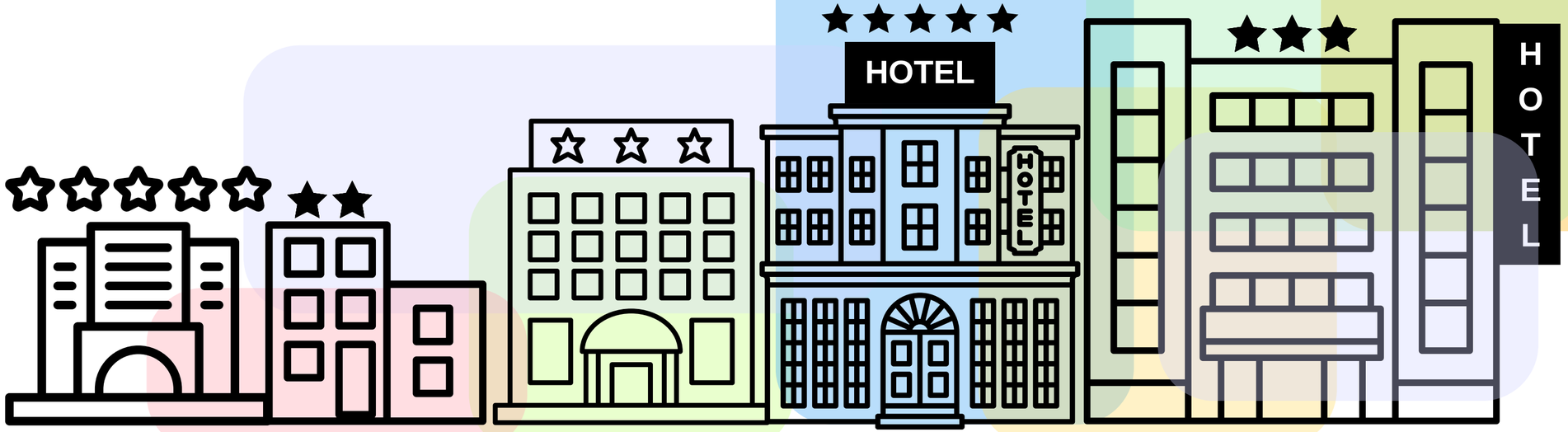
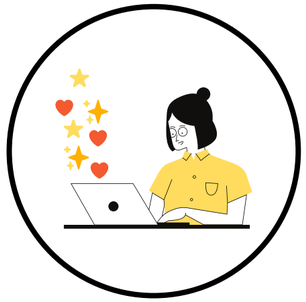
HOTEL REVIEW SENTIMENT ANALYSIS

Laksmi Amalia Wulandiari



Laksmi Amalia Wulandiari

CONTENT

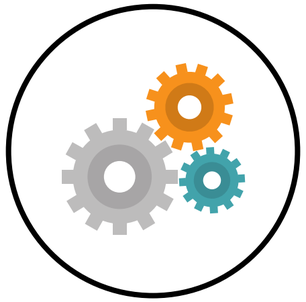


SENTIMENT ANALYSIS

NLTK

DATASET EXPLANATION

EXPLORATORY DATA ANALYSIS



MACHINE LEARNING MODEL

Laksmi Amalia Wulandiari

SENTIMENT ANALYSIS

What is Sentiment Analysis?

Sentiment analysis is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information, widely applied to voice of the customer materials such as reviews and survey responses typically express their opinion or sentiment.

source : https://en.wikipedia.org/wiki/Sentiment_analysis

Laksmi Amalia Wulandiari

NLTK – Natural Language Toolkit

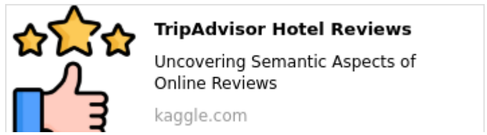
NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

Natural Language Processing with Python provides a practical introduction to programming for language processing. It guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more.

source : https://www.nltk.org/

Laksmi Amalia Wulandiari

DATASET EXPLANATION



Review Column

Rating Column

Total row : 20491
Null values : None
Duplicated values : None

Laksmi Amalia Wulandiari

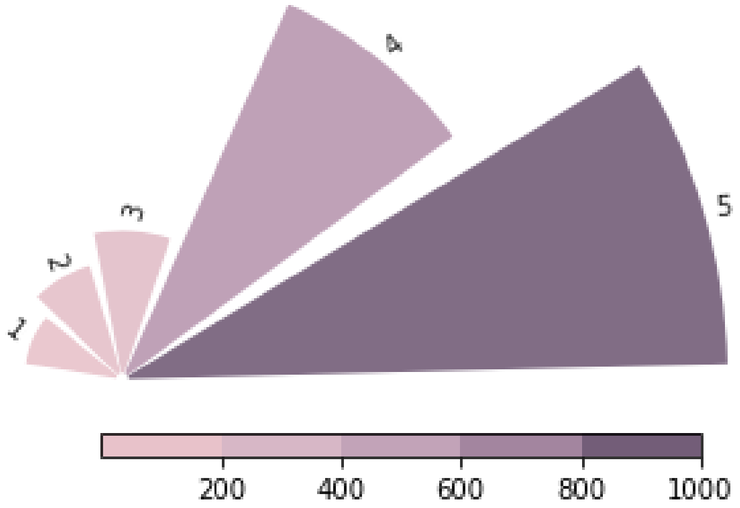
EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing.

source : https://en.wikipedia.org/wiki/Exploratory_data_analysis

Both reviews are dominated with words like **hotel**, **room**, **stay**, **time**. But, in **Positive Reviews** there are much more **positive** and **constructive** words such as **excellent**, **fantastic**, **beautiful**, **wonderful**, and **love**. While **Negative Reviews** contains neutral to **negative words**, such as **bad**, **problem**, **small**, **service**, and **leave**. It means that even though both reviews mostly consist of the **same words**, but it has **different meaning**.

Rating Proportion



Rating proportion is dominated with **5-star** review with almost surpassing 1000 reviews. **5-star** review means that the hotel **providing good services** or **products** to customers. Whilst for **4-star** review means customer/reviewer **enjoyed the service** but may point out some areas for **improvement**, so it is **not a good thing**. For **3-star to 1-star** review, we can considered it as a **bad review**.

Based on above explanation, we can categorize it :

Good Review : ★★★★★

Bad Review : ★
★★
★★★
★★★★

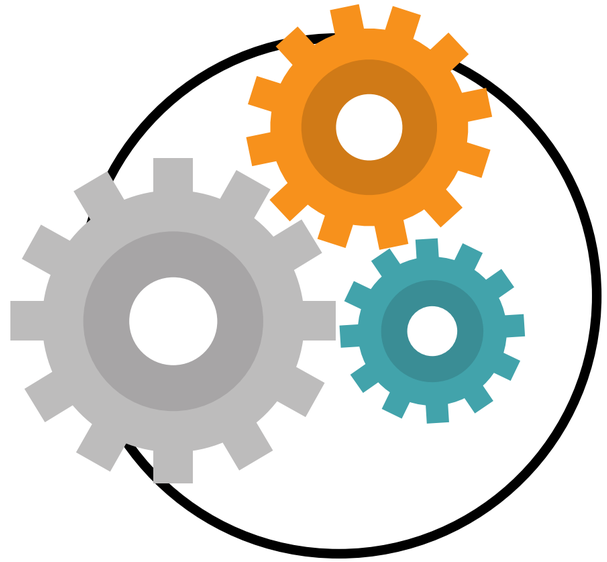
EXPLORATORY DATA ANALYSIS

	Clean_Review	pos
4073	hotel great hotel great money clean good restu...	0.799
2346	hotel hotel gorgeous beautiful clean spacious ...	0.763
12988	great stayed clarendon great stay employee con...	0.746
14475	great star hotel great room pretty small clean...	0.741
5566	excellent great staff glad choose hotel conven...	0.738
18668	fabulous hotel great hotel room amaze large gr...	0.733
8937	excellent great hotel recommend couple perfect...	0.721
11753	adore place clean spacious room extremely nice...	0.717
8122	love regina perfect great location lovely clea...	0.711
12044	lovely room great location stay lovely room sp...	0.709

	Clean_Review	neg
9514	bad hotel awful place dirty room rude staff de...	0.646
19891	bad bad bad hotel verry bad pls stay close ram...	0.645
42	warwick bad good review warwick shock staff ru...	0.458
14744	stay bad resort stay day day long food poison ...	0.447
12123	awful hotel star hotel star room big dirty wal...	0.431
10952	hotel star hell star venue depress calamity ar...	0.431
2408	horrify stayed hotel october filthy terrible f...	0.424
6433	disappointing disappointed stay benjamin origi...	0.399
3979	place pretty bad room damp dirty water line ru...	0.396
3470	bad spa day spend huge family thing steal room...	0.391

Positive Reviews, dominated with words that show high customer satisfaction, such as **great**, excellent, fabulous, and adore. While Negative Reviews dominated with words that show low customer satisfaction, such as bad, horrify, awful, and disappointing. It show that both describing HOTEL but with different adjective that show their satisfaction level.

MACHINE LEARNING MODEL



Machine learning (ML) is a field of inquiry devoted to understanding and building methods that "learn" – that is, methods that leverage data to improve performance on some set of tasks. It is seen as a part of artificial intelligence.

Random Forest commonly-used machine learning algorithm, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

Logistic Regression Method to predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. Here we are using logistic regression for the effective accuracy and the prediction of the data set.

Random Forest

Classification Decision:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	288
1	1.00	0.98	0.99	225
accuracy			0.99	513
macro avg	0.99	0.99	0.99	513
weighted avg	0.99	0.99	0.99	513

This model able to achieve **97%** precision for negative reviews, it means that **97%** negative reviews are correctly classified as negative reviews, while **3%** of negative reviews are classified incorrectly. For positive reviews, it achieved **100%**, it means that all positive reviews are classified as positive reviews and none of the positive reviews are classified incorrectly.

Recall value or True Positive Rate (TPR), predicted how much a machine learning model correctly identifies all relevant cases withing the dataset. In this case, our model able to predicted that **100%** from the dataset are **True Positive** for negative reviews, and **96%** from the dataset are **True Positive** for positive reviews.

F1-score value measures how many times the model made a correct prediction accurately. Our logistic regression model able to predict **99%** correctly for negative reviews and **98%** correctly for positive reviews.

Logistic Regression

Classification Decision:

	precision	recall	f1-score	support
0	1.00	0.99	0.99	350
1	0.99	1.00	0.99	265
accuracy			0.99	615
macro avg	0.99	0.99	0.99	615
weighted avg	0.99	0.99	0.99	615

Precision value equal to **100%** for **0**, means that all negative reviews are classified as negative reviews and none of the negative reviews are classified incorrectly. While precision value for positive reviews equal to **99%**, it means that **99%** of positive reviews are classified as positive reviews and **1%** of the positive reviews are classified incorrectly.

Recall value or True Positive Rate (TPR), predicted how much a machine learning model correctly identifies all relevant cases withing the dataset. In this case, our model able to predicted that **99%** from the dataset are **True Positive** for negative reviews, and **100%** from the dataset are **True Positive** for positive reviews.

F1-score value measures how many times the model made a correct prediction accurately. Our logistic regression model able to predict **100%** correctly for negative reviews and **99%** correctly for positive reviews.

Laksmi Amalia Wulandiari

Thank You

amalia.wulandiari@gmail.com
https://www.linkedin.com/in/laksmiamalia
https://www.github.com/laksmiamalia