April 25, 2022

**Data Science Project**

**Final Report**

Somnath Solaskar|

Ankita Indi

**Customer Review Sentimental Analysis**

# **1.BACKGROUND**

## 1.1 What is Sentimental Analysis?

***“Sentimental Analysis is text mining that finds and extracts subjective information from source material allowing a company to better understand social sentiments of its brand, product or service while monitoring online conversations.”***

On the internet, user reviews and comments about hotels are utilized to plan visits.

As a result, hotel management must have access to this information. We usually introduce a platform that collects user comments, organizes, and categorizes them, and makes data easier to access. Nowadays, a commercial concern or a service-based corporation requires consumer input. The expansion of business necessitates the addition of a wider range of services and solutions.

As a result, the company should be concerned about the evaluations and ratings provided by its customers to expand its business, as they are required for a wider range of services and goods. On online-portals, service-users will express their feelings and provide reviews. We can anticipate the organization's ranking by using opinion mining and sentiment analysis on these details. To generate the ratings in an exact and right manner, one recommender system is necessary. For a hotel business, reviews on a variety of topics such as cleanliness, maintenance, behavior, food, hospitality, room neatness, and hotel employee response, among others, play an important role.

## 1.2 Proposed Solution

## It is feasible to make predictions using simply raw text as input. Extracting meaningful features from this raw data is the most critical step. This is frequently a useful supplement to our data projects, allowing us to extract more learning features and improve the predictive power of our models.

The goal of sentimental analysis is to determine if a review is good or negative. Opinion words, which indicate whether the assessment is good, negative, or neutral, are the most essential markers of sentimental analysis.

We want to know whether each textual review correlates to a positive (satisfied) or negative (dissatisfied) review (the customer is not satisfied).

To make the problem easier to understand, we'll divide them into two categories:

* Overall, negative evaluations have a rating of less than five stars.
* Overall ratings of >= 5 indicate that the review is good.

The goal is to be able to forecast this information using only the review's raw textual data.

Based on above analysis we can perform below operations to make the analysis more informative.

* Analyze the reviews to see how they are feeling.
* Find a link between the nationality of the reviewers and their scores.
* On the dataset, there is a lovely and useful graphic.
* A simple suggestion engine for guests who enjoy a particular hotel feature.

# **2.DATASET DESCRIPTION**

This dataset includes 515,000 customer rating and textual reviews collected from 1493 luxury hotels throughout Europe. The geographic location is provided for additional investigation.

* There are total 17 fields available in the dataset below:
* **Hotel\_Address:** Address of hotel.
* **Review\_Date:** Date when the matching review was posted by the reviewer.
* **Average\_Score:** The hotel's average score is based on the most recent comment from the previous year..
* **Hotel\_Name:** Name of Hotel
* **Reviewer\_Nationality:** Nationality of Reviewer
* **Negative\_Review:** The hotel received a negative evaluation from the reviewer. 'No Negative' should be used if the reviewer does not give a negative review.
* **Review\_Total\_Negative\_Word\_Counts:** Total number of words in the negative review.
* **Positive\_Review:** The hotel received a positive evaluation from the reviewer. 'No Positive' should be used if the reviewer does not provide a negative evaluation.
* **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 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:** Some guests simply gave a score for the service rather than writing a review. This figure reflects the number of valid scores in the system that haven't been reviewed.
* **lat:** Latitude of the hotel
* **lng:** longtitude of the hotel

**2.1 DATASET LINK**

<https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe>

# **3. DATA PRE PROCESSING**

**3.1 Data Cleaning**

We clean the data using the below steps.

We call our function call data which performs a few operations -

* removing of text garbage like numbers, stopples words
* text tokenization and punctation
* lowering text
* Part of Speech tagging
* removing word with one letter
* Tokenize text and remove punctuation
* Remove words that contain numbers
* Remove empty tokens
* Remove words with only one letter

**3.2 Feature Engineering**

We begin by including sentiment analysis elements since we believe that customer reviews are strongly related to how they felt about their hotel stay. We use Vader, which is a sentiment analysis component of the NLTK module. Vader consults a dictionary of words to determine which are positives and which are negatives. It also considers the context of the statements while calculating sentiment scores. Vader returns four values for each text:

* a positive score a neutrality score
* a score for negativity
* a total score that combines all of the preceding score

We will integrate those 4 values as features in our dataset

# **4. EXPLORATORY DATA ANALYSIS**

In this step, the dataset was analyzed to perform the following:

* Distribution of negative reviews
* Average review scores for hotels
* Reviewer’s nationality and Reviewer’s score
* Review sentiment distribution (Positive and negative word count trends)
* Word cloud

**Evaluation Parameters**

1. **AUC Receiver Operating Characteristic** – Performance measurement for the classification problems at various threshold settings.
2. **Average Precision (AP)** – It is a popular metric in measuring the accuracy of object detectors. The mean Average Precision score is calculated by taking mean AP over all classes or overall thresholds.

## 3.1 Distribution of negative reviews

Here we are doing EDA on the complete to check how many good or bad reviews we have. As per the output of below code snippet, we can see that the “Positive reviews” holds approximately 95% of the total reviews and “Negative reviews” holds 5% of it.

Text

Description automatically generated with low confidence

Code snippet – Distribution of Negative reviews

Chart, pie chart

Description automatically generated

*.* Distribution of Negative reviews

## 3.2 Average review scores for hotels

Average review scores for hotels depicts the hotel count, and the average scores for the reviews received. Only 0.1% of the data is considered to generate the output. we are also dropping the duplicated to since the average taken for each column would come same with both the datasets.

Graphical user interface, text, application

Description automatically generated

Code snippet – Average review scores

## Chart, bar chart Description automatically generated

Average review scores

## 3.3 Normalizing and scaling numerical variables

Reviewer’s nationality and Reviewer’s score,we are trying to implement the co-relation between the nationality of reviewer and the scores provided by them. We have taken 20 sample records to implement this.

A picture containing text

Description automatically generated

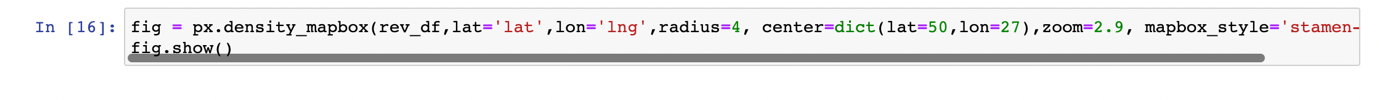
Code output – Reviews per nationality

Chart, bar chart

Description automatically generated

Reviews per nationality

## 3.4 Location on World Map



Code Output-Location on World Map

Map

Description automatically generated

Location on World Map

## 3.5 Reviews sentiment distribution

Here we are visualizing the distribution of total negative word counts vs total positive word counts using scatter plot.

Text

Description automatically generated

Code Output: Sentiment Distribution

Chart, scatter chart

Description automatically generated

Scatter Plot:Sentiment Distribution

## 3.6 Generating word cloud

Here based on the positive and negative reviews received we are finding the words of both responses that help us improve our service using word cloud

Graphical user interface, text, application

Description automatically generated

Code Output: Word cloud

Text

Description automatically generated

Word cloud for Positive Reviews

## 3.6 Highest positive and negative sentiment reviews

Here we display the positive sentiment for positive and negative sentiment (with more than 5 words). The most positive ones do correspond to the pleasant feedback.

Graphical user interface, text, application

Description automatically generated

Highest positive sentiment reviews

Chart, line chart

Description automatically generated

Sentiment distribution for positive and negative reviews

**Text

Description automatically generated**

Sentiment distribution for positive and negative reviews

# **4 MODEL DEVELOPMENT**

## 4.1 Random Forest

The random forest classifier is a supervised learning method for dealing with regression and classification problems. It is one of the most widely used machine learning algorithms due to its high flexibility and ease of implementation.

The False Positive Rate and True Positive Rate of the reviews received are used to assess the accuracy of random forest classifiers.

Compound, pos,doc2vec vector 3, doc2vec vector 1, and the rest of the features listed in the code are a few of the significant features by which we have evaluated accuracy.

Graphical user interface, text, application

Description automatically generated

Code Snippet: Random forest Accuracy classifer

## 4.2 Logistic Regression

A supervised classification approach is logistic regression. For a given collection of features, the goal variable (or output), y, can only take discrete values in a classification operation (or inputs)

We examine the reviews received in Logistic Regression reviews for accuracy, False Positive Rate, and True Positive Rate.

Graphical user interface, text, application

Description automatically generated

*Code snippet Logistic Regression*

## 4.3 XGBoost

**In XGBoost** reviews we evaluate accuracy, the False Positive Rate, True Positive Rate of the reviews received

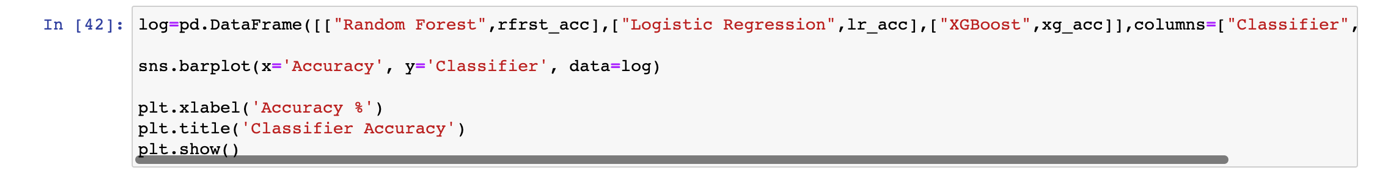
Graphical user interface, text, application, email

Description automatically generated

*Code Snippet:XGBoost*

### 4.4 Comparing Model Accuracy

Here ,we Compare the Model Accuracy for Random forest, Logistic Regression and XGBoost based on the important features .



*Code snippet: Model Accuracy for Random forest,Logistic Regrssion,XGBoost*

Chart, bar chart

Description automatically generated

*Model Accuracy for Random forest,Logistic Regrssion,XGBoost*

# **5. CONCLUSION**

It is feasible to make predictions using simply raw text as input. Extracting meaningful features from this raw data is the most critical step. This is frequently a useful supplement to our data projects, allowing us to extract more learning features and improve the predictive power of our models.

# **6. REFERENCES**

<https://towardsdatascience.com/a-complete-sentiment-analysis-project-using-pythons-scikit-learn-b9ccbb0405c2>

<https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe>