 classes) are shown in the dashboard. A word cloud was created based on the cleaned title of the posts.

6.6 Conclusion

In this chapter, we read between the lines of texts written by travelers from the UK and the USA on the TripAdvisor Austria forum. While analyzing the available user details, it became evident that more then 60% of the users tend to not provide details about their age and gender. Although the forum is mainly used to discuss questions about planning a trip, it is also used as a source of information to support travelers during their decision on a destination. For such questions, the main Austria forum is used more than the sub forums and a lot of the discussions happen at the beginning of the year. The time at which travelers start inquiring about their trip in TripAdvisor was analyzed, and the findings were based on a specific date format in the text. In order to take advantage of all the available information, we recommend including other date formats mentioned in the content on future works. Finally, the extracted findings were visualized in an interactive dashboard. A possibility to update these dashboards in real time, would be through TripAdvsior API, given if is available for the required data. These types of dashboards would reflect a live update on the state of an essential social media channel used by travelers.

[4] Data contains duplicates because some of the users wrote more than one post in the forums



Forum_1	Count of Number of Posts
Vienna Region forums	16000
Salzburg Region forums	8144
Europe forums	7582
Tirol forums	4307
Austria forums	1804
Upper Austria forums	1192
Austrian Alps forums	981
Styria forums	290
Lower Austria forums	237
Vorarlberg forums	229
Carinthia forums	211
Burgenland forums	19
Eben am Achensee forums	12
Total	**41032**

Figure 6.13: TripAdvisor dashboard - Distribution of posts per forum

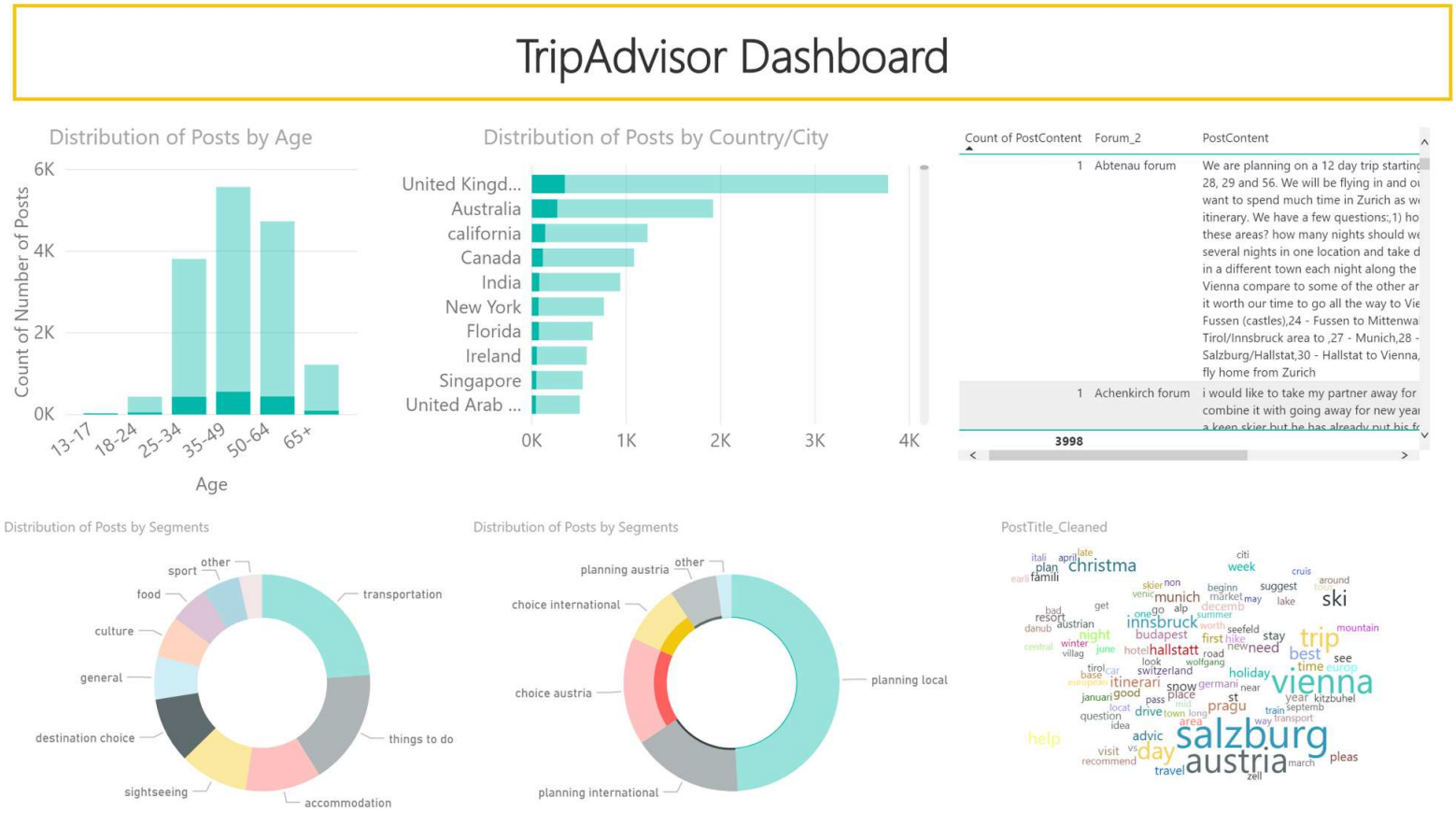


Figure 6.14: TripAdvisor dashboard - Distribution of posts per segment

CHAPTER 7

Time Series Analysis

The aim in this chapter is to analyze the relationship between the following three data sources: historical arrivals data, Google search volume index and extracted data from TripAdvisor and to find out to what extent these data can contribute to the prediction of arrivals to Austria.

7.1 Description of the Data

In this chapter the following data sources are used:

1. Historical Arrivals Data

   As mentioned in chapter 4, this data is about the number of travelers that visited Austria from the USA and the UK. The data is available for the 100 regions of Austria for the period between January 2014 and February 2017.

2. Google Search Volume Index

   This is a relative measurement to see how many searchers were looking for a specific term in Google Search. In our case, terms that relate to a specific province in Austria are considered (e.g., Stephansdom for Vienna, Schloss Hellbrunn for Salzburg, etc.). Google does not provide any absolute numbers for the search queries. The data consists of the search volume index for the period between January 2014 and June 2017 for each of the provinces of Austria. This data was provided by Google through our project partner.

3. TripAdvisor Data

   For each month in the period between January 2014 and February 2017, the number of posts in the Austria forums written by users that had provided information about their home country, was collected. Posts were filtered to consider only the posts that were written by users from the USA and the UK.

7.2 Descriptive Analysis

7.2.1 Correlation

In order to understand the relationship between the historical arrivals and the user generated content, the Pearson correlation coefficient was used. This coefficient represents the strength of the linear relationship between two data sources. It returns a value between -1 and 1. Zero means that there is no correlation, minus values mean negative correlation and positive values mean positive correlation.

Country	Arrivals	Google Search Volume Index	TripAdvisor Data
USA	Austria	0.07	0.34
	Vienna	0.23	0.42
	Tyrol	- 0.21	- 0.06
	Salzburg	0.28	0.19
UK	Austria	0.31	0.27
	Vienna	0.23	0.55
	Tyrol	0.58	0.67
	Salzburg	0.45	0

Table 7.1: Correlation matrix between the arrivals, Google and TripAdvisor Data

The correlation matrix for the arrivals historical data, Google search volume index and TripAdvisor data is calculated in table 7.1. The table shows that the correlation between the time series does not show a similar behavior. Some of the series show strong correlation, for example, for the UK arrivals to Tyrol. We can see a positive correlation of 0.58 with the Google search volume index and 0.67 correlation with the TripAdvisor data. Although a high correlation between input and target variable indicates that the most distributions of the target series can be well explained, it does not necessarily ensure predictive power [HEFL18]. The correlation of the USA arrivals with the other data sources is weak and in some cases negative, for example, for Tyrol the coefficient is -0.21 in Google and -0.06 in TripAdvisor.

The correlation between the data from Google and TripAdvisor was calculated and is shown in table 7.2. On the one hand, it is very weak and sometimes negative for the USA data. On the other hand, it is positive for the UK data, especially for Tyrol.

Country	Google Search Index Volume	TripAdvisor Data
USA	Austria	-0.07
	Vienna	0.04
	Tyrol	0.09
	Salzburg	0.14
UK	Austria	0.54
	Vienna	0.48
	Tyrol	0.80
	Salzburg	0.24

Table 7.2: Correlation matrix between the Google search volume index and the TripAdvisor data

7.3 Methods and Measures

7.3.1 Autoregressive Integrated Moving Average (ARIMA)

Time series have been widely used for the demand forecast in the tourism sector. One of the mainly used methods in the past is the integrated autoregressive moving average models (ARIMAs) developed by Box and Jenkins (1970) [SL08]. This method is used for the forecasting of univariate time series. It is represented by ARIMA *(p, d, q)*. The AR *(p)* stands for autoregression, which is about the past lags of the series considered by the model creation. The I *(d)* stands for integration and denotes the order needed to be taken in order to make a series stationary. A series is stationary when their mean and variance are constant and the auto-correlation between two values is depending only on the time lag and not on the point of time within the series [HEFL18]. The MA stands for moving average *(q)* and is about the order of the error component. A pre-condition for the usage of this method is that the series need to be stationary. This can be achieved by applying the difference or logs of the series. The models are estimated and the model with the lowest information criteria (Akaike information or Schwartz Bayesian criteria) is selected as the most fitting model [SS10].

7.3.2 Seasonal Autoregressive Integrated Moving Average (SARIMA)

The seasonal ARIMA model (SARIMA) is an extension to the ARIMA model which includes information about seasonality in the data. The SARIMA model is denoted with the parameter $(p,\ d,\ q)(P, D, Q)_s$, where the capital letters indicate the seasonal factor and small letters indicate the non-seasonal factor [SS10]. The s represents the number of time stamps for a single seasonal period.

7.3.3 (Seasonal) Autoregressive Integrated Moving Average with Exogenous Input ((S)ARIMAX)

This model is an extension of the ARIMA or SARIMA models to allow including explanatory variables to the model that may improve the accuracy of the forecast. In the paper of Yang et al. [YPS14] the web traffic of destination marketing organizations was used to predict hotel demand. The results showed that ARMAX model outperformed the ARMA models which did not include the exogenous variables. Similarly, Pan and Yang [PY17] included both search queries and web traffic data to forecast weekly hotel demand and they found that ARMAX model outperformed the ARMA models [VLSB18]. However, including search data does not automatically mean an improvement in the demand prediction. Schaer et al. [SKF19] found that established univariate forecasting benchmarks, such as exponential smoothing, consistently perform better than when online information, such as user-generated content is included.

For this thesis, the SARIMA and SARIMAX were used. The scripts for the models were written in the programming language *R* using the *auto.arima* function to fit the best model. The function conducts a search over possible models within the default order constraints. The order of first difference for d is chosen based on the Kwiatkowski-Phillips-Schmidt-Shin *KPSS* test. Since the data on arrivals are seasonal, the order of D is set to 1 by default. For the model selection the Akaike's Information Criteria (AIC) is used. The lower the AIC, the better.

7.3.4 Root Mean Square Error (RMSE)

Root mean square error (RMSE) is one of the most frequently used measures to evaluate the performance in time series forecasting in the tourism filed, as well as other fields [VLSB18]. It is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{7.1}$$

where the residuals are the difference between the actual values and the predicted values, represented as $(\hat{y}_i - y_i)$, and $\hat{y}_i$ is the predicted value for the observation i^{th} while y_i is the target value of the observation. The lower the value of RMSE, the better the prediction.

7.4 Model Building and Evaluation

In this section, the data used for building the models will be described and the models will be presented. The forecasting models have been learned and evaluated for UK and US arrivals to Austria and the following three provinces: Vienna, Tyrol and Salzburg. In figures 7.1 and 7.2 the distribution of the three data sources for visitors from the US and the UK to Austria is demonstrated. The x-axis stands for the month of the year starting from January 2014 (2014.0) up until February 2017. In the first graph the y-axis reflects the number of visitors to Austria (arrivals). The second graph mirrors the results detracted from the Google index, while the third displays the TripAdvisor index.

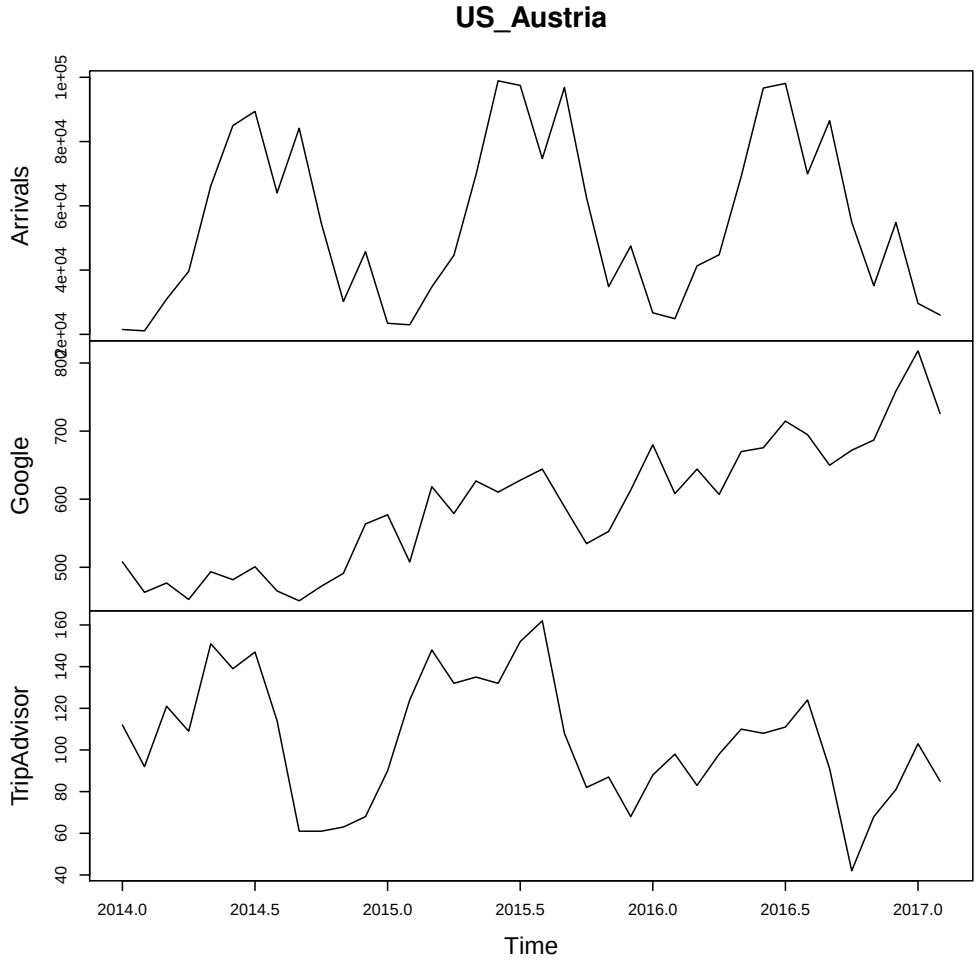


Figure 7.1: Distribution of US arrivals to Austria according to Google and TripAdvisor

The first graph of figure 7.1 demonstrates a clear pattern in the arrivals distribution over

the period of three years. Visitors from the USA tend to come more frequently during summer season. The second graph of the figure was created based on a summary of all Google indexes of all Austrian provinces. It shows a trend in the search behavior of US tourists about Austria using Google. The Google search results do not completely correspond to the arrivals data. This is also reflected in table 7.1 in which the correlation between the Arrivals and the Google Index lies at a very low value of only 0.07. The third graph of figure 7.1 is based on the number of posts created by US users in TripAdvisor in the whole Austria forum. It has a correlation value of 0.34 with the arrival data and a negative correlation with the Google index with a value of -0.07. Thus, it becomes evident that the TripAdvisor data displays higher correlation with the arrivals data in comparison to the data detracted from the Google index.

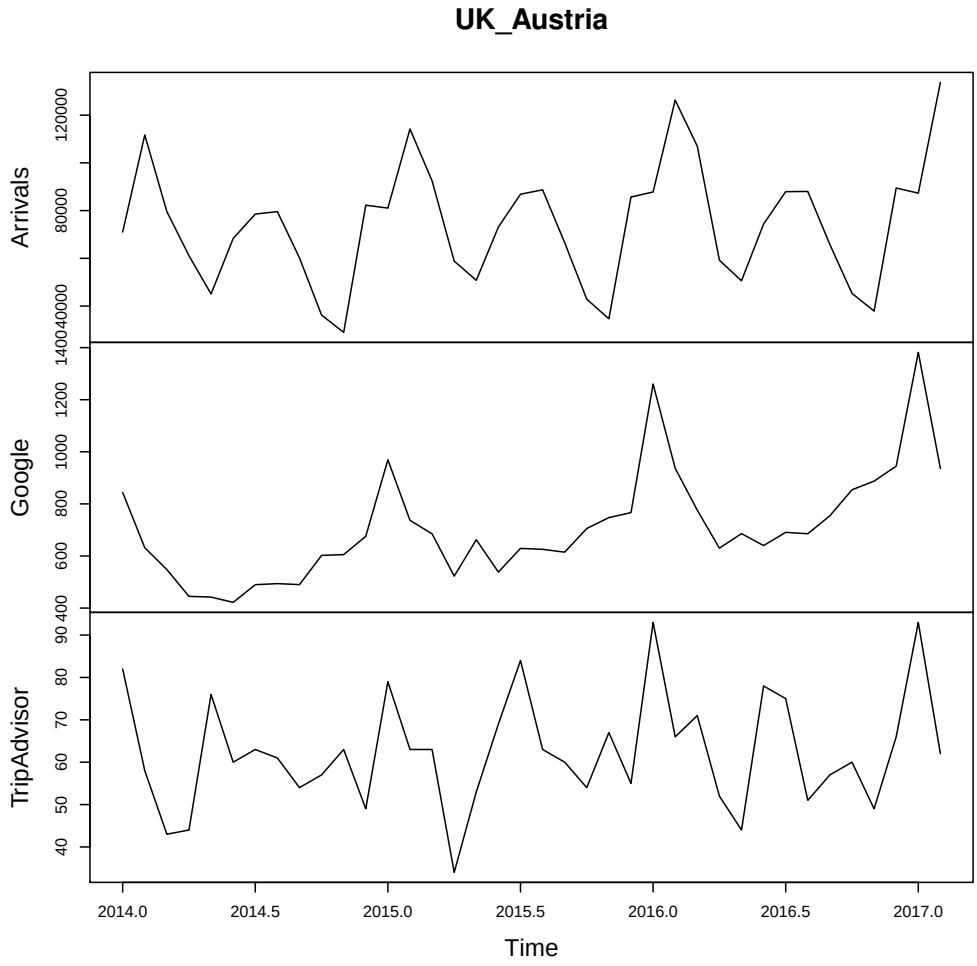


Figure 7.2: Distribution of UK arrivals to Austria according to Google and TripAdvisor

Similar to figure 7.1, figure 7.2 illustrates a comparison of arrivals, Google indexes and TripAdvisor data, however concerning visitors from the UK to Austria. The first graph displays the arrivals distribution showing that visitors from the UK tend to come to Austria in winter rather than in summer. The graph of the arrivals depicts a repetitive pattern over a period of three years. Likewise, in the Google graph a pattern can be established, according to which Google research about Austria through UK visitors takes place at the beginning of the year. This corresponds to UK arrivals, which start to rise in December and reach their high point in February. Considering this interconnection, one could interpret these results as an indication about UK visitors' research behavior prior to their actual trip to Austria. The correlation between the arrivals data and the Google index is higher compared to the previously described US data and lies at a value of 0.31. The TripAdvisor index correlates positively with the Google index with a value of 0.54.

For further insights into specific data distribution of UK and US arrivals to any of the three provinces of Tyrol, Vienna and Salzburg, figures A.1, A.2, A.3, A.4, A.5 and A.6 are attached in the appendix.

As it has already been discussed in previous chapters, travelers start inquiring information about their next holiday several months in advance. This means that the arrivals of a given month are related to the search done by travelers in prior months. Figures 7.3 and 7.4 show the distribution of the UK arrivals to Tyrol according to the Google search and TripAdvisor. The scale on the left side of the y-axis represents the arrivals and the scale on the right side of the y-axis represents the Google search index or the TripAdvisor data. In these figures, the Google Index and TripAdvisor are shifted by a lag of -1. As an example, the arrivals of February are compared with the the Google Search of January. The left graph in figure 7.3 shows the distribution of the arrivals data and Google search and the right graph illustrates the shifted distribution by a lag of -1.

When shifting the data by a specific lag (as shown in figures 7.3 and 7.4), the correlation between arrivals and research data increases (for some of the UK arrivals e.g. for Tyrol Google data).

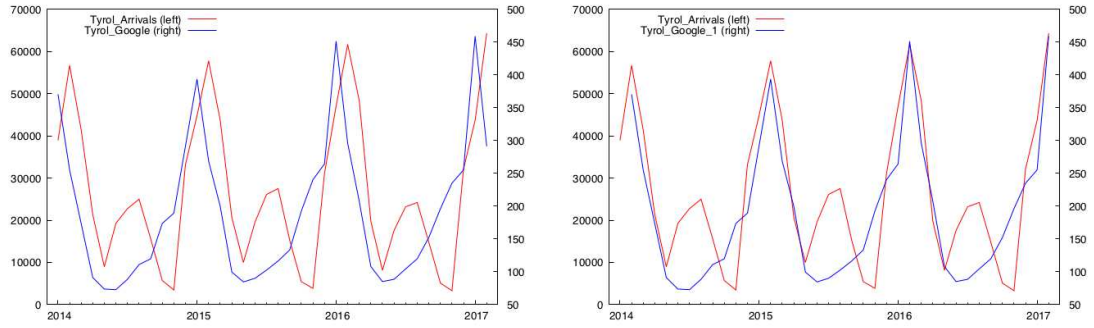


Figure 7.3: Distribution of UK arrivals to Tyrol according to Google search volume index shifted by a lag of -1

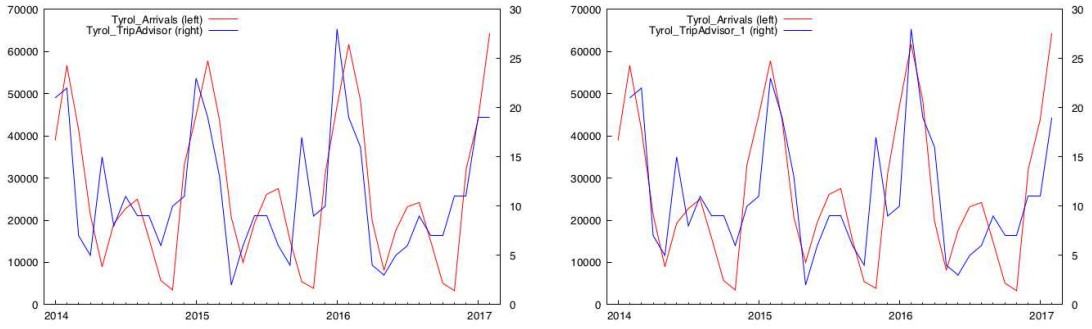


Figure 7.4: Distribution of UK arrivals to Tyrol according to TripAdvisor shifted by a lag of -1

The behavior pattern of the series demonstrates the correspondence between the high points of tourist arrivals and the ones of their research. Figure 7.4 mirrors correspondence even in the low points of both arrivals and Google research from the year 2015 to 2016. Though, the number of TripAdvisor posts in the Tyrol forum is very low compared to the arrivals data, it has a similar pattern. This shows that the UGC in social media can reflect the trend of the arrivals from different countries. We expect more precise results when having sufficient data.

The forecasting of the tourist arrivals was executed with four models solely based on the tourist arrivals data with the Google search and TripAdvisor data as additional inputs (*'exogenous'* variables). To prevent overfitting and to be able to evaluate model performance the data entries were split in training and test set. The model is learned based on the training data and the prediction is formed according to the test data. In this way, the prediction was evaluated with the real tourist arrivals and the prediction error was calculated. The training set consists of the series from January 2014 up until June 2016 and the test set consists of the data points from July 2016 to February 2017. The TripAdvisor data was filtered to select only the posts that were written by UK and US users. Breaking down the data on a monthly basis and over the years clarifies the low number of TripAdvisor posts in some of the forums.

The aim of this analysis is to determine how forecasting future arrivals may be affected by the usage of user generated content as an additional input. This is identified through comparing residuals using Root Mean Square Error (RMSE). The Ljung-Box test was applied to the residuals of the fitted models for each time series. It was used to ensure that the residuals are independent and do not show autocorrelation. The stationary of the data was ensured using the KPSS test.

The outcome of the prediction models is listed in tables 7.3 and 7.4. In these tables, the *country* column shows the home country of the visitors (USA or UK). The *arrivals* column represents the place visited by tourists. It includes Austria and three provinces: Tyrol, Salzburg and Vienna. The column *SARIMA(X) models* represents the four models

with selected values for the parameters *(p, d, q)*$(P, D, Q)_s$ based on automatic search using the *auto.arima* function. The s is 12, because the regular pattern in the arrivals data is repeated every 12 months. This can be detracted from figures 7.1 and 7.2 for the arrivals data. The column *RMSE (train)* shows the root mean square error of the training set of all of the four models. The last column *RMSE (test)* shows the root mean square error for the test set.

All the models are based on the ARIMA model including a seasonality component. The first model *arrivals only* solely is based on historical arrivals data. No exogenous variables are included. The second model with *arrivals + Google* is based on the historical data and the Google search index as a regressor. The third model *arrivals + TripAdvisor* is built according to the historical arrivals data and the TripAdvisor extracted data as a regressor. The fourth model *Arrivals + Google + TripAdvisor* has, beside the historical arrivals, the Google and TripAdvisor data as regressors. The arrivals data derives from historical values and therefore provides a more reliable pattern. The other two data sources depend on the search behavior and the generated content of travelers which is why they vary depending on the available data. An ARIMA model with parameters (1,1,0)(0,1,0)[12] includes a non-seasonal AR term (1), a trend difference order (1), a seasonal difference order (1) and an s of 12 for monthly data that have a yearly seasonal cycle. The results indicate that using the user generated content in forecasting the tourism demand raises the performance of visitors from the USA to Austria, the UK to Austria, the USA to Vienna, the UK to Tyrol and the UK to Salzburg. This is represented by means of the reduction of the RMSE in the test data when the models were enhanced with the Google search volume index, the TripAdvisor data or both. The RMSE for the models of UK visitors to Vienna, USA visitors to Tyrol and USA visitors to Salzburg demonstrate that the forecast without the *"exogenous"* variables delivered better results. According to this output in a number of cases the UGC can contribute to the prediction of tourist arrivals while in others it cannot. This highly depends on the available data.

7.5 Conclusion

In this chapter, the historical arrivals data, the Google search volume index and the data extracted from TripAdvisor were analyzed. The correlation between the three data sources did not show consistent behavior. A number of correlations indicated a positive relationship between the data while others indicated a negative relationship or no relationship at all. This is variable according to the availability of the data. For a number of provinces, the data extracted from TripAdvisor was not sufficient for coming to a conclusion on the arrivals' behavior. Likewise, the Google index showed an increasing trend reflecting the search interest of the travelers. However, the seasonal pattern was not presented. This inevitably had an influence on the prediction models based on the three data sources. The results of the prediction demonstrated that including additional exogenous variables does improve the results in some of the models, nevertheless it does not a imply an overall improvement in the forecasting performance.

Country	Arrivals	SARIMA(X) Models	RMSE(train)	RMSE(test)
USA	Austria	Arrivals only ARIMA(1,1,0)(0,1,0)[12]	2769.88	5464.65
		Arrivals + Google ARIMA(0,0,0)(0,1,0)[12]	2733.25	8449.81
		Arrivals + TripAdvisor ARIMA(1,0,0)(0,1,0)[12]	2735.40	6672.90
		Arrivals + Google + TripAdvisor ARIMA(1,1,0)(0,1,0)[12]	2711.43	**5326.71**
UK	Austria	Arrivals only ARIMA(0,0,0)(0,1,0)[12]	3518.42	4799.39
		Arrivals + Google ARIMA(0,0,0)(0,1,0)[12]	3320.41	4886.882
		Arrivals + TripAdvisor ARIMA(0,0,0)(0,1,0)[12]	3393.49	**4122.45**
		Arrivals + Google + TripAdvisor ARIMA(0,0,0)(0,1,0)[12]	3202.817	4633.405
USA	Vienna	Arrivals only ARIMA(0,0,0)(0,1,0)[12]	1561.29	4397.39
		Arrivals + Google ARIMA(0,0,0)(0,1,0)[12]	1334.55	**3117.61**
		Arrivals + TripAdvisor ARIMA(0,1,1)(0,1,0)[12]	1561.08	3486.46
		Arrivals + Google + TripAdvisor ARIMA(1,1,0)(0,1,0)[12]	1570.10	3309.194
UK	Vienna	Arrivals only ARIMA(0,1,0)(0,1,0)[12]	970.87	**1129.20**
		Arrivals + Google ARIMA(0,1,0)(0,1,0)[12]	929.11	1301.51
		Arrivals + TripAdvisor ARIMA(0,0,1)(0,1,0)[12]	820.29	1530.86
		Arrivals + Google + TripAdvisor ARIMA(1,0,0)(0,1,0)[12]	826.80	1540.22

Table 7.3: Comparison of performance for SARIM(X) models in training and test sets (1)

Country	Arrivals	SARIMA(X) Models	RMSE(train)	RMSE(test)
USA	Tyrol	Arrivals only ARIMA(1,0,0)(0,1,0)[12]	586.03	**1315.06**
		Arrivals + Google ARIMA(0,0,0)(0,1,0)[12]	641.39	1613.75
		Arrivals + TripAdvisor ARIMA(1,0,0)(0,1,0)[12]	569.70	1327.08
		Arrivals + Google + TripAdvisor ARIMA(0,0,0)(0,1,0)[12]	630.46	1695.33
UK	Tyrol	Arrivals only ARIMA(0,0,0)(0,1,0)[12]	1768.29	2828.31
		Arrivals + Google ARIMA(0,0,0)(0,1,0)[12]	1609.48	2498.53
		Arrivals + TripAdvisor ARIMA(0,0,0)(0,1,0)[12]	1721.94	2673.64
		Arrivals + Google + TripAdvisor ARIMA(0,0,0)(0,1,0)[12]	1606.56	**2445.77**
USA	Salzburg	Arrivals only ARIMA(1,0,0)(0,1,0)[12]	914.49	**2032.29**
		Arrivals + Google ARIMA(2,0,0)(0,1,0)[12]	841.20	2752.23
		Arrivals + TripAdvisor ARIMA(1,0,0)(0,1,0)[12]	870.70	2129.15
		Arrivals + Google + TripAdvisor ARIMA(1,0,0)(0,1,0)[12]	869.43	2210.24
UK	Salzburg	Arrivals only ARIMA(0,1,0)(0,1,0)[12]	1595.71	2028.64
		Arrivals + Google ARIMA(1,0,0)(0,1,0)[12]	1435.35	2037.63
		Arrivals + TripAdvisor ARIMA(1,0,0)(0,1,0)[12]	1595.27	2047.36
		Arrivals + Google + TripAdvisor ARIMA(1,0,0)(0,1,0)[12]	1431.63	**1998.82**

Table 7.4: Comparison of performance for SARIM(X) models in training and test sets (2)

CHAPTER 8

Conclusion and Future Work

8.1 Summary

The aim of this thesis was to analyze the content generated in the tourism related social media to help destination marketers and travel professionals find out more about travelers and their behavior regarding the inspiration and planning phase, and to investigate the contribution of this content to the prediction of tourist arrivals. Hence, a literature review was conducted to study the state of the art in the travel planning process, information search, social media, destination marketing and visitors' arrival prediction. Furthermore, a research was conducted to choose the classifiers and performance measures used for the automatic classification of the text.

The data used in this work can be divided into two categories. The first category refers to historical arrivals tourist data for travelers coming from the Unites States of America and the United Kingdom to Austria in the period between 2014 and 2017. The second category consists of the user generated content that was created by travelers. This includes the Google search volume index (search queries about Austria) and the posts in the Austria forum of the travel review site TripAdvisor. Diverse social media and travel review channels were analyzed and the best appropriate data source was TripAdvisor. The Google data was provided by our project partner and the TripAdvisor data was downloaded from the website.

In order to understand the historical tourism data and the potential relationship it can have with the user generated content, descriptive data analysis was conducted and summarized in chapter 4. This analysis showed the arrivals distribution over the years, as well as the most visited provinces by tourists from the USA and the UK. This chapter includes a cluster analysis of the arrivals to the 100 regions of Austria. Four clusters were identified, while Vienna had its own cluster, all clusters were created based on the number of arrivals and the arrivals pattern according to the time in which the regions

were visited (summer, winter, spring, etc.). Subsequently, by that the travel review sites, with focus on TripAdvisor, and the prediction models were explained in chapter 5, 6 and 7,

The posts and the user profiles that started a topic in the forums were used for the analysis. In the beginning, only the title of the post was considered by the analysis, but in some cases it did not provide enough information for classification. Therefore, the whole content of the posts had to be considered. In order to prepare the data to be automatically classified by machine learning algorithms, the labeling and pre-processing of the data were necessary. A set of data was analyzed to find the fitting categorization that can cover different aspects. Three classifications were defined by the expert, including "content-based" with 10 classes, "phase-based" with 3 classes and "detailed phased-based" with 6 classes. Different pre-processing techniques were used, such as text normalization, numbers removal, stop words removal, words stemming, tokenization and term weighting. The data was imbalanced and so the undersampling technique was used to have a balanced data set.

After obtaining a labeled data set, supervised learning was applied to classify the rest of the data (unlabeled data). The algorithms used for the classification are: Support Vector Machines (SVMs), Multinational Naive Bayes and Logistic Regression. The feature extraction was based on count vectorization and TFIDF. The logistic regression algorithm was the best performing algorithm. Cross validation (5-k) was used for the evaluation and the performance scores were not only based on the precision and recall, but also on the confusion matrix. The output of the classification provided an overview of the topics discussed in the Austria forum and its sub forums. The forums are mainly used for asking specific questions for the trip planning, for example, about transport, things to do, accommodation, sightseeing, etc. Furthermore, the forums are also used during the early making decision phase for choosing a holiday destination. Another finding was related to the time at which travelers start actively asking about their trip in the forum. To extract this information, only posts that contained a date were considered. The analysis showed research generally starts, three months prior to the trip, for summer trips, however, earlier.

The sub forums were analyzed for users from the USA and the UK. The Vienna forum was the most frequently used forum with more than 16.000 posts, followed by the Salzburg forum, the Austria forum and the Innsbruck forum. This reflects the same distribution as in the historical arrivals data. The topics distribution was different in each forum, for example, in the Tyrol forum the second most popular topics was "sport". The profiles of the users that started a topic in the forum were investigated. More than 50% of the users did not provide information about their age and gender. Users between the age of 35 and 49 represents the biggest user group and the number of women is higher than the number of men. The number of users from the USA is 7531, 5212 users are from the UK and 10081 are from other countries. The visualization of the graphics was created using the tool *Power BI*. This tool makes it possible to create interactive dashboards. These dashboards, in turn, gather multiple graphics in one place and make it possible to

explore the data.

Time series analysis was used to examine the relationship between the user generated content (Google and TripAdvisor) and the historical arrivals from the USA and the UK. Furthermore, the aim was to determine, how the UGC can contribute to the forecasting of the future arrivals. The TripAdvisor index was created based on the number of posts in each month in the period between January 2014 and February 2017 for Austria, Vienna, Tyrol and Salzburg. The correlation between the arrivals, Google and TripAdvisor was not consistent. In some cases, it was positive and in others, negative or even 0. The seasonal ARIMA method with "exogenous" variables was mainly implemented to compare the different models with and without the UGC as an additional input. The data was divided into a training and a test set. The training set was used to build the model and the test set was used to compare the predicted values. The RMSE was used for the evaluation. Four models were created and the result were compared. The results showed that the prediction performance was increased when adding the UGC as an additional input to the model for some of the time series, such as Austria (USA, UK), Vienna (USA), Tyrol (UK) and Salzburg (UK). For the other time series, only the arrivals showed better performance. This means that adding UGC does not automatically induce better forecasting.

8.2 Improvements and Future Work

The labeled data used by the supervised classification was imbalanced, which limited the performance of the algorithms. This is an area of improvement for the future. More data can be labeled to cover all the classes and to ensure that the data is well-balanced. It is essential to define the classes in a way that avoids an overlap with each other, such as with the classes "planning local" and "planning austria". Since the classification is based on the bag of words approach for finding the words that explain a specific class, the semantic meaning and any information between the lines should be avoided during the labeling of the data.

The number of posts written in the TripAdvisor forum increases on a daily basis. Therefore, it would be interesting to automate the process of the classification, so that analyses and visualizations can be created in real-time. Additionally, other social media sources can be used to add more information and gain more insights.

The time series analysis can be extended to consider the lags of the user generated content to determine if the lagged data can improve the prediction performance. Other models can be used for comparison.

The approach used in this thesis for the content analysis of the user generated content and arrivals prediction can be applied to the forums in TripAdvisor, as well as to other data sources.

APPENDIX A

Distribution of the Data Sources to the Provinces

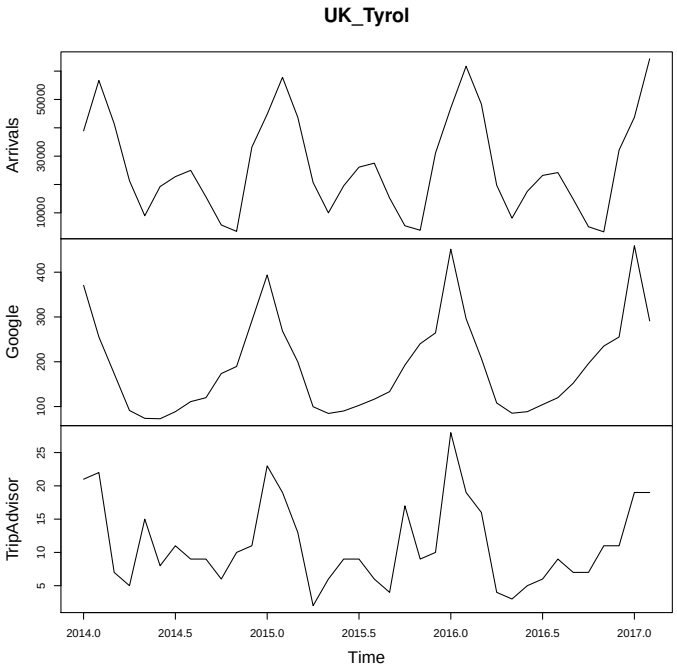


Figure A.1: Distribution of UK Arrivals to Tyrol according to Google and TripAdvisor

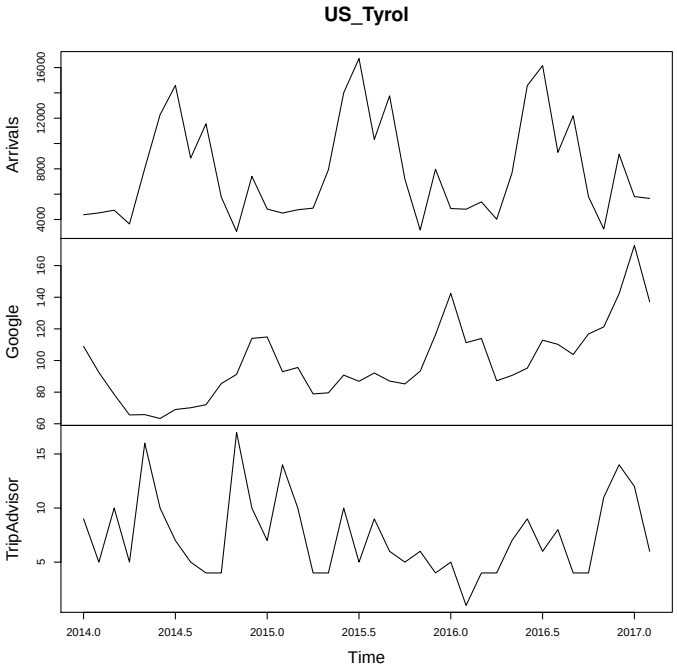


Figure A.2: Distribution of USA Arrivals to Tyrol according to Google and TripAdvisor

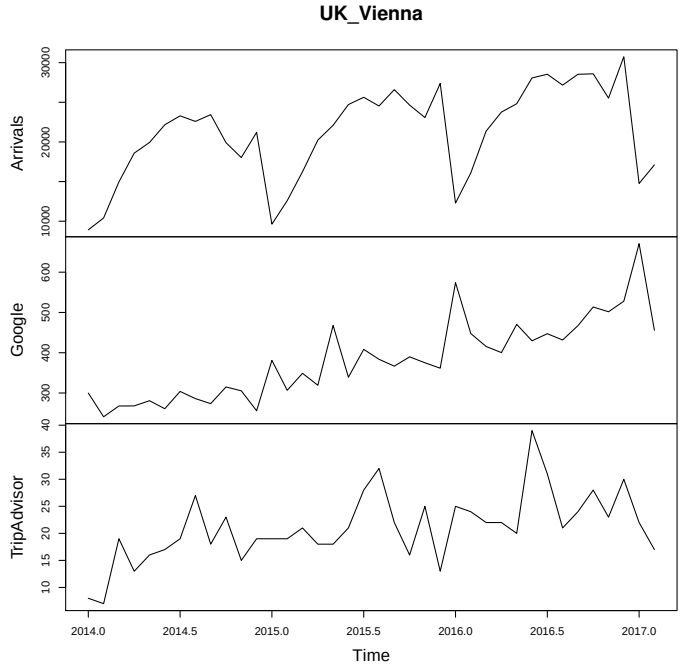


Figure A.3: Distribution of UK Arrivals to Vienna according to Google and TripAdvisor

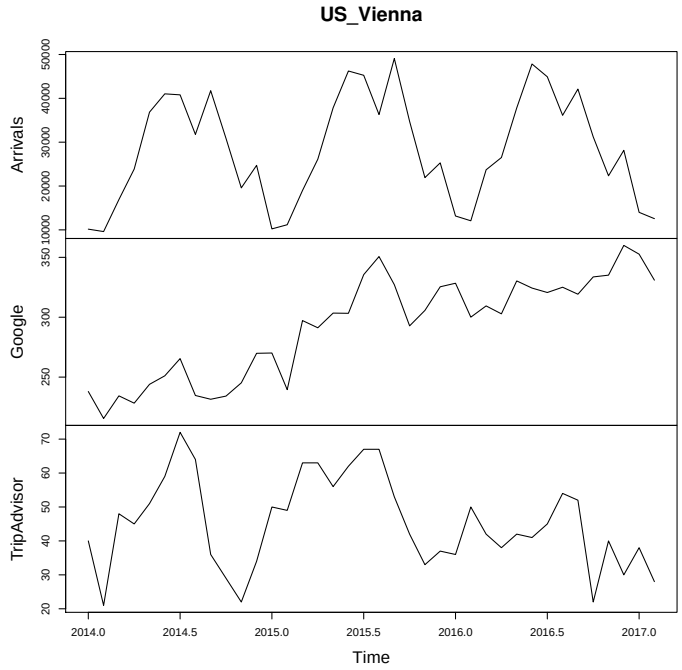


Figure A.4: Distribution of USA Arrivals to Vienna according to Google and TripAdvisor

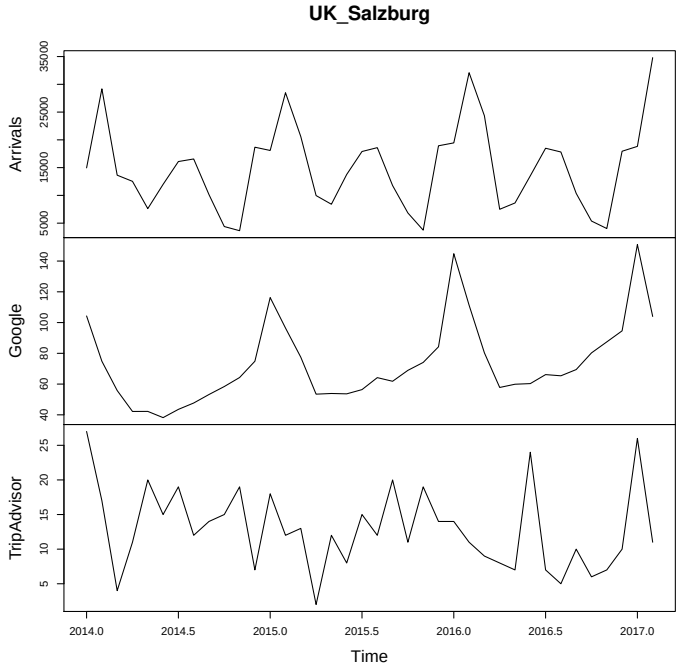


Figure A.5: Distribution of UK Arrivals to Salzburg according to Google and TripAdvisor

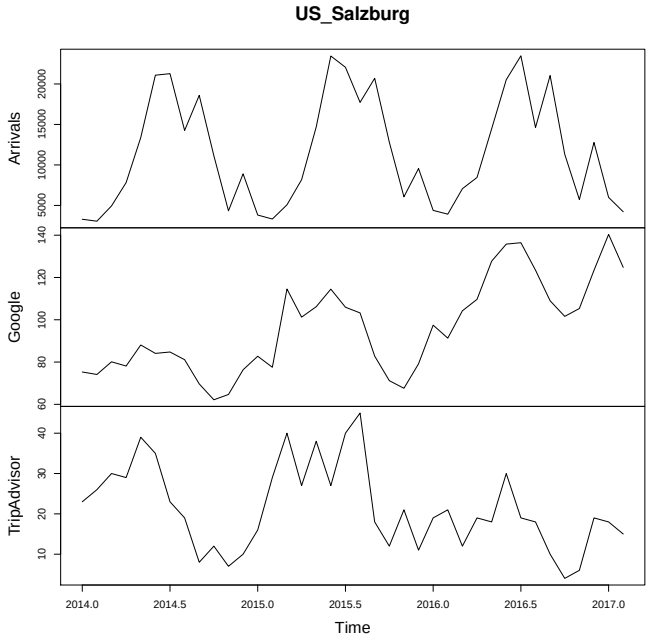


Figure A.6: Distribution of USA Arrivals to Salzburg according to Google and TripAdvisor

List of Figures

List of Tables

Bibliography

[Ake09] Gary Akehurst. User generated content: the use of blogs for tourism organisations and tourism consumers. *Service business*, 3(1):51, 2009.

[ALA+12] Julian K Ayeh, Daniel Leung, Norman Au, Rob Law, et al. *Perceptions and strategies of hospitality and tourism practitioners on social media: An exploratory study*. na, 2012.

[Ana08] Ispas Ana. The tourism destination marketing–a mandatory course for the students of tourism. *ANALELE UNIVERSITĂŢII DIN ORADEA*, page 920, 2008.

[APA+17] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.

[Aus18] Statistics Austria. The information manager. `http://cf.cdn.unwto.org/sites/all/files/pdf/unwto_barom16_03_may_excerpt_.pdf`, 2018. [Online; accessed 03.09.2018].

[BG12] Maria Banyai and Troy D Glover. Evaluating research methods on travel blogs. *Journal of Travel Research*, 51(3):267–277, 2012.

[BK14] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[BSS15] Prosper F Bangwayo-Skeete and Ryan W Skeete. Can google data improve the forecasting performance of tourist arrivals? mixed-data sampling approach. *Tourism Management*, 46:454–464, 2015.

[Buh00] Dimitrios Buhalis. Marketing the competitive destination of the future. *Tourism management*, 21(1):97–116, 2000.

[BWS03] Atreya Basu, Christine Walters, and M Shepherd. Support vector machines for text categorization. pages 7–pp, 2003.

[Cha09] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009.

[CRV] Frano Caleta, Ivan Rezic, and Damjan Vucina. The analysis of preprocessing methods for the purpose of detecting sms spam. *Text Analysis and Retrieval 2018 Course Project Reports*, page 37.

[CV12] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

[Dat18] Dataiku. Data science platform. `https://www.dataiku.com`, 2018. [Online; accessed 15.03.2018].

[DPCB15] Andrea Dal Pozzolo, Olivier Caelen, and Gianluca Bontempi. When is undersampling effective in unbalanced classification tasks? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 200–215. Springer, 2015.

[EA12] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[EBM90] James F Engel, Roger D Blackwell, and Paul W Miniard. Consumer behavior (6th eds.). 1990.

[EBM95] James F Engel, Roger D Blackwell, and Paul W Miniard. Consumer behavior, 8th. *New York: Dryder*, 1995.

[Eko10] Cathy Nanyongo Ekonde. *Tourism destination marketing: a comparative study, between Gotland Island, Sweden and Limbe city, Cameroon.* 2010.

[FB06] Eibe Frank and Remco R Bouckaert. Naive bayes for text classification with unbalanced classes. pages 503–510, 2006.

[FBR12] John Fotis, Dimitrios Buhalis, and Nicos Rossides. *Social media use and impact during the holiday travel planning process.* Springer-Verlag, 2012.

[GL+09] Vishal Gupta, Gurpreet S Lehal, et al. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1):60–76, 2009.

[Goo18] Google. Google trends. `https://trends.google.com/trends/explore?cat=208&geo=US&q=%2Fm%2F0fhp9,%2Fm%2F0b1mf`, 2018. [Online; accessed 03.09.2018].

[GRM15] Wilfried Grossmann and Stefanie Rinderle-Ma. *Fundamentals of business intelligence.* Springer, 2015.

[GY08] Ulrike Gretzel and Kyung Hyan Yoo. Use and impact of online travel reviews. *Information and communication technologies in tourism 2008*, pages 35–46, 2008.

[HEF+17] Wolfram Höpken, Dominic Ernesti, Matthias Fuchs, Kai Kronenberg, and Maria Lexhagen. Big data as input for predicting tourist arrivals. In *Information and communication technologies in tourism 2017*, pages 187–199. Springer, 2017.

[HEFL18] Wolfram Höpken, Tobias Eberle, Matthias Fuchs, and Maria Lexhagen. Search engine traffic as input for predicting tourist arrivals. In *Information and Communication Technologies in Tourism 2018*, pages 381–393. Springer, 2018.

[HLC12] Chaang-Iuan Ho, Meng-Hui Lin, and Hui-Mei Chen. Web users' behavioural patterns of tourism information search: From online to offline. *Tourism Management*, 33(6):1468–1482, 2012.

[Hyd08] Kenneth F Hyde. Information processing and touring planning theory. *Annals of Tourism Research*, 35(3):712–731, 2008.

[iA18] ABA-Invest in Austria. Economic significance of tourism in austria. `https://investinaustria.at/en/sectors/tourism/`, 2018. [Online; accessed 03.09.2018].

[IKT05] M Ikonomakis, Sotiris Kotsiantis, and V Tampakas. Text classification using machine learning techniques. *WSEAS transactions on computers*, 4(8):966–974, 2005.

[Jen00] Jiann-Min Jeng. *Exploring the travel planning hierarchy: An interactive web experiment*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[KFPH04] Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial naive bayes for text categorization revisited. pages 488–499, 2004.

[KG14] Subbu Kannan and Vairaprakash Gurusamy. Preprocessing techniques for text mining. In *Conference Paper. India*, 2014.

[KH10] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[KHRM06] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, 18(11):1457–1466, 2006.

[KJ13] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

[KZP07] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[Lab18] Unwired Labs. Locationiq api: Free and fast geocoding and reverse geocoding service. `https://locationiq.com`, 2018. [Online; accessed 10.02.2018].

[Lan00] Tania C Lang. The effect of the internet on travel consumer purchasing behaviour and implications for travel agencies. *Journal of vacation marketing*, 6(4):368–385, 2000.

[Lea18a] Scikit Learn. Cross validation. `https://scikit-learn.org/stable/modules/cross_validation.html`, 2018. [Online; accessed 06.01.2018].

[Lea18b] Scikit Learn. Logistic regression. `https://web.stanford.edu/~jurafsky/slp3/5.pdf`, 2018. [Online; accessed 26.12.2018].

[Lea18c] Scikit Learn. Support vector machines. `https://scikit-learn.org/stable/modules/svm.html`, 2018. [Online; accessed 26.12.2018].

[LPLH17] Xin Li, Bing Pan, Rob Law, and Xiankai Huang. Forecasting tourism demand with composite search index. *Tourism management*, 59:57–66, 2017.

[LSM08] Choong-Ki Lee, Hak-Jun Song, and James W Mjelde. The forecasting of international expo tourism using quantitative and qualitative techniques. *Tourism Management*, 29(6):1084–1098, 2008.

[M+14] Ipsos MediaCT et al. The 2014 traveler's road to decision. *Google Travel Study*, 2014.

[MBC08] Joana Miguéns, Rodolfo Baggio, and Carlos Costa. Social media and tourism destinations: Tripadvisor case study. *Advances in tourism research*, 26(28):1–6, 2008.

[Mic] Microsoft. Power bi - a business analytics tool. `https://powerbi.microsoft.com`, note = [Online; accessed 15.05.2018], Year = 2018.

[MRMFFR19] Estela Marine-Roig, Eva Martin-Fuentes, and Berta Ferrer-Rosell. A framework for destination image analytics. In *Information and Communication Technologies in Tourism 2019*, pages 158–171. Springer, 2019.

[NMC19] Kenshi Nakaima, Elena Marchiori, and Lorenzo Cantoni. Identification of competing destination brand: The case of okinawa island. In *Information and Communication Technologies in Tourism 2019*, pages 172–183. Springer, 2019.

[NW18] Julia Neidhardt and Hannes Werthner. It and tourism: still a hot topic, but do not forget it. *Information Technology & Tourism*, 20(1-4):1–7, 2018.

[OCo08] Peter OConnor. User-generated content and travel: A case study on tripadvisor.com. *Information and communication technologies in tourism 2008*, pages 47–58, 2008.

[PF06] Bing Pan and Daniel R Fesenmaier. Online information search: vacation planning process. *Annals of Tourism Research*, 33(3):809–832, 2006.

[PRBA15] Francisco C Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems*, 19(3):273–288, 2015.

[PY17] Bing Pan and Yang Yang. Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research*, 56(7):957–970, 2017.

[Rhe12] Carroll Rheem. Empowering inspiration: The future of travel search. *USA: PhoCusWright.[Viitattu 30.1. 2016] http://www. amadeus. com/at/documents/aco/at/de/Empowering_Inspiration_The_future_of_Travel_Search. pdf*, 2012.

[RM03] Toni M Rath and Raghavan Manmatha. Word image matching using dynamic time warping. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2003.

[RM14] D Ramyachitra and P Manikandan. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5(4), 2014.

[Rou87] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[S+17] Alina Stankevich et al. Explaining the consumer decision-making process: Critical literature review. *Journal of International Business Research and Marketing*, 2(6):7–14, 2017.

[SA10] V Srividhya and R Anitha. Evaluating preprocessing techniques in text categorization. *International journal of computer science and application*, 47(11):49–51, 2010.

[SE17] Alexis Sardá-Espinosa. Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, 12, 2017.

[Seb02] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[Ser18] Mete Sertkan. Master Thesis: Classifiying and Mapping e-Tourism data sets. 2018.

[SKF19] Oliver Schaer, Nikolaos Kourentzes, and Robert Fildes. Demand forecasting with user-generated online information. *International Journal of Forecasting*, 35(1):197–212, 2019.

[SL08] Haiyan Song and Gang Li. Tourism demand modelling and forecasting—a review of recent research. *Tourism management*, 29(2):203–220, 2008.

[Smoc17] Scrapinghub and many other contributors. Scrapy framework. `https://scrapy.org/`, 2017. [Online; accessed 03.07.2017].

[SS10] Andrea Saayman and Melville Saayman. Forecasting tourist arrivals in south africa. *Professional Accountant*, 10(1):281–293, 2010.

[Sta18] Statista. Most famous social network sites worldwide as of october 2018, ranked by number of active users (in millions). `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`, note = [Online; accessed 22.11.2018], 2018.

[sth] sthda. Statistic tools for high-throughput data analysis. `http://www.sthda.com/english/articles/29-cluster-validation-essentials/96-determining-the-optimal-number-of-clusters-3-must-know-metho` note = [Online; accessed 28.08.2018], year = 2018.

[T+17] Anna Terttunen et al. The influence of on consumers' travel planning and destination choice. 2017.

[Tri18a] TripAdvisor. About tripadvisor. `https://tripadvisor.mediaroom.com/uk-about-us`, 2018. [Online; accessed 03.09.2018].

[Tri18b] TripAdvisor. Tripadvisor media center. `https://tripadvisor.mediaroom.com/`, note = [Online; accessed 01.05.2019], 2018.

[Tri18c] TripAdvisor. Tripadvisor tripbarometer 2016. https://www.tripadvisor.com/TripAdvisorInsights/wp-content/uploads/2018/01/TripBarometer-2016-Traveler-Trends-Motivations-Global-Findings.pdf, note = [Online; accessed 22.09.2018, 2018.

[UK] TripAdvisor UK.

[UNW16] UNWTO. Unwto world tourism barometer. http://cf.cdn.unwto.org/sites/all/files/pdf/unwto_barom16_03_may_excerpt_.pdf, 2016. [Online; accessed 03.09.2018].

[VIN15] S Vijayarani, Ms J Ilamathi, and Ms Nithya. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.

[VLSB18] Katerina Volchek, Anyu Liu, Haiyan Song, and Dimitrios Buhalis. Forecasting tourist arrivals at attractions: Search engine empowered methodologies. *Tourism Economics*, page 1354816618811558, 2018.

[Wan15] Ying-Chuan Wang. A study on the influence of electronic word of mouth and the image of gastronomy tourism on the intentions of tourists visiting macau. *Turizam: međunarodni znanstveno-stručni časopis*, 63(1):67–80, 2015.

[WF06] Youcheng Wang and Daniel R Fesenmaier. Identifying the success factors of web-based marketing strategy: An investigation of convention and visitors bureaus in the united states. *Journal of Travel Research*, 44(3):239–249, 2006.

[WK+99] Hannes Werthner, Stefan Klein, et al. *Information technology and tourism: a challenging ralationship*. Springer-Verlag Wien, 1999.

[WP11] Youcheng Wang and Abraham Pizam. *Destination marketing and management: Theories and applications*. Cabi, 2011.

[XG10] Zheng Xiang and Ulrike Gretzel. Role of social media in online travel information search. *Tourism management*, 31(2):179–188, 2010.

[YPS14] Yang Yang, Bing Pan, and Haiyan Song. Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research*, 53(4):433–447, 2014.

[Zah18] Fazrian Zahrawani. Clean icon". https://www.iconfinder.com/Utha_design, 2018. [Online; accessed 04.01.2018].