HOTEL REVIEWS

Estrazione di Insight di Business tramite Sentiment Analysis

Data Analytics

Anno Accademico 2022-2023

Urbani Nicolò 856213
Mohamed Nada 857606
Rubini Alessia 851890

SCENARIO DI ESEMPIO

Fonte www.booking.com

IMPATTO DI UNA RECENSIONE

REPUTAZIONE DELL'HOTEL

POSIZIONAMENTO ONLINE

DECISIONE DI PRENOTAZIONE

FIDUCIA DEL CLIENTE

OBIETTIVO

Sentiment Analysis su Dataset HotelReview al fine di estrarre **insights** per prendere decisioni **data-driven**

SCENARIO BUSINESS

Quali aspetti devono essere migliorati ? Quali sono i punti di forza della struttura?

La clientela di **diversa nazionalità** ha diverse preferenze?

SCENARIO CUSTOMER

Il cliente sta esprimendo correttamente quanto vuole intendere riguardo all'hotel?

OBIETTIVO COMUNE

Gli aspetti citati nella recensione sono veramente **coerenti** con quanto espresso nella recensione?
Lo **score è discordante** con quanto espresso nella recensione testuale?

DATASET HOTEL REVIEW - EUROPE

Hotel di lusso in Europa (2015-2017)

17 FEATURES

213 NAZIONI

1492 HOTEL

515,000K

Recensioni in INGLESE

Estratte da BOOKING.COM

515K Hotel Reviews Data in Europe
https:// www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe

DATASET

L' HILTON HOTEL LONDON viene escluso dal train e dal test per utilizzarlo nella DASH al fine di valutare le performance

SCORE 1 (BAD) - 10 (EXCELLENT)

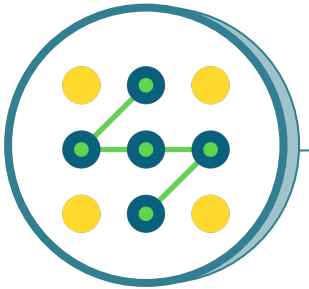
PUNTEGGIATURA ASSENTE

	Hotel_Name	Average_Score	Reviewer_Score	Negative_Review	Positive_Review	Reviewer_Nationality
0	Hotel Arena	7.7	2.9	I am so angry that i made this post available...	Only the park outside of the hotel was beauti...	Russia
1	Hotel Arena	7.7	7.5	No Negative	No real complaints the hotel was great great ...	Ireland
2	Hotel Arena	7.7	7.1	Rooms are nice but for elderly a bit difficul...	Location was good and staff were ok It is cut...	Australia
3	Hotel Arena	7.7	3.8	My room was dirty and I was afraid to walk ba...	Great location in nice surroundings the bar a...	United Kingdom
4	Hotel Arena	7.7	6.7	You When I booked with your company on line y...	Amazing location and building Romantic setting	New Zealand

Utilizzate per Sentiment Analysis

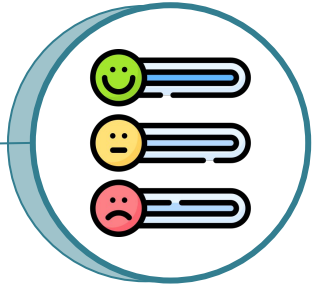
LA SENTIMENT ANALYSIS È UTILE?

FINE-GRADE UNDERSTANDING



INDIVIDUARE PATTERN NASCOSTI

ANALISI RECENSIONI



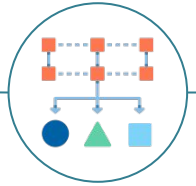
REAL TIME

PROJECT STEP

ANALISI INIZIALE

PREPROCESSING

VETTORIZZAZIONE



MODELLI DI CLASSIFICAZIONE

MODEL EXPLANATION

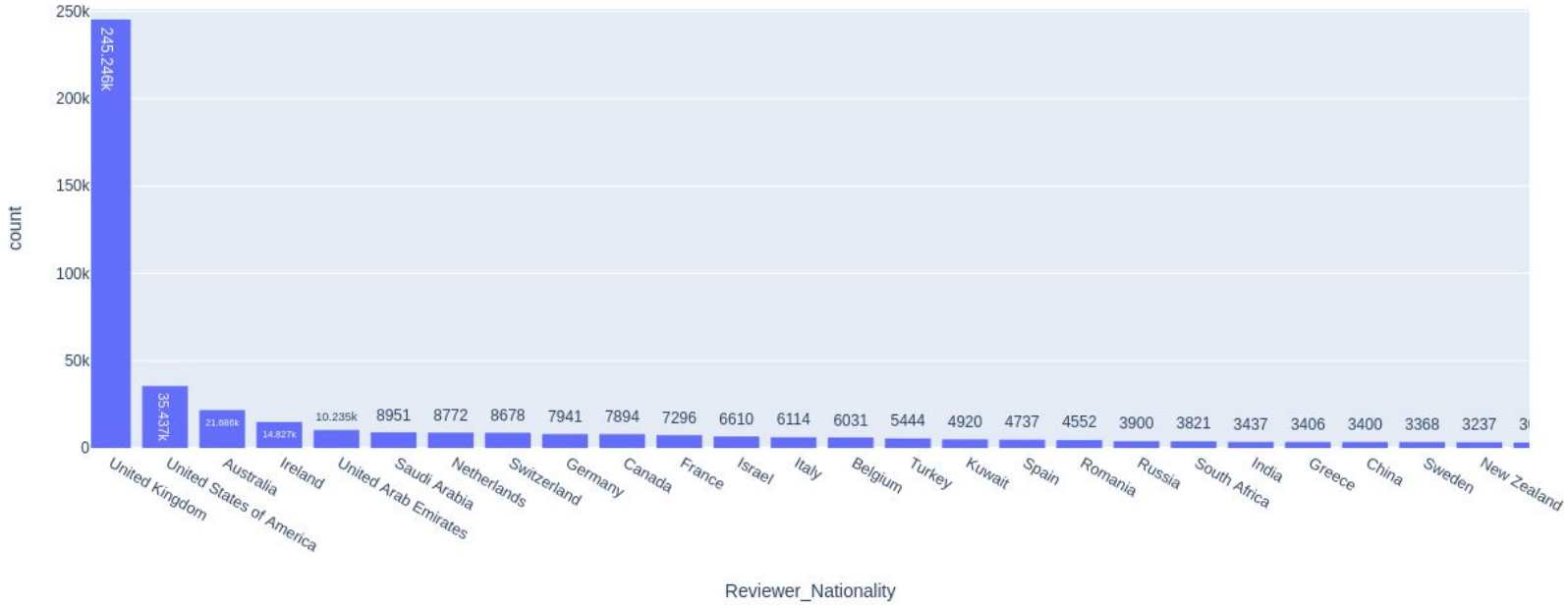
DASHBOARD

01

ANALISI ESPLORATIVA

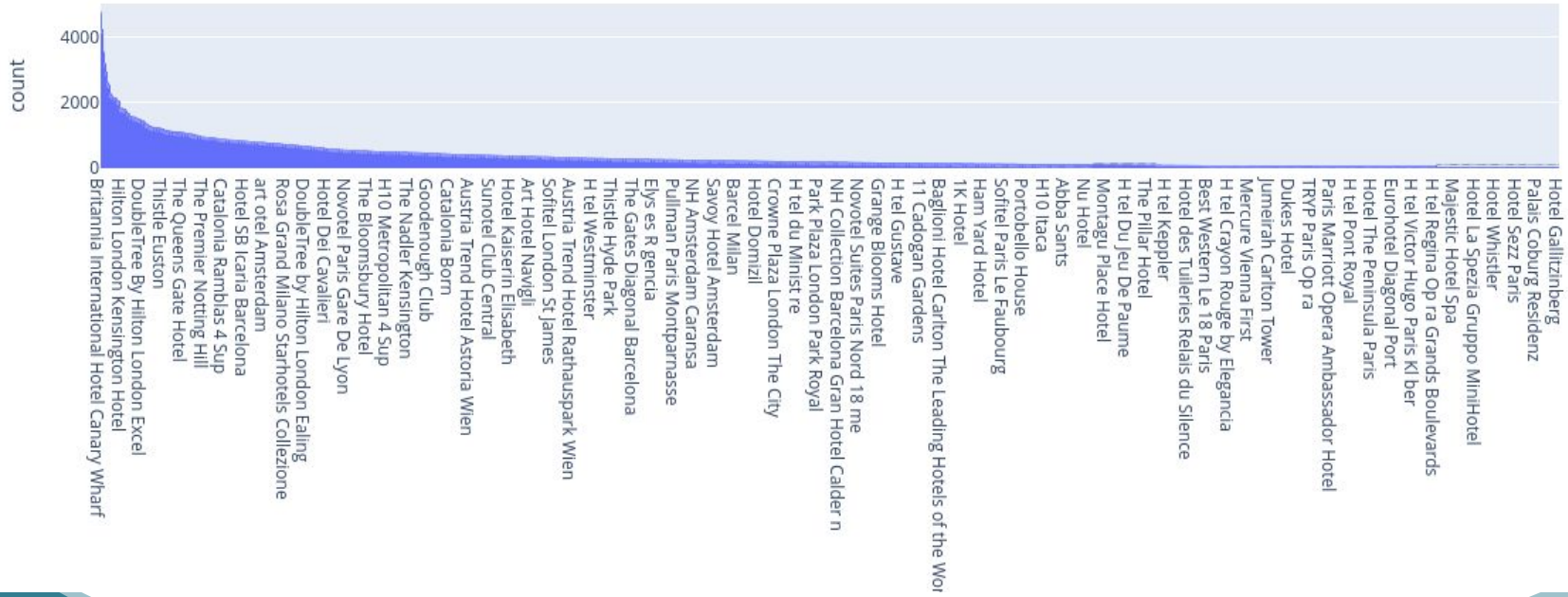
ANALISI ESPLORATIVA

DISTRIBUZIONE NAZIONALITÀ: **227 nazionalità diverse**



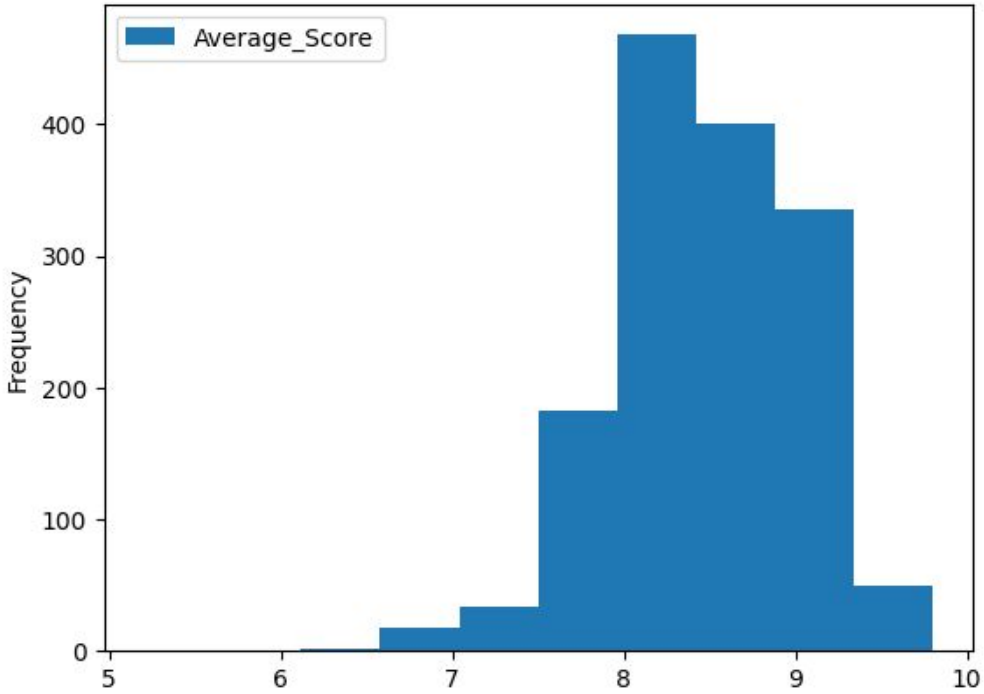
ANALISI ESPLORATIVA

DISTRIBUZIONE HOTEL: **1492 Hotel di Lusso in Europa**



ANALISI ESPLORATIVA

Distribuzione Reviewer Score

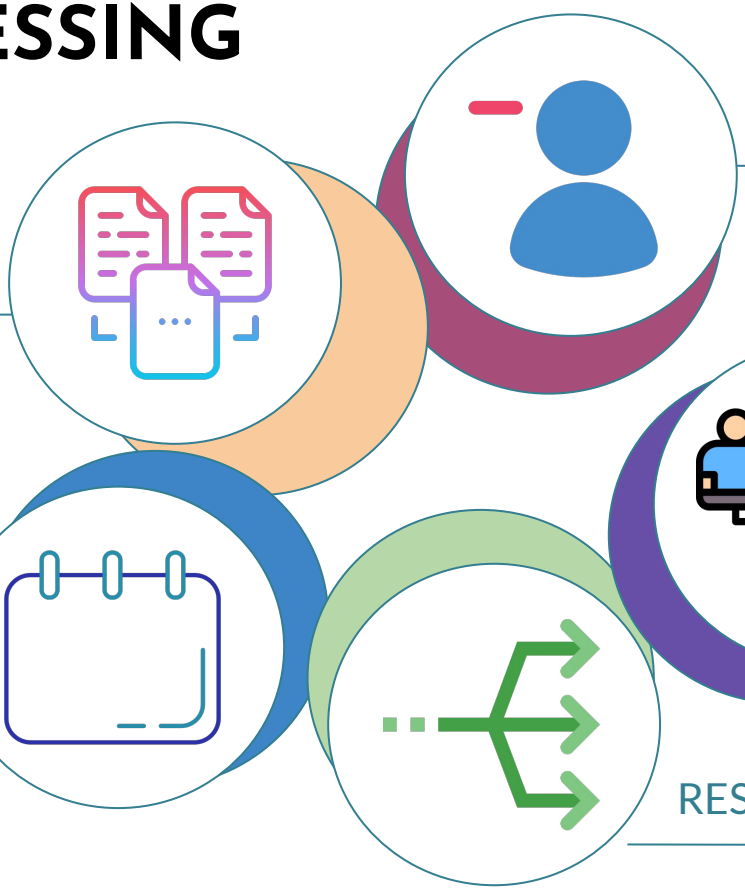


02

PREPROCESSING

Under-Sampling and Data Cleaning

PREPROCESSING



RIMOZIONE
RECENSIONI BREVI

UNIONE

UNDERSAMPLING

RIMOZIONE
RECENSIONI NULLE

RESCORING

RIMOZIONE RECENSIONI BREVI

Rimuovo tutte le recensioni con **meno di 3 parole**, poco significative e spesso poco motivate

NEGATIVE REVIEW
Room, Location

POSITIVE REVIEW
Staff

RIMOZIONE

515738
Recensioni

277520
Recensioni

UNIONE

NEGATIVE REVIEW
Rooms are nice but for elderly a bit difficult...

+

POSITIVE REVIEW
Location was good and staff were ok

REVIEW
Rooms are nice but for elderly a bit difficult...
Location was good and staff were ok...

RIMOZIONE RECENSIONI NULLE

REVIEW
No negative
Location was good and staff were ok...

REVIEW
Location was good and staff were ok...

REVIEW
Rooms are nice but for elderly a bit difficult...
No positive

REVIEW
Rooms are nice but for elderly a bit difficult...

RESCORING

- Utilizzo **2 classi** rispetto alle 10 disponibili
- L'utente difficilmente da una recensione <u>percepisce sensazioni intermedie</u>

Reviewer Score > = 6 → 1

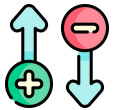
Review Score < 6 → 0

Obiettivo: **opinione netta** del cliente

<u>"We stayed at this hotel for five days"</u> is **<u>neutral</u>**.

"I liked staying here" is **positive**

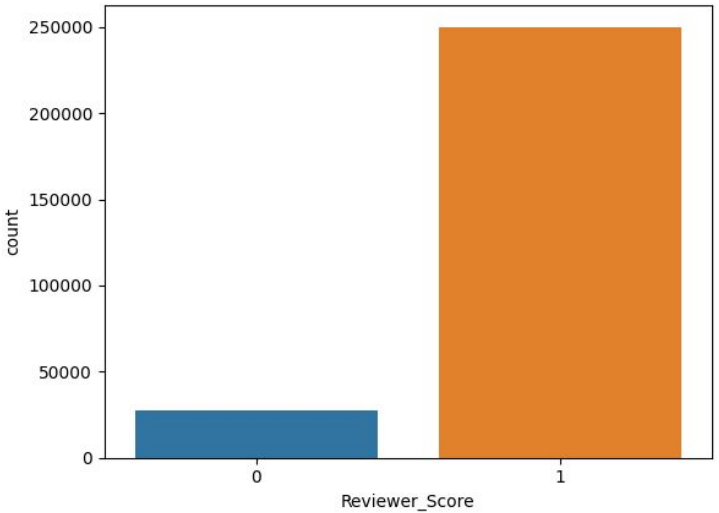
"I disliked the hotel" is **negative**.



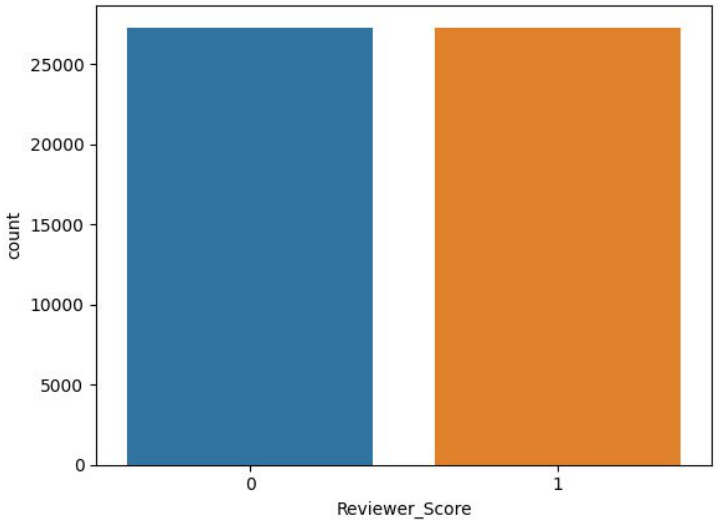
UNDERSAMPLING

Dataset unders-sampling tramite random-undersampling

280.000 ROWS



60.000 ROWS



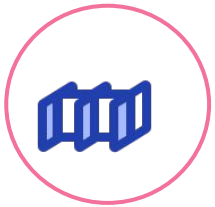
REVIEW CLEANING

Preprocessing sulle recensioni

1

Tokenization

Divisione delle recensioni in tokens



2

Stopwords

Predisposizione delle stopwords

3

POS Tagging

Analisi sintattica del token

4

Lemmatization

Riduzione delle parole ai rispettivi lemmi

5

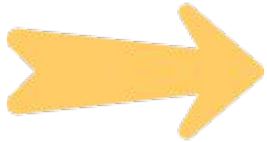
Join

Applicazione di tutte le modifiche

ESEMPIO PREPROCESSING

PRIMA

It was a little dark in lobby dining room but this wasn t really a problem just hard to see Amazing location right at the heart of shopping and tourist sights Staff spoke excellent English and were very helpful

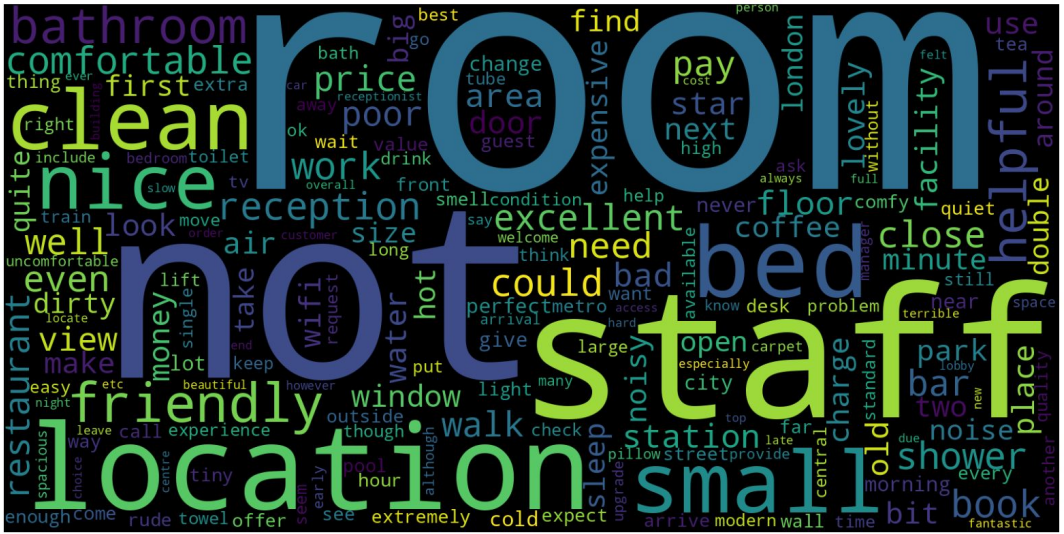


DOPO

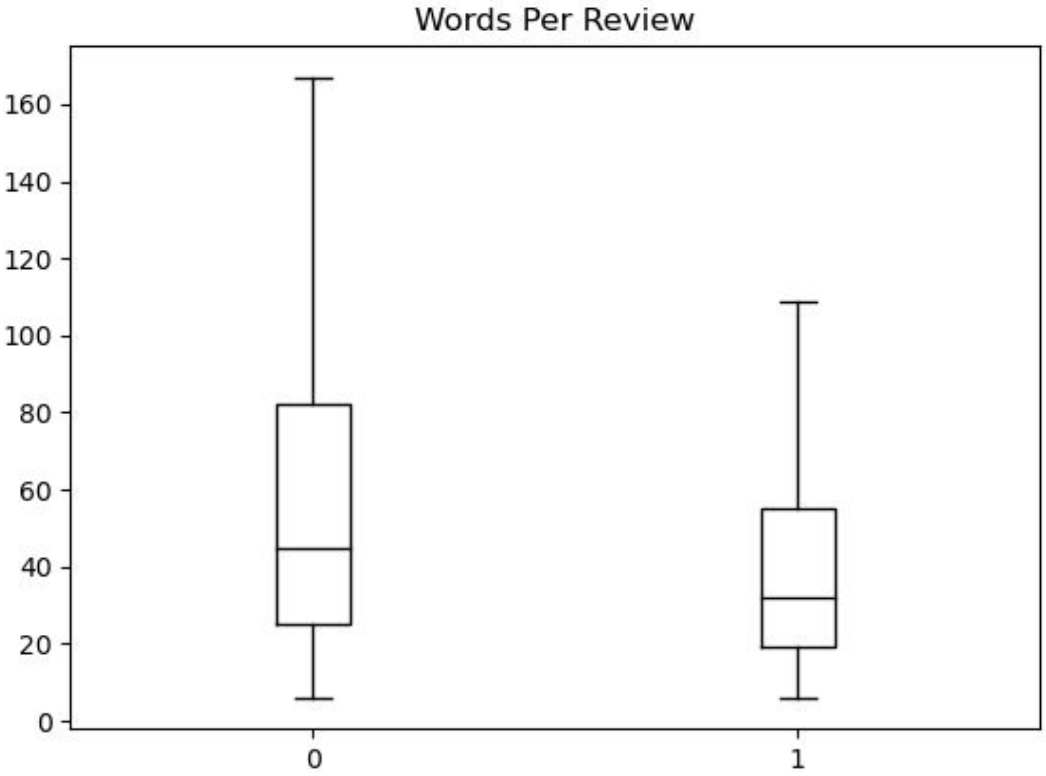
dark lobby dining room problem hard see amaze location right heart shop tourist sight staff speak excellent english helpful

WORD CLOUD

Su tutte le recensioni



WORD PER REVIEW



Recensioni Negative

Tendenzialmente

Più parole

03

VECTORIZATION

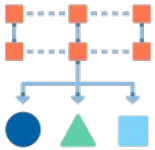
Trasformazione in formato TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency)

Permette una maggiore **Explanation** → TF-IDF riflette l'importanza della parola nel documento

	word_able	word_absolutely	word_ac	word_not	word_accept	word_acceptable	word_access
40012	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
44562	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
46997	0.000000	0.000000	0.057590	0.000000	0.000000	0.000000	0.000000
41957	0.000000	0.057084	0.000000	0.000000	0.000000	0.000000	0.000000
41939	0.000000	0.000000	0.000000	0.000000	0.000000	0.070118	0.000000

1505 COLUMNS



04

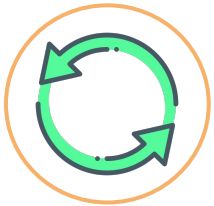
MODELLI DI CLASSIFICAZIONE

Utilizzo di Modelli di Machine Learning per Sentiment Analysis

MODELLI DI CLASSIFICAZIONE

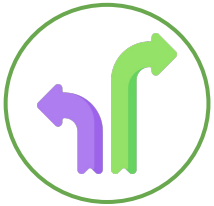
Scelta del modello

Cross validation



Split dei dati

train-test



Logistic regression

Primo modello

Random forest

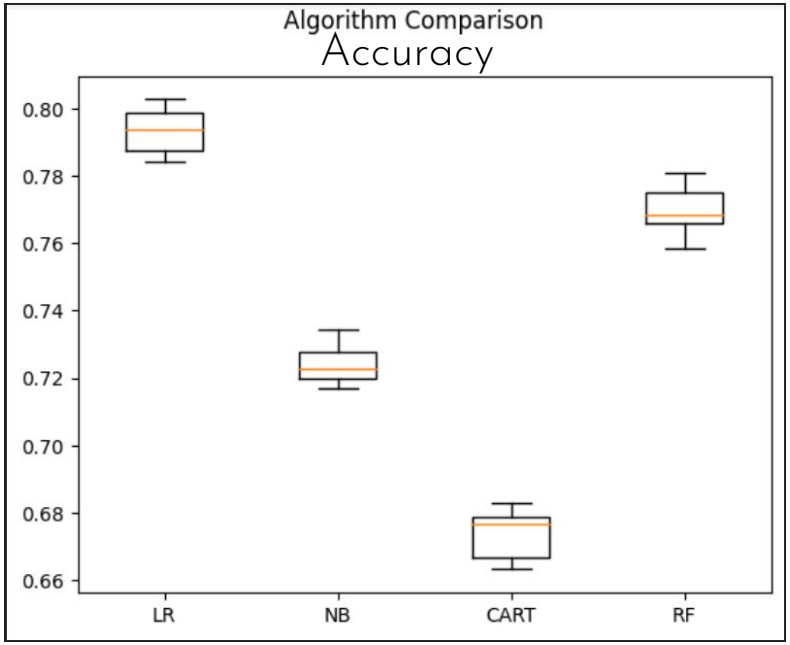
Secondo modello

Lessico

NRC e AFINN

MODELLI DI CLASSIFICAZIONE

10-Fold Cross-Validation

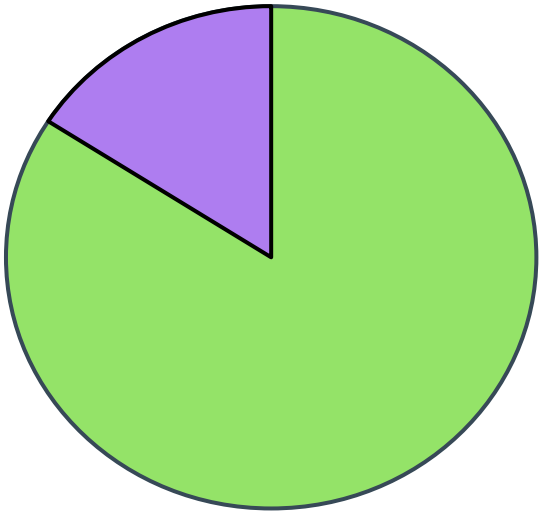


1. Logistic regression
2. Random forest
3. Naive Bayes
4. CART (Decision Tree)

MODELLI DI CLASSIFICAZIONE

20% TEST

Utilizzato per valutare le performance del modello



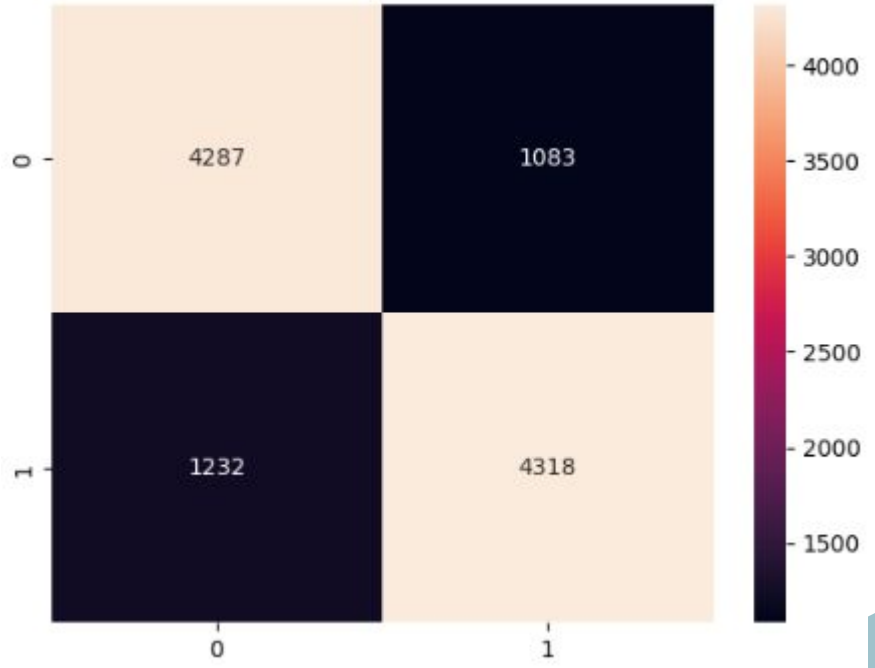
80% TRAIN

Utilizzato per allenare il modello

1 LOGISTIC REGRESSION

Vantaggi: Utile nella Classificazione Binaria

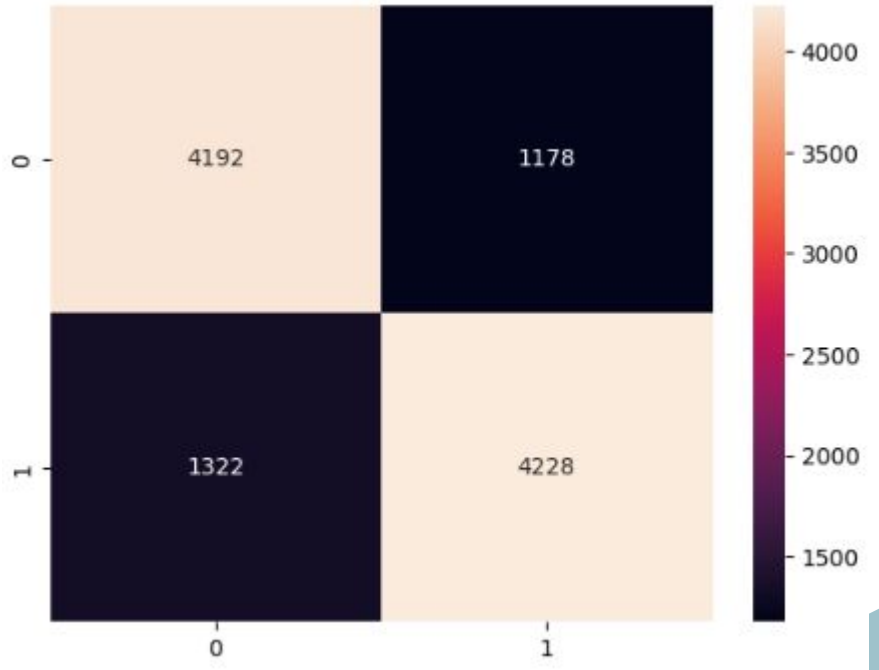
	Precision	Recall	F1-Score	Support
NEGATIVE	0.78	0.80	0.79	5370
POSITIVE	0.80	0.78	0.79	5550
ACCURACY			0.79	10920
MACRO AVG	0.79	0.79	0.79	10920
WEIGHTED AVG	0.79	0.79	0.79	10920



2 RANDOM FOREST

Vantaggi: Buona Interpretabilità, scalabilità, feature importance

	Precision	Recall	F1-Score	Support
NEGATIVE	0.76	0.78	0.77	5370
POSITIVE	0.78	0.76	0.77	5550
ACCURACY			0.77	10920
MACRO AVG	0.77	0.77	0.77	10920
WEIGHTED AVG	0.77	0.77	0.77	10920



3 LESSICO - NRC

Utilizzo di valori binari per indicare se la parola corrispondente appartiene a una delle 8 emozioni principali: rabbia, paura, attesa, fiducia, sorpresa, tristezza, gioia e disgusto.

Vantaggi:

- **categorizzazione del sentiment**
- **lingue multiple**

Dataset originale

	0	1
0	6443	16106
1	2058	21633

Accuracy: 60,7%

Dataset pre-processato

	0	1
0	8965	13078
1	3305	19196

Accuracy: 63,2%

4 LESSICO - AFINN

Utilizzo di valori compresi tra +5 e -5 per determinare quanto un recensione è positiva o negativa.

Vantaggi:

- **semplicità**
- **ampia copertura lessicale**

Dataset originale

	0	1
0	9104	15080
1	1898	23624

Accuracy: 65,8%

Dataset pre-processato

	0	1
0	10455	12428
1	2539	21516

Accuracy: 68,1%

MODELLI DI CLASSIFICAZIONE

MODEL EVALUATION → Il modello scelto sono le **Random Forest**

BUONE PERFORMANCE

Accuracy - Recall - Precision
Equiparabili ≈ 77%

EXPLANATION

Migliore spiegabilità modello
SHAP & LIME

ULTERIORI VALUTAZIONI

Funzionamento reale del modello

05

MODEL EXPLANATION

Differenti tecniche per interpretare i modelli

MODEL EXPLANATION

La spiegabilità del modello tramite SHAP e LIME consente di:

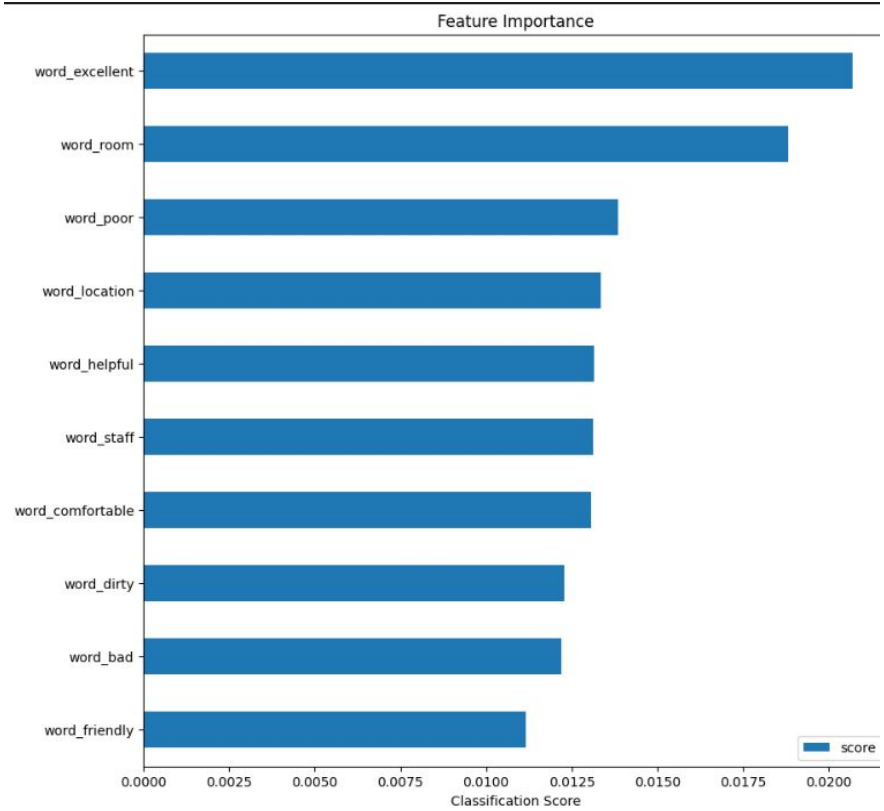
BUSINESS

Analizzare punti di forza e di debolezza della struttura e del servizio clienti

CUSTOMER

Capire maggiormente come le proprie recensioni impattano nella reputazione dell'hotel

RANDOM FOREST- Feature Importance

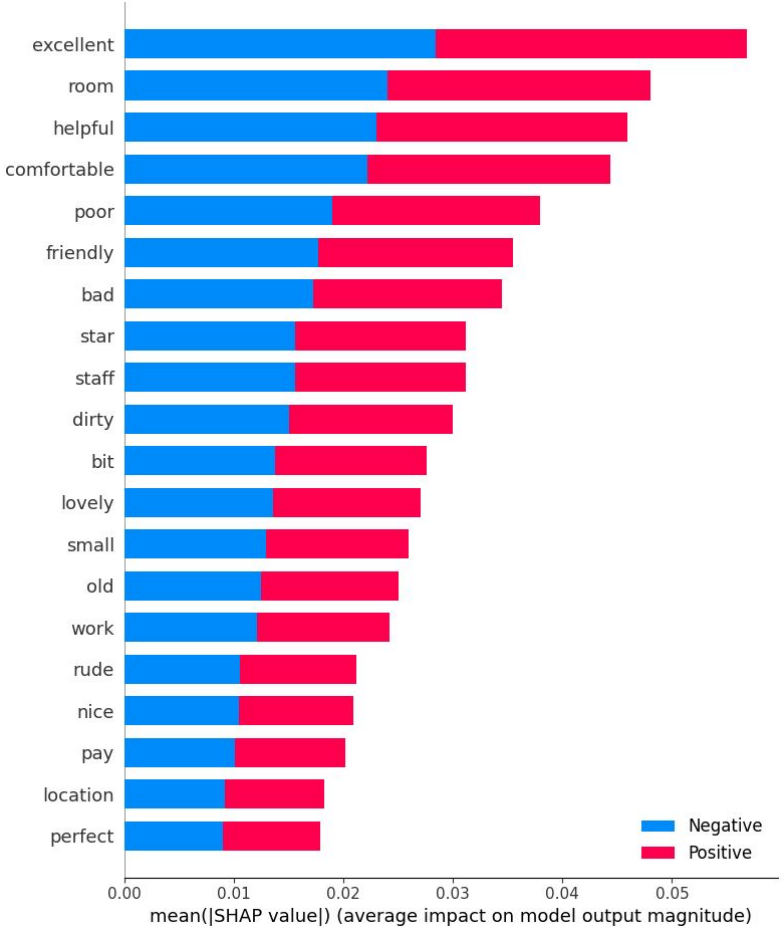


Feature più rilevanti nella Sentiment Analysis

→ **Analisi per CLASSE**

Perchè Recensione è classificata Positiva/Negativa?

MODEL EXPLANATION - SHAP (Train Set)



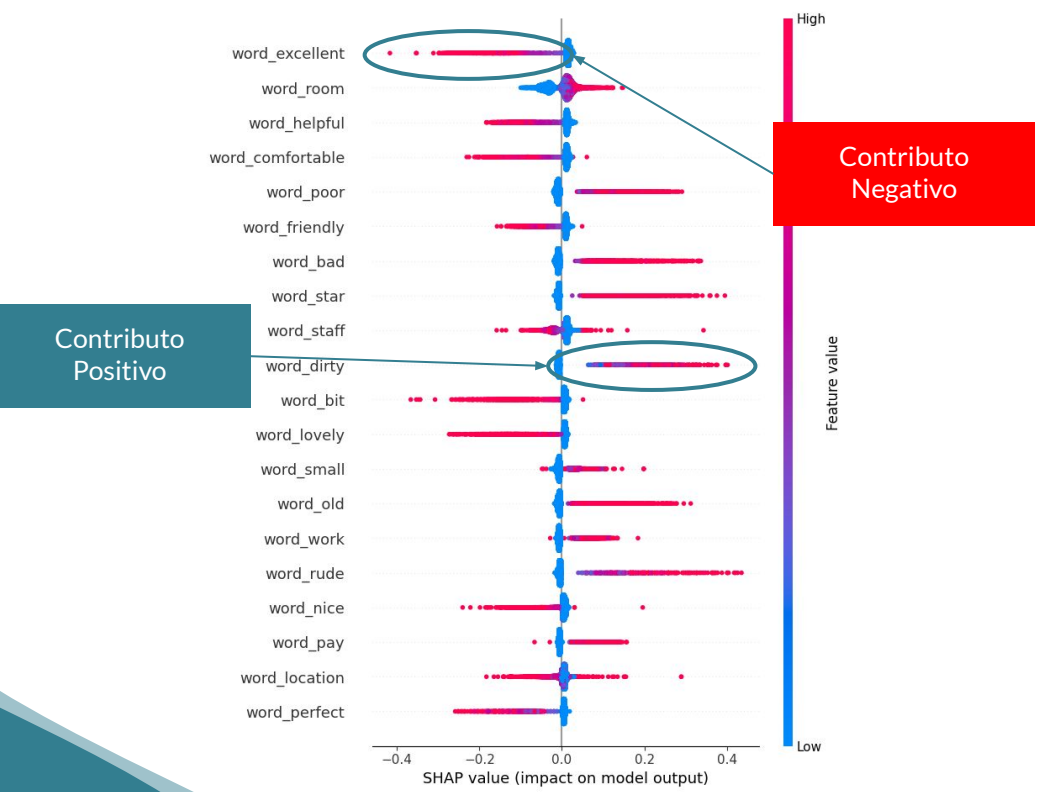
SHAP misura la contribuzione di ciascuna feature su un insieme di dati

- Feature con Maggior contributo (excellent, room ..)
- Circa Stesso Peso nelle classi

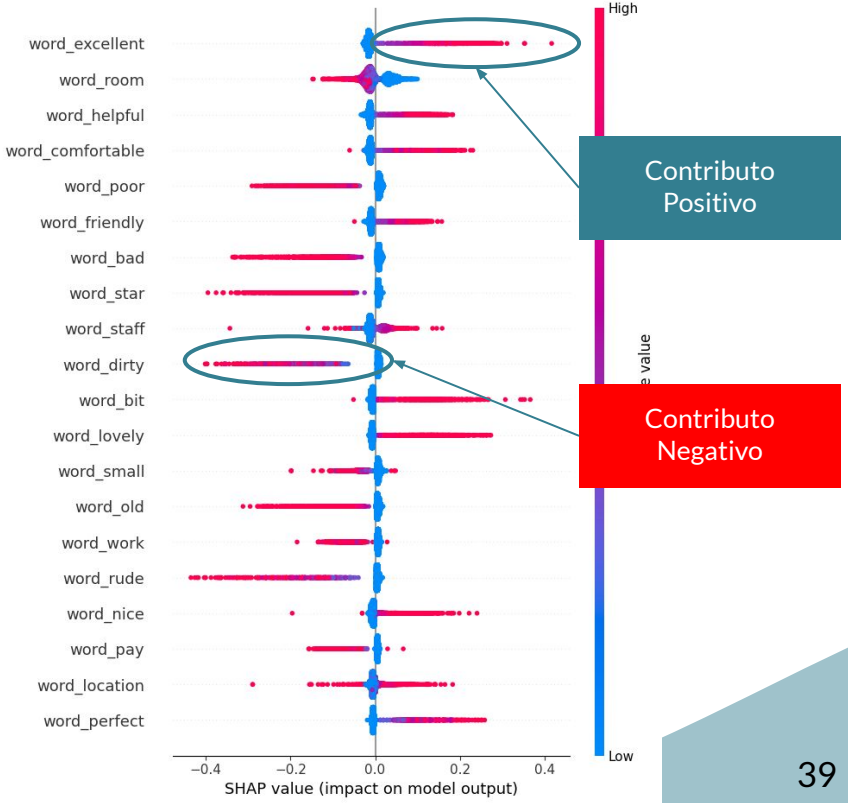
SHAP - Dettaglio

SHAP values per feature e istanza specifica

Negative Review



Positive Review

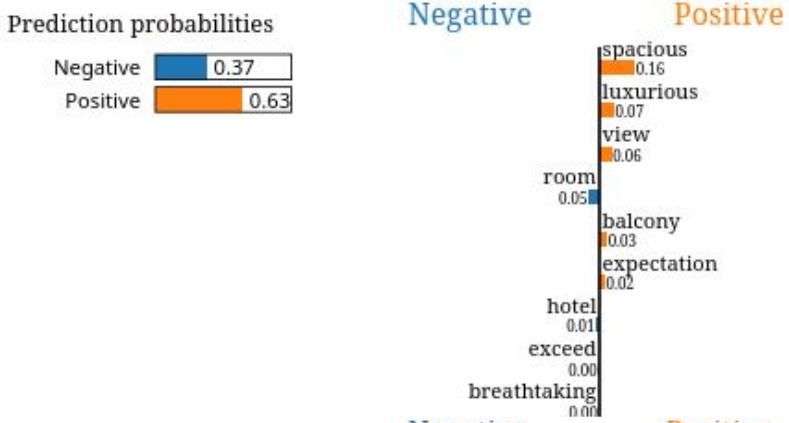


PREDICTION EXPLANATION - LIME

LIME interpretabilità su singoli campioni per identificare quali feature influenzano maggiormente le predizioni

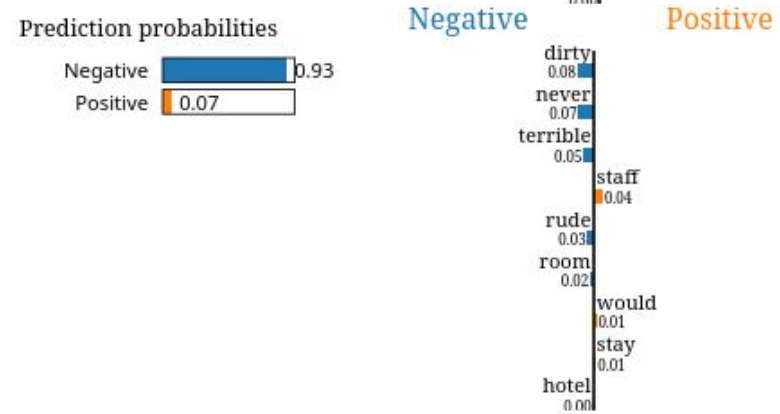
10

This hotel exceeded my expectations. The room was spacious and luxurious, and the view from the balcony was breathtaking.



1,0

The hotel was terrible. The rooms were dirty and the staff was rude. I would never stay here again.



06

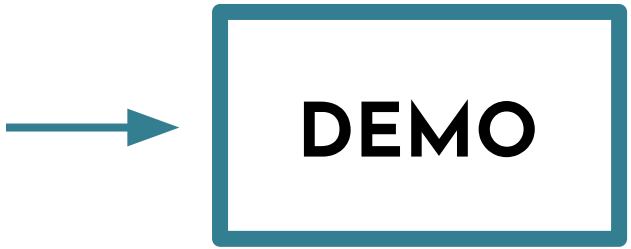
DASHBOARD

Dashboard e analisi finali

DASHBOARD

Obiettivo: fornire uno **strumento di supporto alle decisioni per una struttura alberghiera** che consenta di analizzare approfonditamente le recensioni ottenute

SCENARIO 1: Hilton Hotel London
SCENARIO 2: Recensioni generate casualmente con ChatGPT



07

ANALISI FINALI

Insight e Conclusioni

INSIGHT

Le analisi svolte hanno permesso di estrarre insight utili per prendere **decisioni data-driven**

SCENARIO BUSINESS

Quali aspetti devono essere migliorati ? Quali sono i punti di forza della struttura?

→ SHAP e DASH forniscono parole che sono forti indicatori di **aspetti POSITIVI o NEGATIVI**

La clientela di **diversa nazionalità** ha diverse preferenze?

→ Il tool proposto permette di analizzare al meglio le preferenza della clientela

SCENARIO CUSTOMER

Il cliente sta esprimendo correttamente quanto vuole intendere riguardo all'hotel?

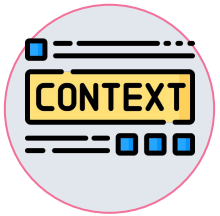
→ Il tool fornisce un **Feedback durante la scrittura** e individuare le parole più efficaci quando scrive una recensione

OBIETTIVO COMUNE

Lo **score è discordante** con quanto espresso nella recensione testuale?

→ Potrebbe esserlo. La sentiment analysis consente di analizzare ogni aspetto della recensione e individuare recensioni discordanti con il Review Score assegnato.

LIMITI



CONTEXT FREE

Utilizzo di word2vec o BERT, per mantenere il contesto, a discapito dell'explainabilility

RISORSE COMPUTAZIONALI DISPONIBILI

Migliorare i modelli e vettorizzazione

↓

FOCUS INSIGHT

AMBITO LUXURY

Lessico specifico utilizzato in un dominio specifico

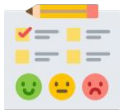
UNICA FONTE DATI

L'utilizzo di più piattaforme di review avrebbe consentito di fornire una visione più realistica

CONCLUSIONI

INSIGHT SIGNIFICATIVI

Dash utile sia lato Business e Customer



VALUTAZIONE E IMPATTO RECENSIONI

Fornire una maggiore consapevolezza dell'impatto delle recensioni online

BUONE PERFORMANCE DEI MODELLI

Maggior parte delle recensioni classificate correttamente

MIGLIORARE MODELLO

Continuo Retraining del modello con nuove recensioni ricevute

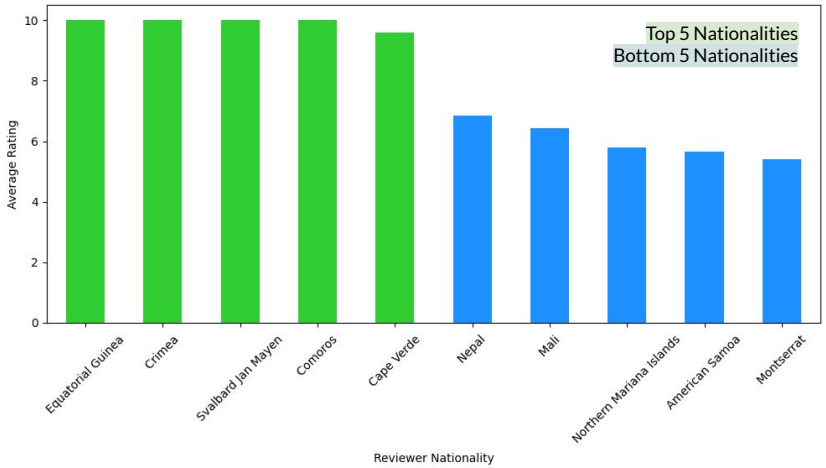
GRAZIE PER L'ATTENZIONE

Data Analytics
Anno Accademico 2022-2023

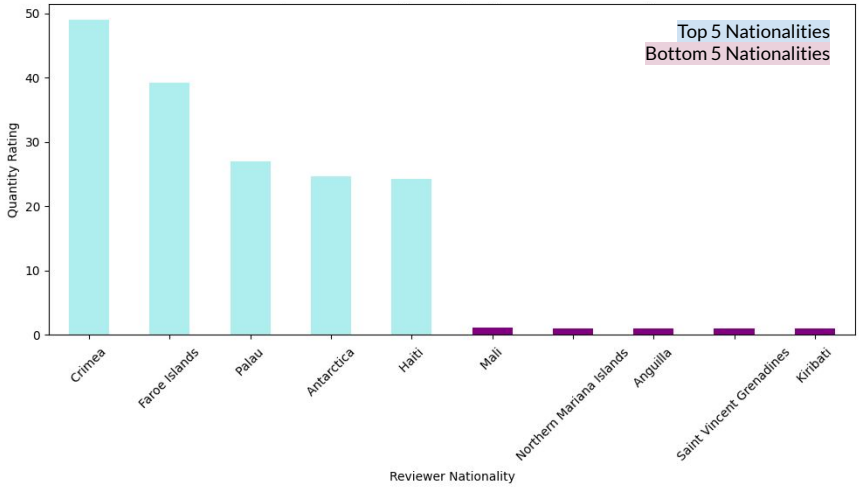
Urbani Nicolò 856213
Mohamed Nada 857606
Rubini Alessia 851890

ANALISI ESPLORATIVA

Score - Nazionalità

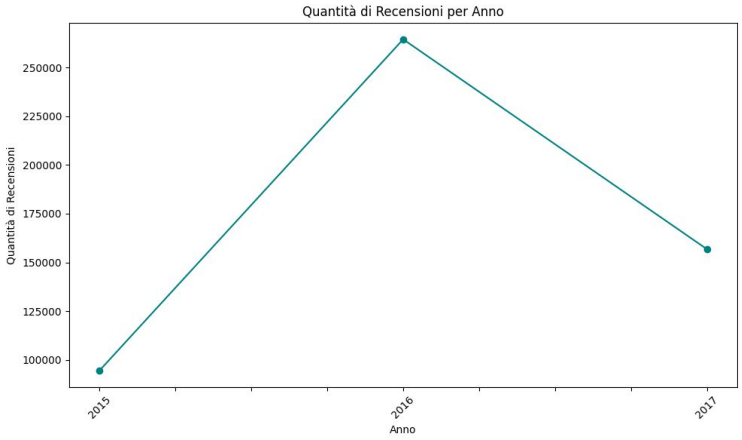


Score - Quantità Recensioni

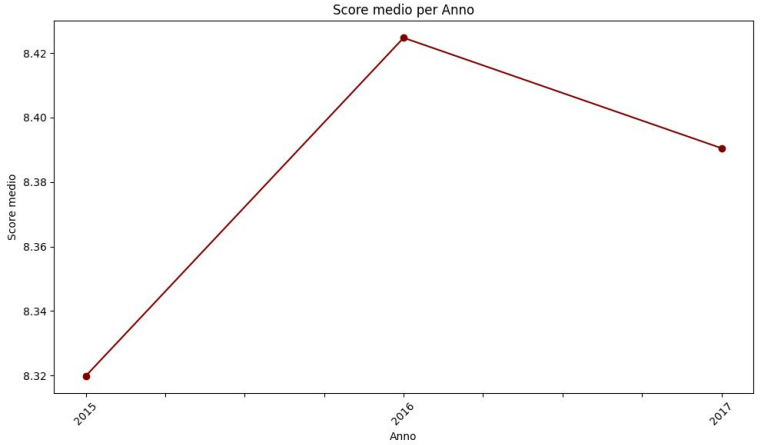


ANALISI ESPLORATIVA

Anno - Quantità Recensioni

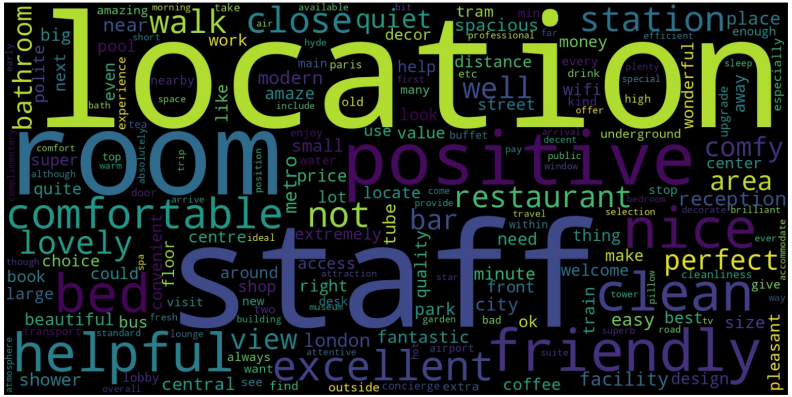


Anno - Score Medio



WORD CLOUD

Recensioni Positive



Recensioni Negative