TripAdvisor Restaurant Review Analysis

An NLP Analysis of Restaurant Reviews For Barcelona City.

The Problem:

Company

TripAdvisor is a website that offers information about places to explore such as- hotels, restaurants, historicals sites etc.

Context

Platforms like TripAdvisor play a crucial role to impact a Customer's Decision and a Restaurant's Reputation. The Reviews posted online offer both opportunities and risk, so it is important to analyze the Sentiment of the reviewers. This might contribute to making Restaurateur's business decisions more precise.

Problem statement

Our goal is to find out the Sentiment behind the reviews on the Restaurants in Barcelona city posted by the reviewers.

Criteria for Success:

Our goal is to find out the features that are influential for the positive or negative reviews for Restaurants in Barcelona city posted by the reviewers.

This is a form implementing Supervised Learning (Sentiment Analysis) based on Keywords (Topics) derived from Unsupervised Learning (Topic Modelling). It is a Classification problem modelled with Topics as Regressors.

The modelling benefits business both ways-

- Topics/ Keywords to provide insight on categorizing the reviews,
- Which of the Topics influence most while leaving Reviews.

Scope of Solution:

Labeling the Topics:

Labeling the Topics according to word frequency, especially when no prior knowledge about data.

Identifying the Topics:

Correct Identification of the Number of Topics.

Corpus definition:

Non-overlapping Keywords and identify the correct stopwords.

Solution:

High Performance Computational Platform.

Gridsearch by HyperParameter Tuning to fit best fit/ update Topics.

Build a Neural Network Algorithm.

Constraints within Solution Space:

Missing Records:

Some of the Dataset had a huge chunk of Records Missing (e.g. New Delhi reviews) due to Web Scraping.

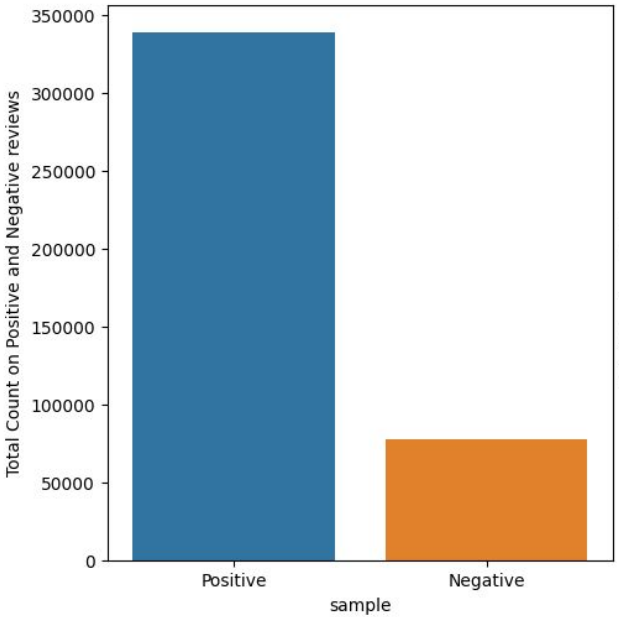
Data Handling:

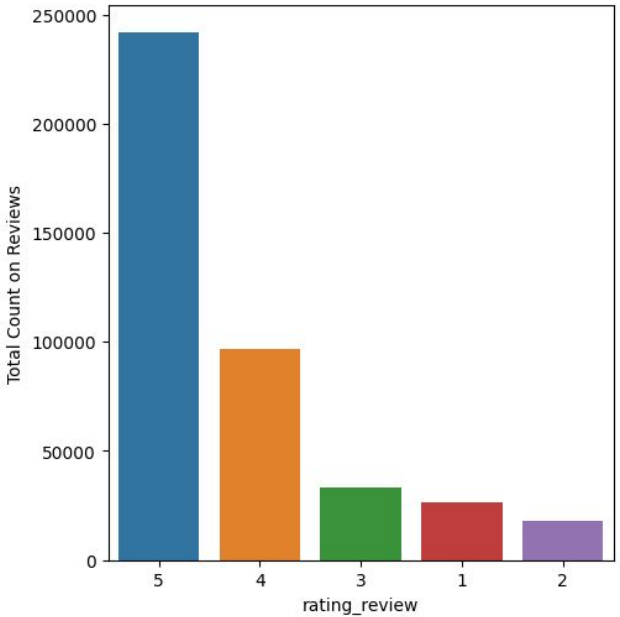
- 6 of the Datasets from 6 different locations.
- Very large Corpus with considering Bigrams and Trigrams.

Constraint on Analysis:

A sample of 20,000 records were chosen from the "Barcelona_reviews" dataset, so this analysis is narrowed down for some specific criteria.

Analysis on the Reviews:





Reviewers were seemed to be interested to leave more positive reviews than negative.

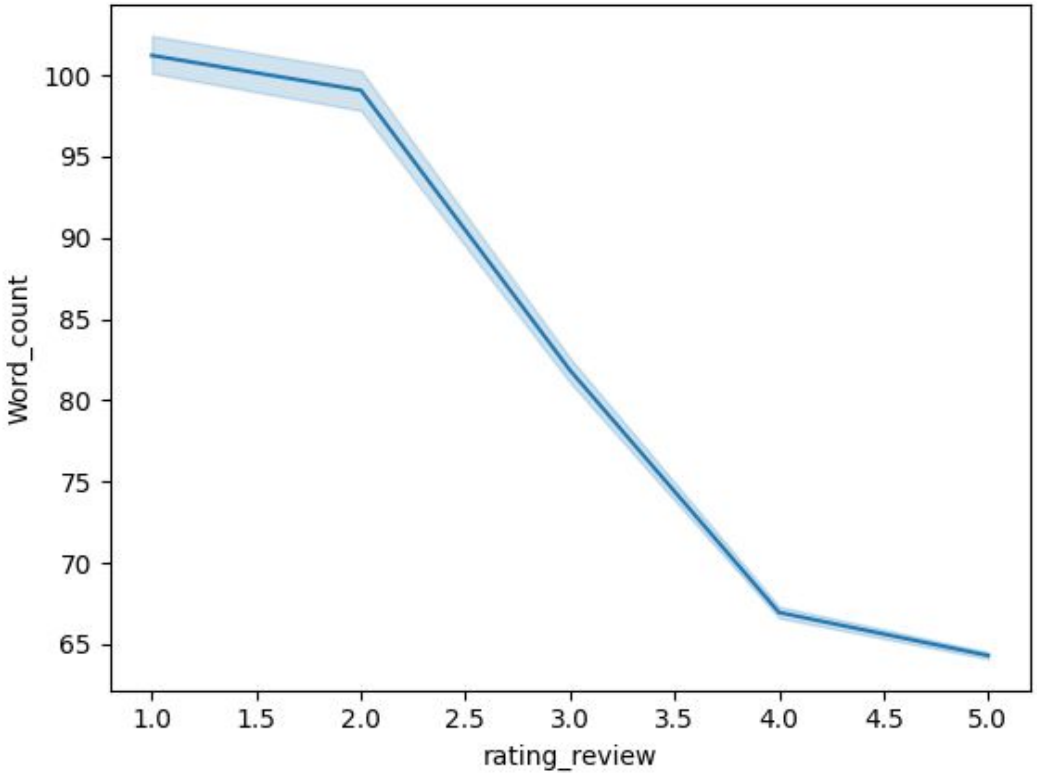
Analysis on the Reviews:

These are Word Clouds for the Positive and Negative Reviews for the Restaurant in Barcelona.

Positive

Negative

Analysis on the Reviews:



Reviews with negative reviews had more words in them.

Analysis on the Reviews:

6 Different Topics were identified to best fit the model.

Analysis on the Reviews:

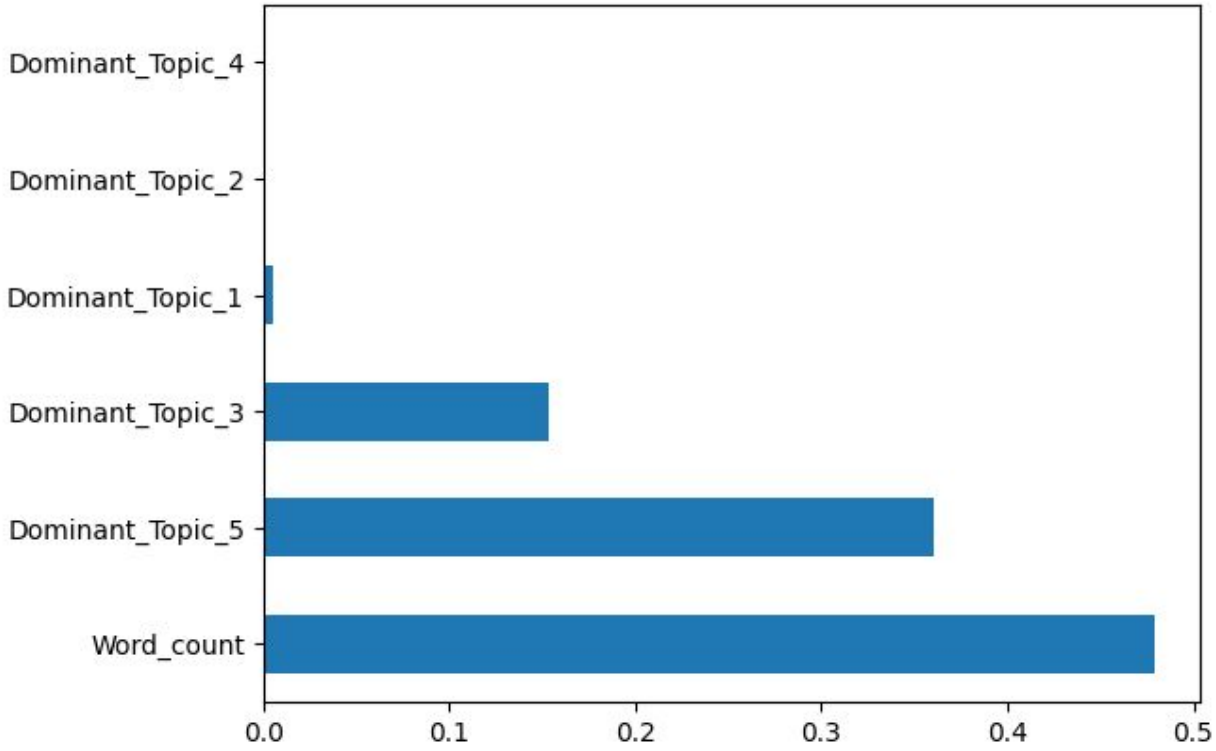
2 Topics were identified to be most Dominant in the Corpus.

‘Topic5’ had most influence, followed by ‘Topic3’

Topic 5:	*good, great, recommend, friendly, nice, really, well, tapa, wine, menu, excellent, also, dish, atmosphere, try, delicious, definitely, love, small, taste, little, quality, tasty, eat, fresh, highly, lovely, bit, choice, different, lot, selection, dessert, cook, tapas, choose, meat, salad, quite, special*

Topic 3:	*get, make, come, order, back, even, eat, waiter, drink, take, want, try, say, friend, see, feel, give, think, pizza, ever, people, leave, last, always, way, ask, know, full, family, absolutely, thing, start, end, owner, decide, away, speak, still, pay, happy*

Analysis on the Reviews:

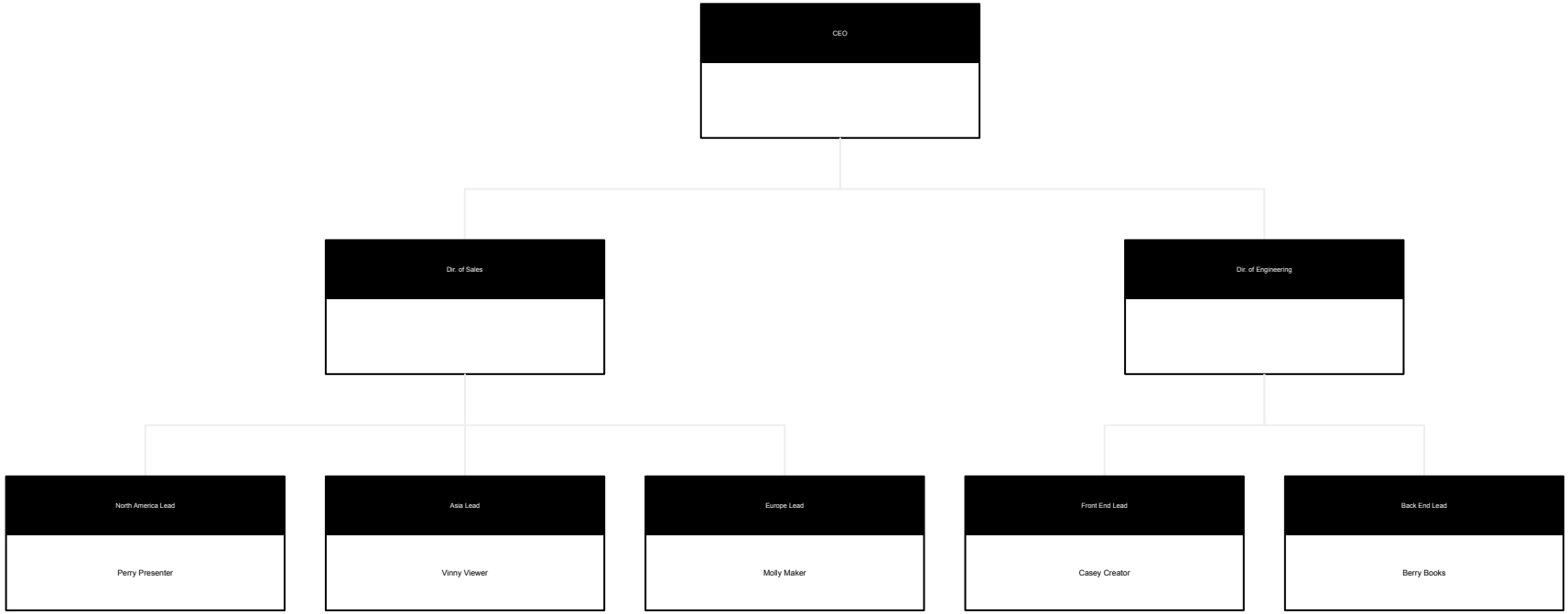


For Predictive Modelling (Classification of the Sentiment), WordCount, Topic5, and Topic3 played as the most important factors while classifying the reviews correctly.

ML Approach on the Sentiment Analysis:

- Logistic Regression and Random Forest Classifier Approach(~59% accuracy) seemed as best fit for this Dataset.
- The Consideration on Performance Metric of the models should be emphasized on F1 score (Precision and Recall), since this business analysis is aimed to target correct classification of the reviews.
- The Topic Identification opens another dimension of interpretation, since this helps more accurately to identify the influential factors for any analysis.

Stakeholders



Questions?

Thank You