ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

A text analytical approach: Predicting and understanding customer satisfaction by making use of customer reviews.

Student: Sofyan El Baouchi

Student Number: 430727

Supervisor: A. Alfons

The World Wide Web has made communication easier than ever. Before this happening
People used to ask acquaintances about their opinion before a purchase has been made. Nowadays one can find someone else's opinion on the internet with a few clicks. The growth of reviews on the internet has made electronic word-of-mouth an important aspect in helping customers with their buying decision. The goal of this research is to classify customer reviews based on their sentiments by making use of data mining techniques. The main focus is to find out how well data mining techniques perform and which aspects are most important for classification. The data that will be used is based on 20 hotels of which 10 of them belong to the expensive ones in London, while the other ones belong to the less expensive hotels in London. The reviews of the expensive hotels have a higher positive sentiment on average than the reviews of the less expensive hotels. The models that have been built all had a high prediction accuracy. The model that is based on Random forest performed the best when it comes to the reviews of the expensive hotels and the inexpensive hotels. What became clear is that the customers of the expensive hotels talk more about the aspects of the hotel, while customers of the less expensive hotels talk more about the non-hotel aspects.

Keywords: *customer reviews, customer opinion, text mining, opinion mining, sentiment analysis, review importance.*

Content

1. Introduction ... 3
1.1. Importance of customer reviews ... 3
1.2. Research Questions ... 5
1.3. Relevance. ... 6
1.4. Contribution ... 6
1.5. Structure of the thesis ... 6
2. Literature review ... 7
2.1. Customer satisfaction ... 7
2.2. Sentiment analysis origin ... 8
2.3. Electronic word-of-mouth ... 9
2.4. Opinion types ... 10
2.5. Predictive models origin ... 10
3. Data ... 11
3.1. Data description ... 11
3.2. Preprocessing & visualization ... 14
4. Methodology ... 25
4.1. Introduction ... 25
4.2. Sentiment analysis ... 25
4.2.1. Sentiment analysis definition & process ... 25
4.2.2. Different levels of sentiment analysis ... 27
Document level ... 27
Sentence level ... 27
Aspect level ... 28
4.3. Logistic regression ... 28
4.4. Variable selection ... 30
4.5. Decision trees ... 31
4.6. Random Forest ... 34
5. Results ... 37
5.1. Sentiment score ... 37
5.2. Predictive models (expensive hotels) ... 42
5.3. Predictive models (inexpensive hotels) ... 45
6. Conclusion ... 47
6.1. Limitations ... 49
6.2. Future research ... 51
Reference ... 53

1. Introduction

1.1. Importance of customer reviews

Since the advent of the World Wide Web, the internet has become very important for customers, but also for organizations. Many years ago, before reviews were available on the internet, people asked their family and friends for their opinion. Nowadays, the opinions of others can easily be found on the internet (Liu, 2012). With an ever-growing popularity of e-commerce on the internet, the total number of customer reviews and expert opinions keeps growing. The growth of reviews has made electronic word-of-mouth (eWOM) an important aspect in helping customers with their buying decision (Tang, 2017). Customers write about the feeling they get when they buy a product, but also the value they give to a product (Kulmala et al., 2013). With this information, a potential customer can have an indirect opinion about a product without the direct knowledge they otherwise would have when they would have bought the product themselves. Customers don't depend anymore on the information of a product that has been provided by the producer of the product (Tang, 2017).

Companies ask their customers to give a review about what they have bought in order to manage customer opinions. Customer's behavior and choices are based on the perception and beliefs of others. When customers want to make a decision, they seek out the opinions of other customers. This is due to the fact that customers find it difficult to assess the quality of a product or a service (Park & Nicolau, 2015). Therefore, having reviews available as a company will help other potential customers to make a decision when they want to buy any kind of product (Hu & Liu, 2004). This is not only important for customers, but also for companies because it helps them to better understand the buying behavior of their customers. By having a better knowledge of what a customer likes or dislikes, a company can give good customer service which is one of the key factors to maintain and improve the loyalty of customers (Reibstein, 2002).

The importance of reviews for customers has been discussed in several studies. The following empirical studies give an indication of how important these reviews are for both the customers and organizations:

- 92% of online customers make use of customer reviews before they make a buying decision (Ludwig, et al., 2013).

- Internet users who are looking for more information about a product in-store often skip the retail employee and search for information on their phone, according to a survey which has been conducted by eMarketer (2019). About sixty-nine percent of the respondents indicated that they would look for reviews on their smart phone instead of asking information to an employee. This shows how offline resources have been replaced by word-of-mouth e-marketing.

- The research of Park & Nicolau (2015) showed that 86% of online travelers find consumer reviews websites helpful when making decisions about which hotel they want to book.

- The research of Pang & Lee (2008) even indicate that purchase decisions of more than three-fourth of the consumers are influenced by product reviews.

- According to the research of Park and Han (2007), customers buying intentions increases with the quality and quantity of online reviews.

The abovementioned shows how important customer reviews are for customer purchase decisions. The importance of reviews for customers makes it even more important for organizations to analyze the reviews in order to find the different sentiments that can be found in the several reviews. Analyzing this data can lead to new business opportunities and competitive advantage.

One way to analyze reviews of customers is to make use of sentiment analysis. Sentiment analysis is also called opinion mining and is a natural language processing (NLP) technique which is used to analyze the opinions and sentiments of people towards different kind of entities such as: products, organizations and events (Liu B. , 2012). Both the terms sentiment analysis and opinion mining appeared in 2003 (Nasukawa & Yi, 2003; Dave et al., 2003). Sentiment analysis can be seen as a text mining technique which is used to classify reviews based on its contextual polarity, which is either positive or negative (Pang & Lee, 2008). Sentiment analysis has become an active research area because it has many challenging problems but it gives also the possibility to apply this analysis to different industries (Liu B. , 2012).

1.2. Research Questions

As has been mentioned in *chapter 1.1*, the WWW has offered companies a chance to understand their customers better based on the data they leave behind on the internet. Understanding the customer's dissatisfactions, but also their satisfactions, leads to more insights about how the customers perceive the brand/company. By having a better knowledge of what a customer likes or dislikes, a company can improve their customer service. This research is meant for the booking site of which the data is extracted from, because it helps them to better understand how machine learning can help to gather useful information based on their provided data. Therefore, the research questions that will be answered in this research are the following:

- **How well can machine learning techniques predict whether a customer is happy/unhappy based on its review?**
- **Which features are most important for a good/bad rating, and how are these features related to the sentiments of the customer reviews of the expensive hotels and inexpensive hotels?**

The first research question needs to be stated because it is important to know how well machine learning techniques can predict before a conclusion can be made about which features have an influence on a good or a bad rating. If the accuracy of the predictions is high, then we may have more confidence about the method. The more confident we get about the method, the more accurate the conclusions about the features will be. Therefore, after each model has been trained and tested, the model with the highest accuracy needs to be chosen.

After the best method has been chosen, it is possible to give an answer on the second research question. The model can show which features have the most influence on the rating that has been given by a customer. This gives us the opportunity to find out what makes customers satisfied, and at the other hand, what makes them dissatisfied. Are these customers more interested in the service that a hotel offers, the food that a hotel offers or the service that a hotel offers? What do these customers expect from an expensive or an inexpensive hotel? In other words, this information will help to understand what customers find important about a hotel.

1.3. Relevance.

Nowadays, there are millions of online users who share their judgements and read about the judgements of others towards a certain entity. As has been discussed in chapter 1.1, there are many online users who first make use of reviews before they make a buying decision. They even seem to have more trust in the reviews than in the employees of the store when it comes to gathering information. Since the reviews are of big importance for customers, it is of bigger importance for organizations. Therefore, it is useful for organizations to analyze the data in order to have a better understanding of their customers. The results of this research may be interesting for hotels in London. Understanding the opinion of customers may help them to provide a better service quality in the future.

1.4. Contribution

The main goal of this research is to find out which kind of features of a review are most important for a high or a low rating, and also how these features are related to each other. The research will also summarize and visualize the reviews as has been done in other studies, but it will differ in the way that the dataset is divided. This research has set a focus on classifying whether a customer is happy or unhappy based on their review.

The dataset that will be used consists of 20 hotels which are all based in London. 10 of these hotels belong to the most expensive ones in London and the remaining 10 hotels are the ones that belong to the least expensive hotels in London. The main contribution to the literature is that this research will look whether there are differences in sentiments within these two categories of hotels. This gives more insights in what customers like or dislike about the cheaper hotels, relative to the expensive hotels in London. With this information, organizations might understand better how well they perform on different aspects, which can help to improve the quality of the service of the hotel.

1.5. Structure of the thesis

The structure of the thesis will be as follows:

- **Literature review (Chapter 2):** this chapter consists of an explanation of the history of sentiment analysis and the domains in which this technique has been applied. As an addition, some other papers with the same goal will be discussed. Also, there will be an explanation of what customer satisfaction is, what electronic word-of-mouth is and what kind of opinion types exist.

- **Data (Chapter 3):** gives an explanation of the data that will be used. As an addition, some graphs will be shown to see the underlying structure of the reviews. Graphs with term frequencies will show which words occur most often, but also which words occur the least in reviews.

- **Methodology (Chapter 4)**: the methods that will be used in order to make predictive modelling possible will be discussed in this chapter.

- **Results & Evaluation (Chapter 5):** the results and evaluation will be discussed. Graphs and tables will be used in order to give a better understanding of the results. The different machine learning methods will be compared to each other based on their predicting accuracy. The best model amongst the others will be chosen.

- **Conclusion (Chapter 6):** lastly, this chapter will discuss the conclusion followed by some limitations of the research. As an addition, some recommendations for future research will be discussed.

2. Literature review

2.1. Customer satisfaction

One important goal of marketing is to have a good knowledge of what a customer likes or dislikes in order to influence its buying behavior. By understanding your customer, a company can give good customer service, which is a key factor to maintain and improve the satisfaction of a customer (Reibstein, 2002). The customers decision making process consists of a couple of steps according to Engel et al. (1995):

- **Recognition of need:** a customer realizes at some point that it needs a product when he/she compares it current situation with a situation in which the product is in his possession.
- **Searching for information:** a customer may ask friends or family about their opinion about a product before it will make a buying decision. Customers may also search on the internet to find some information about the product of interest.
- **Evaluation of alternatives:** consumers will analyze different alternatives so that they can ultimately choose the option that offers the best response to their need.
- **Purchase:** the customer comes at a point in which he makes the purchase of a product or service.
- **Consumption:** the customer utilizes the purchased good.
- **Evaluation of purchase:** after the purchase and consumption, the customer evaluates to which extend the purchased option satisfied them.

2.2. Sentiment analysis origin

Sentiment analysis is a field of study that has been analyzed by many researchers for many years. Before the year 2002, there has not been done a lot of research in the field of sentiments and opinions even though natural language processing has already been used many years before (Liu B. , 2012). Pang et al. (2002) and Turney (2002) were the first who did a research in which sentiment classification has been performed. Pang et al. (2002) performed this technique on reviews of movies, while Turney (2002) performed this technique on reviews of automobiles, banks, movies and travel destinations. However, the term sentiment analysis has been mentioned a year later for the first time in the paper of Nasukawa & Yi (2003). Since then, much research has done in the field of sentiment analysis. This is mainly because there is more opinion data available on the web (due to social media and review sites) than was the case before the year 2000 (Liu B. , 2012).

Sentiment analysis has been widely used in many domains. Liu et al. (2007) have made use of blogs in order to predict sales performance for movies. Bai (2011) performed sentiment analysis on online movie reviews of IMDB, in which the research its goal was to predict consumer sentiment based on several features such as unigrams and bigrams. When it comes to the food industry, Pai et al. (2013) researched the sentiments of customers who have bought fast food at Mc Donald’s, KFC and other fast food stores. Sentiment analysis has also been performed by Kang et al. (2012) for restaurants that don’t sell fast food, in which naïve Bayes algorithms have been used.

Other applications on which sentiment analysis has been performed are: news articles (Steinberger et al., 2011), Book reviews (Hu et al., 2012), Reactions on Social media such as YouTube and Flicker (Gupta et al., 2013), Digital cameras & MP3 reviews (Chen & Tseng, 2011) and news about stock markets (Hagenau et al., 2012).

The existing literature that perform sentiment analysis on hotel reviews mainly focus on machine learning techniques that help to analyze whether reviews are helpful to customers (O’Mahony & Smyth, 2009; Martin & Pu, 2014 ; Hu & Chen, 2016). Other authors have investigated the influence of customer reviews on hotel room sales (Ye et al., 2009; Ye et al, 2011; Anderson 2012).

The studies that mainly focus on the visualization and summarizing of sentiments differ in their used techniques. Akhtar et al. (2017) made use of sentiment analysis and summarized these sentiments in word frequency graphs. An even more interesting way to visualize customer reviews, has been done by making use of google maps (Bjørkelund, Burnett, & Nørvåg, 2012).

2.3. Electronic word-of-mouth

Electronic word of mouth (eWOM) has been defined as (Litvin et al., 2008): “*all informal communications directed at consumers through internet-based technology related to the usage or characteristics of particular goods and services*”. The growth of online reviews has made eWOM an important aspect in helping customers with their buying decision (Tang, 2017). Potential customers make use of reviews in order to have an indication about the quality of the product or service of interest based on the experience of others. These reviews lead to independence of customers since they do not depend anymore on the information of a product provided by the producer of the product (Tang, 2017). The reviews about a product or service can be found on many opinion platforms such as Tripadvisor, Yelp and Trustpilot. Customer reviews are also provided by many big companies such as Amazon, AliExpress, Uber and Airbnb.

2.4. Opinion types

According to Liu (2012), there exist four different types of opinions. The four types are *regular opinions, comparative opinions, explicit opinions* and *implicit opinions***.** The ***regular opinion*** itself can be divided into two types, namely ***direct opinion*** and ***indirect opinion***. The definition of a *direct opinion* is "***A direct opinion refers to an opinion expressed directly to an entity or an entity aspect***" and the definition for an *indirect opinion* is "**An opinion that is expressed indirectly on an entity or aspect of an entity based on its effects on some other entities**" (Liu B. , 2012).

An example of direct opinion is the following: "*The battery of my phone is perfect*". This opinion is one which directly gives an opinion about an aspect of the phone. An example of indirect opinions is the following: "*Since I have used my medicine, my body felt good*". This customer used a medicine which effected the body positively. This is an indirect opinion about the medicine that a customer used.

The second type of opinion, the ***comparative opinion***, has the following definition: "**A comparative opinion expresses a relation of similarities or differences between two or more entities or a preference of the opinion holder based on some shared aspects of the entities**" (Liu B. , 2012). An example of this is the following: "I like the burgers of KFC more than the ones of McDonalds".

Explicit opinion has the following definition: "**explicit opinion is a subjective statement that gives a regular or a comparative opinion**" (Liu B. , 2012). Example: "your tv is of such a great quality". ***Implicit opinion*** has the following definition: "**an implicit opinion is an objective statement that implies a regular or comparative opinion. Such an objective statement usually expresses a desirable or undesirable fact**" (Liu B. , 2012). Example: "I've bought these shoes a week ago and the sole is already falling off".

2.5. Predictive models origin

Several predictive models have been used in order to prognose the features that are most important for a good or bad review rating, but also to find out how well machine learning techniques predict whether a customer is happy/unhappy. One method that has been used in many applications is logistic regression. Pierre Francois Verhulst was the first who came up with the logistic function in his paper published in 1838 (Verhulst, 1838), followed by a paper of 1845 (Verhulst, 1845) and (Verhulst, 1847). The term logistic has first been mentioned in the paper of Pearl and Raymond (1922). However, Wilson was the first who published an application of the logistic (1943).

Another method that is often used for this analysis purpose is decision tree. According to Wei Yin Loh (2011), the earlier developments of the decision trees started with Fisher’s (1936) paper about discriminant analysis. The first classification tree algorithm that has been published is THAID (Fielding & O'Muircheartaigh, 1977; Messenger & Mandell, 1972). Years later, Breiman et al. (1984) introduced the CART and Quilan (1993) came with his own method called C4.5. C4.5 is based on the entropy measurement of impurity while CART is based on the Gini index which both will be discussed in chapter 4.

One method that improves the decision tree is the Random forest. The earlier developments of Random forest has first started with Ho’s proposed method (1995). In his paper, Ho introduced a method in which multiple trees are built by random selection. Furthermore, Amit and Geman (1997) proposed a method in which small number of variables are randomly chosen for each node. A couple of years later, Breiman (2001) developed the random forest by combining the methods of Ho (1995) and Amit & Geman (1997) with his own method of bagging sampling approach (1996).

3. Data

3.1. Data description

The dataset that will be used for this research is derived from an open source platform (Kaggle). The dataset consists of customer reviews of hotels that are based in London. The dataset that will be used consists of 20 hotels which are all based in London. 10 of these hotels belong to the most expensive ones in London and the remaining 10 hotels are the ones that belong to the least expensive hotels in London. The dataset exist of customer reviews in different languages. All the reviews will be taken into consideration. Removing the other languages is difficult and can lead to inaccurate predictions.

The dataset gives the opportunity to: compare the overall ratings between the different budget categories of hotels, compare the count of words of the reviews, analyze the most frequent used words, find upcoming word trends, find out positive and negative words and to perform sentiment analysis. The dataset consists of a total of 27330 different customer reviews. The dataset has been divided into training/validation/test sets with the following ratio 60/20/20. The following seven variables can be found in the dataset:

- ***Property name:*** this is the name of the hotel which has been booked by the specific customer.

- ***Review rating:*** it indicates what value a customer gives for the perceived service at the hotel that has been booked. The higher the rating, the better the experience was and vice versa.

- ***Review title:*** this gives a short description of what the review of the customer is about. This might be positive, negative or neutral. It could even be empty.

- ***Review text:*** this is the longer description of the review of the customer itself. The review contains the judgement and emotions of the customer towards an entity.

- ***Location of the review:*** the location where the customer is based when the review has been posted. It contains the city and country of the customer.

- ***Date of review:*** the date that the review has been posted. It indicates the precise day, month and year.

- **ID**: each review/customer has its own ID-number.

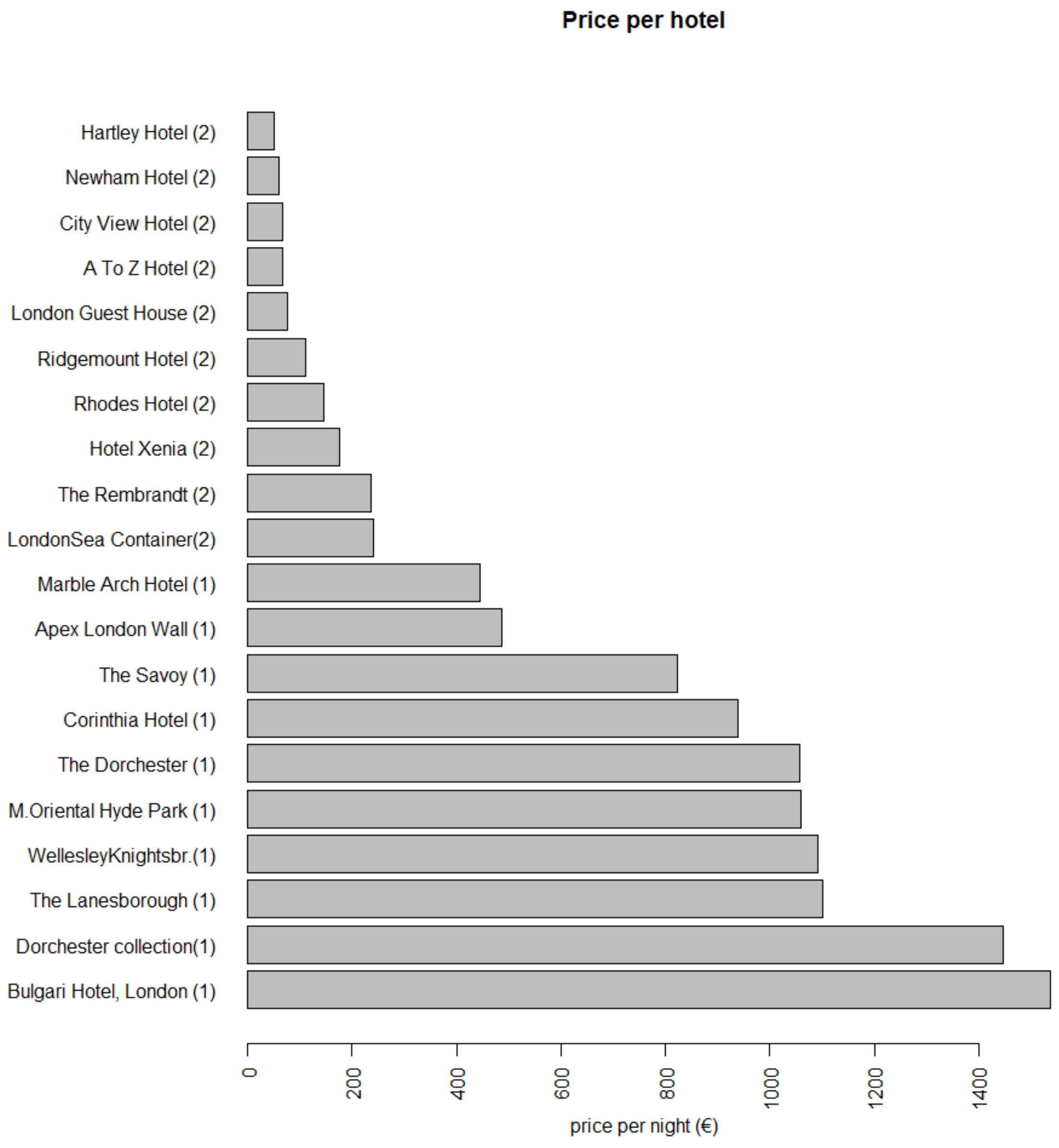

Figure 1: Price per night, per hotel.

Figure1 shows the ten most expensive and the ten least expensive hotels with their corresponding booking price. All the hotels have been arranged in order from highest to lowest price. As can be obtained from this table, the most expensive hotel in this dataset is the Bulgari hotel, followed by the 45 Park Lane – Dorchester Collection. The least expensive hotel is the Hartley hotel which has a price of €50. The price of the ten most expensive hotels ranges from €445 to €1538, while the price of the least expensive hotels ranges from €50 to €240. The average price of the ten most expensive hotels is €999, and the average price of the ten least expensive hotels is €123. 14.757 reviews have been shared by the customers of the ten most expensive hotels, while 12.572 reviews have been shared by customers of the ten least expensive hotels. The number of reviews for both categories seems to be almost equally divided.

3.2. Preprocessing & visualization

In order to make visualization and analysis of the data possible, it is of big importance to preprocess the data. Preprocessing the data has typically three steps that should be taken: remove stop words, apply stemming and term weighting (Solka, 2008). The first two steps, and also some additional ones which will be explained later, need to be taken because the data that has been obtained consists of reviews which may have some aspects in it that are hard to handle for data mining algorithms. For instance, reviews might consist of special characters which could be an emoticon, such as ☺, which indicates that a customer is happy. Human beings understand these characters but computers don’t unless you teach them. The following parts need to be handled:

1. Find sad emoticons such as ☹, and replace them with the emotion expressed in words.
2. Remove happy emoticons such as ☺, and replace them with the emotion expressed in words.
3. Find time AM in numbers, and express this in words (time_AM).
4. Find time PM in numbers, and express this in words (time_PM).
5. Find general time in numbers, and express this in words (time).
6. Remove all –
7. Remove all “
8. Remove all ;
9. Remove excess spaces
10. Remove excess .
11. Remove all numbers
12. Remove all special characteristics such as © and letters such as à.

13. Remove all words with one or 2 letters.

14. Remove capitals.

15. Remove stop words (except for no, not and never).

After all unnecessary words and special characters have been removed, stemming needs to be applied. Humans know that words such as poor and bad have the same meaning. This is not the case for computers. This is the reason why stemming is used. After the data has been cleaned and stemming has been applied, the data will be reduced in size. This is due to the fewer unique words that are now occurring in the text.

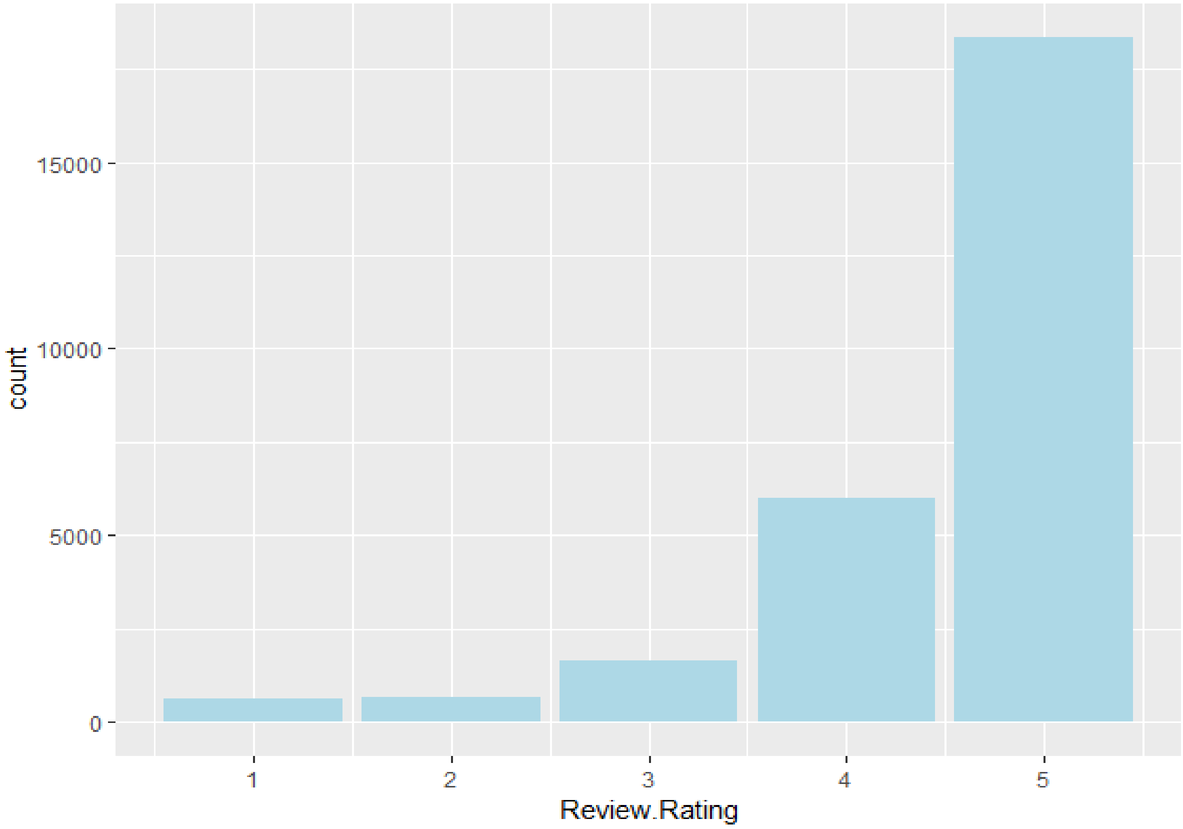

Figure 2: distribution of review ratings.

The review ratings have a non-normal distribution as can be observed from *Figure 2*. The ratings have a J-shaped distribution which has been discussed in chapter 2. There are way more reviews with a 4-star rating and a 5-star rating than reviews with a rating of 3 and lower. Customers will be divided into happy and unhappy based on their review rating. Happy customers are the ones who gave a rating of 4 and higher, while unhappy customers are the ones who gave a rating of 3 and lower. This will still lead to a J-shaped distribution with more happy customers than unhappy customers. The J-shaped distribution means that the data is imbalanced. One manner to overcome this problem, is to make use of an oversampling technique which is based on putting a higher weight on the minority class (unhappy customers).

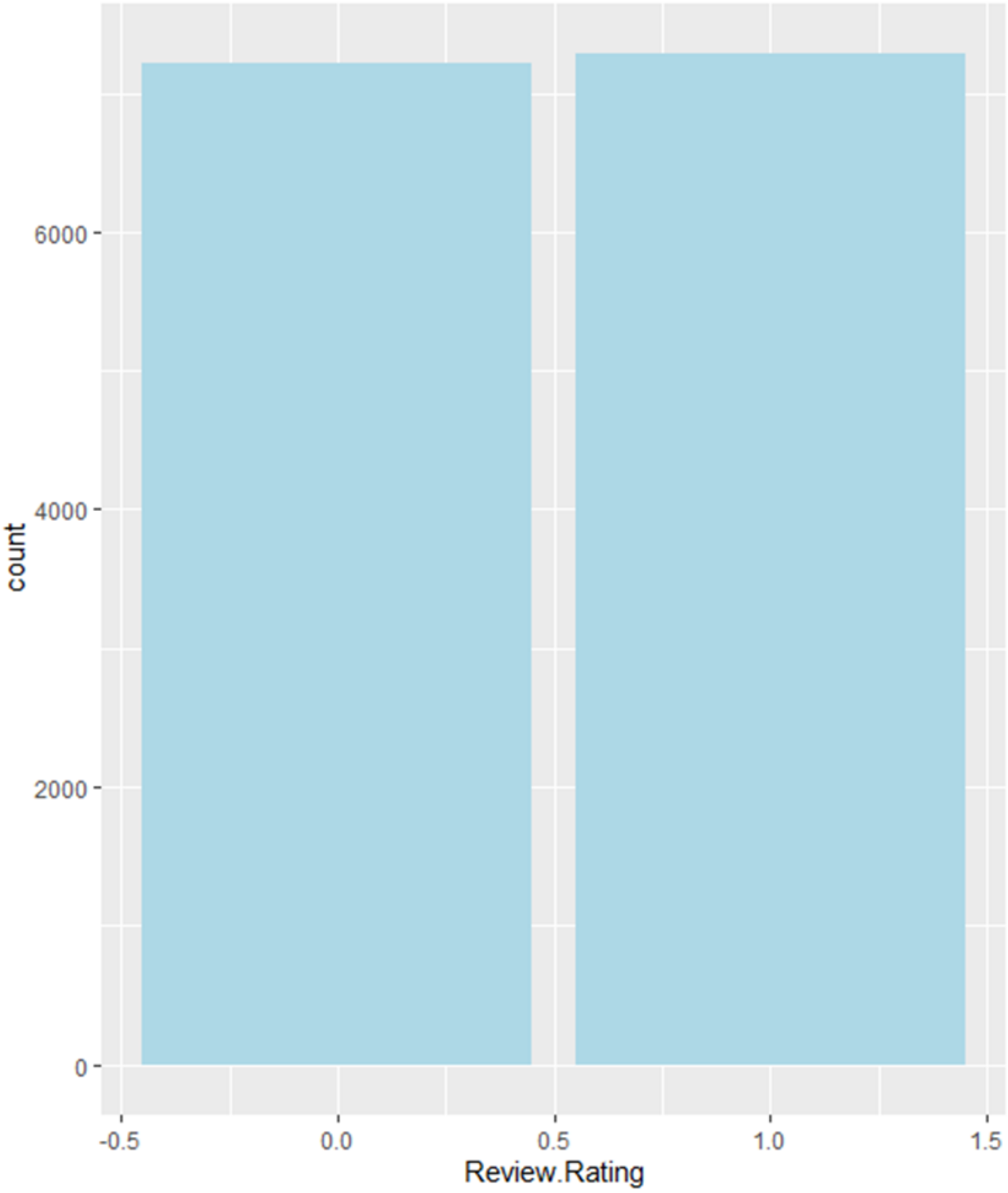

Figure 3: happy and unhappy customers (most expensive hotels).

In the case of this research, the minority class will get a higher weight of being picked for the new sample than the majority class (Liu A. C., 2004). This will be done based on sampling with replacement. *Figure 3* shows how the distribution of happy and unhappy customers looks like. The oversampling technique has led to an equal distribution of happy and unhappy customers, which means that the data is not imbalanced.

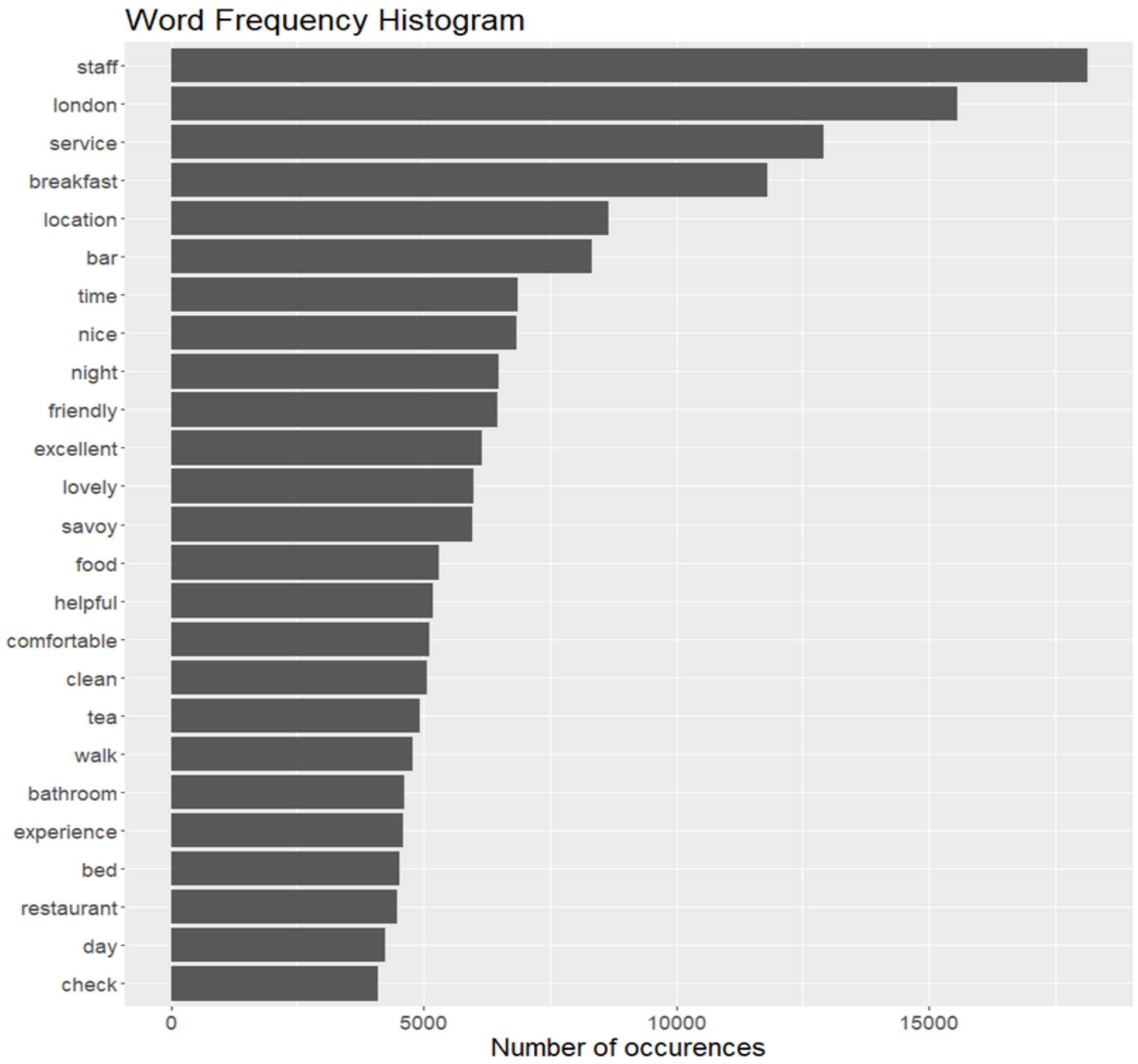

Figure 4: top 25 most occurring words in reviews

With the cleaned data, it is possible to give a clear visualization of the most relevant words. The more a customer talks about some aspect in a review, the more relevant it could be to the hotel. The word “hotel” was the most frequent word in reviews, with a count of over 40.000 times. This word has been deleted since this word is not informative by itself. As can be seen from *Figure 4*, the words that occur more than 10.000 times are “staff”, “London”, “service” and “breakfast”.

Something that might be more interesting to look at, is how words of these reviews are structured. More insights can be gathered when words are visualized in a Multidimensional scaling (MDS) map. MDS is a set of statistical techniques that tries to find hidden structures in multidimensional data. According to Agrafiotis et al. (2001): “ MDS is a collection of statistical techniques that attempt to embed a set of patterns described by means of a dissimilarity matrix into a low-dimensional display plane in a way that preserves their original pairwise interrelationships as closely as possible”.

The distance between two objects i and j in a Euclidean space is defined as:

$$d_{ij}(X) = \sqrt{\sum_{s=1}^{P} \left(x_{is} - x_{js} \right)^{2}}\ .$$

P indicates the number of dimensions. x_{is} is the value of the i^{th} row and k^{th} column (James et al., 2017). In order to perform MDS on the data, the SMACOF package will be used.

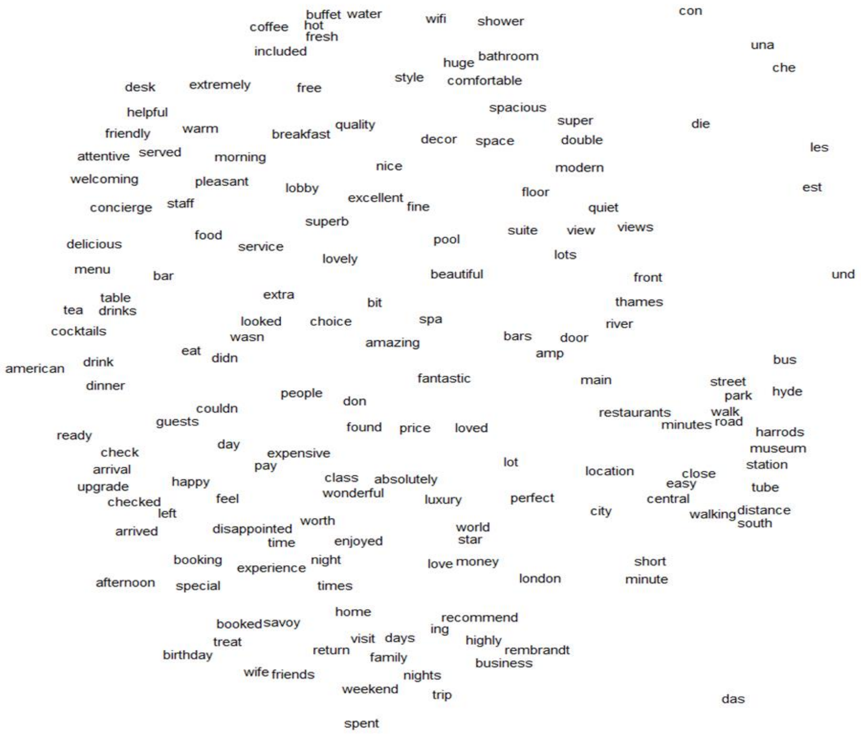

Figure 5: MDS map with window of 2

Figure 5 shows how the MDS map for this dataset looks like. With MDS, it is possible to count how often words are within a window of N-grams. A window means the reach in which words occur together. This has manually been chosen over a window of three and a window of six, because a window of two gives more interpretable results over the others. On the top, it can be seen that the words “coffee”, “hot”, “fresh”, “buffet” and “included” are near each other which is about a buffet and what it included. On the top left ,the words “attentive”, “friendly”, “helpful”, “desk”, “welcoming”, “warm” and “served” are all near each other which might mean that customers are very happy with the service of the employees who work at the desk.

On the middle right side of the lower-half of the MDS-map, the words “bus”, “museum”, “street”, “park”, “road” etc., are all words that occur together which is pretty normal because these customers might have been talking about the activities that can be done outside the hotel. It seems that customers are positive about the lobby, the pool and their suite since positive words such as “excellent”, “lovely”, “nice”, “superb” and “fine” are all near to these aforementioned words. This can be observed when looking at the upper half, slightly above the center of the MDS-map. Other words that are near each other are “Check”, “arrival”, “left” and “upgrade” which all have something to do with the check-in and check-out process.

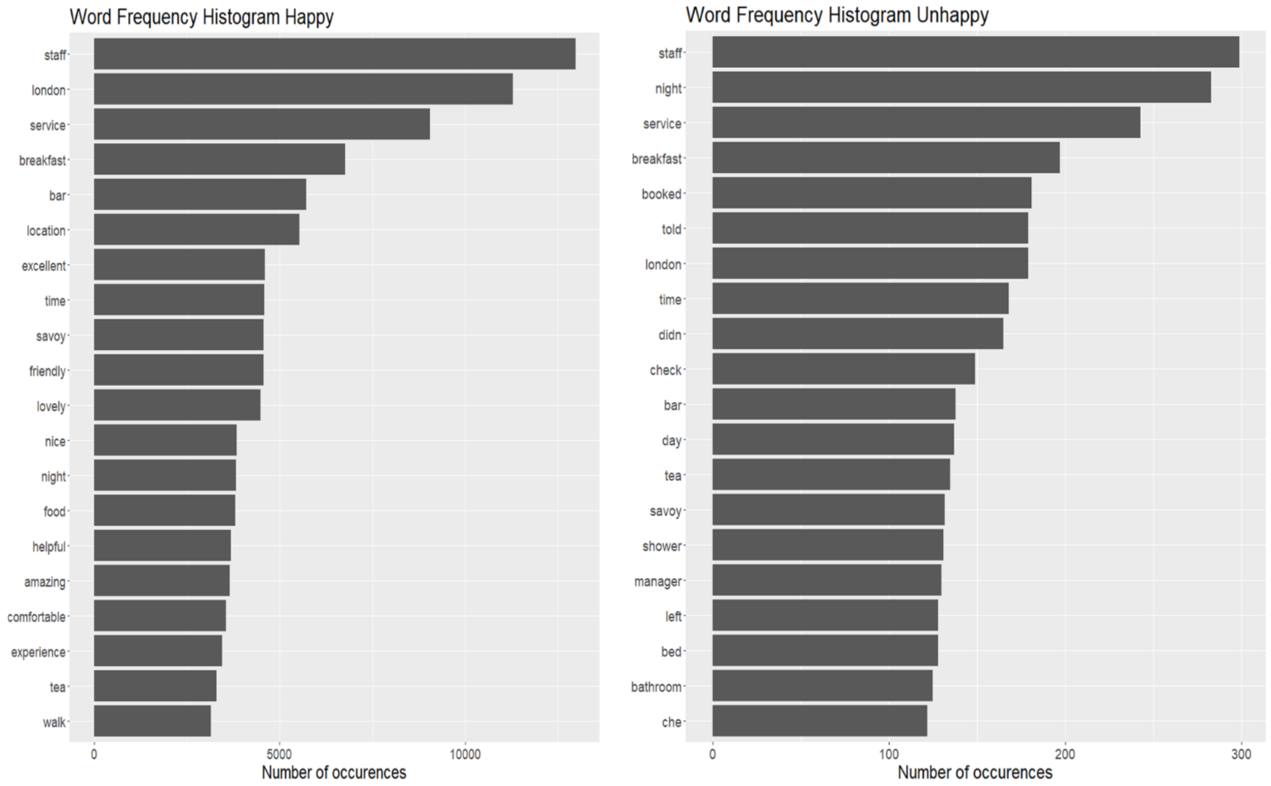

Figure 6: term frequency of happy and unhappy reviews

Even more insights can be gathered when looking at the word frequency of words that occur in reviews in which customers are the happiest, and reviews in which customers are most unhappy. Happy customers are the ones who gave a rating of 4 and higher, while unhappy customers are the ones who gave a rating of 3 and lower. *Figure 6* shows the word frequency of the happy and unhappy customers. Happy customers are very satisfied with the bar, the breakfast, the service that the hotel offers and the staff of the hotels.

On the other hand, it can be noticed that unhappy customers also have some bad reviews about the staff, the breakfast and the service. These same aspects of the reviews in happy and unhappy reviews might occur because this dataset consists of twenty hotels in which the ten most expensive and the ten least expensive hotels have been taken into consideration. Not every hotel has the same staff and the same service and the same quality when it comes to the offered breakfast. Also the bathroom, the shower and the manager are occurring most often in negative sentences.

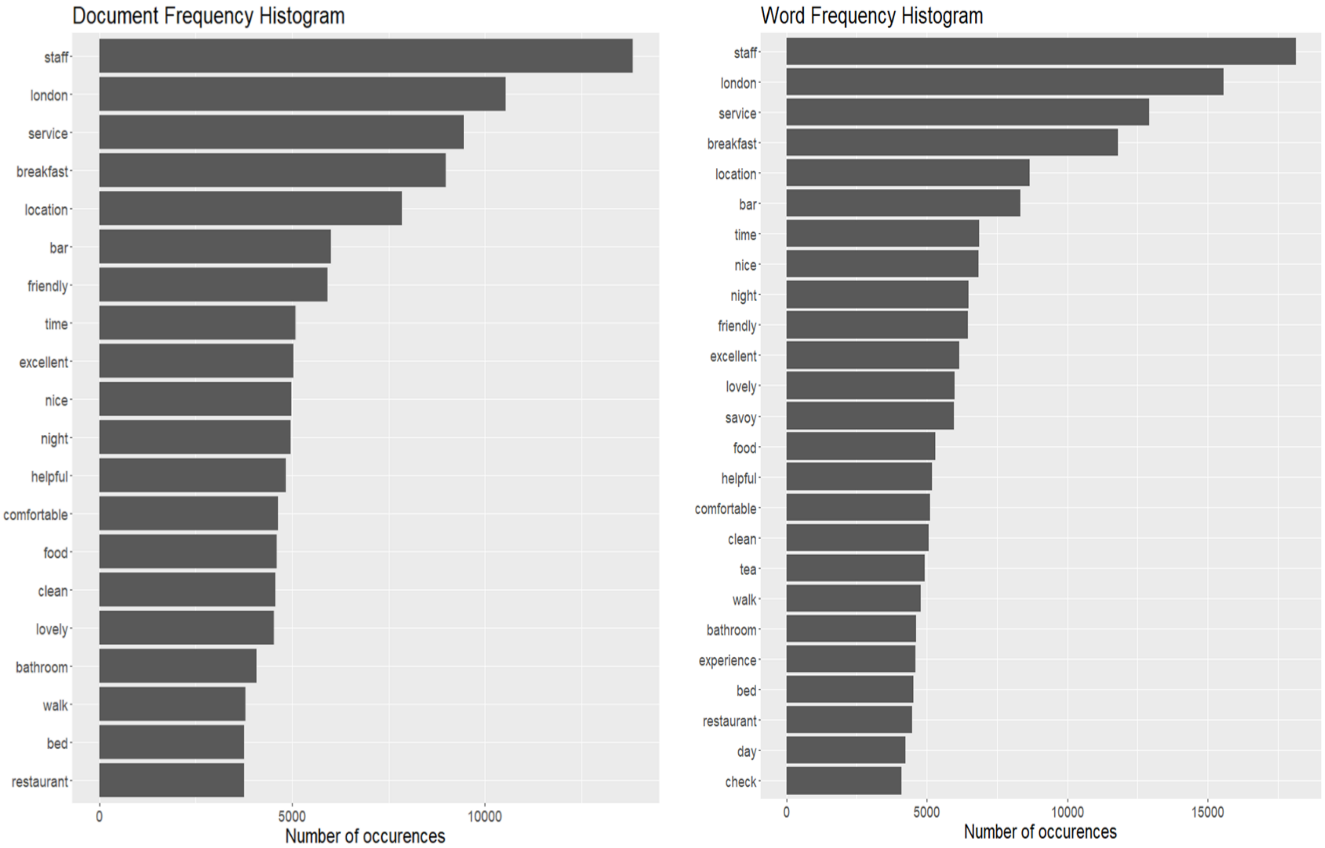

Figure 7: Document frequency and term frequency

As has been noticed in *Figure 6*, some words occurred in both the reviews of happy customers and unhappy customers. Therefore, having a look at the relativity between reviews of happy and unhappy customers will be more insightful since this way of visualizing words only looks at relative differences between these reviews. The relative difference can be gathered when making use of TF-IDF (= Term Frequency – inverse Document Frequency). Basically, this is a manner to classify words as being more relevant and unique for a given rating in documents (Zhang et al., 2011).

Term frequency is about how many times a word (w) occurs in a document (d), while ***document frequency*** is the number of times a word occurs in a document (d) (Neto et al., 2002). TF-IDF of word w in document d is defined as the term frequency of word w in document d (denoted by *TF(w,d)),* multiplied by the Inverse document frequency of word w (denoted by *IDF (w))*. The formula looks as follows:

$$TF\text{-}IDF\ (w,d) = TF(w,d) * IDF\ (w).$$

The higher TF *(w,d),* the more often this word occurs in a given document (d). Also, the higher the *IDF (w),* the less often this word occurs in a document and vice versa (Neto et al., 2002). *Figure 7* shows the document frequency and the word frequency histogram. The term frequency in *Figure 7* is the same as in *Figure 4*. It has been put together just to see how the term frequency and word frequency differ. Some words such as “staff”, “london”, “service”, “breakfast”, “location”, etc., do occur in both histograms and are even on the same place. The main difference is that these words don’t have the same number of occurrences. For instance, the word “breakfast” occurs more than 10.000 times when it comes to its term frequency, which counts how many times this word occurs in all reviews in total. On the other hand, this word occurs less than 10.000 times when it comes to its document frequency, which indicates that this word occurs in less than 10.000 reviews.

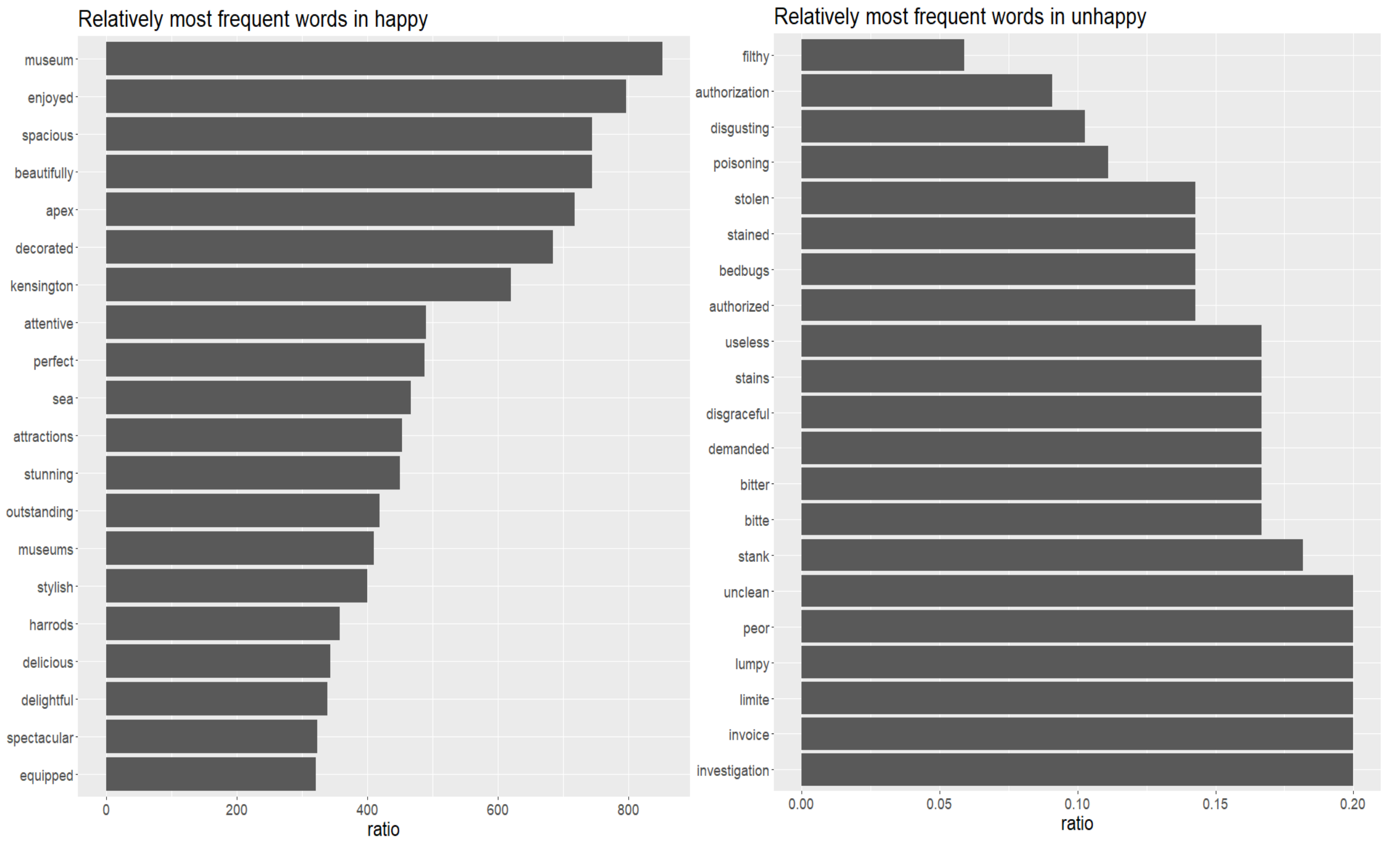

Figure 8: relatively most frequent words in happy and unhappy reviews

Figure 8 shows the relative difference between words that occur in happy and unhappy reviews. All the words in the histogram for happy reviews have a high TF-IDF value which means that these words are most frequent for happy reviews and less frequent in unhappy reviews. On the other hand, the histogram for unhappy customers has a low TF-IDF value which means that these words are most frequent for unhappy reviews, but less frequent for unhappy reviews. As can be seen from the histogram of words that occur in positive reviews, there are many positive words that have been used such as “enjoyed”, “beautifully”, “attentive” and “perfect”. Happy customers also talk most often about the sea and the museum, relative to the unhappy customers. It might be that some hotels have a sea and a museum nearby which is an aspect that are highly valued by customers.

The words that occur in unhappy reviews are not very positive relative to the happy reviews, which is pretty obvious. Unhappy customers use non-positive words such as “filthy”, “disgusting”, “stolen”, “poisoning” and “useless”. Also some aspects of the hotel have been spoken about in a negative sentence such as “bedbugs” and “invoice”. It might be that the customers had a stay in a hotel in which their bed had bedbugs and was not very clean. These unhappy customers might also have had some problems with their payment of the reservation of the hotel room since the word invoice scores relatively high in unhappy reviews.

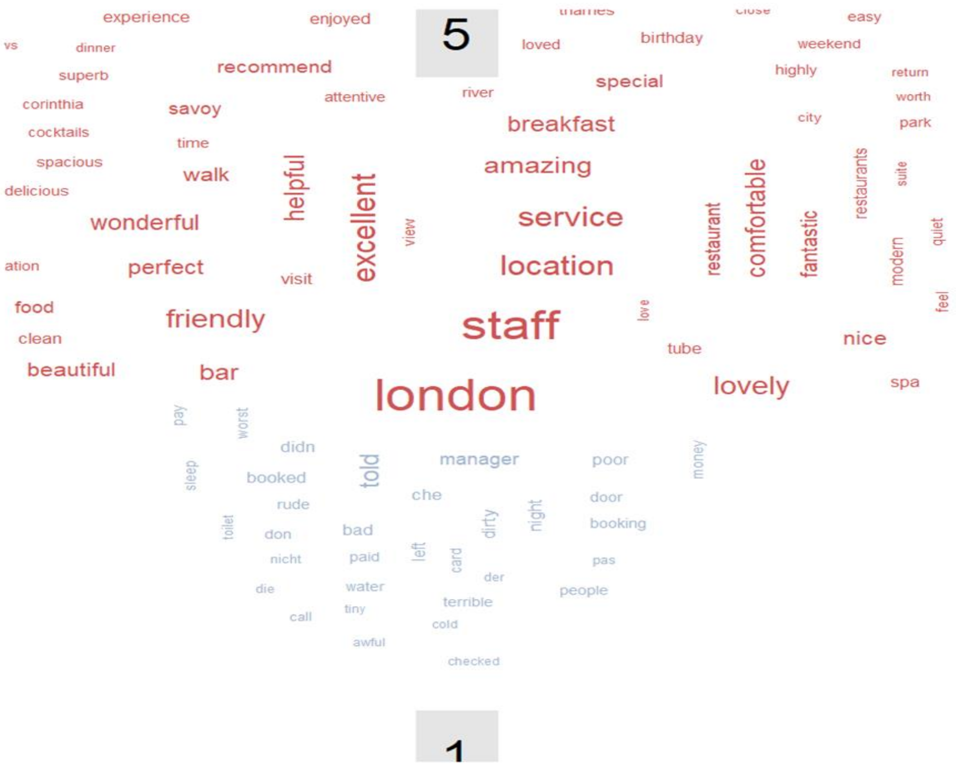

Figure 9: word cloud of happy and unhappy reviews

The best way to visualize the relative difference between unhappy and happy reviews, is by making use of a word cloud. *Figure 9* is an example of how a word cloud looks like. As might be seen, the word cloud has blue words on the bottom half. These words belong to the reviews of unhappy customers who gave a rating of 3 and lower. On the top half, the words are red. These words belong to the happy customers who gave a rating of 4 and higher. The bigger the size of the word, the more important this word is to the reviews of these customers.

Again, the happy customers use very positive words in their reviews such as “excellent”, “helpful”, “comfortable” and “fantastic”. They are also positive about some aspects of the hotel such as the spa, the service, the bar and the staff. Even the words “return” and “recommend” occur in these reviews, which might be because the customers indicate that they recommend the hotel and they would like to come back because they had a good experience with the hotel. The unhappy customers make use of non-positive words such as “poor”, “rude”, “terrible” and “worst”. Some aspects of the hotel that they did not like, had something to do with the manager, the toilet and the card.

4. Methodology

4.1. Introduction

There are many data mining techniques that can be used in order to make classification of reviews possible. However, this research will focus on four interrelated methods. First, sentiment analysis will be discussed followed by Logistic regression. The third method that will be used is Decision tree. Lastly, Random forest will be discussed.

4.2. Sentiment analysis

As has been mentioned in *chapter 1.4.*, the ever-growing data availability has led to many applications of sentiment analysis that have been published in several domains. Applications have been performed on the domain of movies, the food industry, news articles, social media, books, gadgets, stock markets etc. this all indicates how important sentiment analysis has become in the recent years. It is needed to better understand how this method works in order to better understand why this method suits this domain. This subchapter will discuss which different types of opinions exist, the definition of sentiment analysis and the different levels of sentiment analysis. Lastly, the main challenges of this method will be explained.

4.2.1. Sentiment analysis definition & process

Sentiment analysis is a machine learning method that is usually used to analyze reviews of customers. According to Fang & Zhan (2015) a sentiment can be either a judgement, ideas, emotions or opinions which are prompted by feeling. Sentiment analysis is also called opinion mining and is a natural language processing (NLP) technique which is used to analyze the opinions and sentiments of people towards different kind of entities such as: products, organizations and events (Liu, 2012). It is a text mining technique which classifies reviews based on its contextual polarity, which is either positive, negative or neutral (Pang & Lee, 2008).

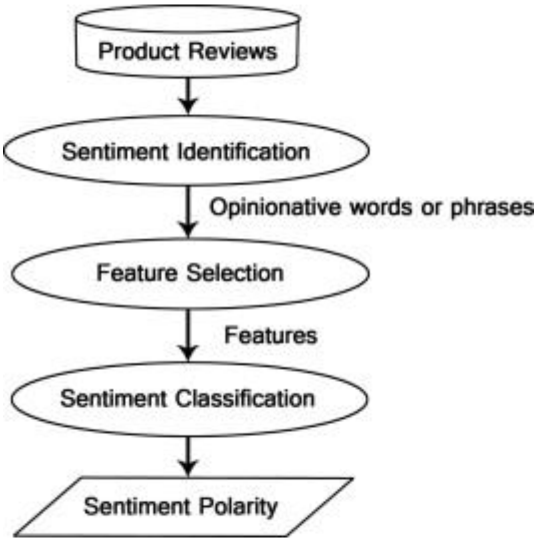

Figure 10: sentiment analysis process (Medhat et al., 2014)

Figure 10 shows the process of sentiment analysis. The basic idea behind this is that words in a document (reviews in this case) will be compared to a lexicon, which is a sentiment dictionary. For each review, the sentiments need to be identified. Hereafter, features need to be selected. The most common selected features are the following:

- ***Term frequency:*** this feature counts the frequency of word n-grams. n-grams are words or letters that follow each other in a given text (Meng & Chu, 1999). In this case the n stands for the number of words that follow each other in the given text. A unigram is therefore one single word, while a bigram consists of two words that follow each other up. The count of words can be in terms of one and zero, in which one means that the word has occurred in the document and 0 that it did not. The count of words can also be done in terms of how many times the words occur in total in a document (Medhat, Hassan, & Korashy, 2014).

- ***Parts of speech:*** this manner of feature selection tries to find words of different parts of speech, because it can be seen as one of the important aspects of customer opinions. This method looks at each word independently and checks whether a word is a noun, an adjective a verb etc. in short, every word gets its own tag (Medhat, Hassan, & Korashy, 2014).

- ***Opinion sentences and words:*** these words can also be used as a feature because they show an opinion about an entity. For example the words good and bad. They both give an opinion about an object (Medhat, Hassan, & Korashy, 2014).

- ***Negations:*** negations are words that lead to a whole different meaning of a word/sentence. Let's consider the following sentences: "you are not good to me" & "you are good to me". The word "not" leads to a sentence that has the opposite meaning. This makes this feature an important one when it comes to performing sentiment analysis (Medhat, Hassan, & Korashy, 2014).

After feature selection, each word that occurs in the dictionary will be classified as positive, negative or neutral. This dictionary has been compiled by Hu and Liu (2004). Each positive word in a document will lead to a higher sentiment score, because each positive word will add one point (+1). Each negative word lowers the sentiment score of a document, because each negative word will lower the overall score with one point (-1). Neutral words will not have any effect on the overall score (Hu & Liu, 2004).

4.2.2. Different levels of sentiment analysis

There are three levels of polarity classification. The three levels of classification are the sentence level, the document level and the aspect level (Fang & Zhan, 2015). The three levels will be discussed in the following chapters.

Document level

The document level considers the whole text as one unit, and checks whether this unit of text is either positive, negative or neutral. In this case, each customer's review will be seen as one document and each document will be classified independently. This level considers that each document has only one opinion about an entity (Boudad et al, 2017). Therefore, this level is not suited for documents that consists of opinions about more than one entity.

Sentence level

On the sentence level, each sentence in a document will be classified separately. The goal is to find out what the sentiment and opinion of each sentence is. Sarcasm occurs much often in reviews, which makes analyses more difficult. One of the main challenges of sentiment analysis is that it has problems with handling sarcasm. Performing sentiment analysis on the sentence level can deal with this problem (Boudad et al, 2017).

Aspect level

Both the sentence level and the document level don't look at what customers like or dislike about an entity. The aspect level on the other hand looks directly to the opinions that belong to each aspect of the entity. The idea behind this approach is that every judgement of a customer has also an object for which the judgement is meant (Liu B. , 2012). The following example will give a better explanation. The sentence "My room was great, but the service was awful" has one positive and one negative opinion. Each of these opinions belong to two different aspects of a hotel, which is the entity. With this kind of analysis, a summary about the opinions of each entity with their corresponding aspects can be made (Liu B. , 2012).

4.3. Logistic regression

Binary logistic regression is a statistical technique which is used when the response variable has two classes. Logistic regression seeks to model the probability that an event occurs based on the value of the independent variables. The reason that logistic regression has been chosen over probit regression is because the logistic regression makes it possible to interpret an odds ratio. The odds ratio is equal to the probability that an event occurs divided by the probability that an event does not occur (Kleinbaum et al., 2002). The mathematical form of the odds ratio is:

$$Odds = \frac{P(d)}{1-P(d)},$$

where $P(d)$ is the probability of the occurrence of event d. For example: the probability of an event occurring is equal to 0,20 and the probability of an event not occurring is equal to 0,80. The odds ratio will then be 1/4, which means that the odds are 4 to 1 that the event will not occur. Logistic regression tries to classify observations by making estimations of the probability that an observation is in one of the two categories. The logistic model consists of the index of $y = \beta_0 + \beta_1 x_1 + . \ldots + \beta_q x_q$, which is a linear model. This index (y) consists of independent variables (x's) and coefficients β_0 to β_q which are unknown parameters.

Furthermore, the probability of an event can also be defined as: $P(d \mid x_1, x_2, \dots, x_q)$. Where d stands for the outcome of an event occurring, given the x's. With the defined index and probability, it is possible to define the logistic model. The logistic model tries to find the probability that the response variable is equal to 1 for a given observation. The logistic model is defined as:

$$P(d) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\dots+\beta_q x_q)}}.$$

In order to model the probability, the logistic function is needed. The formula of the logistic function is as follows:

$$\mathrm{f}(y) = \frac{1}{1+e^{-y}}\ .$$

This function shows in which boundary the logistic model lies. The probability of an event occurring can be determined by inserting the linear function y into the logistic function. This formula makes it clear that $\mathrm{f}(y)$ ranges in between 0 and 1. As an example: when y is equal to minus infinite, then the logistic function $\mathrm{f}(y)$ is almost equal to 0. Furthermore, when y is equal to infinite, then the logistic function $\mathrm{f}(y)$ is almost equal to 1. The logistic function is, unlike the linear function, S-shaped since its values are ranged in between 0 and 1. This can also be observed in *Figure 11*.

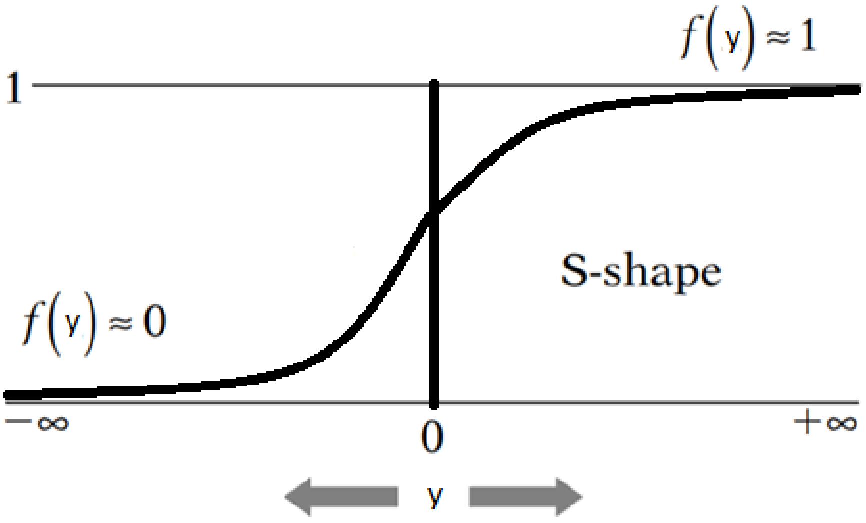

Figure 11: Logistic function

The unknown parameters β_0 to β_q still need to be estimated. This will be done based on data obtained for the x's and its outcome y. In short, the unknown parameters will be estimated based on the data set. In order to estimate the unknown coefficients of the model, the likelihood function (L) will be used. The likelihood function is formulated as (Kleinbaum et al., 2002):

$$L(\beta) = \prod_{i=1}^{m_1} P(d_i) \prod_{i=m_1+1}^{n} (1 - P(d_i)) \, .$$

The observed data can be divided into m_1 happy customers and $n - m_1$ unhappy customers. $P(d_i)$ is the probability that the data has been obtained for a happy customer, while the probability of obtaining the data for the unhappy customers is then equal to $1 - P(d_i)$. The logistic model needs to be substituted into the likelihood function and after that it needs to be maximized. The likelihood function will then look as follows:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{exp\left(\beta_0 + \sum_{j=1}^{q} \beta_j \, x_{ij}\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{q} \beta_j \, x_{ij}\right)} \right)^{Y_i} \left(\frac{1}{1 + exp\left(\beta_0 + \sum_{j=1}^{q} \beta_j \, x_{ij}\right)} \right)^{1-Y_i} .$$

In short, the likelihood function tries to estimate values for all unknown coefficients by maximizing the likelihood function. Putting these estimates in the logistic function will lead to a value of near one for each individual who is happy and a value close to zero for each individual who is unhappy (Kleinbaum et al., 2002).

4.4. Variable selection

Variable selection is of big importance for the logistic regression. Having too many variables in the model can lead to overfitting, while using too few can lead to underfitting. Forward-backward selection is a feature selection method that can be used to find the best model and corresponding variables. Forward selection means that the model begins only from the intercept in which no other variable has been taken into account (James et al., 2017). From there on it adds one predictor variable at a time. Backward selection makes use of the full model with all predictor variables being considered. From there on it removes one predictor variable at a time.

Forward-backward is the hybrid approach of forward selection and backward selection. It also starts with a null model and adds one variable at a time, but after a variable has been added, one can also be deleted when it does not give any improvement to the model. This method tries to remove and add variables based on an information criteria (James et al., 2017). The information criteria can be one of the following (Chen et al., 1999):

- AIC = $-2\log(\max(L)) + 2q$;
- BIC = $-2\log(\max(L)) + \log(n)q$.

The number of parameters that need to be estimated is indicated by q. Furthermore, n stands for the sample size and $\max(L)$ stands for the maximum likelihood. The best model is the one with the lowest AIC and BIC. AIC will be used as the information criterion to find the most optimal model.

4.5. Decision trees

With decision trees, it is possible to build classification or regression models which is visualized in a tree structure. Each decision tree consists of a root node, internal nodes, branches and leaf nodes. From *Figure 12*, it can be observed that each internal node tests an attribute, each branch is an answer on a given test and each leaf node gives the outcome after all tests have been answered.

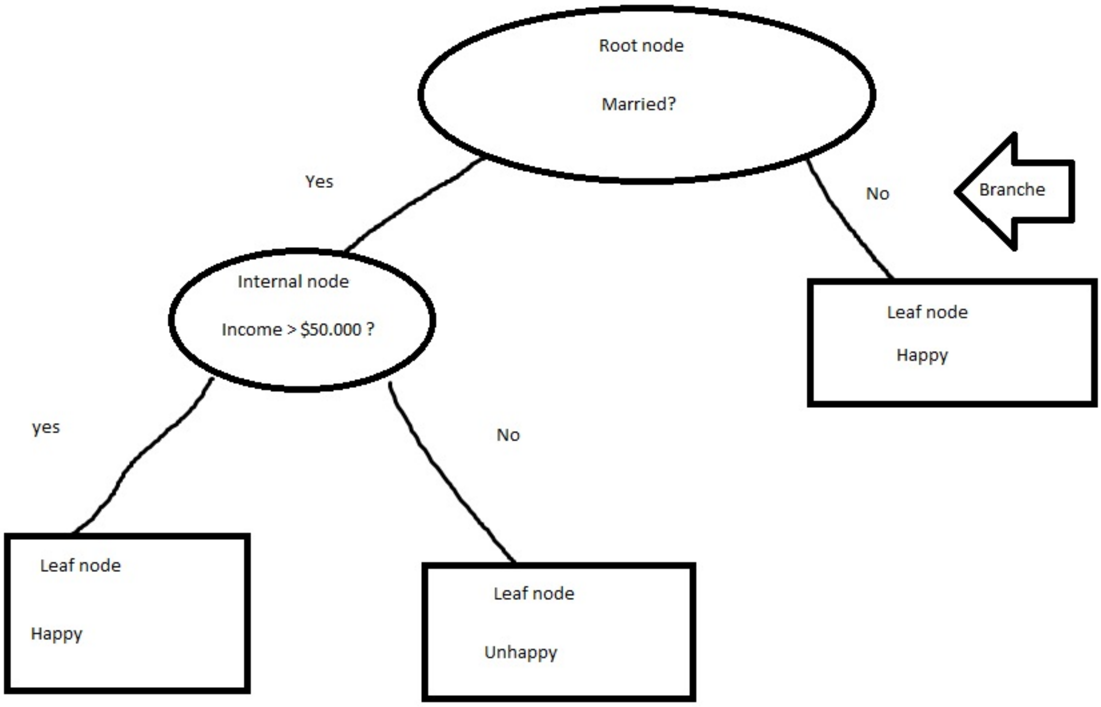

Figure 12: Decision tree example

The way decision trees are visualized is easy to interpret, which is one advantage of this method. The disadvantage of this method is that a small change in the data can cause a big change in the outcome of the decision tree. This means that this method suffers from a high variance. In order to build decision trees, it is needed to consider a couple of steps. The first step is to make use of top-down approach, which means the tree starts at the top. This method is based on recursive binary splitting. Decision trees try to divide observations in different regions. The algorithm is designed to maximize the precision of classifications for each split. In order to make this possible, the objective function is needed which is the classification error rate. According to James et al. (2017), the classification error rate is defined as:

$$E = 1 - \max_{k}(\hat{P}_{gk}).$$

$\hat{P}_{gk}$ stands for the proportion of the training observations that belongs to region g and a given class k. The classification error rate is the part of the training observations that has been put into the wrong region. Other methods to grow a tree are Gini index and cross-entropy. These methods measure the total variance amongst all K classes. The Gini index is defined as:

$$G = \sum_{k=1}^{K} \left(\hat{P}_{gk}(1 - \hat{P}_{gk}\right).$$

If the Gini index has a value of zero, it means that the node consists of observations from a single category. In short, the lower this index, the more pure this node is. The cross-entropy is defined as:

$$C = -\sum_{k=1}^{K} \hat{P}_{gk} log \hat{P}_{gk}.$$

If the value of $\hat{P}_{gk}$ is near to one and zero, then the cross-entropy will be near to zero. The lower the cross-entropy, the more pure the nodes are. The Gini index gives the probability that a sample has been misclassified, while the Entropy measurement indicates the information gain/loss. Since the Gini index is easier to interpret, this one will be considered for each split.

Decision trees can be tuned by making use of the complexity parameter (cp). Another manner of tuning decision trees is by making use of bagging which will be discussed in the next subchapter. The cp controls the size of the tree by checking whether the cost of adding another variable is above the value of cp. If this is the case, the tree building will stop. In *Figure 13*, the optimal cp-value is the point in which the error is at its lowest point. The optimal cp-value is equal to 0,011.

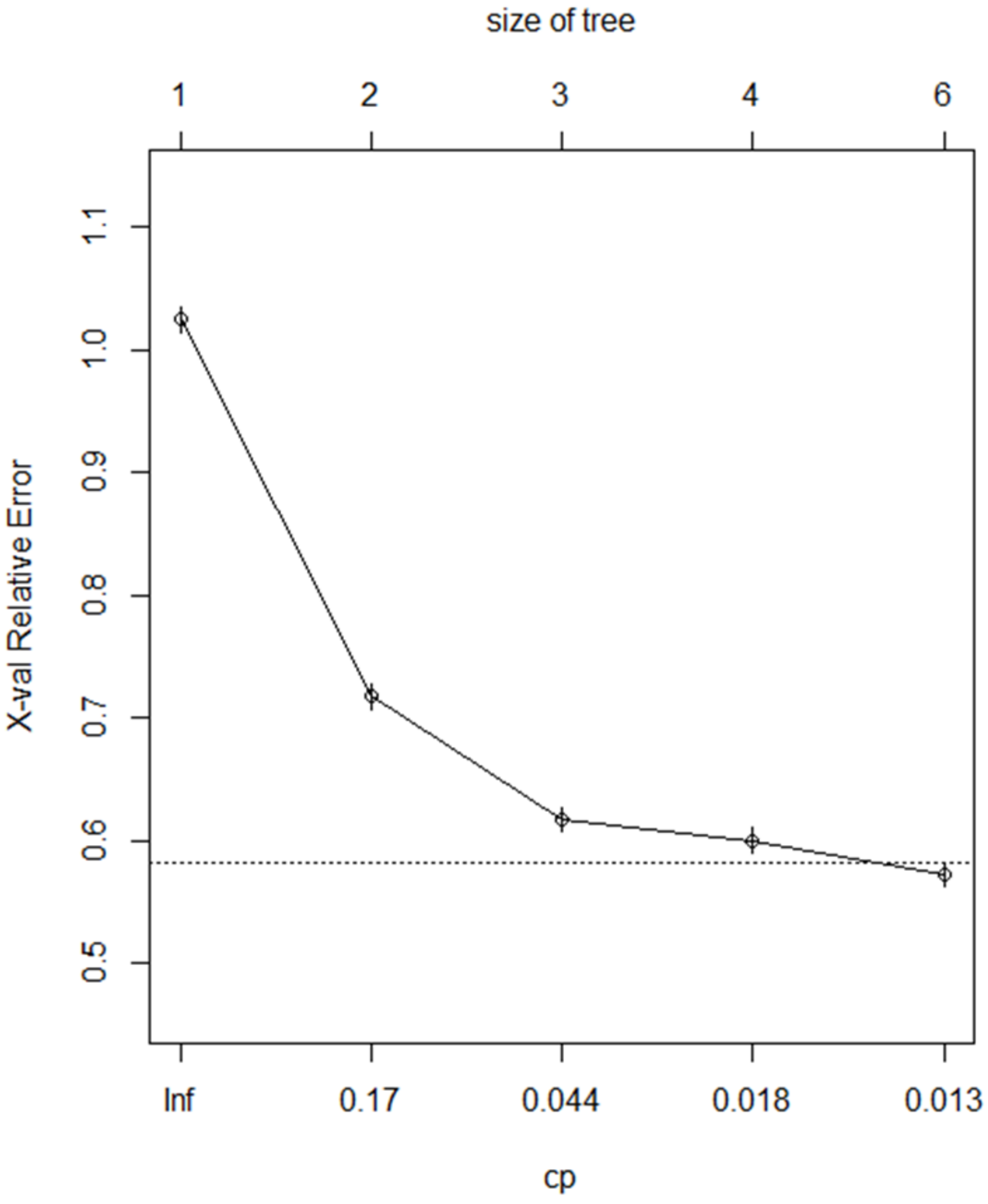

Figure 13: optimal cp-value

4.6. Random Forest

Random forest is a method that is mainly used to improve the prediction power of decision trees. As already has been mentioned, the disadvantage of decision trees is that a small change in the data can cause a big change in the outcome of the decision tree. This means that this method suffers from a high variance. Random forest is a method that reduces the variance of decision trees by building several decision trees on bootstrapped training samples (James et al., 2017). With bootstrapping, observations will be randomly selected from the dataset and it will be added to a sample. Some observations can occur more than ones in a bootstrapped sample. Averaging the predictions will lead to the final model with a lower variance.

Random forest also reduces the variance of decision trees by making use of random selection of variables. For each split of a decision tree, a random sample of these variables will be considered of the total number of available predictors. This means that random forest only allows a subset of the predictors to be considered at each split, rather than the full set of the predictors. By randomly choosing features for each split, the built trees will become less correlated with each other. The Gini index is used to as a selection measure for the Random forest classifier. As has been discussed in chapter 4.5, the Gini index is formulated as the following:

$$G = \sum_{k=1}^{K} \left(\hat{P}_{gk}(1 - \hat{P}_{gk})\right).$$

Furthermore, it is possible to tune the algorithm by choosing the most optimal number of variables considered at each split and the optimal number of grown trees. The optimal number of trees can be chosen based on Figure 14.

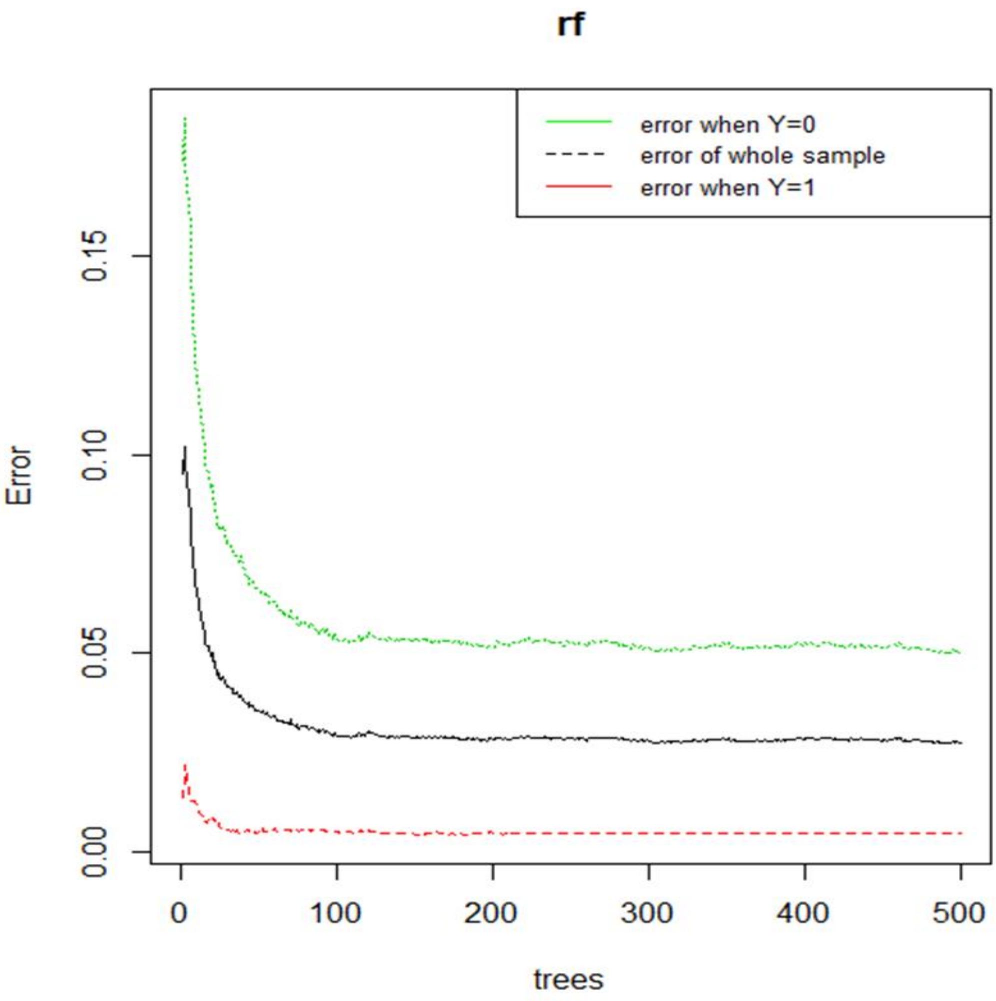

Figure 14: optimal number of trees

The red line in *Figure 14* indicates the error when Y=1, while the green line indicates the error when Y=0. The black line represents the entire sample (Muschelli et al., 2014). The trees are tuned by making use of training and test set. The optimal number of trees is the point in which the black line does not decline anymore. To be sure, the number of trees that will be chosen is 100. According to James et al. (2017), the error can be calculated as:

$$1 - \frac{accurate\ predictions}{total\ predictions}\,.$$

In other words, it is the probability that an observation has been misclassified.

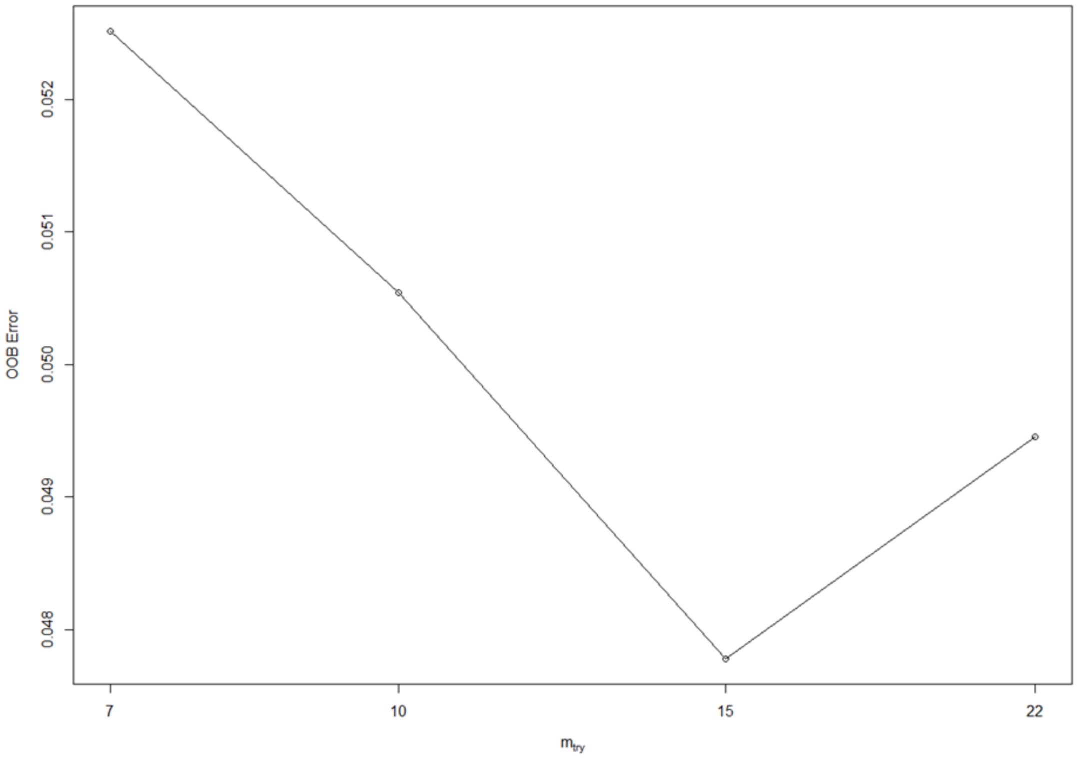

Figure 15: optimal number of considered variables at each split

The number of variables that will be considered at each split can be determined by looking at *Figure 15*. The optimal number of variables being considered at each split can be found at the point in which the error is at the lowest point, which means that the optimal number of variables considered at each split is 15.

5. Results

5.1. Sentiment score

In order to obtain the results, sentiment analysis has been ran on the dataset. Sentiment analysis is important for this research because it helps to understand which kind of sentiment information is present in the reviews, which sentiments are positive, negative or neutral and the strengths of these sentiments.

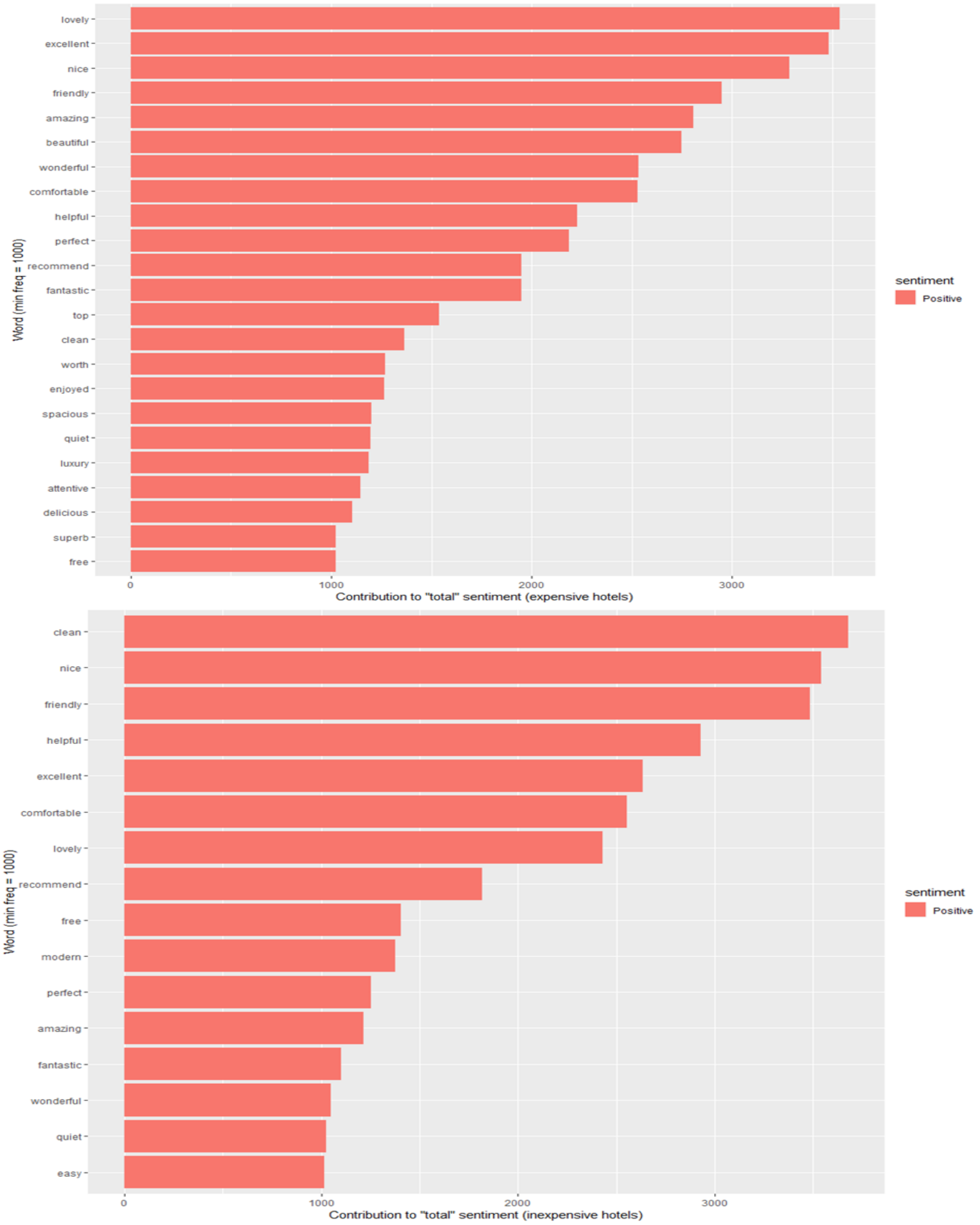

Figure 16: Most occurring positive sentiments in each category

Figure 16 shows the positive sentiments that occur the most in both the reviews of the inexpensive and the expensive hotels. When it comes to the difference in positive words used in reviews in both categories, one big difference is the word “clean” that has the highest sentimental score for the inexpensive hotels. Customers might have had a feeling that the inexpensive hotels would not be very clean due to the cheaper price they’ve paid. They might therefore be more satisfied when they find out that the hotel is clean, which could have led to the frequently mentioned word. It can also be observed that the reviews of the customers of the expensive hotels are much stronger than the positive ones used in the reviews of the inexpensive hotels. For instance, the seven most occurring positive words in the reviews of the inexpensive hotels consists of words such as “clean”, “nice”, “friendly”, “helpful”, “comfortable”, “excellent” and “lovely”, while the reviews of the expensive hotels consist of words such as “lovely”, “excellent”, “nice”, “friendly”, “beautiful”, ”wonderful” and “amazing”. Words such as “excellent”, “perfect” and “wonderful” are more frequently used in the reviews of the expensive hotels. Therefore, it becomes clear that the positive sentiments of the reviews of the expensive hotels are all much stronger than the ones mentioned in the reviews of the inexpensive hotels.

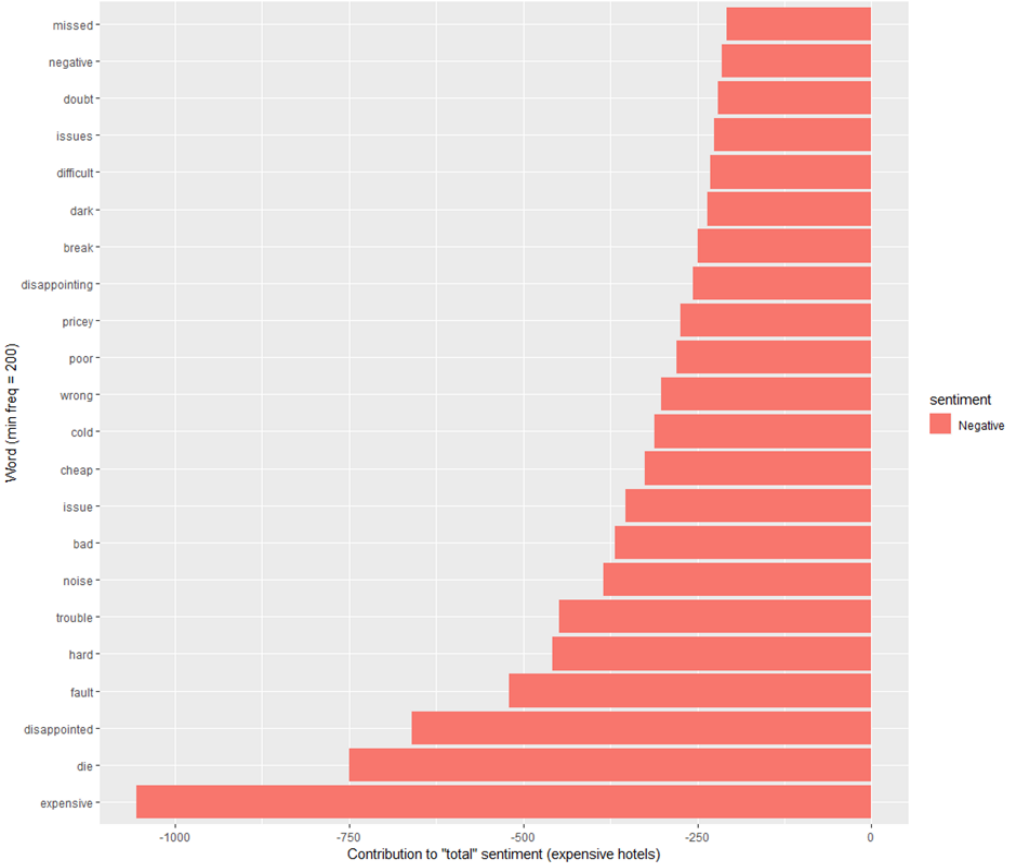

Figure 17: Most occurring negative sentiments (expensive hotels)

There are two important negatively used words in *Figure 17* which indicate well how some customers think about the expensive hotels. The words "expensive" and "disappointed" are much often used and have a high negative overall sentiment score in the reviews of the expensive hotels. The word "expensive" speaks for itself, since its known that these hotels belong to the most expensive ones. At the other hand, the word "disappointed" might be about the fact that customers have had high expectations of the hotel, but their expectations have not been fulfilled at the end.

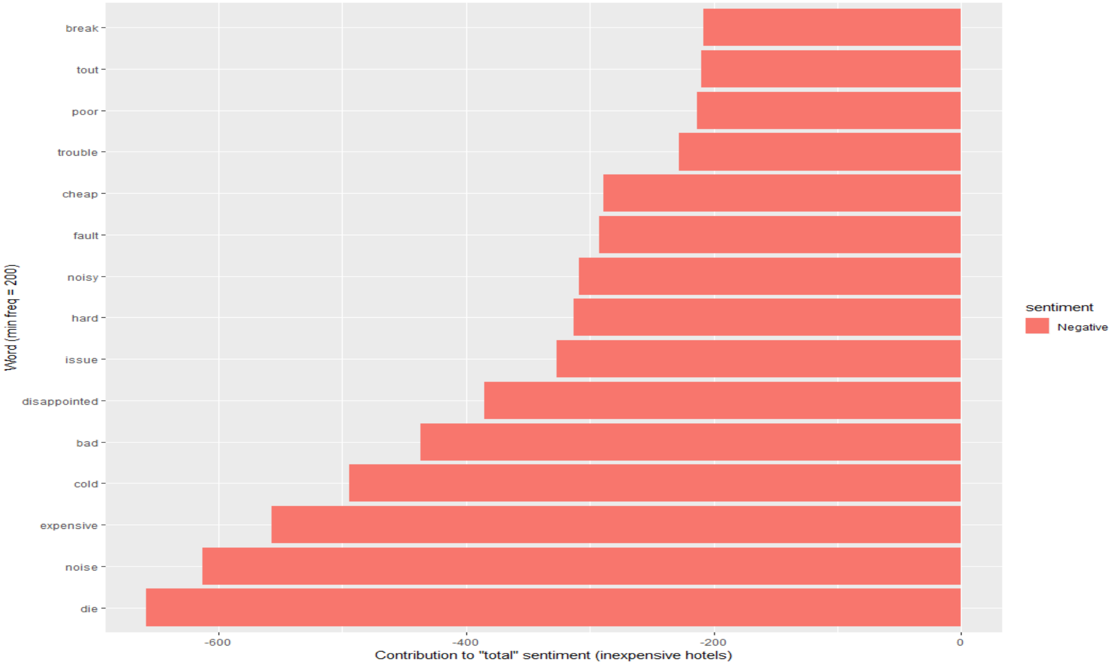

Figure 18: most occurring negative sentiments (inexpensive hotels)

According to *Figure 18*, the most occurring negative words in reviews of the inexpensive hotels are "noise" and "expensive". When having a closer look at the reviews, it became clear why customers mentioned the word expensive. In the inexpensive category, there are still some hotels that are pretty expensive compared to the others. When the customers mentioned the word expensive, it had most of the time to do with the Mondrian London at Sea Containers hotel and the Rembrandt hotel, which both belong to the less cheaper hotels. It was also the case that customers mentioned expensive, but not in a sentence in which the hotel had something to do with it. It was mainly about the fact that the city London itself is not very cheap. It became also clear that the word "die" occurs much often in reviews, which indicates a negative sentiment. However, when looking closely to the sentences in which this word occurs, it became clear that this word occurred much often in reviews of German customers. It is clear that “die” is a word that has a negative sentiment in English language, while in German it just means “the”.

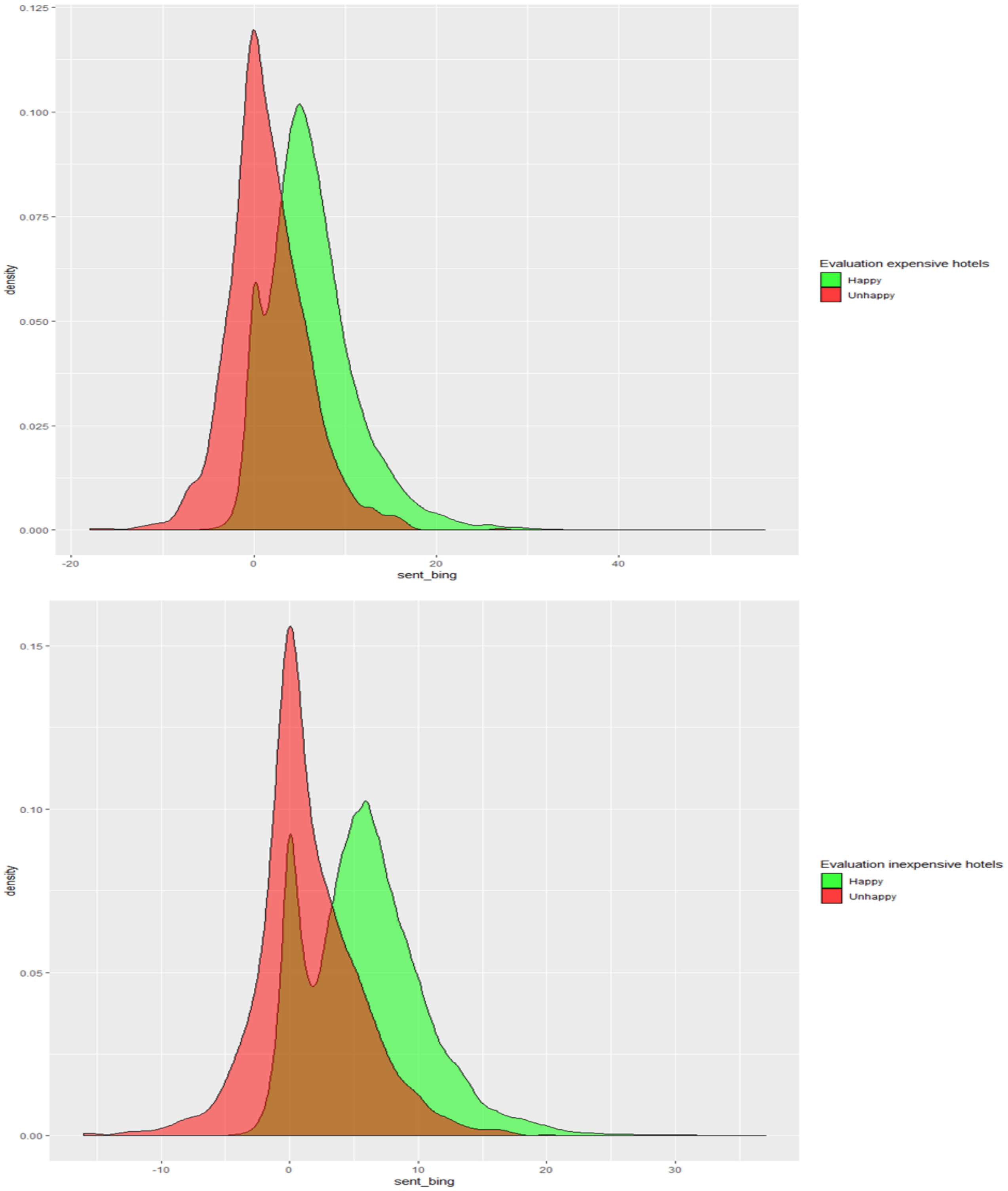

Figure 19: sentiment score distribution based on Bing method

Figure 19 shows how the sentiment scores of reviews of the happy and unhappy customers has been distributed. Happy customers are the ones who gave a rating of 4 and higher, while unhappy customers are the ones who gave a rating of 3 and lower. The scores have been determined by making use of the dictionary that has been made by Hu & Liu (2004). With these plots, it can be seen whether the sentiment scores make sense. It is expected that the customers who are happy will have a higher sentiment score than the ones who were unhappy. The figures should show that the reviews of the unhappy customers have a lower sentiment score than the ones of the happy customers. It’s clearly visible that the reviews of the unhappy customers are more distributed on the left side and have a lower sentiment score on average than the reviews of the happy customers.

The reviews of the expensive hotels have an average sentiment score of 6.38 for the happy customers and 1.49 for the unhappy customers. On the other hand, the reviews of the inexpensive hotels have an average sentiment score of 5.93 for the happy customers and a score of 1.48 for unhappy customers. On average, the happy customers of the expensive hotels leave a more positive review behind than the ones that were a happy customer of the less expensive hotels.

In order to make it possible to classify whether a customer is happy or unhappy based on its review, some features need to be selected which are needed for the model. In total, four features have been made. The first feature that will be used are the factors. There are in total 20 factors which have been obtained by performing Principal Component Analysis (PCA) on the Document Term Frequency matrix (DTM). The second feature that will be used are the emotions. In total, there will be 10 different emotions such as “anger”, “joy”, “sadness” and “disgust”. The third feature are the unigrams, which contains 50 words. These are the ones that occur most often in reviews. The last feature are the bigrams, which has in total around 30 bigrams in each category (inexpensive/expensive). A unigram is one word in a given text while bigrams are two words that follow each other up in a given text (Meng & Chu, 1999).

5.2. Predictive models (expensive hotels)

There have been made several different models based on decision trees, Random forest and Generalized linear regressions. It is important to choose the right model based on some performance indicator. In this case, the accuracy will be used to find out which model has the best performance.

Model	**Description**	**accuracy**
Glm.Allfeatures	All factors+ all words + all emotions+ all bigrams	73,52%
Glm.nodictionary	Glm.All without emotions	55,13%%
Glm.factorsonly	Only factors	53,10%
Glm.emotionsonly	Only emotions	75,63%
Glm.unigrams	Only unigrams	52,23%
Glm.bigrams	Only bigrams	52,46%
Glm.unibi	Bigrams and unigrams	54,02%
Glm.posneg	Negative and positive words	73,61%
Glm.fwbw	Forward-backward selection choses the optimal number of variables.	73,89%

Table 1: accuracy for each logistic model

The model with the highest accuracy will be considered as the best model. It can be observed from *Table 1* that the three best performing models are the ones with only emotions, only positive and negative words and the model that is based on forward backward selection. Because the differences are very small, it is needed to make use of another performance indicator, which is the AIC.

Model	**AIC**
Glm.fwbw	10341
Glm.emotionsonly	10645
Glm.posneg	11262

Table 2: model performance based on AIC

The best logistic model will be chosen based on AIC. The best model is the one with the lowest AIC-score. According to *Table 2*, the best model is the one based on forward backward selection.

Model	**Description**	**accuracy**
Glm.fwbw	Forward backward selection	73,43
DT	Decision tree technique	72,28%
RF	Random forest technique	96,14%

Table 3: accuracy score of each model

The best performing logistic model needs to be compared to the models that are based on decision tree and random forest. This will also be done based on the accuracy, which can be found in *table 4*. As can be observed, random forest performs the best amongst the other methods.

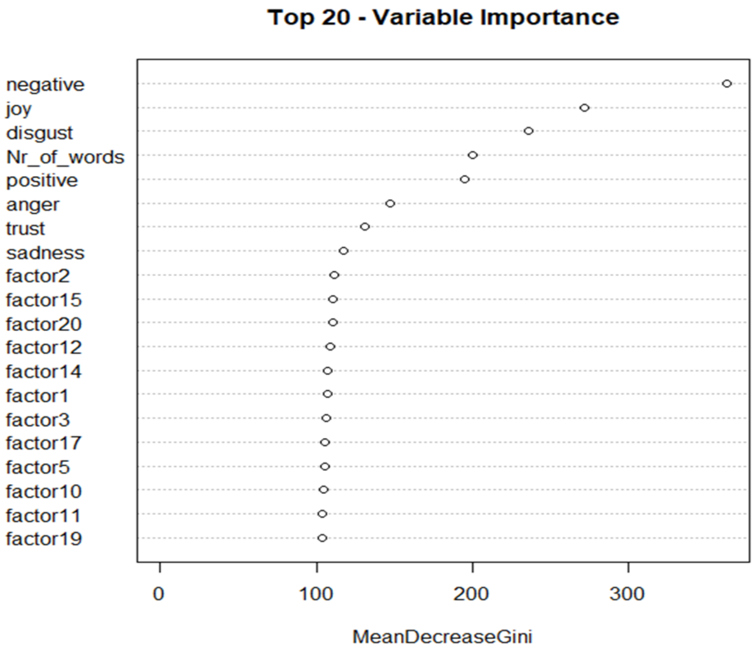

Figure 20: Variable importance

Since random forest is the most accurate model, it is important to better understand which variables are most important. The most important variables can be found by making use of mean decrease Gini (MDG). MDG is the average decrease of node impurity of a variable, which is then normalized by the total number of trees (Calle & Urrea, 2010). This measure shows how important the independent variables are when it comes to estimating the dependent variable. The higher the value of this measure for a variable, the more important this variable is. *Figure 20* shows that the most important variables are mainly the emotions and the factors. Four of the top five most important variables are emotions. The number of words occurring in a review has also a big importance. More than half of the top twenty most important variables are factors. All factors have approximately the same variable importance. Since factors have such a high importance in this model, a closer look has to be taken in order to better understand what kind of words are important for the reviews that belong to the most expensive hotels.

Something that became clear when a closer look has been taken into the factors, is that no negative words occurred. Most of these factors had a very positive sentiment. For instance, factor 6 consists of words such as “amazing”, “excellent”, “service” and “butler”. Having a closer look at the word “butler” gave insights that customers were very impressed and satisfied with the service of the butlers. Factor 14 and 1 are related to factor 6 since these factors are all about how satisfied customers are with the service that the hotels offer. Furthermore, Factor 2, 15 and 20 where about whether the hotel is located nearby a park or station. Factor 12 indicated that customers loved the fact that some hotels offered luxury spa treatments. Factor 19 has positive words occuring in it such as “amazing”, “nice”, “restaurant” and “food”. The reviews that contain these words were very positive about the restaurants, and very satsfied with the food that the restaurants offers.

5.3. Predictive models (inexpensive hotels)

Model	Description	Accuracy
Glm.Allfeatures	All factors+ all words + all emotions+ all bigrams	71,29%
Glm.nodictionary	Lm.All without emotions	54,04%
Glm.factorsonly	Only factors	52,96%
Glm.emotionsonly	Only emotions	70,70%
Glm.unigrams	Only unigrams	52,33%
Glm.bigrams	Only bigrams	47,55%
Glm.positive_negative	Negative and positive words	70,34%
Glm.unibi	Unigrams and bigrams	63,005%
Glm.fwbw	Forward-backward selection choses the optimal number of variables.	70,70%

Table 4: accuracy for each logistic model

It can be observed from *table 4* that the three best performing models are the ones with only factors, only bigrams and the model that is based on forward backward selection. These models have all the same performance, therefore it is needed to make use of another performance indicator in order to find the better one amongst these models.

Again, it is supposed to choose the model with the lowest AIC-score. According to *table 5*, the best performing model is the one with all variables being considered.

Model	Description	AIC
Glm.all	All features	8238
Glm.fwbw	Forward backward selection	8284
Glm.onlyemotions	Only emotions	8302

Table 5: model performance based on AIC

Again, the best performing logistic model needs to be compared to the models that are based on decision tree and random forest. This will be done based on the accuracy, which can be found in *table 6*. As can be observed, the model that is based on random forest has an accuracy of 88,77% which is the best performing model amongst the others.

Model	**Description**	**Accuracy**
Glm.all	Generalized linear model with all variables	73,08%
DT	Decision tree	70,37%
RF	Random forest technique determines the features	88,77%

Table 6: accuracy of each model

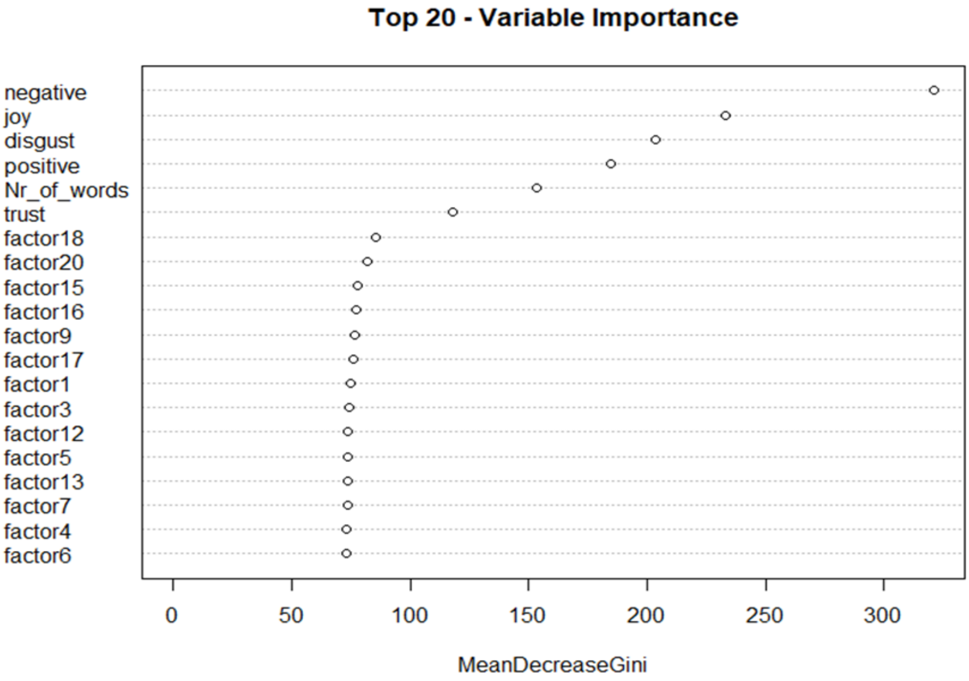

Figure 21: Variable importance

The random forest technique has the same important variables as the previously discussed one, which are the emotions and factors. It has the same 5 variables in the top 5. Since factors are also of big importance for this model, it is important to have a closer look at the important ones to better understand what is most important in these reviews.

Factor 18 consists of words such as “clean”, “friendly”, “comfort” and “staff”. Customers that use these words usually have not more to say about the room than that it is clean and comfortable. Next to that, they also indicate also that the staff is friendly. This might be because customers could not find other aspects of the hotel to give a positive review about. Factor 20 is about the breakfast that hotels offer. Factor 15 consists of words such as “bar”, “drink”, “cocktail” and “view”. These words are all related to each other because customers had a positive review about the bar of the Mondrian hotel. This hotel has a bar on the rooftop with a nice view. Factor 19 consists of words such as “station”, “walk”, “London”, “Paddington” and “excellent”. These words all related to each other because many customers had a review about the Rhodes hotel which is nearby the Paddington station. Factor 13 is related to factor 19, because factor 13 is more about the location of the hotel and whether there is a museum or a station nearby. Factor 3 consists of words such as “price” and “location”. Zooming in on these words makes clear that these words occur in reviews in which customers indicate that the price of the hotel rooms are reasonable based on the location.

6. Conclusion

The advent of the World Wide Web has become very important for both customers as organizations. Nowadays, many customers write about the value they give to a product, which helps other potential customers with their buying decision. With this information, a potential customer can have an indirect opinion about a product without the direct knowledge they otherwise would have when they would have bought the product themselves. The information that has been provided by customers on the internet, helps organizations to better understand how they value their product/service. Therefore, analyzing this information has become very popular in the recent years. Organizations make use of sentiment analysis to better understand what the goods and the bads are of their company.

This research has set a focus on the hotel industry. There has been done a lot of research on this subject, but not much has been done when it comes to comparing the sentiments between categories of hotels. This research focusses on finding out what different sentiments can be found in reviews of the ten most expensive hotels and least expensive hotels in London. Two main questions have been stated for this research:

- **How well can machine learning techniques predict whether a customer is happy/unhappy based on its review?**
- **Which features are most important for a good/bad rating, and how are these features related to the sentiments of the customer reviews of the expensive hotels and inexpensive hotels?**

In chapter 5, it became clear that machine learning techniques can predict very well how happy/unhappy a customer is based on their review. The accuracy for the reviews of the expensive hotels ranged in between **72,28%** and **96,14%** which is very low. The model based on Random Forest seemed to perform the best amongst the others. The accuracy for the hotel reviews of the inexpensive hotels ranged in between **70,37%** and **88,77%**. In both categories, the models that are based on Random Forest performed the best and were most accurate.

Based on these models, it also becomes clear which features where most important when it comes to predicting the happiness of customers. Emotions such as negative, joy, disgust and positive had the highest variable importance, followed by several factors. Even te number of words of a review is very important for the model. Something that became clear when a closer look has been taken into the factors that belong to the expensive hotels, is that no negative words occurred. The most important factor contained words such as “amazing”, “excellent”, “service” and “butler”. Having a closer look at the word “butler” gave insights that customers were very impressed by the service and satisfied with the butlers of the hotels. The customers were also very positive about the restaurants, and very satsfied with the food that the restaurants offers. Customers loved the fact that some hotels offered luxury spa treatments.

The cheaper hotels had also the same emotions that were important for the models. The main difference is that these emotions were more important for this category than for the less cheaper hotels. The two most important factors were about a bar on the rooftop with a nice view and stations that are located nearby the hotels. Customers also used often words such as “clean”, “friendly”, “comfort” and “staff” which usually means that they could not find a lot of aspects to give a positive review about. Other factors also indicated that customers talk much often about museums in london. Furthermore, customers also talk about the price of the hotelrooms. They indicate often that the price seems reasonable based on the location.

Based on the reviews of the customers and the models that are based on the reviews of both the expensive hotels and inexpensive hotels, it can be concluded that the reviews of the expensive hotels have more positive words in it. Even the words in these reviews had a much stronger positive weight. Customers of this category were happier on average (*Figure 19*). Next to that, these customers talk more about the aspects of the hotel such as how good the food of the restaurant is, how satisfied they are with the luxurious spa of the hotel and the good service that is provided by the butler. This all fits the expectation of the service and facilities that an expensive hotel should provide to its customer. Customers of the cheaper hotels talk more about the non-hotel aspects such as the walk between the station and hotel. These customers also talk about how friendly the staff is and how clean the rooms are. They also mention that the ratio of the price and quality seems reasonable.

6.1. Limitations

According to Saifee & Jay (2013), there are nine challenges that come along with sentiment analysis:

- ***Coreference:*** sentiment analysis has problems when identifying what a noun phrase refers to. Example: *“Remember the burger we ate at the McDonald’s previous week?* ***That*** *was delicious.”.* It can be challenging for sentiment analysis approaches to identify what “that” refers to.

- ***Time change:*** the reviews might have problems when it comes to the year in which a review has been published. It might be possible that a reviewer had a positive review about a product in 2012, let’s say iPhone 5, but this might not be the case when the reviewer is asked to give his opinion about the iPhone 5 today. People might change their opinion or judgement towards a product over time.

- ***Sarcasm:*** sarcastic sentences in a review might lead to a big challenge, because the review might have the opposite meaning of what has been written. Example: *"The pasta was pretty much better than nothing at all".* The sentence will be classified as positive, because the sentence consists of more positive than negative words. Humans know directly that this is a sarcastic sentence, but machines have difficulty to understand it.

- ***World knowledge:*** machine learning techniques don't have the same world knowledge as human beings.

- ***Object:*** some sentences don't mention all the objects which contain an opinion about it. Example: "You should better read the book rather than going to the cinema". This sentence consists of a positive opinion about the book, but also a negative opinion about the movie which is not directly mentioned.

- ***Synonyms:*** some words in reviews look different, but have the same meaning. As an example we consider the words "screen" and "display". These words are written differently, but they are synonyms. These words should be combined, as if it is one word, in order to make better classifications according to Saifee & Jay (2013).

- ***Plot twist:*** some sentences can be classified as positive, while they should be classified as negative. For instance: "The screen is great, the sound is pretty good, but I'm not a fan of this phone". The overall classification will be positive because the screen and the sound of the phone where judged positive. But when you look at the overall judgement of the phone itself, it should be classified as negative.

- ***Negation:*** sentiment analysis has some challenge when it comes to negations. Let's consider the next two sentences: "I want more" & "I want no more". The whole sentence changes from positive to negative just because of one word. It is important to find out which of these negations are most important when performing sentiment analysis. Some words are more important than others and should therefore be taken into consideration, while others can be left out of the analysis.

- ***Fake reviews:*** the last challenge of sentiment analysis are the fake reviews. Fake reviews can lead to inaccurate conclusions because these reviews where not the real judgement of the customers. Some people might have some bad intentions when it comes to a product or a service of a company, probably because this company is their competitor and sells the same kind of products. There have been many of these cases that have been reported in the news in which a company wrote fake reviews about its competitor (Liu B. , 2012). Therefore, detecting fake reviews is of big importance.

6.2. Future research

This research mainly focused on reviews of customers of hotels in London. London is a big city in which many tourists of all over the world come to see the iconic busses and big buildings such as the big ben. Therefore, many reviews are based on other languages than English which might have an influence on the results. Some words in other languages might have been classified as a word with a negative character, while this has not been the case. For instance, the word “die” is negative in English while in German is just means “the”. Therefore, we recommend for future research the use of a technique that helps to analyze multi-language reviews.

One of the aforementioned limitations of reviews is that there are many fake reviews on the internet. Fake reviews occur due to the fact that an organization hires people to posts fake reviews on one website where they are allowed to do so because this website does not verify a customer whereas the other website does. Elmurgni & Gherbi try to detect fake reviews by making use of several machine learning techniques such as naïve bayes and Support Vector Machine (Elmurngi & Gherbi, 2017). It might be interesting for future research to detect fake reviews by making use of machine learning techniques.

Opinion changes over time, which is one limitation of this research. The last recommendation for future research is to take the date into consideration and divide the dataset based on different years or months within a year. The date of a review could give some new insights because one might find out how opinions change over time. Cheng et al. proposed a method in order to detect how opinions change over time (Cheng, Ke, & Shiue, 2011).

Reference

Agrafiotis, D. K., Rassokhin, D. N., & Lobanov, V. S. (2001). Multidimensional scaling and visualization of large molecular similarity tables. *Journal of Computational Chemistry*, 488-500.

Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect based sentiment oriented summarization of hotel reviews. *Procedia computer science*, 563-571.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 1545-1588.

Anderson, C. (2012). The Impact of Social Media on Lodging performance. *Cornell hospitality report*, 1-12.

Bai, X. (2011). Predicting consumer sentiments from online text. *Decision Support Systems*, 732-742.

Bjørkelund, E., Burnett, T. H., & Nørvåg, K. (2012). A study of opinion mining and visualization of hotel reviews. *In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, 229-238.

Boudad, N., Faizi, R., Thami, R. O., & Chiheb, R. (2017). Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal.*

Breiman, L. (1996). Bagging predictors. Machine learning. 123-140.

Breiman, L. (2001). Random forests. Machine learning. 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. CRC Press.

Calle, M. L., & Urrea, V. (2010). Letter to the editor: stability of random forest importance measures. Briefings in bioinformatics. 86-89.

Chen, C. C., & Tseng, Y. D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, 755-768.

Chen, M. H., Ibrahim, J. G., & Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society: Series B*, 223-242.

Cheng, L. C., Ke, Z. H., & Shiue, B. M. (2011). Detecting changes of opinion from customer reviews. In 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery . *(FSKD) (Vol. 3, pp. 1798-1802). IEEE.*

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international conference on World Wide Web* , 519-528.

De Albornoz, J. C., Plaza, L., Gervás, P., & Díaz, A. (2011). A joint model of feature mining and sentiment analysis for product review rating. In European conference on information retrieval. *Springer, Berlin, Heidelberg.*, 55-66.

De Leeuw, J., & Mair, P. (2011). Multidimensional scaling using majorization: SMACOF in R.

Elmurngi, E., & Gherbi, A. (2017). Detecting fake reviews through sentiment analysis using machine learning techniques. *IARIA/data analytics, 65-72.*

Engel, J. F., Blackwell, R. D., & al., e. (1995). Consumer Behavior. *Forth Worth, Dryden Press*.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*.

Fielding, A., & O'Muircheartaigh, C. A. (1977). Binary segmentation in survey analysis with particular reference to AID. (The Statistician). *Journal of the Royal Statistical Society*, 17-28.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics. 179-188.

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. *In ENTER* , 460-470.

Gupta, S. K., Phung, D., Adams, B., & Venkatesh, S. (2013). Regularized nonnegative shared subspace learning. *Data mining and knowledge discovery*, 57-97.

Hagenau, M., Liebmann, M., Hedwig, M., & Neumann, D. (2012). Automated news reading: Stock price prediction based on financial news using context-specific features. In 2012 45th Hawaii International Conference on System Sciences (pp. 1040-1049). . *IEEE*.

Ho, T. K. (1995). Random decision forests. *In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.*

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. *ACM*, 168-177.

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision support systems*, 674-684.

Hu, Y. H., & Chen, K. (2016). Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. . *International Journal of Information Management*, 929-944.

Huang, S., Ward, M. O., & Rundensteiner, E. A. (2005). Exploration of dimensionality reduction for text visualization. In Coordinated and Multiple Views in Exploratory Visualization . *IEEE.*, 63-74.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to statistical learning.* Springer Texts in Statistics.

Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 6000-6010.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression. *New York: Springer-Verlag*.

Koch, L. (2019). *Two-Thirds of Shoppers Check Phones In-Store for Product Information, Skipping Store Associates.* Retrieved from https://www.emarketer.com/content/two-thirds-of-internet-users-check-phones-in-store-for-product-information-skipping-store-associates

Kulmala, M., Mesiranta, N., & Tuominen, P. (2013). Organic and amplified eWOM in consumer fashion blogs. *Journal of Fashion Marketing and Management: An International Journal*, 20-37.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, 458-468.

Liu, A. C. (2004). The effect of oversampling and undersampling on classifying imbalanced text datasets. . *University of Texas at Austin.*

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167. 1-167.

Liu, Y., Huang, X., An, A., & Yu, X. (2007). ARSA: a sentiment-aware model for predicting sales performance using blogs. *Liu, Y., Huang, X., An, A., & Yu, X. (2007, July). ARSA: a sentiment-aware model for predicting sales performance using blogs. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 6*, 607-614.

Loh, W. Y. (2011). Classification and regression trees . *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14-23.

Martin, L., & Pu, P. (2014). Prediction of helpful reviews using emotions extraction. *In Twenty-Eighth AAAI conference on artificial intelligence*.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 1093-1113.

Meng, F., & Chu, W. W. (1999). Database query formation from natural language using semantic modeling and statistical keyword meaning disambiguation. . *Computer Science Department. University of California.*

Messenger, R., & Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. . *Journal of the American statistical association*, 768-772.

Muschelli, J., Betz, J., & Varadhan, R. (2014). Binomial regression in R. In Handbook of Statistics . *Elsevier*, 257-308.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *In Proceedings of the 2nd international conference on Knowledge capture*, 70-77.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *In Proceedings of the 2nd international conference on Knowledge capture*, 70-77.

Neto, J. L., Santos, A. D., Kaestner, C. A., Alexandre, N., & Santos, D. (2000). Document clustering and text summarization.

O'Mahony, M. P., & Smyth, B. (2009). Learning to recommend helpful hotel reviews. . *In Proceedings of the third ACM conference on Recommender systems*, 305-308.

Pai, M. Y., Chu, H. C., Wang, S. C., & Chen, Y. M. (2013). Electronic word of mouth analysis for service experience. *Expert Systems with Applications*, 1993-2006.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. *Association for Computational Linguistics*, 79-86.

Park, D. H., Lee, J., & Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International journal of electronic commerce*, 125-148.

Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research*, 67-83.

Pearl, & Raymond. (1922). The biology of death . *Philadelphia: Lippincott*.

Phillips, P., Barnes, S., Zigan, K., & Schegg, R. (2017). Understanding the impact of online reviews on hotel performance: an empirical analysis. *Journal of Travel Research*, 235-249.

Quinlan., J. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Reibstein, D. J. (2002). What attracts customers to online stores, and what keeps them coming back? *Journal of the academy of Marketing Science*.

Saifee, V., & Jay, T. (2013). "Applications and Challenges for Sentiment Analysis: A Survey". *International Journal of Engineering Research & Technology (IJERT)*.

Solka, J. L. (2008). Text data mining: theory and methods. Statistics Surveys. *Statistics Surveys*, 94-112.

Steinberger, J., Lenkova, P., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., & Vázquez, S. (2011). Creating sentiment dictionaries via triangulation. In Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis. *Steinberger, J., Lenkova, P., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., ... & Vázquez, S. (2011, June). Creating sentiment dictionaries via triangulation. In Proceedings of the 2nd workshop on computational approaches to subjectivity and s*, 28-36.

Tang, L. (2017). Mine your customers or mine your business: the moderating role of culture in online word-of-mouth reviews. *Journal of International Marketing*, 88-110.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics. *Association for Computational Linguistics.*, 417-424.

Van Eck, N. J., Waltman, L., Dekker, R., & van den Berg, J. (2010). Van Eck, N. J., Waltman, L., Dekker, R., & van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology*, 2405-2416.

Verhulst, P. F. (1838). Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys*, 113-126.

Verhulst, P. F. (1845). Mathematical researches into the law of population growth increase. Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles. 1-42.

Verhulst, P. F. (1847). Deuxième mémoire sur la loi d'accroissement de la population. Mémoires de l'académie royale des sciences, des lettres et des beaux-arts de Belgique, 20, 1-32. 1-32.

Wilson, & Worcester. (1943). The determination of L.D.50 and its sampling error in bio-assay. *Proceedings of the national academy of sciences*.

Ye, Q., Law, R., & Gu., B. (2009). The impact of online user reviews on hotel room sales. *international journal of hospitality management*, 180-82.

Ye, Q., Law, R., Gu, B., & Chen., W. (2011). the influence of user-generatd content on traveler behavior: an empirical investigation on the effects of E-word-of-mouth to hotel online bookings. *computers in human behavior*, 634-39.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 2758-2765.