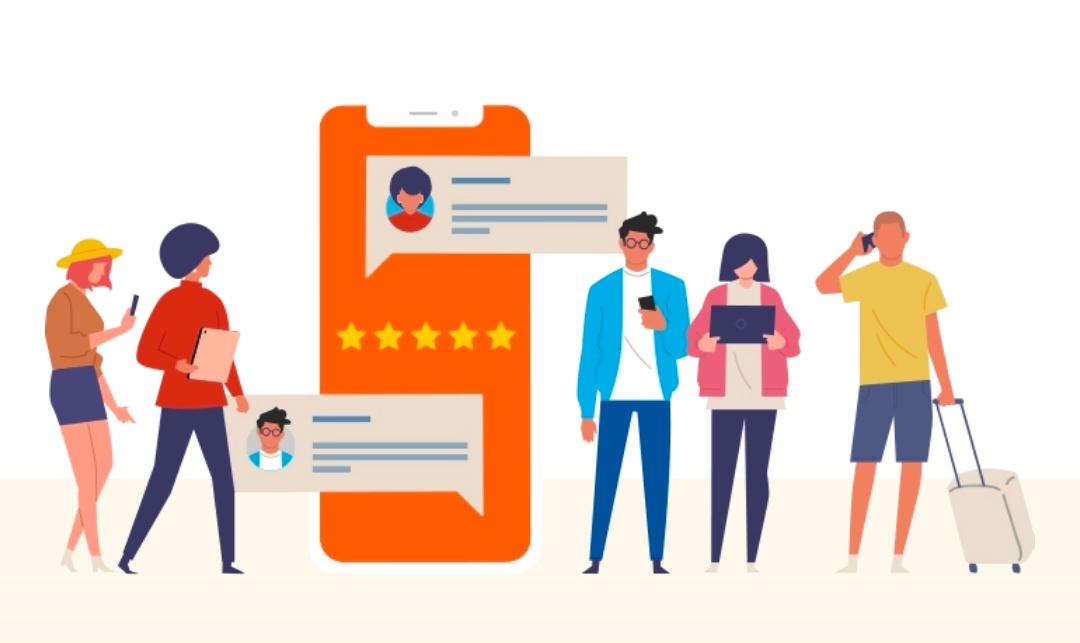
Text

Description automatically generated

**THE HOTEL REVIEW ANALYSIS**

**CAPSTONE PROJECT**

****

**Submitted By**

**Lalit Kumar (E21009)**

**Neha Afreen (E21010)**

**Vishwak Balaji (E21026)**

**Yogesh Babu (E21028)**

**Shameel Zeeshan (E21029)**

**IIIrd Term of Post-Graduation Programme**

**in**

**Data Science**

**Project Guide**

**Prof. Gourab Nath**

**Praxis Business School**

**(Bangalore Campus)**

**Year 2021 - 2022**

**ABSTRACT**

The hotel industry is an important part of the hospitality and tourism infrastructure and a strategic part of growth story. Hotels are primarily viewed as a service industry with intangible areas of guest experience and service levels. The research objective of this project is to better understand the hotel guest satisfaction and the areas that hotel management can change, in order to get better results. For this purpose, an analysis of hotel guest satisfaction ratings based on attributes such as Location, Sleep quality, Rooms, Service quality, Value for money and Cleanliness was performed. Further, text analysis of customer reviews was also performed to better understand the positive and negative sentiments of hotel guests. We focused on identifying the attributes that differentiate one hotel from another, and then using these attribute insights to make recommendation to hotel management, on how they can improve their operations, guest satisfaction and generally differentiate themselves from their competition. Data from an online website, MakeMyTrip and goibibo, was used to analyse and compare customer ratings and reviews on five hotels. Statistical data analysis techniques were used to identify the key attributes that are most important in choosing hotels and are critical to focus on in order to ensure guest satisfaction expectations are met. Based on text analytics, the key results from this study indicated that hotel guests look for a good room and a hotel with a pool and good service. Based on the ratings analysis, the most important attributes for guest satisfaction turned out to be Rooms, Value for money and Location.

**Keywords:** Makemytrip, Goibibo, text analytics

**ACKNOWLEDGEMENT**

We would like to express our special thanks of gratitude to our professor **Gourab Nath**, who gave us the golden opportunity to do this wonderful project on the topic **“The Hotel Review Analysis”,** which also helped us in doing a lot of research and we came to know about so many new things, despite of his busy schedules he gave us different ideas in making this project unique. We are really thankful to you sir.

We would also like to express our gratitude to all professors and faculty members of Data Science programme at **Praxis Business School**, that have thought us a lot of priceless and valuable knowledge in the field of Data Science.

Secondly, we would also like to thank our parents and friends who helped us in in gathering different information, collecting data and guiding us time to time in making this project and also helped us a lot in finalizing this project within the limited time frame. They have always been inspiration in life and have supported in every part of life.

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Fig. No.** | **Description** | **Page No.** |
| **2.1** | **snapshot of data extraction** | **2** |
| **5.1** | **snapshot of binary conversion** | **5** |
| **6.1** | **snapshot of association rule** | **6** |
| **7.1** | **snapshot of max support finding** | **7** |
| **8.1** | **snapshot of compactness pruning algorithm procedure** | **8** |
| **9.1** | **snapshot of p support** | **10** |
| **10.1** | **snapshot of polarity** | **10** |
| **11.1** | **snapshot of deployment page** | **11** |
| **11.2** | **comparison using Common feature** | **12** |
| **11.3** | **comparison using the features present in the hotel only** | **12** |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Description** | **Page No.** |
| **5.1** | **binary conversion** | **5** |

**Business Intuition**

**1. Business Landscape of Domain:**

The business domain of our project is Hotel Industry.The hotel industry is an important part of the hospitality and tourism infrastructure and a strategic part of growth story. Hotels are primarily viewed as a service industry with intangible areas of guest experience and service levels. The research objective of this project is to better understand the hotel guest satisfaction and the areas that hotel management can change, to get better results.

For this purpose, an analysis of hotel guest satisfaction ratings based on attributes such as Location, Sleep quality, Rooms, Service quality, Value for money and Cleanliness was performed. Further, text analysis of customer reviews was also performed to better understand the positive and negative sentiments of hotel guests. We focused on identifying the attributes that differentiate one hotel from another, and then using these attribute insights to make recommendation to hotel management, on how they can improve their operations, guest satisfaction, and generally differentiate themselves from their competition.

Data from an online website, MakeMyTrip and goibibo, was used to analyse and compare customer ratings and reviews on five hotels. Statistical data analysis techniques were used to identify the key attributes that are most important in choosing hotels and are critical to focus on to ensure guest satisfaction expectations are met. Based on text analytics, the key results from this study indicated that hotel guests look for a good room and a hotel with a pool and good service. Based on the ratings analysis, the most important attributes for guest satisfaction turned out to be Rooms, Value for money and Location.

**2. Business challenge we intend to solve:**

Now a days, most travellers use online sources to review hotels and other tourism operations during their decision-making process. Online reviews or the electronic word-of-mouth are at present a trustworthy source of information which travellers refer to for forming a better picture of the destination they plan to visit. Whenever a person wants to book any hotel very first thing, they do is to investigate the reviews given by the customers and accordingly make their decision. But one might look at 10-15 reviews, reading all the reviews is time consuming and tedious task.

With this project we intend to come up with the application which help customer by providing summary of all reviews, which is used to understand which hotel is better than the others. This project has tried to fill that gap and help the user select two hotels and compare those on his/her preferred features.

**3. The data that has been collected/used :**

The source of data is from data scraping & we majorly collected data from the popular website like MakeMyTrip, Goibibo & another website. We have used python, beautiful soup & primarily we have used tool called web scraper for scraping review, rating & hotel name. we have scraped around 60,000+ review from 80 + hotels.

**4. The solution to resolve the challenge :**

From Data we collected We are supposed to pre-process it for example remove punctuations, stop words, make every word as small letters and remove emoji’s etc. and then we take all nouns from that so that we can compare that nouns polarity in that sentence to any other polarity of the same word from other hotel.so once we take the nouns we do a binary conversion so that we can fit an association rule in it to get all the frequent repeating nouns and once we do that we take the maximum support of the duplicate nouns. And now we have all the nouns, and we must filter the useful ones for that we use two types of pruning compactness pruning and redundancy pruning. Compactness pruning is to find those two set nouns which are meaningful when you place in a sentence together. And redundancy pruning is used to take the single meaningful nouns by looking at how many time that nouns have repeated by itself so that we can conclude that those nouns when placed in a sentence has a meaning of its own. And once we get all the feature we must find the intensity of the polarity of the sentence where the feature is present. Like that we get all the polarity. Similarly, we do that with another hotel and just plot a side-by-side bar plot for all the polarity of the common feature present between those 2 hotels.by this way we can see that the user can select the feature which is essential to him and compare between 2 hotels so that he could decide which one is better.

**CONTENT**

**Abstract ii**

**Acknowledgement iii**

**List of Figures iv**

**List of Tables v**

**Business Intuitions vi**

Business Landscape of Domain………………………………………...vi

Business challenge we intend to solve………………………………….vi

The data that has been collected/used………………………………….vii

The solution to resolve the challenge…………………………………..vii

**1 Introduction 1**

**2 Data Extraction 2**

**3 NLP Pre-Processing 3**

**3.1** Procedure 1: …………………………………………………………………..3

**3.2** Procedure 2: …………………………………………………………………..3

**3.3** Procedure 3: …………………………………………………………………..3

**3.4** Procedure 4: …………………………………………………………………..3

**3.5** Procedure 5: …………………………………………………………………..4

**3.6** Procedure 6: …………………………………………………………………..4

**3.7** Procedure 7: …………………………………………………………………..4

**3.8** Procedure 8: ……………………………………………....…………………..4

**4 Why Nouns? 5**

**5 Binary Conversion 5**

**6 Association Rule 6**

**6.1** Association Rule Mining …………………………………………...………...6

**6.2** Support and Confidence ……………………………………………………...6

**7 Find Max Support 7**

**8 Compactness Pruning 8**

**9 Redundancy Pruning 10**

**10 Polarity 10**

**11 Results 11**

**12 Conclusion & Future Work 13**

**1. Introduction:**

Online reputation has become a major factor in choosing of a hotel. Hotel Reviews are gold mine of customer insights for any hotel businesses. Also, it's importance increases by many folds since majority of the future customers rely on peer reviews while finalizing theirs stay. In our project we have extracted ratings and reviews of the hotels based in Goa.

Now a days, most travellers use online sources to review hotels and other tourism operations during their decision-making process. Online reviews or the electronic word-of-mouth are at present a trustworthy source of information which travellers refer to for forming a better picture of the destination they plan to visit. Whenever a person wants to book any hotel very first thing, they does is to look into the reviews given by the customers and accordingly make their decision. But one might look at 10-15 reviews, reading all the reviews is time consuming and tedious task.

Customers generally give reviews about food, staff services, location, etc. So people who are more interested in good food, or want a good location can book the hotels accordingly. Consumers rely on several sources of information when deciding on purchase. These sources include personal recommendations from friends and family, company websites and other related communication materials.

With this project we intend to come up with the application which help customer by providing summary of all reviews, which is used to understand which hotel is better than the others.

In this project we are doing sentiment analysis of the reviews given by the customers in different platform like make my trip and goibibo

Sentiment analysis is the field that tries to give machine the ability to understand the emotions of the user. It uses NLP to determine whether data is positive, negative, or neutral. It is often performed on textual data to help business monitor brand and product sentiment in customer feedback and understand customer needs.

For example: - Suppose one customer have given review "Resort is great", from the word great we can figure out overall sentiment of this review is positive.

Our goal -

* Comparison of overall rating.
* Comparison of the review counts
* Analysis of frequently used words.
* Sentiment Analysis to find out positive and negative words.

Sentiment Analysis for overall scores for polarity and emotions.

**2. DATA EXTRACTION:**

Data extraction is the mean process in this project. There are various ways of data extraction are available, yet we able to do one of its types which is more suitable for us to scrape the data, we needed in smooth and easier manner. The tool is **‘web scraper’** and it is a chrome extension. Using this tool, we did scrape “Reviews, Ratings, Hotel Names” of the following hotels form the city called ‘Goa’. We scraped from different website’s such as **Makemytrip.com** and **Goibibo.com**. And collected **6047** entries, over **10** hotels with different time frame. Then, later we cleaned the retrieved data using python for the better understanding. In down below you can see the final piece of the data for the further process.

Text

Description automatically generated

**Fig:2.1 snapshot of data extraction**

And this data will further process by **‘NLP Pre-Processing’** method.

**3. NLP PRE-PROCESSING:**

This is an important step to do, whenever we are working with text data. In this text data you can see, there will always be a lot of un-important information present. And this pr-processing helps us to clean the noise/un-important information to work better in future process. This pre-processing was done by eight procedure, which we going to see briefly.

**3.1 Procedure 1: Remove Punctuation**

Punctuation will never give information and it is just denoted in a sentence to help us read, write and speak in efficient manner. And removing these will helps us to process better.

Eg: ‘Money’ 🡪 Money

**3.2 Procedure 2: Converting Text into Lower Case**

Mostly lower-case text gives a familiarity for the user to read and understand than upper-case text. And lower-case text gives more distinctive shape than upper-case. That is the reason why we are converting all the text into lower-case.

Eg: Money 🡪 money

**3.3 Procedure 3: Remove Extra-Space**

We are removing all the extra spaces present in the text corpus, except the single spaces between each word.

Eg: ‘ Worth for Money ‘ 🡪 ‘Worth for Money’

**3.4 Procedure 4: Remove Emoji**

We are removing all the emoji present in the text corpus because emoji’s will be giving an emotional and meaningful message in some places, but not in all the places. That’s the reason why we are removing the emoji.

Eg: The resort is very good 😍 🡪 The resort is very good

Breakfast is not available 😇 🡪 Breakfast is not available

**3.5 Procedure 5: Tokenization**

Tokenization is **essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units**, such as individual words or terms. Each of these smaller units are called tokens. Here we did sentence tokenization to split the text corpus.

Eg: The resort is good, breakfast is good. Staff service is too slow 🡪 [[ The resort is good breakfast is good], [Staff service is too slow]]

**3.6 Procedure 6: Remove Less Than Two Words**

In this text corpus two letter words and single letter words doesn’t give a meaning, when they were considered separately. So, we are removing the two letter words and single letter words.

Eg: The resort is good 🡪 The resort good

**3.7 Procedure 7: Extracting Common Noun Tags**

This extraction is one of the most important because in this project we are considering only common nouns as features and further process will be continued with the help of this common noun.

Eg: The resort is very good *🡪 resort (Common Noun)*

**3.8 Procedure 8: Converting Text into Singular Text**

Here we are trying to convert the text corpus into a single form, which is singular.

Eg: Houses 🡪 House, Cities 🡪 City

**4. WHY NOUNS?**

Nouns are the probable features we are assuming that most frequently occurring nouns are candidate features and will be used for sentiment analysis. We are extracting nouns using POS(Part of Speech) tagger.

To extract the frequently occurring nouns and nouns phrases we use association Rule Mining algorithm or apriori algorithm to mine this association.

Eg: N N N N

The **resort** is very good. **Breakfast** is fine. No **response** from the **reception** team.

**5. BINARY CONVERSION:**

After extracting the nouns from the reviews, the feature (noun words) is then converted into binary matrix form representing the occurrence of that noun in that each review. The binary converted nouns are then used for association rule mining to understand the relationship of two or more features(nouns).

Eg: N N N N

The **resort** is very good. **Breakfast** is fine. No **response** from the **reception** team.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Resort** | **Breakfast** | **Response** | **Reception** |
| **The resort is very good** | **1** | **0** | **0** | **0** |
| **Breakfast is fine** | **0** | **1** | **0** | **0** |
| **No response from the reception team** | **0** | **0** | **1** | **1** |

**Table: 5.1 binary conversion**

**A picture containing text, light, wire, screenshot

Description automatically generated**

**Fig: 5.1 snapshot of binary conversion**

**6. ASSOCIATION RULE:**

* **6.1 Association Rule Mining:**

Association rule mining is the data mining process of finding the rules that may govern associations and causal objects between sets of items. So, in each transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together.

Table

Description automatically generated

**Fig: 6.1 snapshot of association rule**

* **6.2 Support and confidence:**

Support is an indication of how frequently the items appear in the data. Confidence indicates the number of times the if-then statements are found true. We are using association rule mining to understand which word or features in the reviews are coming together and more frequently. For our case we have filtered out the rules by considering support value of 0.03 i.e 30%. We have rules of greater length also, but we are restricting ourselves to rules of less than 2 words.

**7. FINDING THE MAX SUPPORT:**

Here we are trying to find the max support value for different feature as well as, we are trying to remove duplicate values like (resort, room / room, resort) Which were present in antecedents and consequents. This process helps us to process further for the betterment of the text corpus we are having.

Table

Description automatically generated

**Fig: 7.1 snapshot of max support finding**

**8. Compactness Pruning:**

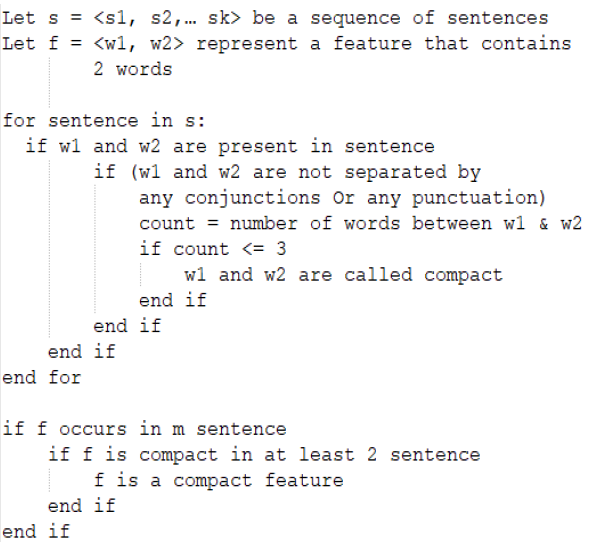
From max support we get all the frequent item set, but not every item set can be considered as a feature, so we use two types of pruning to remove the redundant nouns.

There are feature which has two words in it and few of them won’t make sense when placed together in a sentence, so to remove those kind of feature we can use compactness pruning. You can consider a feature compact when both the words are separated in a sentence by at most 3 words.

Now let’s see how the algorithm works from the figurer below.

Let’s take s to be a sentence and I have to check for 2 words if they are compact or not in that sentence.so first I will check whether those words are present in that sentence or not and once we confirm they are in the sentence we find the order of the words placed by looking their index as the order of the words are important. After that we calculate the number of words between them, if the number of words between them is less than 4 then we can call it a compact feature.

But we can’t stop the process just for checking it in one sentence we have to check it for every sentence and it should satisfy for at least 2 sentence. If it does we can conclude that It is a compact feature



**Fig: 8.1 snapshot of compactness pruning algorithm procedure**

**For Example:**

**Feature: [Hotel, Room]**

‘Only the room cleaning uncle was good rather than other all of the staff was a bit annoyed I know on what but the room and facility was awesome good pool nice hotel clean room beautiful room.’ - Compact Feature

‘We reached hotel at midnight nice room but bad smell inside the room.’ – Compact Feature

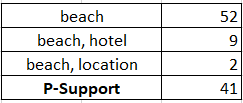
‘This hotel has nice big spacious room.’ – Not Compact Feature

**9. Redundancy Pruning**

This method is used to eliminate those single feature who has no meaning in a sentence by itself. This is done by calculating the p-support of the feature.

P-support is the number of the feature repeats in a paragraph by itself without considering the superset of the feature.

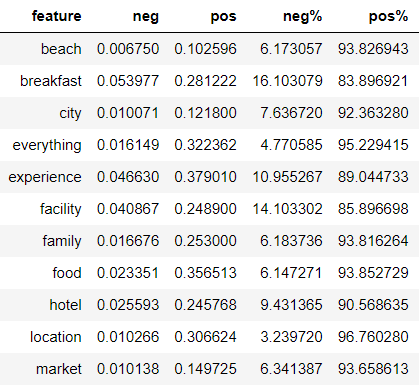
For Example For The feature Beach We can see the count of the subsets of ‘Beach’ , and the total count of beach in the reviews.so the P-support of the feature beach would be, (52-(9+2)).



**Fig: 9.1 snapshot of p support**

**10. Polarity:**

Now we have all the Features, so our next step is to get the semantic orientation of each opinion word. We focussed both on the polarity and the strength of the polarity. For that we have used the VADER (Valence Aware Dictionary and Sentiment Reasoning) .sentiment analysis tool. This tool returns us the intensity of the polarity strength as value between -1 to +1.



**Fig: 10.1 snapshot of polarity**

**11. Results:**

Now we have the strength of the polarity for each feature, therefore we can compare the feature which are common in both the hotel. First you must create a data base for each hotel and find the polarity of every feature using the above algorithms and once you do that give the user the option to select the two hotels to compare.



**Fig: 11.1 snapshot of deployment page**

Once the user selects the two hotel you should find the common features from the hotel and compare their polarity using side by side bar plot. Which gives the user to compare their priority feature by themselves and choose the hotel according to it.

The user should also know about the feature which are only present in the specific hotel so we have to take that feature which is only present in that hotel and display their strength of polarity as well this will help the user to see if there some feature which are provided in the hotel than the other one.

Below figures are an example of comparison of two hotels.

Fig11.2 – comparison using Common feature

Fig 11.3 – comparison using the features present in the hotel only

**Chart, bar chart

Description automatically generated**

**Fig: 11.2 comparison using Common feature**

Chart, bar chart

Description automatically generated

**Fig: 11.3 comparison using the features present in the hotel only**

**12. Conclusion & Future Work:**

* **Conclusion:**

In world which is fast paced and rapidly changing in terms of how we use technology brings us to a point where a person can not go through hundreds of reviews for different hotels to compare and understand which hotel is better. This project has tried to fill that gap and help the user select two hotels and compare those on his/her preferred features.

* **Future Work:**
  + - * + In this Project all we did is to consider the single feature nouns to compare with all other reviews present in the review corpus. But in the feature, we can use double feature nouns to compare in the review corpus to get even more better results.
        + In this project we are only able to compare between two hotels. But in the future, we can also try with more than two hotels to get even more better/greater suggestion to the customer.
        + Also, we can use Clustering analysis to find or define the feature comparison into a separate corpus.