MEF UNIVERSITY

BIG DATA ANALYTICS ON HOTEL REVIEWS

Capstone Project

Burcu Demir

ISTANBUL, 2021

MEF UNIVERSITY

BIG DATA ANALYTICS ON HOTEL REVIEWS

Capstone Project

Burcu Demir

Advisor: Prof. Dr. Özgür Özlük

ISTANBUL, 2021

MEF UNIVERSITY

Name of the project: BIG DATA ANALYTICS ON HOTEL REVIEWS
Name/Last Name of the Student: Burcu Demir
Date of Thesis Defense: 24/01/2021

I hereby state that the graduation project prepared by Burcu Demir has been completed under my supervision. I accept this work as a "Graduation Project".

24/01/2021

Prof. Dr. Özgür ÖZLÜK

I hereby state that I have examined this graduation project by Burcu Demir which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

24/01/2021

Prof. Dr. Özgür Özlük

Director of Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Dr. Özgür ÖZLÜK	.....

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Burcu Demir	24/01/2021	

EXECUTIVE SUMMARY

BIG DATA ANALYTICS ON HOTEL REVIEWS

Burcu Demir

Advisor: Prof. Dr. Özgür ÖZLÜK

JANUARY, 2021, 31 Pages

This analysis aims to get a regression model of the reviews and the score by the guests to observe the effects of the content of the reviews on scores. The content of the reviews is also suitable for a sentiment analysis. These analyses are useful indicators of the hotel sector to catch the market direction positively.

In this analysis, clustering hotel-based reviews and customer segmentation based on the reviews will be the key point. Nationality of the guests will be helpful information of the guests to get them into the segmentation pool.

The guest who wants to stay in the best hotel in Europe while their trip could choose the best hotel. They can conclude that selection by meeting their needs.

Key Words: Regression Model, Sentiment analysis, clustering, segmentation

ÖZET

OTEL YORUMLARI ÜZERİNDEN BÜYÜK VERİ ANALİTİĞİ

Burcu Demir

Proje Danışmanı: Prof. Dr. Özgür ÖZLÜK

OCAK, 2021, 31 Sayfa

Bu analizde, otel müşterilerinin yorum içeriklerinin puanlama ve değerlendirme üzerindeki etkilerinin gözlemlenebildiği bir regresyon modeli ortaya çıkarmaktır. Yorum içerikleri aynı zamanda duyarlılık analizi için de uygundur. Bu analizler, otel sektöründekilerin piyasayı olumlu yönde yakalamasını sağlayacak sonuçlar sunar.

Bu analizde, otel müşterilerinin yorumlarının kümelenmesi ve müşterilerin buna göre gruplandırılması en önemli nokta olacaktır. Konukların uyrukları otel misafirlerinin gruplanmasında kullanılacak faydalı bir bilgidir.

Konuklar Avrupaya yapacakları seyahat öncesinde kalacakları otelin seçimini yaparken en iyi olanı göz önünde bulunduracaklar. En iyi ve seyahat sırasında ihtiyaçları olanları karşılayabilecekleri otelin seçimini kolayca yapabilecekler.

Anahtar Kelimeler: Regresyon Modeli, Duyarlılık analizi, kümeleme, gruplama

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.INTRODUCTION

Big data analysis serves to obtain meaningful results by analyzing large volumes of data. This big data is collected from a wide variety of sources such as social networks or sales transactions. The purpose of this project is analyzing data to reveal patterns and connections in this data to provide valuable information about the users who created them.

With the use of technology in daily life, the extent of networking through social media has become more important. So that social media analytics could be used to get some clues about the interactions between people. Social Media Analysis (SNA) is interested in the relations between social groups to get helpful information. The patterns had been consisting of these who-questions. (Elgendy, Nada & Elragal, Ahmed.,2014, pg.220)

Tourism including the hotel industry consists of approximately 10% of the world economy and big data collected at booking systems sites could help to organize a competitive strategy for cost-effective hotel management. Nowadays, consumers use websites to do their daily life actions. They give importance to the user experience and tend to criticize every dissatisfaction through their consuming journey. With the growth of tourism and the development of BI applications the hotels faced a need to study a large amount of data to make quicker actions in the sector. The hotels want to invest in BI technologies in the two decades because of its power on decisions. Both the hotels actions to satisfy the pleasures and attract customers are supported by the BI technologies. Besides that, the finance is also governed by the data which is processed by the BI technologies. (Rodrigues, Sousa and Brochado, 2020)

Consumer reviews are useful particularly such as hotels, airlines, and healthcare. The reviews have characteristics that can be interpreted as good or bad service. The nominative opinions of the customers turn into quantifiable contents to comparison. (Mankad, Shawn & Han et al.,2016, pg.126). Before booking a hotel for their trip, the reviews give some clues to people in choosing the right hotel. However, they trust the bad reviews then the good ones. Good results are more trustworthy while the number of reviews is higher. (Gavilan, et al.,2017)

In this research, the guest's attitudes will be analyzed by sentiment analysis and methods with various classification. Text preprocessing and vector representation will be explained by Word2vec and Bag of words. Various experiments will be carried out with the help of classification algorithms on this data set. In the conclusion part, the results that were obtained will be discussed.

1.1. Sentiment Analysis: Literature Survey

Sentiments are interpreted as opinions or thoughts of people about the environment. These thoughts are individual views. These subjective sentiments can be analyzed by machine learning techniques such as supervised and unsupervised. The aim of the sentiment analysis is converting the people's thoughts into meaningful global interpretations. While the negative and positive reviews are categorized, both the seller and buyer get advantages of these categorization. The general process of the sentiment analysis starts with the collection of the data, in this project gain the data from booking.com website. The data content is important for the company, in this research the hotels want to learn the satisfaction of their guests. So that the review about their hotels is extracted with keywords. (S. Shayaa et al., 2018)

Sentiment analysis has been done especially the source of text, images, or audios in the machine learning process. The analysis had been concluded as positive or negative classifications. As it was explained below, the sectors where the sentiment analysis is used are academic and commercial companies. (Boiy, E., Moens.,2009, p. 526)

In this project, the hotels could use the negative and positive impacts of the hotel reviews on their sales strategy. Nowadays, social media has a considerably huge effect on society choices,

so that these reviews have been important for the hotels. The data had been retrieved from Booking.com, the user feedback. In preprocessing processing of the sentiment analysis, the texts are taken and segmented into categories. There are different types of feature vector selection, unigrams, stems, negation, or discourse features. (Boiy, E., Moens.,2009, p. 526)

It could be used unigrams that had been represented that the words or tokens that are taken from sentences. In hotel reviews, the words could be “good” or “bad”. Because of the guests’ attitudes, if a visitor had not been unpleasant of the hotel, the customer tends to write the word “bad” in their hotel review on the website. In this sentiment analysis, ‘nothing’ and ‘everything’ are used to classify the negative and positive reviews from the guest’s reviews.

1.2. Bag of Words Method: Literature Survey

It is known that Bag of Words was used by e-mail service providers such as Yahoo and Hotmail. The algorithm structure used in the bag of words model is a useful and simple method for detecting many spam and scam transactions. Currently, the bag of words model is used with various arrangements in advanced systems.

Bag of Words is the other way of representative aspect of the text in sentiment analysis. Binary vectors have been retrieved from documents to representation, followed by the calculation in frequency of these words in documents. A feature vector is created Nx1, N represents the most frequent word in these documents. This vector will be trained in the machine learning process.

An example of the bag of word process is the sentence 1 "The best hotel in USA" and sentence 2 "The Aston hotel is very dirty". Then a built dictionary contains words ("The", "best", "hotel", "in", "USA", "Aston", "is", "very”, “dirty"). The above sentence has 9 different words so that each sentence is described as follows, sentence 1 [1,1,1,1,1,0,0,0,0] and sentence 2 [1,0,0,0,0,1, 1,1,1]. It is created Document Vectors, the elements in this document are words or phrases in the vocabulary. (Farisi, Sibaroni and Al Faraby, 2019). Each entry in the lists indicates the number of corresponding entries in the list. This vector does not preserve the order of representational words in source sentences. However, term frequencies are not necessarily the best text representation, common words such as ‘one’, ‘and’ always have the highest frequency of terms

in the text. So that a larger number of word frequencies does not mean that the corresponding word is more important.

The bag of words model is an unordered document representation, it cannot reveal the verb of the sentence. So that n-gram is used to split the sentences into words by n set orderly. In this example, "The movie is amazing" split into “The movie”, “movie is”, “is amazing”. As a result, it can be easily found that the amazing is the verb of this sentence. (Bofang Li, et al., 2016)

In the Bag of Words model explanation, in these two examples, the bag of words model is started to be operated by frequency. To generate the feature vector, it could be used the TF-IDF weighting scheme or Word2Vec.

1.3. Word2vec Method: Literature Survey

Embedding approach word2vec is one of the methods that represents the words into numbers as vectors. Natural Language Process through deep learning uses word2vec to develop a word vector that consists of text. It trains text followed by vector representations of them. Clustering the vectors which are like each other along a vector space to numbering them by their similarities. However, prediction is built through this model which is formed by the word2vec method. (H. Yousaf et al., 2020)

The sentiment analysis towards hotel reviews in English language, the words labelled as positive and negative by word2vec model. The hotel reviews have various aspects such as service, location, or room. These aspects will be useful in positive and negative labelling in training the data. There are two types of architecture in word2vec model; CBOW method predicts the current word from the surrounding context words; on the contrary, continuous skip-gram architecture uses the current word to predict surrounding words. Both use the vector to semantic coding from words. (Aydogan, Murat & Karci, Ali., 2019)

Meaning relations between words cause clusters of words that belong to word vectors according to their meaning relationships. Vectors representing the word correctly, provide logical results to be obtained. Semantic results can be obtained from the cosine similarities of the new vector that is obtained by adding and subtracting vectors with Word2vec. To give an example of

semantic results that can be produced by arithmetic operations; king – man + women result in queen.

1.4. Term Frequency-Inverse Document: Literature Survey

The Term Frequency-Inverse Document Frequency is another way to represent the sentences based on words' context. It provides a statistical labelling to a word by the importance of them in sentences or documents. The frequency of the visibility of a word is the aspect of these calculations. However, this method could be having some limitations as it could not detect semantic meaning from words, so it could be supported by naïve bayes. (Qaiser, Shahzad & Ali, Ramsha., 2018)

A specific keyword has high frequency in a document while a particular keyword has high frequency in a document and documents containing the keyword have low frequency among all documents. Consequently, using the TF-IDF calculated by Equation below. (Kim, SW., Gil, JM., 2019)

TF = Number of Word in the Related Document

IDF = log (Total Amount of Documents/ Number of Documents a Word Occurs)

TF-IDF = TF*IDF

TF-IDF is a statistical criterion designed to indicate how important a word is in a document space. Term Frequency is the frequency of a term in the document. It indicates how many times the word is used in the document. Inverse Document Frequency is the measure of how much information the word contains. It indicates whether the word is common or rare in all documents. The product of the Term Frequency value and the IDF value gives the value of this term in the document. It could be understood which subject the document is predominantly about, and how heavily the term that is searched for in the search process is mentioned in the document. The fields that the classification method TF-IDF is used are search engines on websites, text summarization or removing the stop words.

2. ABOUT THE DATA

In this research, the objective of the analysis with the use of a leading hotel website, booking.com obtained guests reviews in Europe. An engaging feature of the reviews in Europe is that it contains different types of guests and provides variety in the data at a level of global.

The review date is between the years 2015 and 2017. Average score is also between the range of 0-10. In the columns both the native and positive reviews, all guests do not fill both types of feedback to the hotel. For example, a guest can give positive feedback and 'No Negative' or vice versa. Some guests have both good and bad experiences about the hotel by expressing their thoughts using both positive and negative fields.

Reviewer score is between the range of 1-10. On the other hand, reviewers give some tags to express their thoughts on why the hotel should be chosen for different purposes, such as leisure or business trip.

2.1. Features

This dataset contains 515K customer reviews and scoring of 1493 luxury hotels in Europe, detailed below.

- Hotel Address: Address of hotel
- Review Date: Date when reviewer posted.
- Average Score: Average Score of the hotel based on last year.
- Hotel Name: Name of Hotel
- Reviewer Nationality: Nationality of Reviewer
- Negative Review: Negative Review the reviewer gave to the hotel, if the reviewer has no negative review this column will be "No Negative."
- Review Total Negative Word Counts: Total number of words in the negative review.
- Positive Review: Positive Review the reviewer gave to the hotel, if the reviewer has no negative review this column will be "No Positive."

- Review Total Negative Word Counts: Total number of words in the positive review.
- Reviewer Score: Score the reviewer has given to the hotel.
- Total Number of Reviews Reviewer Has Given: Number of Reviews the reviewers have given in the past.
- Total Number of Reviews: Total number of valid reviews the hotel has
- Tags: Tags reviewer gave the hotel. ex. ' Leisure trip ', ' Family with young children '.
- Days since review: Duration between the review date and taken date of the data.
- Additional Number of Scoring: Some guests who just made a scoring rather than a review. This number indicates how many valid scores without review of the hotel.
- Lat: Latitude of the hotel
- Lng: Longitude of the hotel

2.2. Exploratory Data Analysis

Observed that 526 reviews are duplicate, and we removed them. After removing Duplicates, data contains 17 features and 515.212 data points.

It could be obviously seen that there are some missing values in Latitude and Longitude columns after checking the missing values from the data. Total number of the missing values in the dataset is 3.268 classified by hotel name.17 different hotels have not been known the location. The higher number of missing values is belonging to Fleming s Selection Hotel Wien City aside by 658 rows. After removing null values from the data, the data became the proper data format for the data analysis.

Outliers are values that deviate excessively from other observations or samples on the data. All data are separated by simple differences, but generally have similar values. But there are some inputs that are very conspicuous and can affect all other data entered. The features of the data are described by their mean, standard deviation, and quartile ranges. Standard deviation represents the spread of a data set. If the standard deviation is small, the data are scattered close to the mean.

Conversely, if the standard deviation is large, the data are scattered far from the mean. Average Score and Reviewer Score has the minimum variance by 0.54 and 1.63. Scores are generally out spread 8.3 points. The score data is homogenous. The total number of reviews has the highest standard deviation with 2323. Total number of valid reviews the hotel has on booking.com has ranged extensively. While a hotel has the total number of reviews as 16.500, the other one has 2. 500.However, the model is based on the content of the reviews and the sentences, so that the outliers do not affect the analysis. If the model is based on the hotel names and the reviews that they have done, the outliers might affect the result.

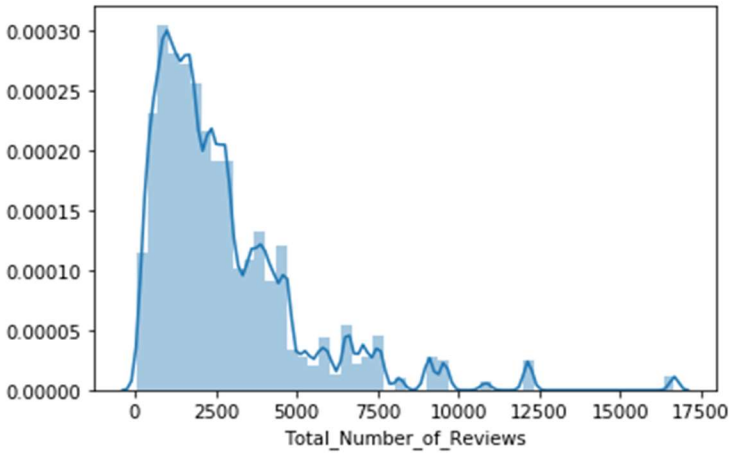

Figure 1: The total number of reviews column values to the probability per unit

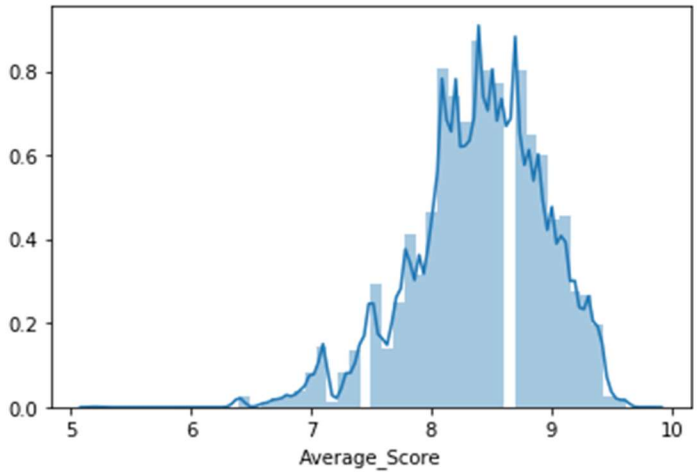

Figure 2: The total average Score column values to the probability per unit

2.2.1. Hotel Name Analysis

The data contains 1.475 different hotel reviews, Britannia International Hotel Canary Wharf is the highest number of reviews by 4.789. It could be concluded as this hotel is the most known hotel in Europe. It is placed in London; it could be interpreted as London is the most popular city in Europe to make a trip. However, it could not assume that this hotel is the best for a trip, unless analyzing the context of the reviews. The top 20 reviewed Hotels have been described as below.

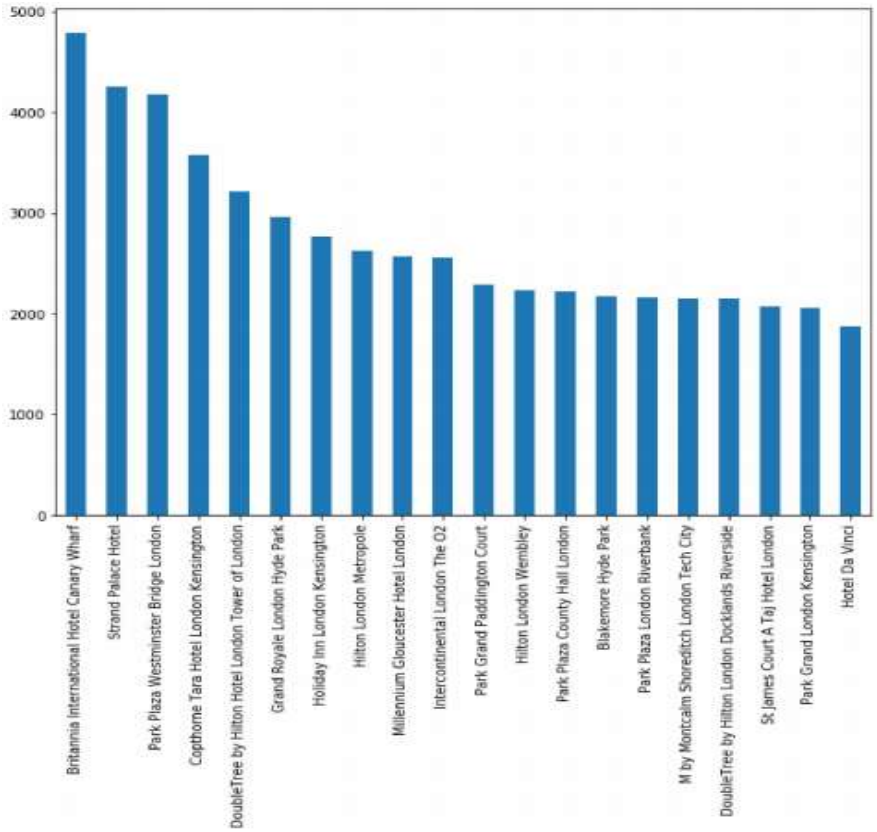

Figure 3: The count of reviews with respect to Hotel Name Analysis

2.2.2. Average Score Analysis

Hotels' average score lies in the range of 8.0 and 9.1 range as seen below. The graph could not be described as normal distribution, the accumulation is on the higher scores. It could be interpreted as the great majority of the reviewers gave scores the range between 8 to 9. Most of them slightly above the average score point in normal distribution however, not the higher score. It is acceptable for an evaluation of a service.

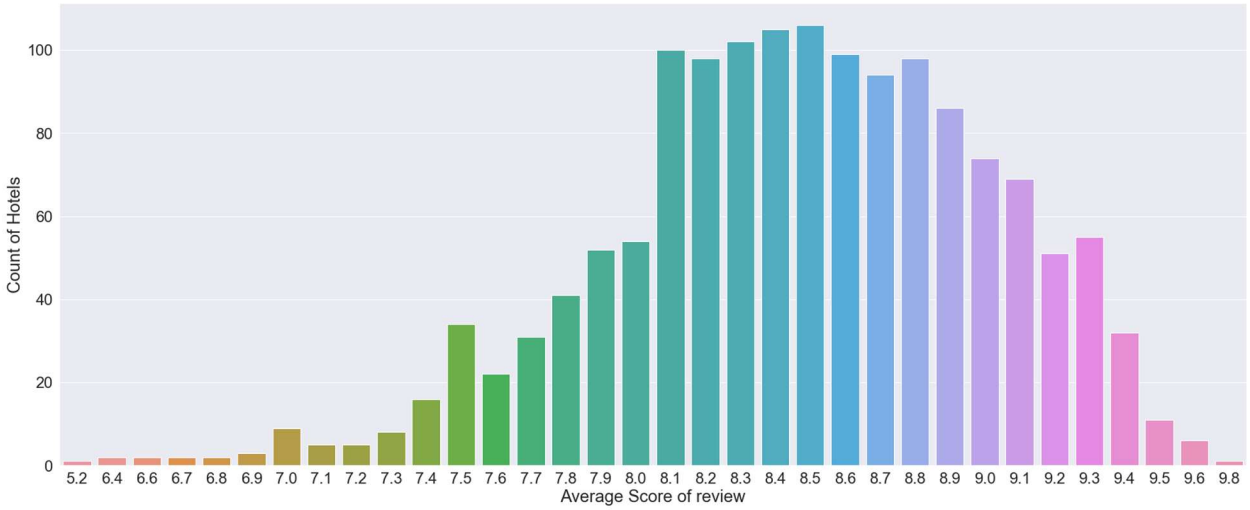

Figure 4: The average scores given by guests according to hotels through normal distribution.

Nationality of the visitors is analyzed by their given scores to Hotels as below, the distribution is ranged between 5.2 and 9.8. It cannot be interpreted as a normal distribution between the scores and the nationality of the reviewers.

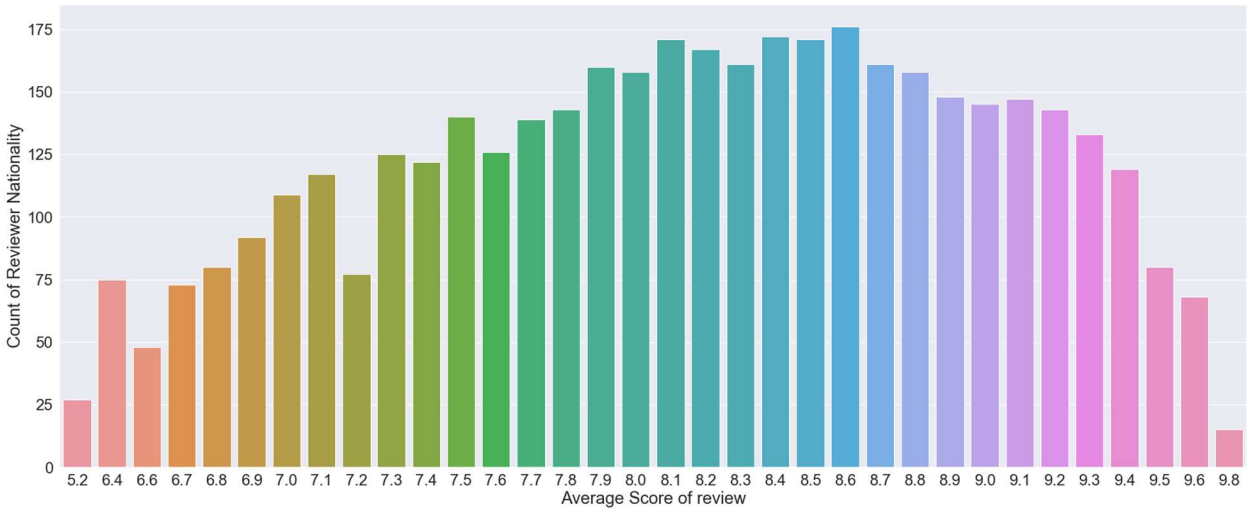

Figure 5: The average scores given by guests grouping in their nationality through normal distribution.

2.2.3. Review Date Analysis

The data contains reviews of hotels from 8th month in 2015 to 8th month in 2017. Summer seasons have got the highest count of reviews. Most Reviews are given in August beyond all over the years by approximately 60K reviews. Most of the reviews ranged from 7th to 9th month especially in 2016. This period of the year is known as the touristic season all over the word, so after their trips tourists reviewed their hotels. However, there is a stability in reviews through all seasons in years.

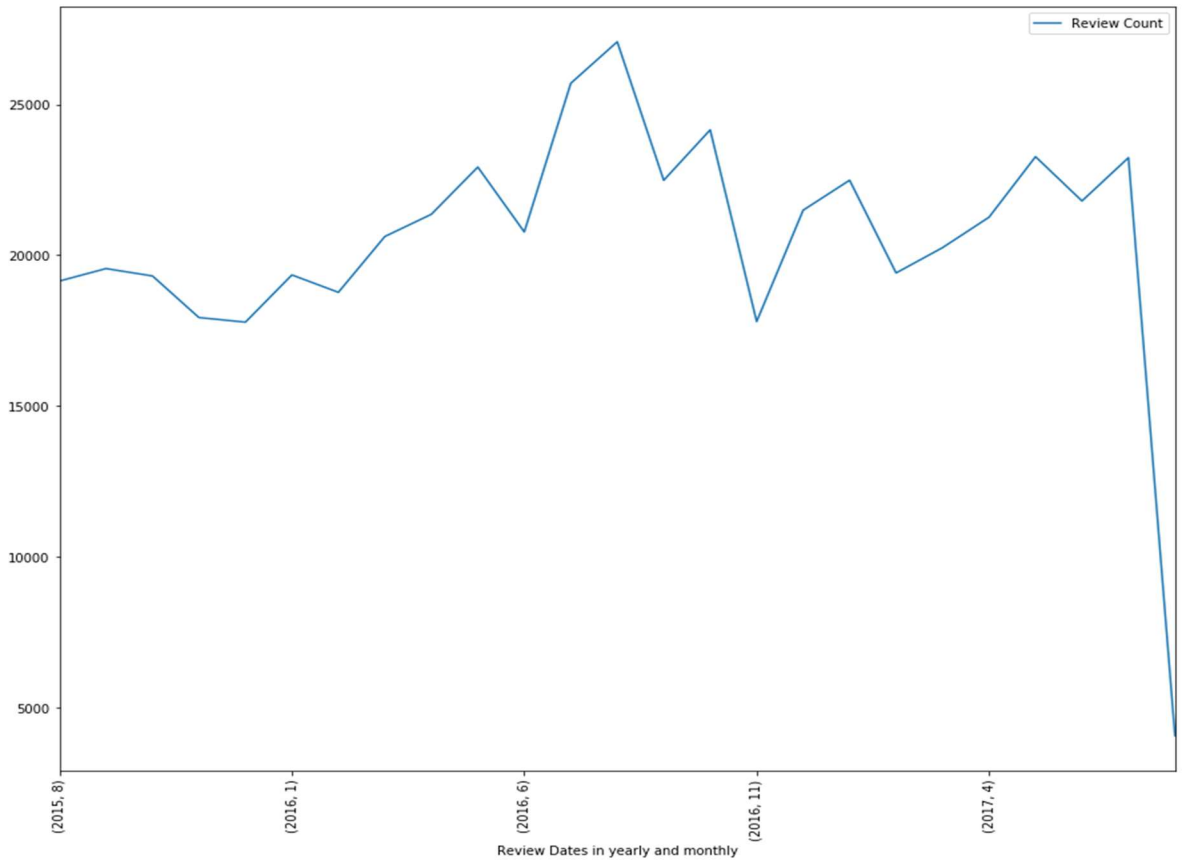

Figure 6: The count of reviews in yearly and monthly

2.2.4. Relation Between Average Review Score Through Reviewer's Nationality

Reviewer's nationality has been stored as countries in the raw data, this column and the mean of the score has been calculated. The reviewer's nationalities which have more than 100 reviews has taken from this data to avoid anomaly in result. We should have similar counts of reviews from these reviewers to get more accurate results in EDA analysis. Both ascending and descending functions had been used to generate plots as below.

The best hotels scored ranged 8.5 and 8.8. The reviewers from Puerto Rico tend to give higher scores, followed by Panama and the USA. On the contrary, from 103 different nationalities the worst hotels according to reviewer scores scored between 7.6 and 8. The reviewers from Bangladesh and Iran give lower scores to hotels. It could be interpreted as the nations which are

mostly Muslims are less satisfied from hotels in Europe. The differences between the cultures especially based on religious aspects could cause this. On the other hand, the tourists from continental countries are more pleased from European hotels. They do not tend to be resistant to changes and enjoy touristic activities.

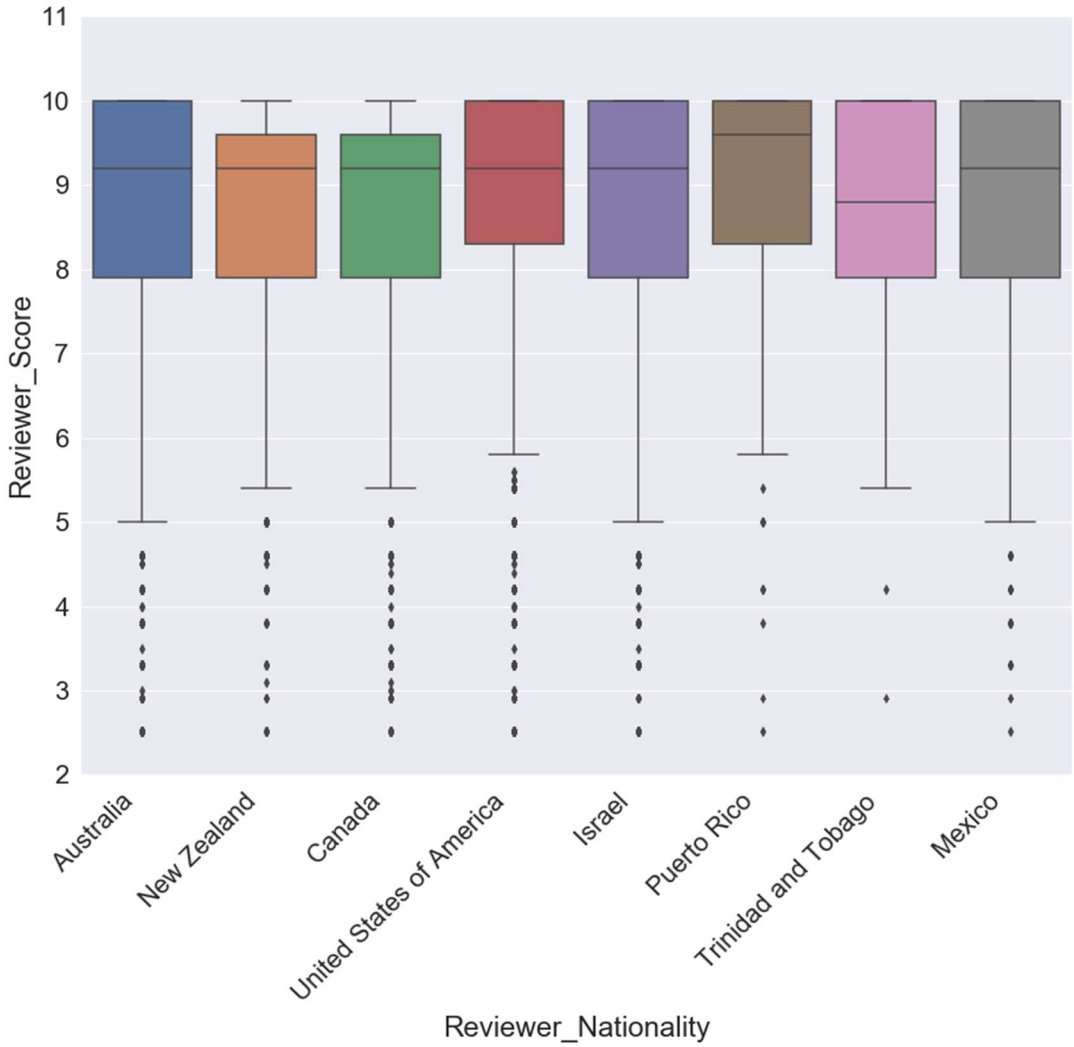

Figure 7: The scores given from guests according to their nationalities (Best average scores)

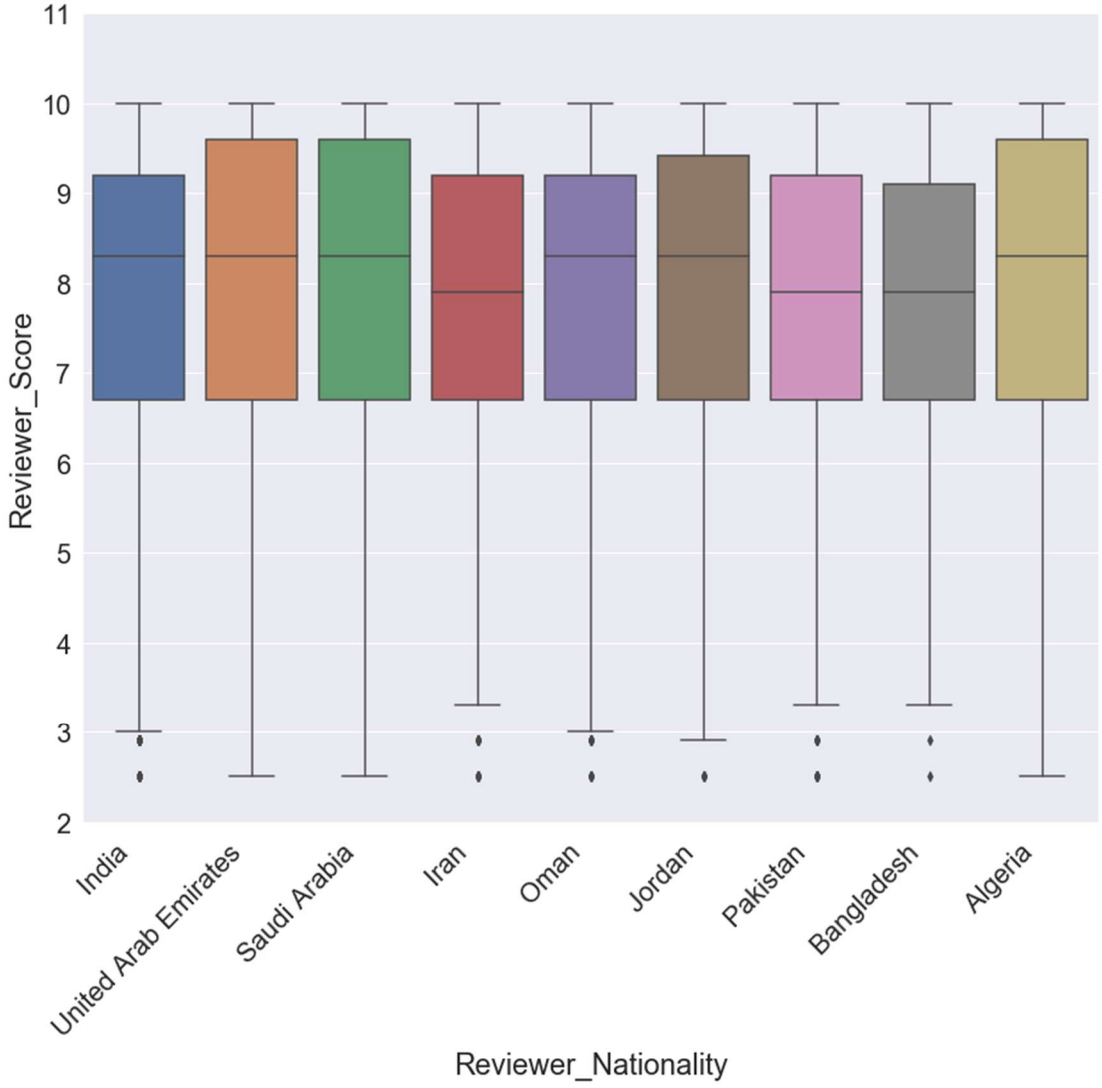

Figure 8: The scores given from guests according to their nationalities (Worst average scores)

2.2.5. Relation Between Scores and Country of Hotels

There are 6 different countries' hotel reviews that are considered in exploratory data analysis. The data contains the latitude and longitude of the hotels in coordinate dimensions. In the first step, the latitude and longitude values are calculated to find where the hotels are placed in word; AT(Austria), ES(Spain), NL(Netherlands), FR(France), IT(Italy), GB(England). Austria has the most average score based on their hotels in European countries. The hotels meet the requirements of guests to comfort and pleased them in Austria more than others. The hotels in Austria appeal to more luxury customers in terms of their economic power. They give importance to less guests and more customer satisfaction. Spain also has a huge potential of both summer and

winter trips for the guests. It has fantastic historical buildings and atmosphere which pleased the guests on its own. So that the hotel sector is more improved than the other countries in Europe.

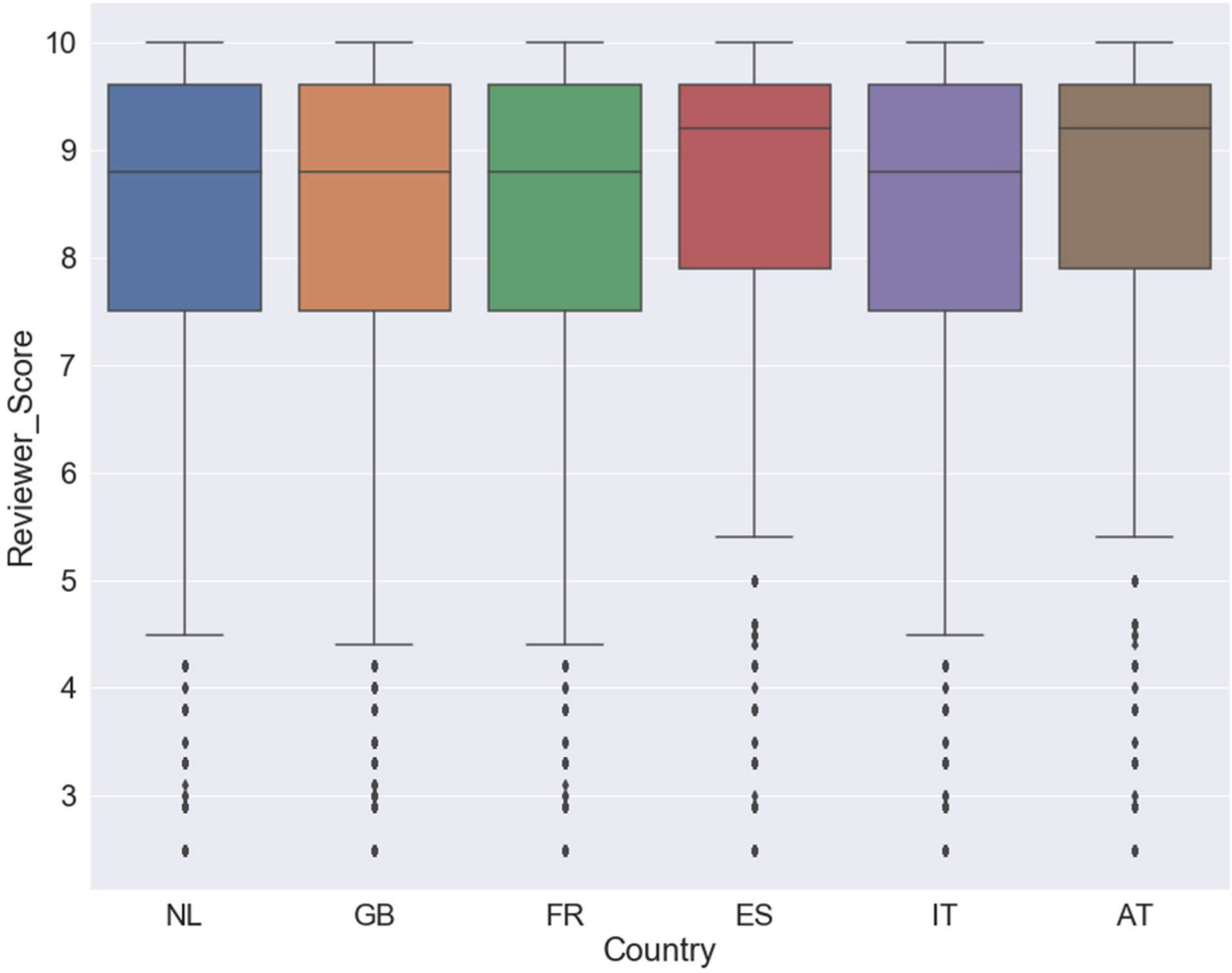

Figure 9: The boxplot of reviewer scores by country of the hotels

2.2.6. Relation Between Scores and Word Counts

Total negative and positive word counts represent the counts of negative and positive words in a review. It counts the words such as “good” and “beautiful” as positive words, on the contrary the words like “useless” and “dirty” as negative words. To the correlation map, the relation between the scores that are given by guests and the reviews that are given by them is normalized between 0 to 1. Reviewer score and negative word counts in review have negative relations. This means that the score is as low as the count of negative words in a review. The count of positive words is as much as the review’s score. They have a positive relationship between them. The

strength of relations negative and positive are respectively 0.38 and 0.22. This implicates that the guests who use negative words in their reviews tend to give lower scores to hotels then the reviewers who use positive words in their reviews give higher scores.

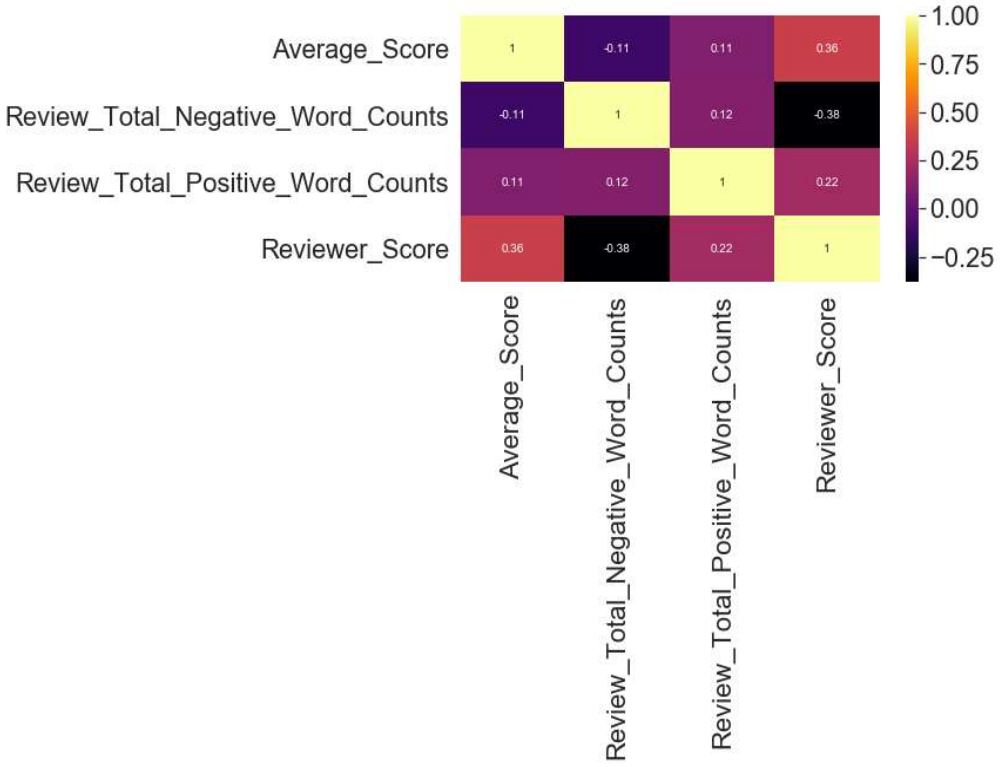

Figure 10: The correlation map between world counts and scores.

2.2.7. Positive and Negative Word Counts Analysis

"Review Total Positive Word Counts" column is taken into analysis as below. This column is calculated from Positive reviews in the dataset to a different column. The count of the unique words has been retrieved from the reviews. 0 words meaning that they are completely Negative reviews seems Positive reviews.

Number of completely Negative reviews in the dataset is 35.695. A piece of data has been shown as below.

	Positive_Review	Negative_Review
8	No Positive	Even though the pictures show very clean room...
32	No Positive	Our bathroom had an urine order Shower was ve...
98	No Positive	Got charged 50 for a birthday package when it...
121	No Positive	The first room had steep steps to a loft bed ...
134	No Positive	Foyer was a mess Only place to relax was the ...
146	No Positive	We booked a 3 night stay in a suite On arriva...
169	No Positive	Nothing One Of The Receptionist she did a rac...
172	No Positive	Hotel under sonstruction which we weren t awa...
202	No Positive	Renovation around the hotel sometimes can sta...
209	No Positive	Not given the room type we had booked and pre...

Figure 11: A piece of completely Negative reviews in the dataset

"Review Total Negative Word Counts" column is taken into analysis as below. This column is calculated from Negative reviews in the dataset to a different column. The count of the unique words has been retrieved from the reviews. 0 words meaning that they are completely Positive reviews seems Negative reviews.

Number of completely Positive reviews in the dataset is 126.902. A piece of data has been shown as below.

	Positive_Review	Negative_Review
1	No real complaints the hotel was great great ...	No Negative
13	This hotel is being renovated with great care...	No Negative
15	This hotel is awesome I took it sincirely bec...	No Negative
18	Public areas are lovely and the room was nice...	No Negative
48	The quality of the hotel was brilliant and ev...	No Negative
53	Beautiful setting in a lovely park room very ...	No Negative
55	The hotel is lovely and the staff were amazin...	No Negative
59	Basically everything The style of the hotel i...	No Negative
75	The whole hotel was very clean the staff were...	No Negative
78	Hotel was really nice staff were very friendl...	No Negative

Figure 12: A piece of completely Positive reviews in the dataset

2.2.8. Calculation of Negative and Positive Reviews

It is retrieved from the review context to classify the negative and positive reviews. The words “No Positive” or “Nothing” are considered as negative reviews, the word “everything” in negative review is also interpreted as negatively. It is labeled as 1 to review that if the review contains ‘everything’, if it has the word ‘nothing’ it is assigned 0. While we classify the counts in this content; %36 out of them has negative meaning from all positive and negative reviews. On the contrary, considering the words” No Negative” or “Nothing” from negative reviews, “everything” from positive reviews resulted %29 of them considered as positive.

After classifying these positive and negative reviews, the hotels that have the higher positive review ratio are visualized below. The top ten hotels are in London with the higher reviewer scores ranged between 8.6 to 9.3.

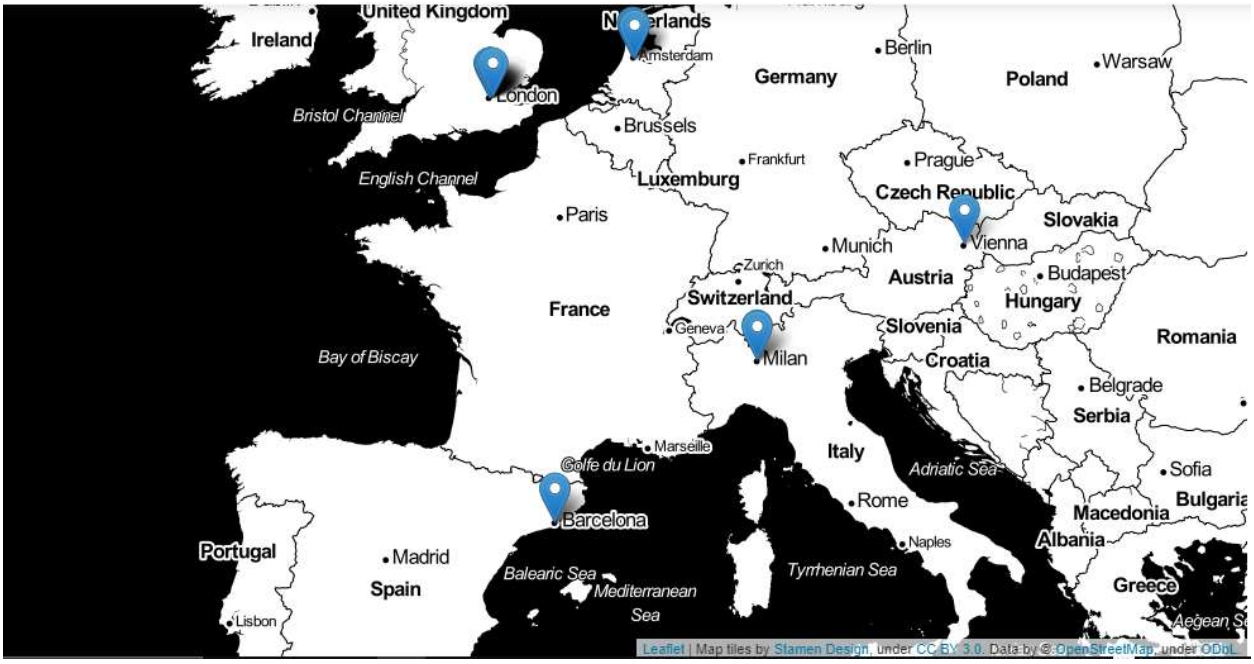

Figure 13: The top ten hotels which has the highest reviewer scores on the map.

3.PROJECT DEFINITION

3.1. Problem Statement

From the thousands of reviews that are given by guests through the internet, it is apparent that the tourism sector wants to gain an advantage over competitors. It is important to understand and react to the changes in their demands earlier than competitors.

In this study, the focus is on guest reviews of hotels located within the borders of Europe and listed on Booking.com.Booking.com is the world's largest community-based online resource of choice as a social media in the field of tourism with one million of properties in one database.

Sentiment analysis mainly focuses on views that express or imply positive or negative emotions. Especially, the positive or negative classification of comments is important in this research. Sentiment studies are performed with two different methods: dictionary-based method and machine learning method. Logistic Regression, Support Vector and Naive Bayes are preferred methods of machine learning process in this analysis.

3.2. Methods, Tools and Techniques

Social media comments of accommodation facilities serving in the Europe region within the scope of the research collected via the Booking.com platform. This raw data is obtained from Kaggle. Machine learning to find the emotional intensity of comments Support Vector Machine, Logistic Regression and Multinomial Naive Bayes methods were used.

In the preprocessing process, exploratory phases, and earlier phases of preparing the data are done by R language in Anaconda Spyder. One hot encoding is used to the preprocessing data frame for the sentiment analysis. It is frequently used in the encoding process though deep learning. It is the transformation of n observations into a binary variable as it is mentioned positive (1) or negative (0).

Stemming and the removing stop words such as ‘a’ or ‘the’ in English sentences are the further steps that should be implemented in the data. The stop words library is downloaded to be removed from the guest’s reviews. To make the matrix size not to be too large, first punctuation marks and numeric characters have been removed and the remaining characters converted to lowercase. Then, with the stop words application, words that do not affect the meaning of the sentence too much were removed from the comments and have the same root. Words (use or

useful) are downloaded to their roots by stemming method. Stemming is the second method to get accurate results from sentiment analysis. It reduces the related words from the same syllable. After pretreatment, the total size of the created document term matrix is 1.023.888.

In conclusion, the number of words that are retrieved from positive reviews is 4.907.136, the negative is 4.552.070. The most common positive words top ten are sorted descending as 'staff', 'location', 'room', 'hotel', 'good', 'great', 'friendly', 'breakfast', 'helpful', 'nice'. The common negative words are 'room', 'negative', 'hotel', 'breakfast', 'small', 'staff', 'nothing', 'rooms', 'would', 'could'.

The following step, the positive reviews and negative reviews are converted into text files to convert them into NumPy arrays. Comments must be digitized to be processed by the machine learning algorithm so that these NumPy arrays will be used in this process. All unique words in the comments are found then each given a numerical ID. The frequency value of the word is represented in the matrix for each comment. Term Frequency-Inverse was used for the calculation in frequency values of the words. For any term to be found in the matrix, it is limited to the first 1000. These values are added to the GridSearchCV module in the Scikit-learn library.

The most preferred text mining tool to find the emotional intensity of the collected comments from classifying algorithms, Linear Regression, Confusion Matrix and Naive Bayes were used. Confusion matrix shows the current situation in the data and the number of correct and incorrect predictions of our classification model. Below is the 2x2 confusion matrix of our methods.

Table 1: The confusion matrix values from the models.

Method	Tunning	TP	FP	FN	TN
BOW: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)]	140088	13656	9371	144052
BOW: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.1,.9]]	122907	30837	4597	148826
BOW: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.2,.8]]	130691	23053	5746	147677
BOW: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[2,10]]	128725	25019	5367	148056
BOW: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.22,.78]]	131862	21882	5980	147443
BOW: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.25,.75]]	132994	20750	6288	147135
BOW: Multinomial Naive Bayes	'alpha': [i for i in range (1,100,10)],'fit prior': [False]	141213	12531	9792	143631
BOW: Support Vector Machine	'alpha': [float(i)/10 for i in range (1,10,1)],'learning rate': ["optimal"]	138429	15315	22492	130931
BOW: Support Vector Machine	alpha': [float(i)/10 for i in range (1,10,1)],'learning rate': ["invscaling"]	140906	12838	32135	121288
BOW: Support Vector Machine	alpha': [float(i)/10 for i in range (1,10,1)],'learning rate': ["pa1"]	140532	13212	27545	125878
BOW: Support Vector Machine	alpha': [float(i)/10 for i in range (1,10,1)],'learning rate': ["pa2"]	139308	14436	24114	129309
BOW: Support Vector Machine	alpha': [float(i)/10 for i in range (1,10,1)],'learning rate': ["constant"]	140464	13280	27407	126016
BOW: Support Vector Machine	'alpha': [float(i)/10 for i in range(1,10,1)],'max_iter':[10]	131657	13352	27202	126221
BOW: Logistic Regression	'alpha': [float(i)/10 for i in range (1,10,1)]	134873	18871	20778	132645
BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,200,10)	134074	19670	19859	133564
BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,1000,10)	135201	18543	21332	132091
BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,2000,100)	134345	19399	20141	133282

BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,500,100)	135404	18340	21766	131657
TF-IDF: Multinomial Naive Bayes	'alpha': [i for i in range (1,100,10)]	140094	13650	9465	143958
TF-IDF: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.1,.9]]	99848	53896	1963	151460
TF-IDF: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.2,.8]]	119235	34509	3479	149944
TF-IDF: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[2,10]]	114167	39577	2958	150465
TF-IDF: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.22,.78]]	121657	32087	3775	149648
TF-IDF: Multinomial Naive Bayes	alpha': [i for i in range (1,100,10)],'class prior': [[.25,.75]]	125014	28730	4326	149097
Word2Vec: Logistic Regression		145859	7885	14134	139289

Referring to the confusion matrix values from table 1, the words have marked negative for negative reviews, and positive for positive reviews. The purpose of the model obtained from the machine learning method is to predict both positives and negatives. If a situation that normally exists as positive in the prediction process is predicted as positive, TP is estimated. If the existing condition was negative and the prediction was negative, TN is estimated. In other words, a wrong situation was correctly predicted as wrong. If the existing condition is negative but the prediction system predicts as positive, the first type of error FP case occurs. If the existing state is positive and the estimator predicts negatively, FN second type error occurs. (Santra, A & Christy, Josephine., 2012)

The ROC curve is an important performance measure for classification problems. The ROC is a probability curve and the area under it, AUC, represents the degree or measure of separability. In the ROC curve, there is FPR (False Positive Rate) on the X axis and TPR (True Positive Rate) on the Y axis. (Hajian-Tilaki, Karimollah., 2013)

As the remaining under the curve increases, the discrimination performance between classes increases. This means that the higher the AUC, the better the model has been classified. The methods that are used have similar AUC values so that they are classified successfully.

Accuracy is the ratio of the correct estimates made in the system to all estimates. Precision indicates success in a positively predicted situation, on the other hand recall shows how successfully positive situations are predicted. It is clearly concluded that the Multinomial Naive Bayes has the most successful prediction to the labelling the reviews as positive or negative with the higher accuracy recall and precision values.

Table 2: The methods' results based on the hyperparameter changes in models.

Method	Tunning Detail in Parameters	Accuracy	Recall	Precision	AUC
BOW: Multinomial Naive Bayes	class prior' is not set	92.50%	93.89	91.34	0.97
BOW: Multinomial Naive Bayes	class prior': [[.1,.9]]	88.46%	97	82.83	0.97
BOW: Multinomial Naive Bayes	class prior': [[.2,.8]]	90.62%	96.25	86.49	0.97
BOW: Multinomial Naive Bayes	class prior': [[2,10]]	90.11%	96.5	85.54	0.97
BOW: Multinomial Naive Bayes	class prior': [[.22,.78]]	90.93%	96.1	87.07	0.97
BOW: Multinomial Naive Bayes	class prior': [[.25,.75]]	91.20%	95.9	87.64	0.97
BOW: Multinomial Naive Bayes	fit prior': [False]	92.73%	93.61	91.97	0.97
BOW: Support Vector Machine	learning rate': ["optimal"]	87.69%	85.33	89.52	0.94
BOW: Support Vector Machine	learning rate': ["invscaling"]	85.36%	79.05	90.42	0.94
BOW: Support Vector Machine	learning rate': ["pa1"]	86.73%	82.04	90.5	0.94
BOW: Support Vector Machine	learning rate': ["pa2"]	87.45%	84.28	89.95	0.94
BOW: Support Vector Machine	learning rate': ["constant"]	86.75%	82.13	90.46	0.94
BOW: Support Vector Machine	max_iter':[10]	86.80%	82.26	90.43	0.94
BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,10,1)]	87.09%	86.45	87.54	0.94
BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,200,10)	87.13%	87.05	87.16	0.94

BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,1000,10)	87.02%	86.09	87.69	0.94
BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,2000,100)	87.13%	86.87	87.29	0.94
BOW: Logistic Regression	alpha': [float(i)/10 for i in range (1,500,100)	86.94%	85.81	87.77	0.94
TF-IDF: Multinomial Naive Bayes	class prior' is not set	92.47%	93.83	91.34	0.98
TF-IDF: Multinomial Naive Bayes	class prior': [[.1,.9]]	81.81%	98.72	73.75	0.98
TF-IDF: Multinomial Naive Bayes	class prior': [[.2,.8]]	87.63%	97.73	81.29	0.98
TF-IDF: Multinomial Naive Bayes	class prior': [[2,10]]	86.15%	98.07	79.17	0.98
TF-IDF: Multinomial Naive Bayes	class prior': [[.22,.78]]	88.32%	97.53	82.34	0.98
TF-IDF: Multinomial Naive Bayes	class prior': [[.25,.75]]	89.24%	97.18	83.84	0.98
Word2Vec: Logistic Regression	-	93.08%	91.8	94.19	0.98

Hyperparameter optimization is the process of finding the most suitable hyperparameter combination according to the success metric specified for a machine learning algorithm. Overfitting and underfitting balance can be achieved by balancing model complexity with hyperparameter optimization. Additionally, the problem of over-learning caused by the flexibility of the model can be solved with the limitations introduced with hyperparameters. For this reason, different methods have been developed for hyperparameter optimization. GridSearcCV and RandomizedSearchCV are among these methods. GridSearchCV method is used in this analysis. The most successful hyperparameter set is determined by a separate model which is established with all combinations that are tested.

Hypermeter parameters could be selected to get a more accurate model. Firstly, the hyperparameters and their values are defined in a dictionary structure. Then the GridSearchCV method imported from the Sklearn library is called and the necessary parameters are specified. Class priority (prior probabilities of the classes, if specified the priors are not adjusted according to the data.) and fit prior (using data for prior class probability) values are altered manually to analyze different results from the Multinomial Naive Bayes classifiers.

Learning Rate is another hypermeter parameter that is altered manually in SGDClassifier. There are 5 different values that can be performed as the development rate in the algorithms on the train data set. Generally, the lower values of the learning rate make the model learn slowly, so that a more accurate result is got. Alpha is another parameter that controls how the model is affected by the coefficients or learns each step after updated. It is like the learning rate in logistic regression model.

4.RESULTS

That is determined the subjective attitude into a text expresses feeling, sentiment analysis is one of the important research areas of natural language processing. The purpose of sentiment analysis is to determine the class of a particular text (positive, negative). When sentiment analysis is evaluated from the perspective of management information systems, it is extremely useful for businesses. For example, businesses always have products or want to know public or consumer opinions about its services. Same potential customers without purchasing a product or using a service want to learn the opinions of existing customers first. In addition, businesses with emotion analysis.

They can create a short-term marketing campaign to meet their demands, feel by applying the analysis, and make their campaigns even more suitable for their target audience. Therefore, written texts are of great importance for businesses.

Text representations in emotion classification the most used method for creation is Bag of Words. It considers whether the words are in the document without taking the semantic context of the words. On the other hand, Word2vec word vectors have also been used, and the emotion classification success effects have been studied.

For this purpose, in the study, a data set of hotel reviews in English language was used and the suggested methods. The effect on texts in languages is also evaluated. Linear regression, Naive Bayesian and Support Vector Machine were applied as classification algorithms.

While the hotel review words are analyzed by three different segmentation algorithms in various tunning methods. Different class priority parameters are chosen in Naïve Bayes algorithm. Without using TF-IDF in BOW model, the accuracy, recall, and precision values are higher than the model with TF-IDF. However, the model's AUC is higher with TF-IDF by 0.98. It is interpreted as the model with TF-IDF is successfully classified.

In support vector machine, the learning rate value optimal has the highest accuracy among other parameters. It concludes that the model with optimal value of learning rate provide the model learning better. Logistic regression got similar accuracy rates with different Alpha values.

Word2Vec model has the highest accuracy rate by 93.08% among other models that is used. It follows by 92.73% accuracy in BOW model using Multinominal Naïve Bayes. The parameter fit prior is out of the analysis so that the model does not consider the class prior probabilities. In conclusion, the Multinomial Naive Bayes algorithm in BOW without using the information on the more important words or less important gets the highest accuracy rates. For this model, the important of the words do not have much impact on the positive or negative hotel reviews.

5.SOCIAL AND ETHICAL ASPECTS

It is especially useful when sentiment analysis is evaluated from the perspective of management information systems. For example, businesses always want to know public or consumer opinions about its products or services. On the other hand, potential customers want to learn the opinions of existing customers before purchasing a product or using a service. Most companies are social and active on media platforms. In contrast, a social media platform is not a platform where only businesses can promote their brands or services, it is also a place filled with information about how it is perceived by customers. Hence sentiment analysis helps to optimize marketing strategies and offers opportunities to businesses.

In addition, businesses can create a short-term marketing campaign to meet their demands by sentiment analysis. They can make their campaigns even more suitable for their target audience. Especially, decision support systems that are implemented by sentiment analysis help to make faster and more efficient decisions.

REFERENCES

[1] Elgendy, Nada & Elragal, Ahmed. (2014). Big Data Analytics: A Literature Review Paper. Lecture Notes in Computer Science. 8557. 214-227. 10.1007/978-3-319-08976-8_16.

[2] Mankad, Shawn & Han, Hyunjeong & Goh, Joel & Gavirneni, Srinagesh. (2016). Understanding Online Hotel Reviews Through Automated Text Analysis. Service Science. 8. 124-138. 10.1287/serv.2016.0126.

[3] Rodrigues, J., Sousa, M. and Brochado, A., 2020. A Systematic Literature Review on Hospitality Analytics. *International Journal of Business Intelligence Research*, 11(2), pp.47-55.

[4] S. Shayaa et al., "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges," in IEEE Access, vol. 6, pp. 37807-37827, 2018, doi: 10.1109/ACCESS.2018.2851311.

[5] H. Yousaf et al. / Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 5, 1282-1287 (2020)

[6] Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

[7] 515K Hotel Reviews Data in Europe. (2017, August 21). Retrieved from https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe

[8] Boiy, E., Moens, M. A machine learning approach to sentiment analysis in multilingual Web texts. Inf Retrieval 12, 526–558 (2009)

[9] Gavilan, Diana & Avello, Maria & Martinez, Gema. (2017). The influence of online ratings and reviews on hotel booking consideration. Tourism Management. 66. 53-61. 10.1016/j.tourman.2017.10.018.

[10] Farisi, A., Sibaroni, Y. and Al Faraby, S., 2019. Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier.

[11] Kim, SW., Gil, JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019). https://doi.org/10.1186/s13673-019-0192-7

[12] Santra, A & Christy, Josephine. (2012). Genetic Algorithm and Confusion Matrix for Document Clustering. International Journal of Computer Science Issues. 9.

[13] Hajian-Tilaki, Karimollah. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian journal of internal medicine. 4. 627-635.

[14] Aydogan, Murat & Karci, Ali. (2019). Turkish Text Classification with Machine Learning and Transfer Learning. 10.1109/IDAP.2019.8875919.

[15] Bofang Li, Zhe Zhao, Tao Liu, Puwei Wang, and Xiaoyong Du. 2016. Weighted neural bag-of-n-grams model: New baselines for text classification. In Proceedings of COLING 2016. pages 1591–1600.