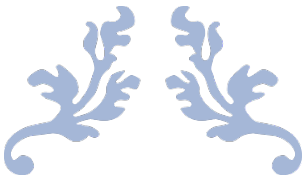


UNIVERSITÀ DI PISA

AI Power in Hotels Reviews Sentiment Analysis

Master's Degree in Artificial Intelligence and Data Engineering
Business and Project Management

Academic Year 2022/2023

Project developed by

Basma Adawy & Francesco De Vita

Table of Contents

Introduction

This paper presents a project that aims to use natural language processing (NLP) techniques to analyze online reviews that customers write about hotels on the web. The project wants to explore customer sentiments related to the reviews and provide useful insights to hotel managers.

The paper is structured into three main parts that address three business-relevant questions:

- **Can sentiment analysis help to categorize online reviews and measure customer satisfaction with different hotels?**

  This first part applies sentiment analysis technique on online reviews to obtain feedback about the customers' attitude towards the hotels they stayed at. This process can help hotel managers to get a first impression of customer satisfaction or dissatisfaction with their services using text mining classification methods.

- **Can consumer happiness with various hotels be analyzed and categorized from internet reviews using sentiment analysis?**

  This second part uses another NLP technique called keyword extraction to dive deeper into the online reviews. This process can help hotel managers to extract the most relevant words from the reviews and understand better the aspects of their services that are more or less appreciated. With this analysis, hotel managers can make more informed decisions in the future to meet customer expectations.

- **Can AI be useful to solve problems related to Hotels Business?**
  From the reviews we analyzed revealed that AI technology can be used to improve hotel services, operations, and revenues in various ways. In this section, we will discuss how different hotel brands have implemented AI solutions to enhance their customer experience, optimize their pricing and revenue management, and increase their customer loyalty. We will also examine some of the benefits and challenges of using AI for hotels.

Data

Dataset Description

The dataset used in this project, to answer those questions and develop the analysis, could be downloaded by Kaggle from this link: https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe

This dataset contains more than 500 thousand reviews about some of Hotels from 6 countries in Europe.

The following fields are available in this dataset:

- **Hotel_Address:** Address of hotel.
- **Review_Date:** Date when reviewer posted the corresponding review.
- **Average_Score:** Average Score of the hotel
- **Hotel_Name:** Name of Hotel
- **Reviewer_Nationality:** Nationality of Reviewer
- **Negative_Review:** Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should 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 does not give the negative review, then it should be: 'No Positive'.
- **Review_Total_Positive_Word_Counts:** Total number of words in the positive review.
- **Reviewer_Score:** Score the reviewer has given to the hotel, based on his/her experience.
- **Total_Number_of_Reviews_Reviewer_Has_Given:** Number of Reviews the reviewers has given in the past.
- **Total_Number_of_Reviews:** Total number of valid reviews the hotel has.
- **Tags:** Tags reviewer gave the hotel.
- **days_since_review:** Duration between the review date and scrape date.
- **Additional_Number_of_Scoring:** There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
- **lat:** Latitude of the hotel
- **lng:** longtitude of the hotel

The following pictures show a sample of dataset rows:

	Hotel_Address	Additional_Number_of_Scoring	Review_Date	Average_Score	Hotel_Name	Reviewer_Nationality	Negative_Review	Review_Total_Negative_Word_Counts	Total_Number_of_Reviews
0	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	8/3/2017	7.7	Hotel Arena	Russia	I am so angry that i made this post available...	397	1403
1	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	8/3/2017	7.7	Hotel Arena	Ireland	No Negative	0	1403
2	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	7/31/2017	7.7	Hotel Arena	Australia	Rooms are nice but for elderly a bit difficul...	42	1403

Positive_Review	Review_Total_Positive_Word_Counts	Total_Number_of_Reviews_Reviewer_Has_Given	Reviewer_Score	Tags	days_since_review	lat	lng
Only the park outside of the hotel was beauti...	11	7	2.9	[' Leisure trip ', ' Couple ', ' Duplex Double...	0 days	52.360576	4.915968
No real complaints the hotel was great great ...	105	7	7.5	[' Leisure trip ', ' Couple ', ' Duplex Double...	0 days	52.360576	4.915968
Location was good and staff were ok It is cut...	21	9	7.1	[' Leisure trip ', ' Family with young childre...	3 days	52.360576	4.915968

Data Visualization

Data visualization is a useful tool that can turn data into knowledge and action. Data visualization can explore and analyze data to learn about it and summarize data to show its main points.

Hotels average score map

This map can be helpful to the hotel manager to show him the hotels in different areas and their average score, from this he can understand the market around him and his competitors.

We divided the hotels into 3 categories, depending on the average score into:

- Good Hotels which represent the green color
- Medium Hotels which represent the yellow color
- Bad Hotels which represent the red color

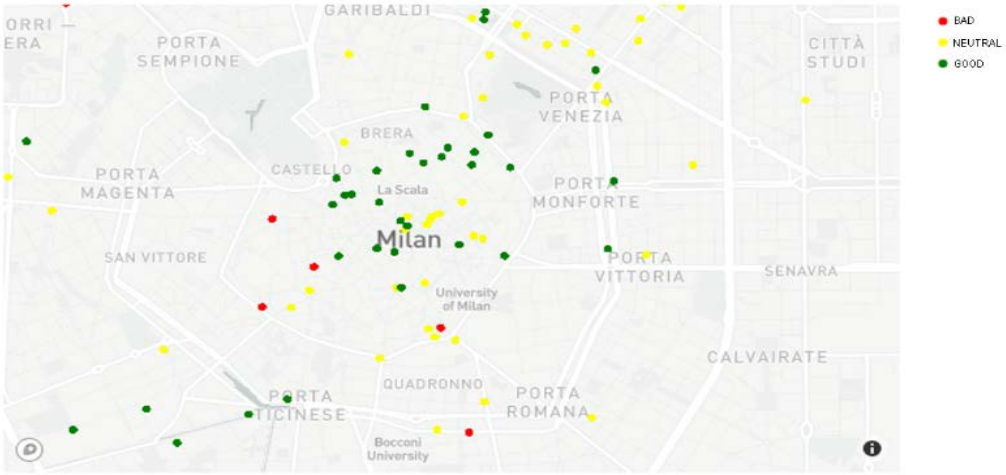


Figure 1 - Hotels average score map

Reviewer score distribution

The reviewer give a score from 1 to 10 to rate the Hotel

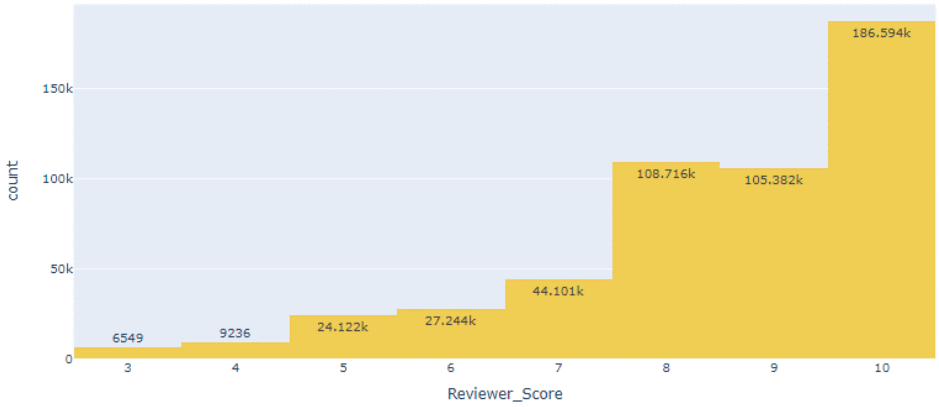


Figure 2 - Reviewer score distribution

Reviewer Sentiment distribution

This graph is helpful to understand outliers in Data, because usually if the difference between Review Total Positive Word Counts and Review Total Negative Word Counts is approximately 0 it means that the Review Sentiment should be Neutral.

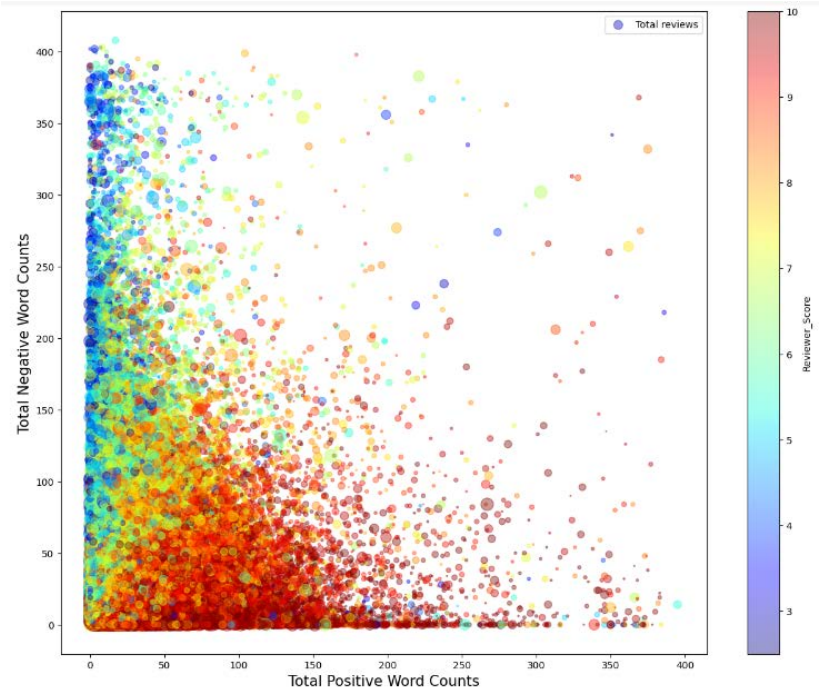


Figure 3 - Reviewer Sentiment distribution

Year-Month review distribution

This representation is for all hotels to help the hotel managers to know which are the months with the higher number of people who reviewed the hotel after their stay.

In the Fig. 5 we analyzed 2016 as year because in the other years there is missing data.

This analysis can be done also for each hotel.

Figure 4 - Year-Month review distribution

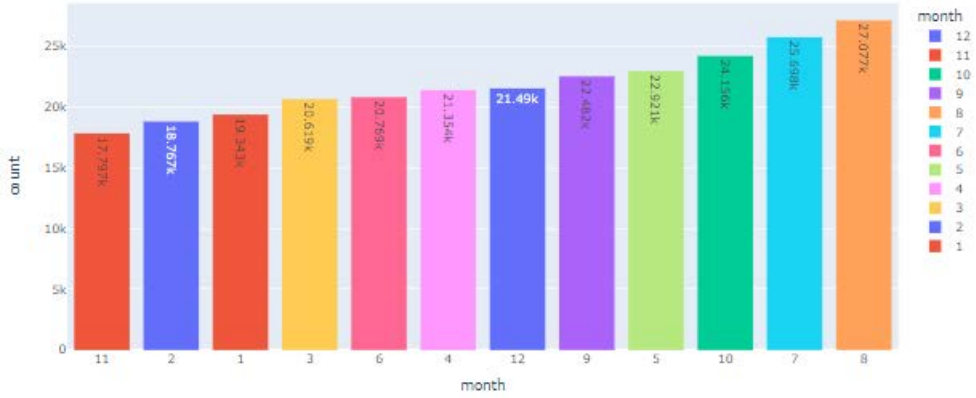


Figure 5 - 2016 Reviews distribution

Top 5 Hotels with most number of reviews

This graph shows which are the Top 5 hotels in number of reviews in which we will use in competitors analysis

Figure 6 - Top 5 Hotels with most number of reviews

Hotels Review Count

This Analysis is for each hotel from the Top 5 Hotels with most number of reviews to represents the count of positive, negative and total Reviews.

Fig. 7 is an Example, you can find others in the Notebook

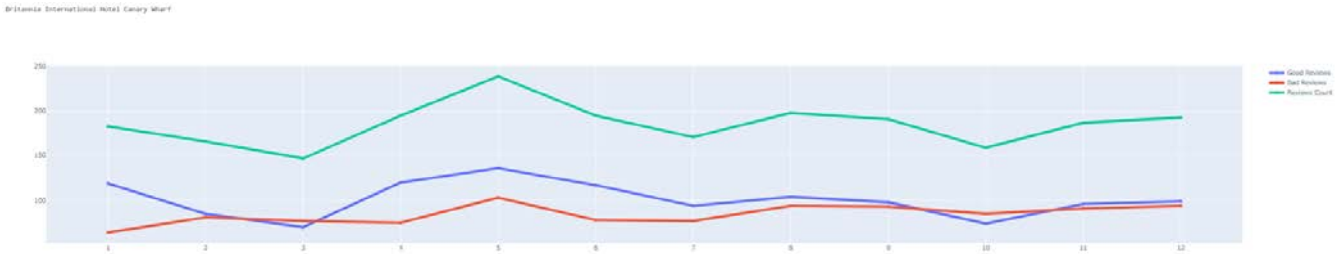


Figure 7 – "Britannia International Hotel Canary Wharf" Review Count

Positive and Negative Reviews Wordcloud

This type of graphs are useful to Summarize the main topics of the reviews ,highlight the most common words and compare the word usage or sentiment.

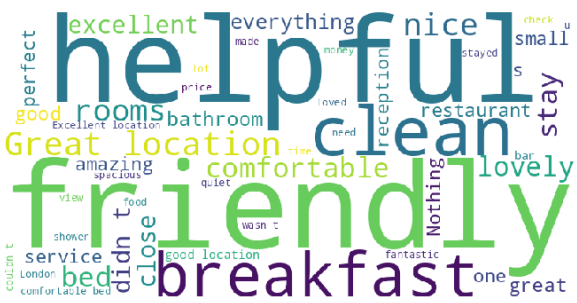


Figure 8 - Positive Reviews Wordcloud

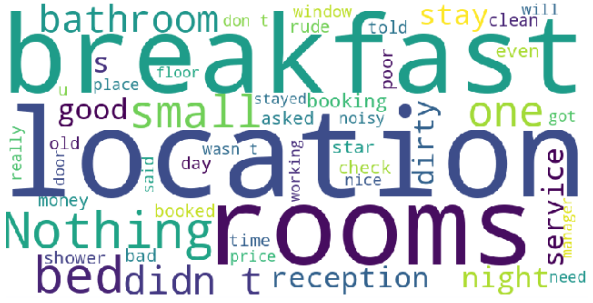


Figure 9 - Negative Reviews Wordcloud

Data Cleaning

The dataset downloaded was "dirty" and therefore subjected to a thorough cleaning procedure to be able to make a more accurate analysis of the text left in the comments.

Deleted NaN

Drop the rows which contain not existing values.

```
reviews_df.dropna(inplace=True)
reviews_df.isna().sum()
```

Duplicated Rows

Drop the rows which contain repeated values.

```
print("Duplicated rows before: ",reviews_df.duplicated().sum())
reviews_df.drop_duplicates(inplace=True)
print("Duplicated rows after: ",reviews_df.duplicated().sum())
```

Delete “No Positive” and “No Negative” values in Review Columns

```
print('No Positive values before:', end=" ")
print(reviews_df['Positive_Review'].value_counts().get('No Positive', 0))
print('No Negative values before:', end=" ")
print(reviews_df['Negative_Review'].value_counts().get('No Negative', 0))

reviews_df['Negative_Review']= reviews_df['Negative_Review'].replace("No Negative" , "")
reviews_df['Positive_Review']= reviews_df['Positive_Review'].replace("No Positive" , "")
```

Merge Together Positive_Review and Negative_Review Columns

```
reviews_df['Review'] = reviews_df['Negative_Review'].astype (str).str.cat (reviews_df['Positive_Review'], sep=' ')
```

Convert to Lower Case

```
def ConvertToLower(text):
    temp = text.lower()
    temp = re.sub("@[A-Za-z0-9_]+","", temp)
    temp = re.sub("#[A-Za-z0-9_]+","", temp)
    temp = re.sub(r'http\S+', '', temp)

    return temp
```

Removing Stopwords

They are often removed from text for NLP tasks, such as text analysis, text classification, sentiment analysis, information retrieval, etc. This can reduce noise and focus on important words.

```
stop_words = stopwords.words('english')
reviews_df_1['Review'] = reviews_df_1['Review'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stop_words)]))
```

Stemming

It’s a technique that reduces words to their root form by removing prefixes and suffixes.

We choose Stemming instead of Lemmatization because it returns better accuracy in the classification problem.

```
ps = PorterStemmer()
corpus = []
for i in range(0, len(review_features)):
    review = re.sub('[^a-zA-Z]', ' ', review_features['Review'][i])
    review = review.split()
    review = [ps.stem(word) for word in review if not word in stop_words]
    review = ' '.join(review)
    corpus.append(review)
```

Classification

Preparing Data for Classification

Define Classes (0 = Bad, 1 = Neutral, 2 = Good)

```
def multiclass(x):
  if x<=3.5:
    return 0
  elif (x<=6.5 and x>=5.5):
    return 1
  elif x==10:
    return 2
  else:
    return 3

reviews_df_1["Review_Type"] = reviews_df_1["Reviewer_Score"].apply(multiclass)
reviews_df_1.drop(reviews_df_1[reviews_df_1.Review_Type == 3].index, inplace=True)
```

TF-IDF Vectorizer

It's a technique that converts a collection of text documents into a matrix of numerical features that represent the importance or relevance of words in each document. TF-IDF stands for Term Frequency-Inverse Document Frequency.

```
tfidf_vectorizer = TfidfVectorizer(max_features=5000,ngram_range=(2,2))
X= tfidf_vectorizer.fit_transform(review_features['Review'])
```

SMOTE over Sampling

SMOTE is a technique that creates synthetic samples for the minority class in imbalanced classification datasets, help to balance the class distribution and improve the performance. Works by finding the nearest neighbors of the minority class samples and generating new samples along the line connecting them.

```
smote = SMOTE(random_state=42)
X_res, y_res = smote.fit_resample(X, y)
```

Figure 10 - Classes Distribution before Sampling

Data Splitting

Used to split the Data into training and test set to train the models on the training set and test it on the test set.

```
X_train, X_test, y_train, y_test = train_test_split(X_res, y_res, test_size=0.25, random_state=0)
```

Models Testing

We applied various models for multiclass classification, which is a type of machine learning problem where the target variable has more than two possible classes. Some of the models we used are:

- Logistic Regression: This is a linear model that predicts the probability of each class using a logistic function. It can handle binary or multiclass classification problems.
- Multinomial NB: This is a probabilistic model that applies the Bayes' theorem with the assumption of independence among features. It can handle discrete data with multiple classes.
- Linear SVC: This is a support vector machine model that uses a linear kernel function to separate the data into different classes. It can handle binary or multiclass classification problems.
- Random Forest: This is an ensemble model that combines multiple decision trees to create a more accurate and robust prediction. It can handle both classification and regression problems.
- XGBoost: This is an optimized implementation of gradient boosting, which is a technique that builds a series of weak learners and improves them by minimizing a loss function. It can handle both classification and regression problems.

In Fig. 11 is represented the accuracy of each model, as we can see, the best model is the Random Forest Classifier, its accuracy score is 0.88.

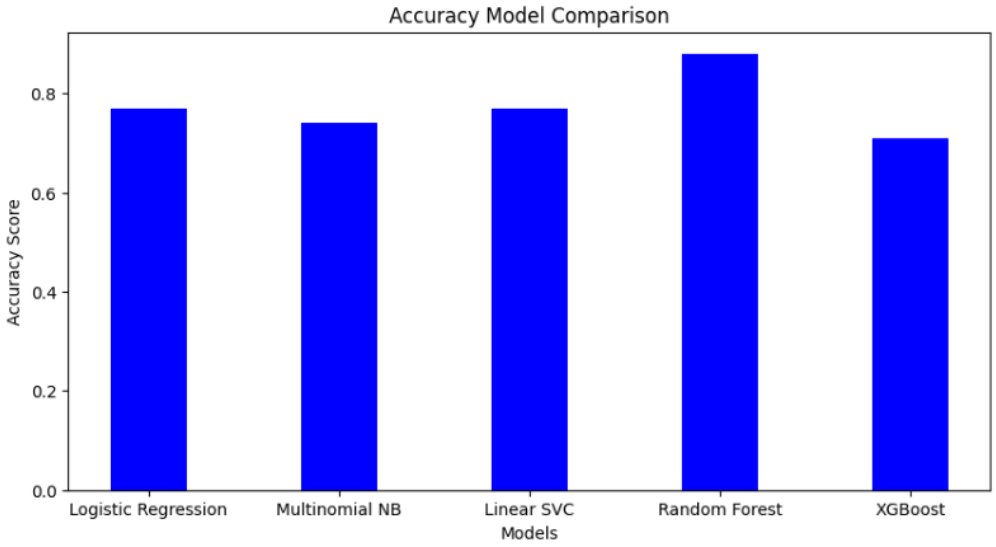


Figure 11 - Accuracy Models Comparison

roBERTa Model

*r*oBERTa is a type of deep learning model that uses a neural network to learn from a large amount of text data. It is based on another deep learning model called BERT, but it has some improvements and optimizations that make it more powerful and efficient. RoBERTa can be used for various natural language processing tasks, such as sentiment analysis.

The output of the RoBERTa model is a vector, that represents the positive, neutral and negative Sentiment in a text.

roberta_neg	roberta_neu	roberta_pos	Review	Review_Type
0.003532	0.042326	0.954142	The fact stay longer Staff great room beautifu...	2
0.003017	0.031512	0.965472	Having breakfast included price would perfect ...	2
0.967604	0.029713	0.002684	The staff Rude incompetent staff	0

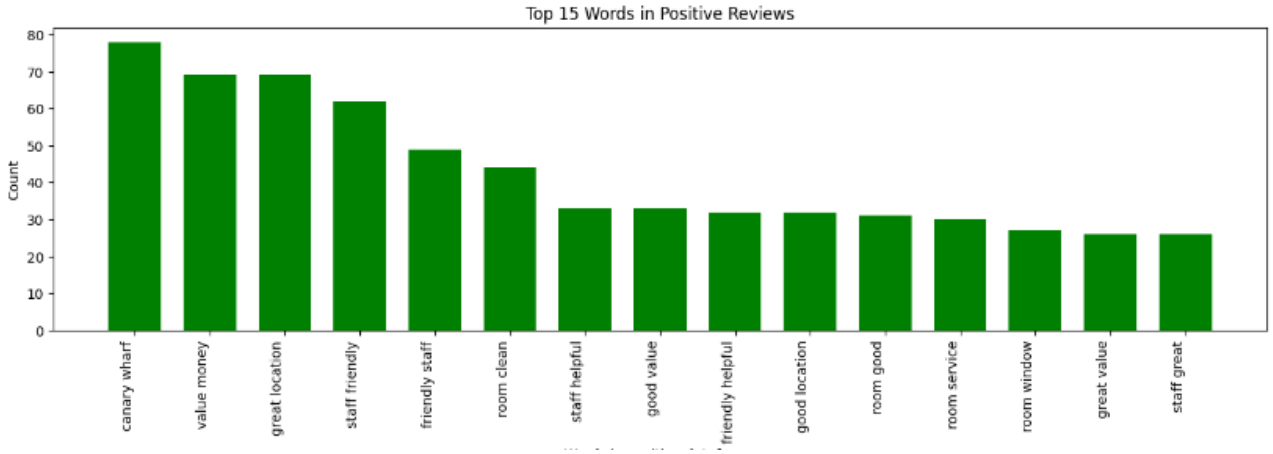
This model can help the hotel managers to know the customer emotions from their Review without training the model, because it is a pre-trained model.

Competitors Analysis

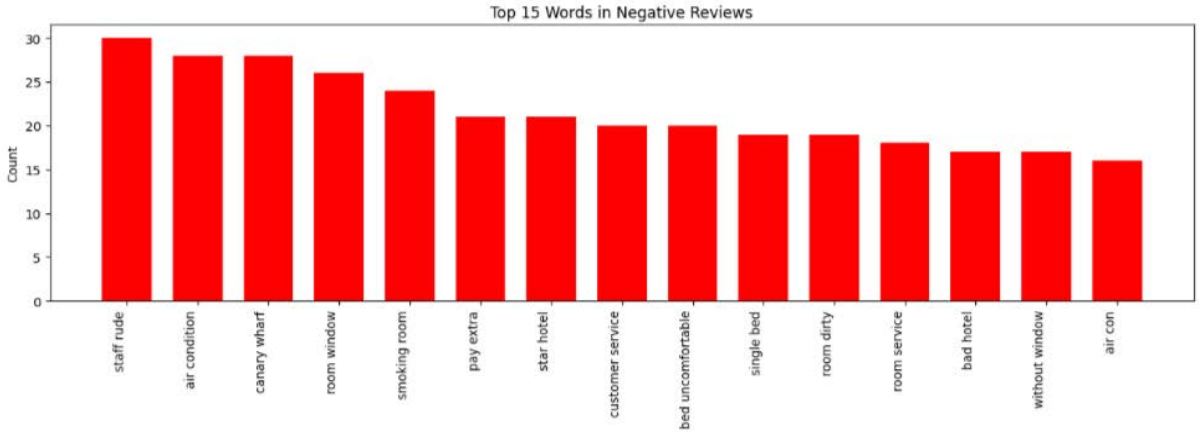
This analysis is to present the 5 hotels with their strength and weak points.

Britannia International Hotel Canary Wharf

From this Bi-Gram graph the hotel Manager can understand that the strong points of his hotel are the great value for money, the location and the staff.



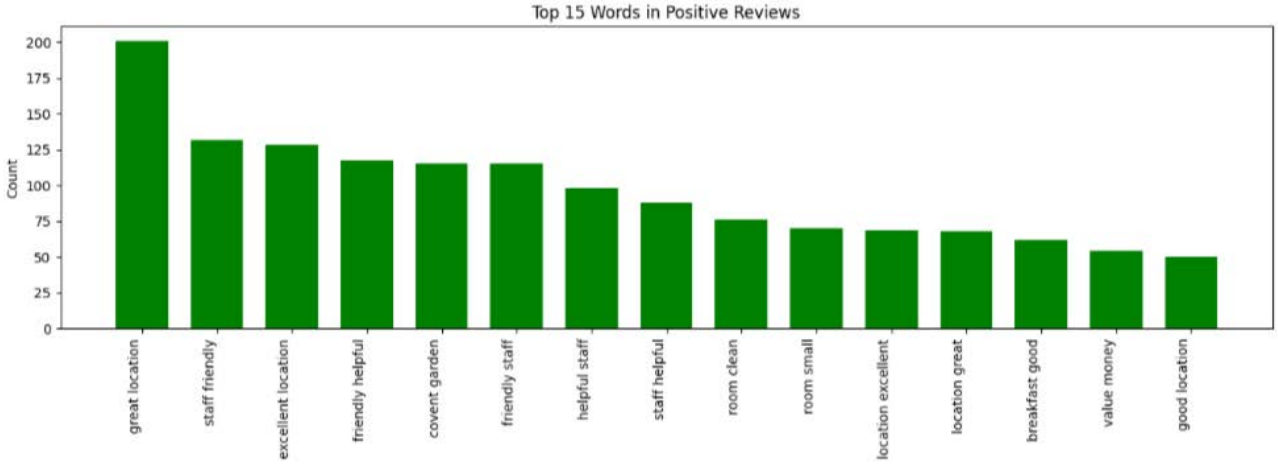
Some customers have criticized the staff and the air conditioning.
For the rude staff, the hotel manager can Apologize to the customers who experienced poor service and offer them a discount or a free upgrade.
Organize a staff training session on customer service skills and emphasize the importance of being polite, respectful, and helpful to guests and Monitor the performance and provide feedback and coaching to improve their service quality.
For the faulty air conditioning, the hotel manager can Apologize to the customers who suffered from uncomfortable temperatures and offer them a compensation, such as a refund, a room change, or a voucher.
Check the thermostat settings and make sure they are correct and consistent with the guests' preferences Conduct regular maintenance checks on the air conditioning system to prevent future problems.



Strand Palace Hotel

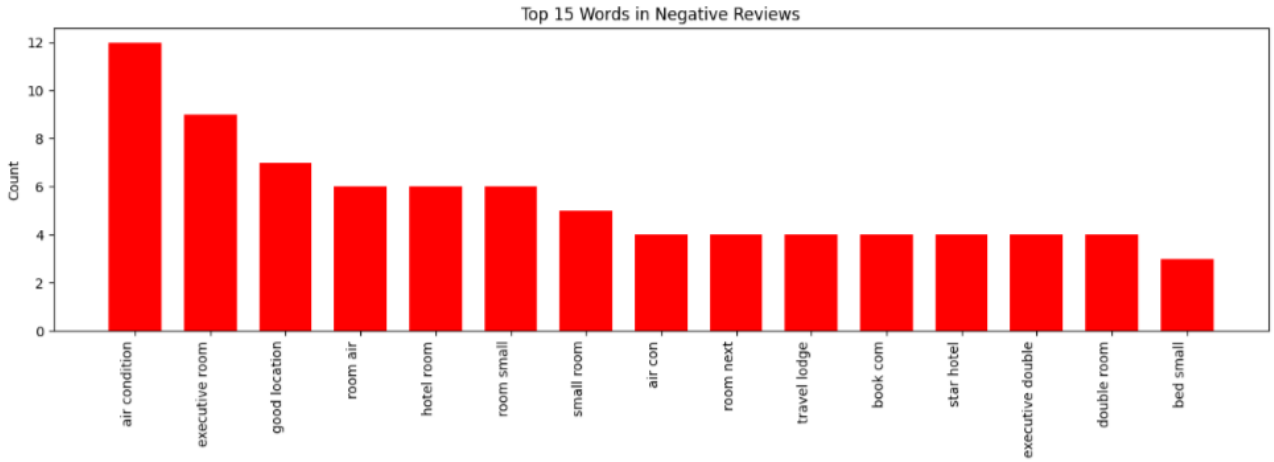
From this Bi-Gram graph the hotel Manager can understand that the strong points of his hotel are the great Location and the staff.

The hotel manager should give rewards to the staff because the service was excellent and met the expectations of the customers.



The hotel weaknesses are the Air Conditioning, room size and the executive room.

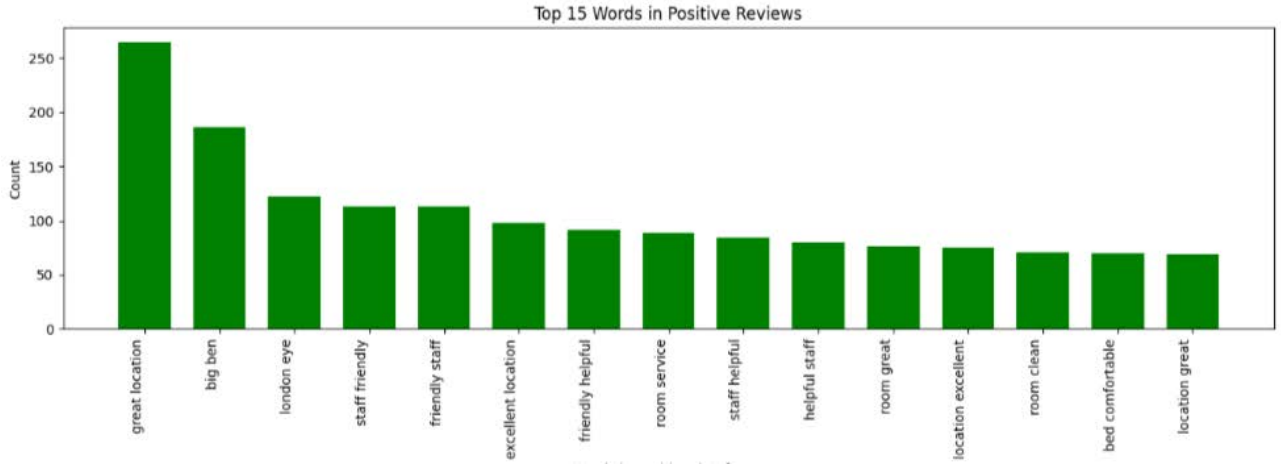
For the small room size, the hotel manager can use AI for creative design and layout techniques to make the rooms appear more spacious and inviting, such as using mirrors and light colors.
Highlight the benefits of staying in a smaller room, such as lower price, coziness, efficiency, and environmental friendliness.
Conduct market research and benchmarking analysis to see how the hotel's room size compares to its competitors and adjust the pricing strategy accordingly.
For the executive room the hotel manager should Ensure that the executive room meets the expectations of business travelers and offers more value than standard rooms, such as a larger bed, a seating area, a desk, a minibar, a coffee maker, and a safe. Solicit feedback from executive room guests and implement improvements based on their suggestions and preferences.



Park Plaza Westminster Bridge London

From this Bi-Gram graph the hotel Manager can understand that the strong points of his hotel are the great Location and the View.

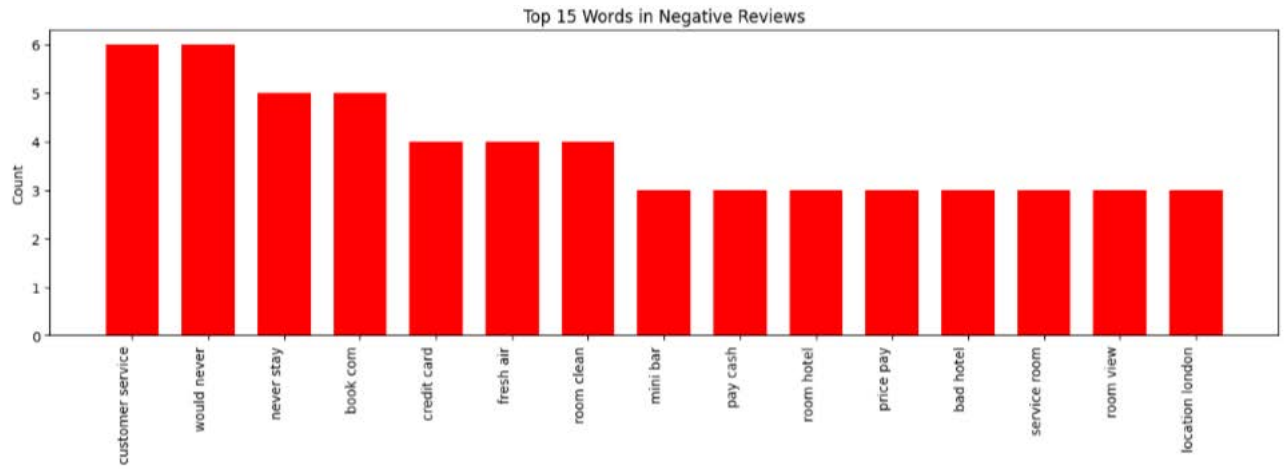
To increase the value of this hotel, the hotel manager should Set fair and dynamic prices for rooms with a view, Showcase the view in marketing materials and use emotional appeal. Create memorable and unique experiences related to the view.



The hotel's weaknesses are the Customer Service, Air Conditioning and Payment Methods.

For the Customer service the hotel manager should Conduct a staff training session on customer service skills and emphasize the importance of being respectful and helpful to guests, he can also use technology to enhance the customer service, such as customer service chatbots, contactless check-in, checkout, and guest messaging. These can make the guests' lives easier and reduce the workload of the staff.

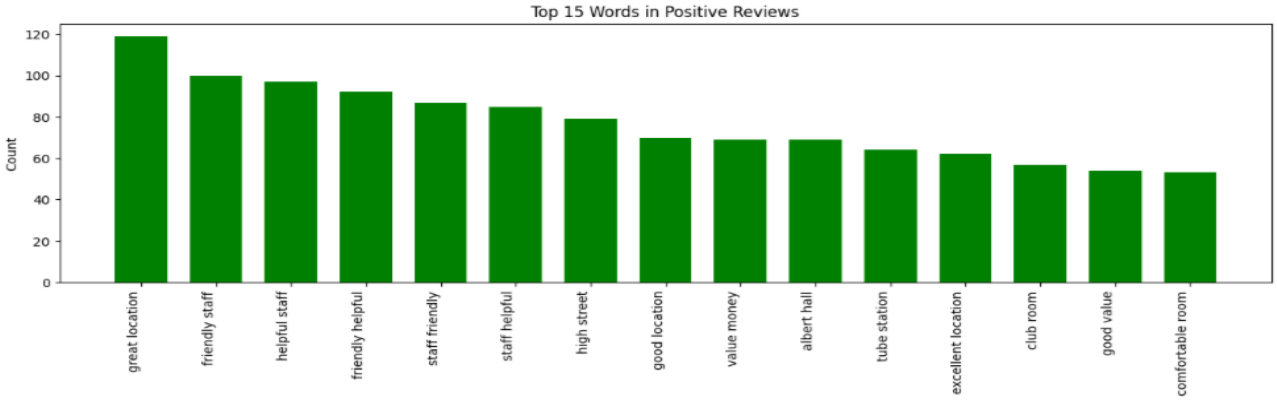
For the payment methods, the hotel manager can Offer different payment options to cater to the preferences and needs of different guests, such as credit card, debit card, cash, e-wallets, and mobile payments. Choose a reliable and secure payment gateway that can process various payment methods efficiently and safely. Inform the guests about the available payment options.



Copthorne Tara Hotel London Kensington

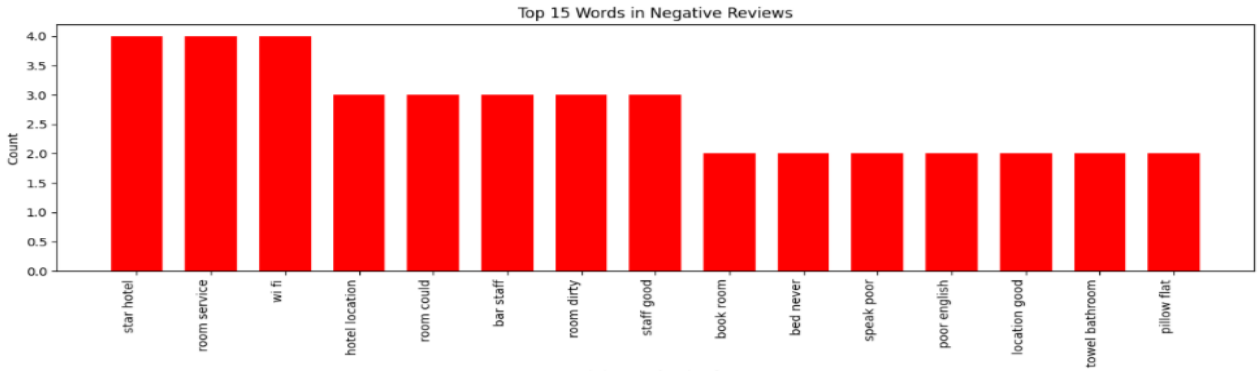
From this Bi-Gram graph the hotel Manager can understand that the strong point of his hotel are the great Position and the View, as the couples of words “albert hall”, “high street” and “tube station” suggest.

The hotel manager can establish a collaborative relationship with the Albert Hall, which is a renowned theater in London, and offer his guests the opportunity to purchase tickets for the theater plays at a lower price than the regular market value.



The hotel weaknesses are the Room Service, Wi-fi and Clean of Rooms.

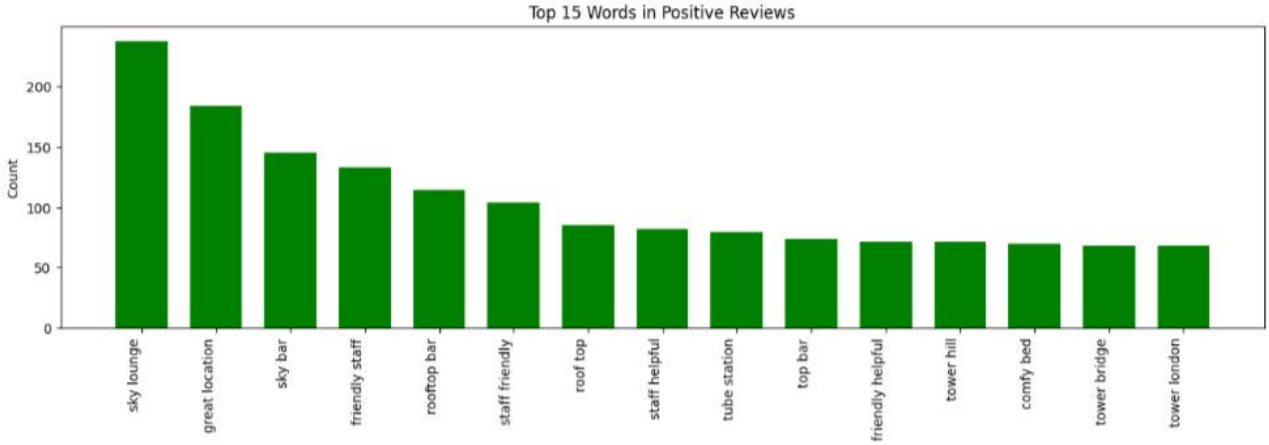
For the room service, the hotel manager can: Create a room service guide that includes the menu, prices, delivery times, payment options, and contact details. Keep the menu fresh and appealing and cater to common dietary requirements. Open feedback channels for guests to share their opinions and suggestions on room service.
For the Wi-Fi, the hotel manager can Inspect the Wi-Fi signal strength and coverage in different areas of the hotel and install additional access points or repeaters if needed. Call a professional technician to fix any technical problems.
Conduct regular maintenance on the Wi-Fi.
For the cleanliness of rooms, the hotel manager can Ensure that the housekeeping staff follow the best practices and standards of cleanliness and hygiene. Schedule and track cleaning of guest rooms and common areas using a service optimization solution that can identify who, when, and how often specific areas are cleaned. Solicit feedback from guests and inspect rooms regularly to ensure quality control.



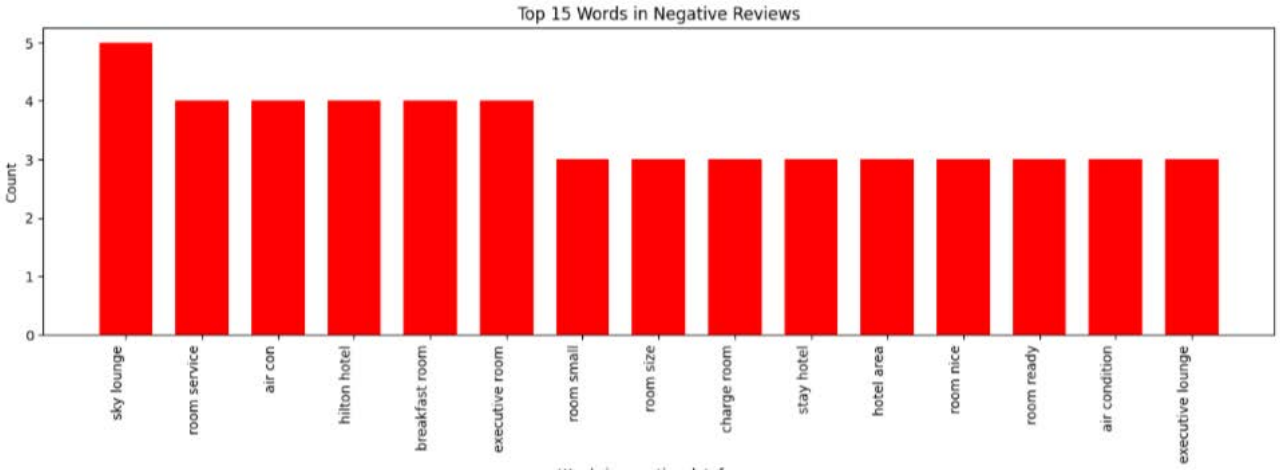
DoubleTree by Hilton Hotel London Tower of London

From this Bi-Gram graph the hotel Manager can understand that the strong points of his hotel are the great Location, Sky Lounge and the Rooftop.

To increase the value of a hotel with a sky lounge and a rooftop, the hotel should Set fair and dynamic prices for rooms with access to the facilities. Showcase the facilities in marketing materials and use emotional appeal. Provide excellent service and amenities to guests with access to the facilities.
Create memorable and unique experiences related to the facilities.



The hotel's weaknesses are the Room Service, Air Conditioning and Breakfast.



Hotels and AI

AI is a powerful technology that can improve the hotel industry in various ways. It can automate tasks, such as bookings and check-ins, that would otherwise take up a lot of time and resources. It can also personalize services, such as room preferences and recommendations, based on the guests' needs and preferences.

AI can also analyze data, such as guest feedback and online reviews, to predict market trends and customer behavior. This can help hotel managers and marketers optimize their revenue management and marketing strategies based on data-driven insights.

It increases efficiency and profits by reducing operational and administrative costs, improving energy management, and enhancing customer satisfaction and loyalty. AI can provide guests with a comfortable and convenient space that meets their expectations and enhances their experience.

Facial Recognition

One of the technologies used by the hotel is facial recognition. This technology uses AI to scan and identify the faces of guests and match them with their profiles. This can speed up the check-in process and enhance security. Some of the advantages of facial recognition are faster and more convenient check-in, improved security and verification, and personalized services based on guest profiles. The disadvantages are privacy and ethical concerns, potential errors or biases, and high cost and maintenance.

AI-ChatBots

Another technology used by the hotel brand is AI Chatbots. This technology uses AI to create a chatbot that can interact with guests and provide information and assistance. This can improve customer service and engagement. The advantages of AI-powered robot are 24/7 availability, friendly and consistent service, multilingual capabilities, and increased efficiency and productivity but they lack human touch and empathy, limited functionality and flexibility, high cost and maintenance, and possible technical glitches or malfunctions.

Conclusion

Reviews are a powerful tool to assess and form an opinion on a company, as they reflect the experiences and perspectives of both customers and businesses. This study aims to provide a valuable service for both parties, by using visualization and competitor analysis to evaluate the strengths and weaknesses of each hotel in the market. This way, customers can make an informed choice based on their preferences and expectations, and hotel managers can improve their services and facilities, as well as design more effective business strategies and advertising campaigns that highlight the advantages of their hotel.

Recent advances in AI technologies have the potential to increase the value of companies, but they also pose some challenges and risks. Depending on the target market of the company, some customers may not trust or adopt new technologies easily, so choosing AI solutions is not always the best option. It is necessary to evaluate each use case carefully and consider the human factors involved. That is why, at present, humans cannot rely entirely on AI totally, but need to work together with it to achieve the best outcomes.