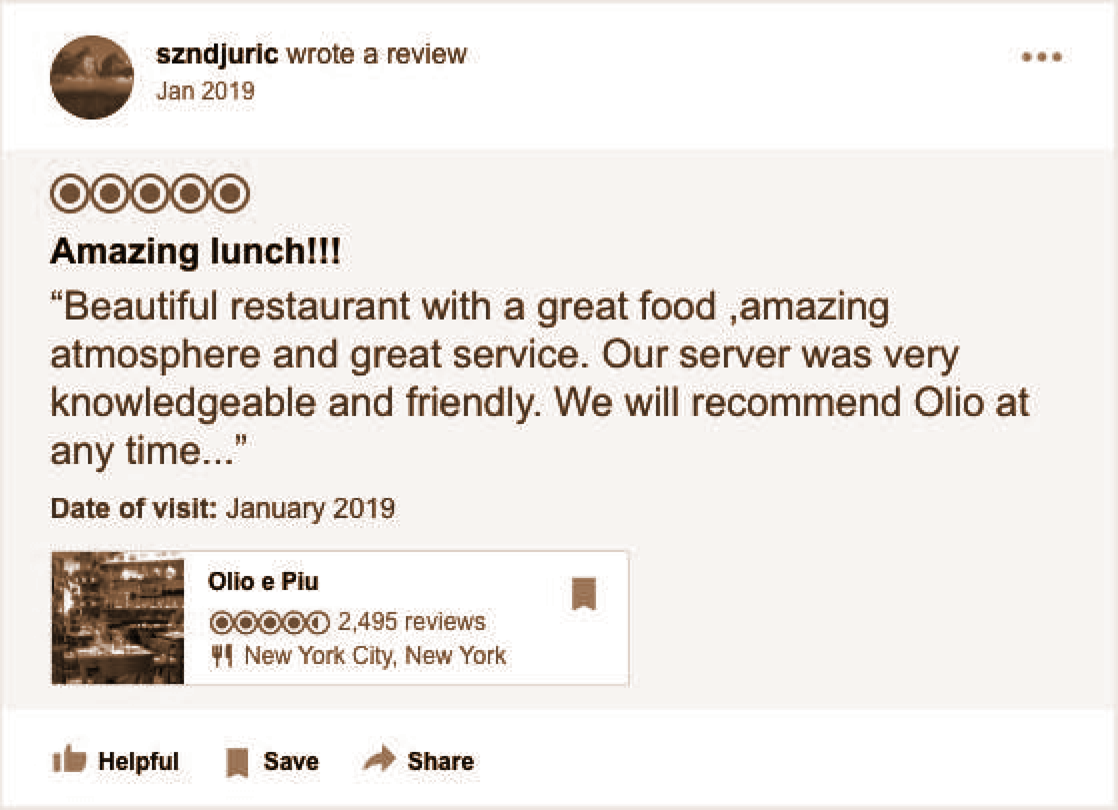
Date

December 2023

An NLP Project Report On Analyzing Sentiments of Restaurant Reviews provided by TripAdvisor

A Capstone Project Report Prepared by Mahmuda Yasmin for DATA SCIENCE CAREER TRACK Certification by Springboard.

Context:



szndjuric wrote a review

Jan 2019

Amazing lunch!!!

“Beautiful restaurant with a great food ,amazing atmosphere and great service. Our server was very knowledgeable and friendly. We will recommend Olio at any time...”

Date of visit: January 2019

Olio e Piu

2,495 reviews

New York City, New York

People have to eat to live, and for that people who love to dine at a restaurant rely on restaurant reviews . In this modern era of food industry and internet, reviews left on the internet for a restaurant has a significant impact on their business, but also opens the opportunity to find hidden information on new opportunities for consumer business.

TripAdvisor has a rating based review system which helps the visitors get an idea on the review on restaurants. As mentioned in the data source at Kaggle -

“ In the social platform of TripAdvisor, users are linked to restaurants by means of reviews posted by them. Using the information of these interactions, we can get valuable insights for forecasting, **proposing tasks related to recommender systems, sentiment analysis, text-based personalisation or text summarisation, among others.** Furthermore, in the context of TripAdvisor there is a scarcity of public datasets and lack of well-known benchmarks for model assessment. **We present six new TripAdvisor datasets from the restaurants of six different cities: London, New York, New Delhi, Paris, Barcelona and Madrid.”**

For this analysis , we will be considering the dataset for ‘Barcelona’. In this NLP project, we will apply Unsupervised Learning (Topic Modelling) to categorize the Sentiment on reviews.

Objective:

The goal of this study focuses on - “What were the Customers Happy or Disappointed?”, “What did the Customers Like or Disliked most?”.

We will try to find out what factors may contribute to the Positive and Negative Reviews and provide supporting analysis to build an action plan to categorize the keywords that might contribute for the reviews.

Our Scope to find Solution for the following:

- What factors contribute most for the restaurant reviews?
- Is there any way to predict a review from a general interpolation?

- What are the contributing keywords that help to identify the sentiments in a review?

About the Dataset:

The source of the data is here. As mentioned in this Kaggle website- this dataset contains 6 tables for 6 different cities (London, New York, New Delhi, Paris, Barcelona and Madrid), in CSV format. For our analysis, we will consider the 'Barcelona' Dataset.

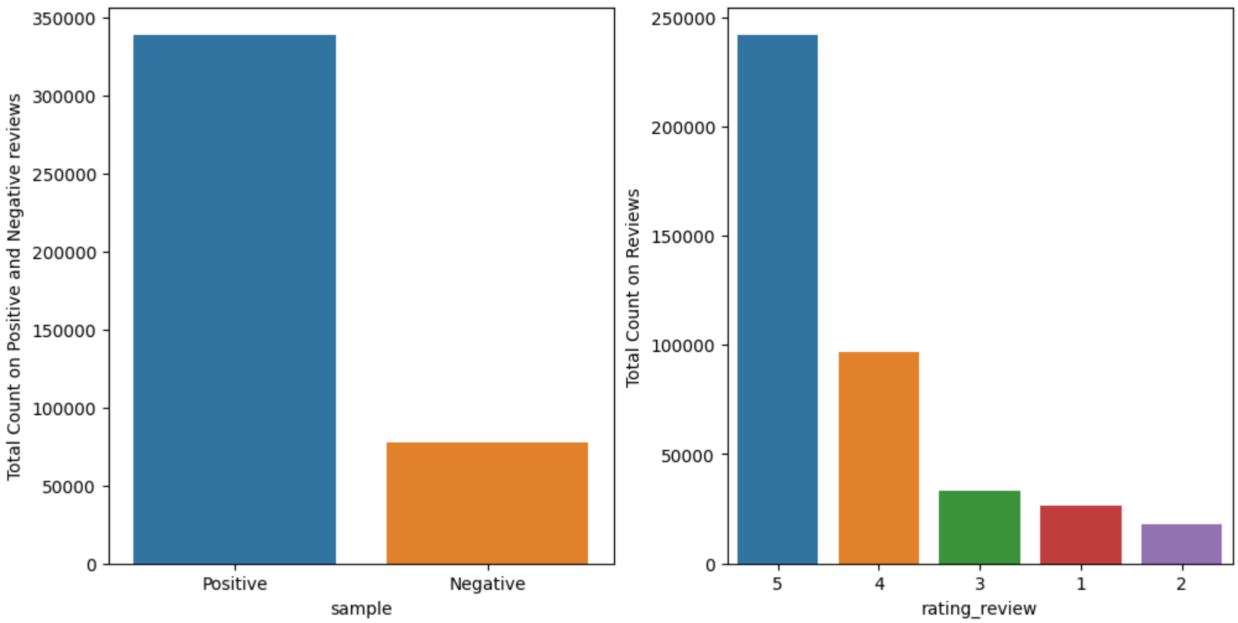
- The 'Barcelona' table (426641 rows, 13 columns) contains information on all 426641 customer reviews from from allover in Barcelona
- Each record represents one customer review, and contains details about restaurant name, rating review (1 to 5), sample (positive or negative), full review, date, city etc.

Our interest was focused on the Sentiment Analysis of the reviews in the Dataset. The "sample - positive or negative" Column is our event of interest. So, this is a classification problem based on Natural Language Processing. To get a more precise idea, we will apply Topic Modelling to identify the groups (of keywords) and try to find if there is any association between the groups/ topics with the sentiment.

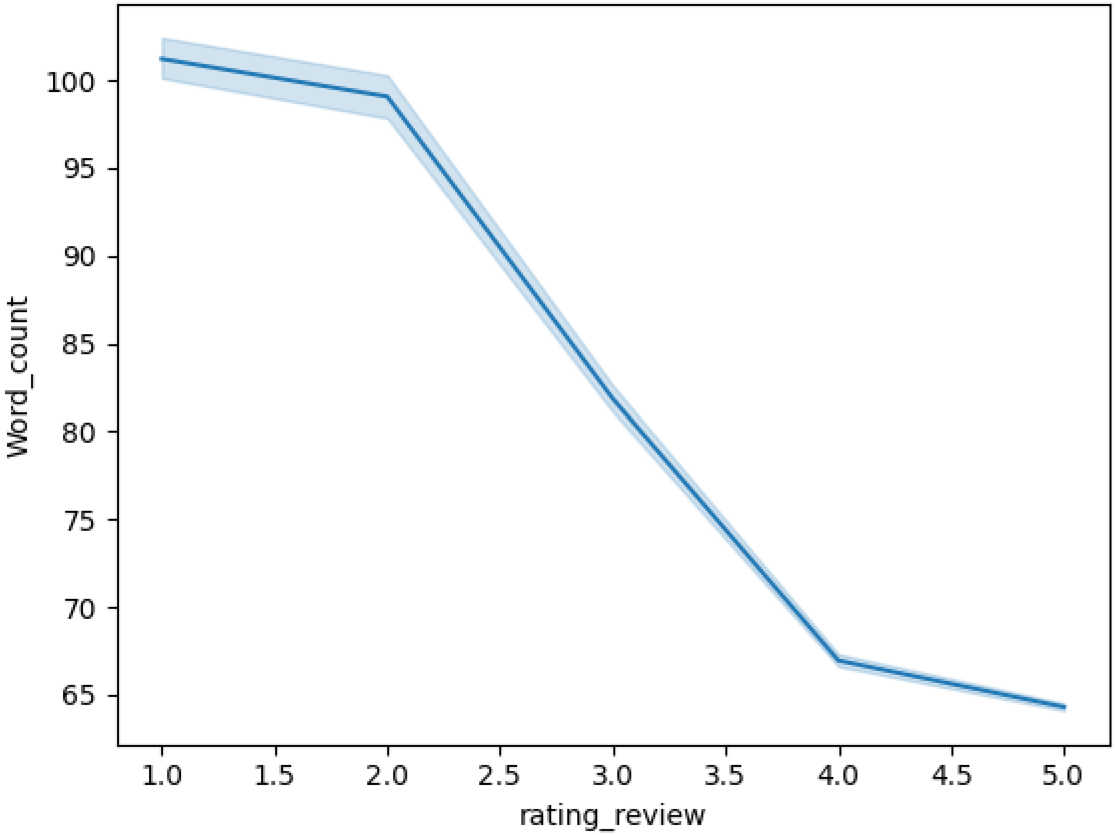
Exploring the Dataset:

The exploration of the rating review revealed some interesting findings. The reviewers seemed to be interested in leaving positive reviews rather than negative.

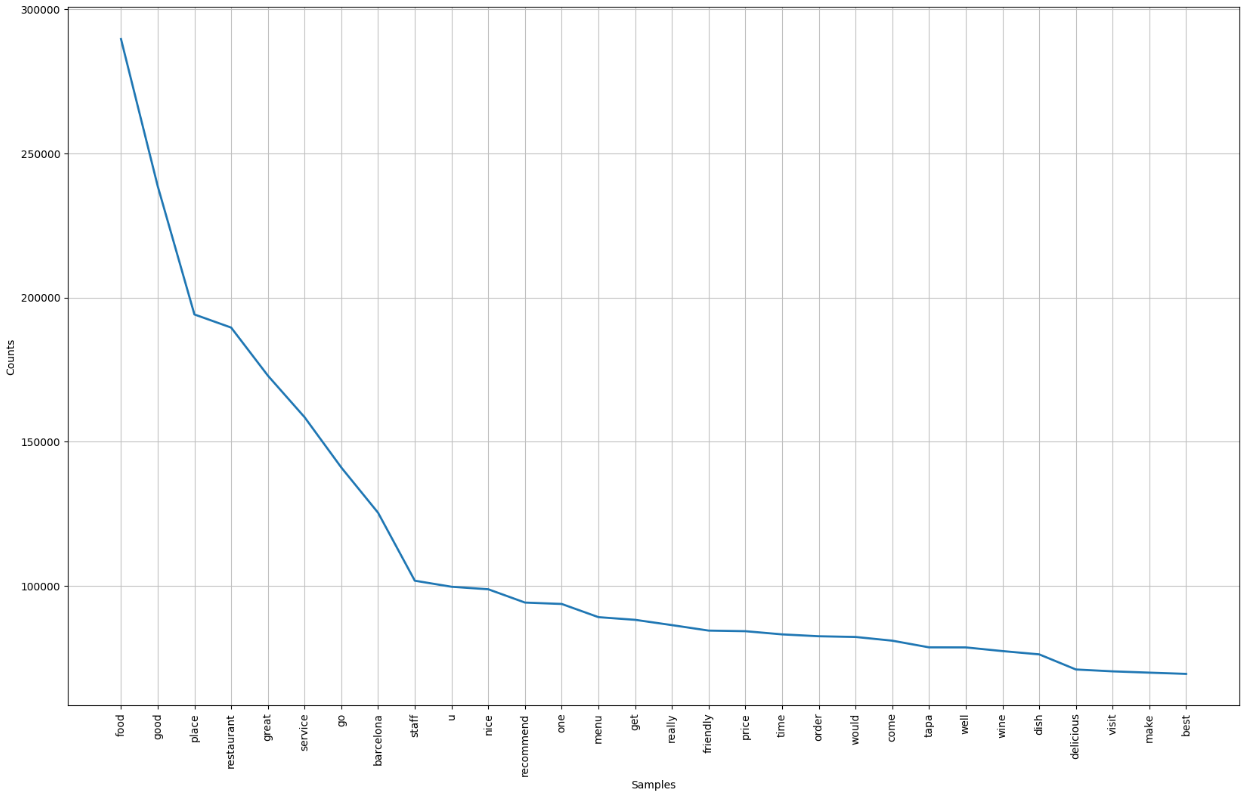
This was identified as a fact for the Imbalanced Class Classification Problem for predictive analysis.



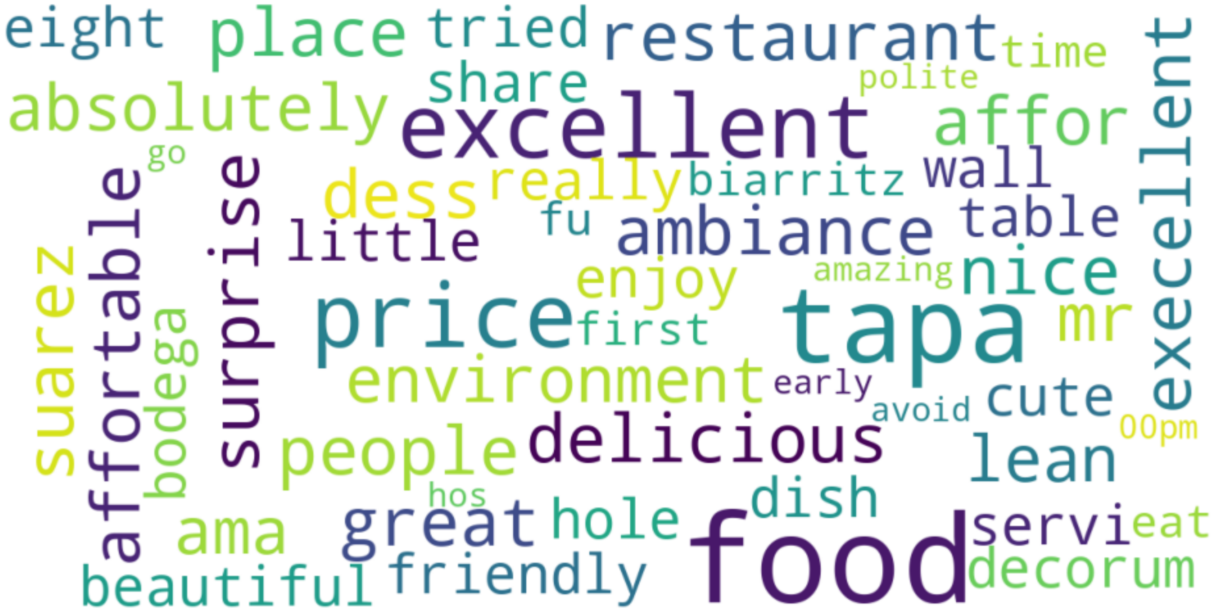
On a different note, the Negative Reviews seemed to have higher word count i.e. reviewers tend to use more words while leaving negative reviews than positive.



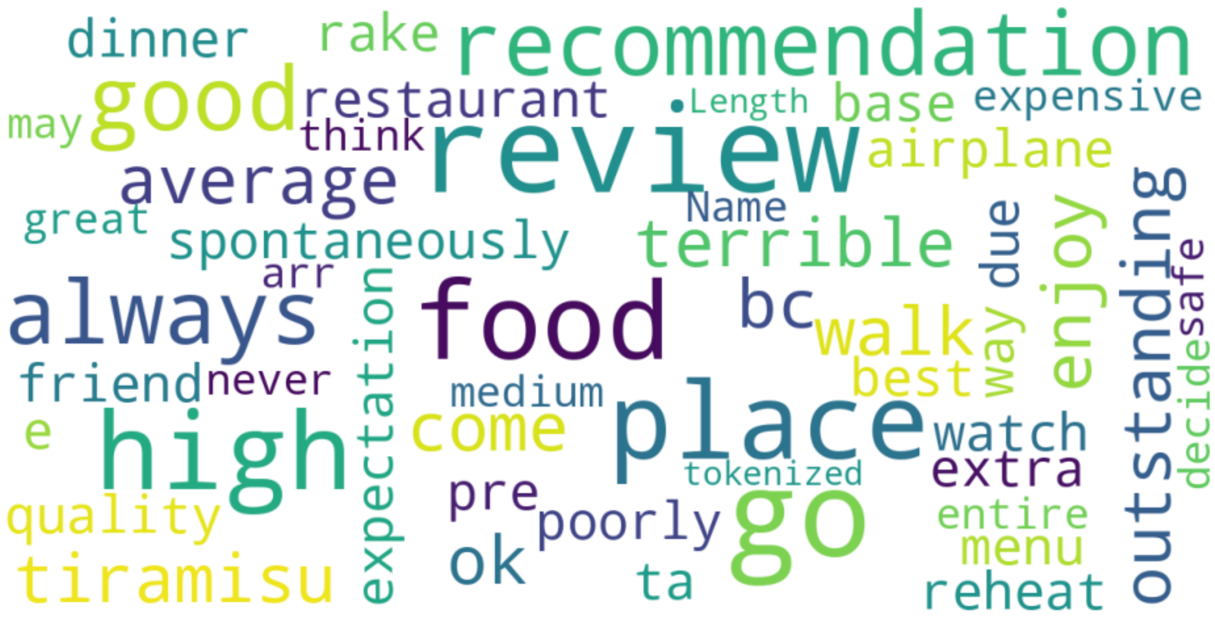
Following is the frequency plot of the highest occurring keywords to get an idea on the high frequency words in the reviews.



The word cloud for positive and negative reviews can be visualized as-



Positive



Negative

Data Preprocessing:

1. Missing Values:

The Data Frame had 2 rows of missing values. Since it was not very significant, the missing records were dropped.

2. Drop Unnecessary Columns:

Following columns were dropped: parse_count,restaurant_name,review_id, title_review, review_preview, date, city, url_restaurant, author_id.

3. Converting String to Float:

While processing the data, the ‘rating_review’ column I had to convert from string type to float and integer.

4. Text Preprocessing:

The basic preprocessing step for NLP on the reviews_full column I applied is following:

   a. Set Stopwords using NLTK stopwords corpus and added: 'food', 'place', 'restaurant', 'go','barcelona';
   b. Lowercase words,
   c. Remove punctuation,
   d. Parts of Speech Tagging,
   e. Lemmatize,
   f. Tokenize.
   g. Create Bigrams (Trigrams were not considered due to computational performances)

5. Select a Sample of the dataset:

Due to lack of high performance logistic support, I took a sample of 20000, considering 1:1 selection from both Positive and Negative reviews which created a balanced classification for Topic Identification.

6. From the final BagofWords, I filtered out words that occur less than 10 by numbers in the documents, or more than 90% of the documents

7. Topic Modeling:

I considered the LDA approach of Topic Modelling (details on LDA in here) and used the ‘Genism’ python library for topic modeling.

The Topics identified for this dataset were:

Topic 0: *paella, seafood, beautiful, music, pleasant, chef, fast, english, awesome, beach, light, rice, lucky, pack, cozy, travel, flavor, world, vegetarian, comfortable, rush, vibe, environment, knowledgeable, gothic_quarter, slightly, occasion, relaxed, mussel, fairly, sea, create, guide, prompt, freshly, beautifully, lobster, match, rib, strongly*

Topic 1: *service, staff, time, visit, meal, experience, night, lunch, day, serve, great, first, wonderful, course, fantastic, helpful, enjoy, main, return, amazing, couple, cocktail, trip, thank, glass, din, set, prepare, next, work, group, need, starter, kitchen, watch, second, help, lovely, however, seem*

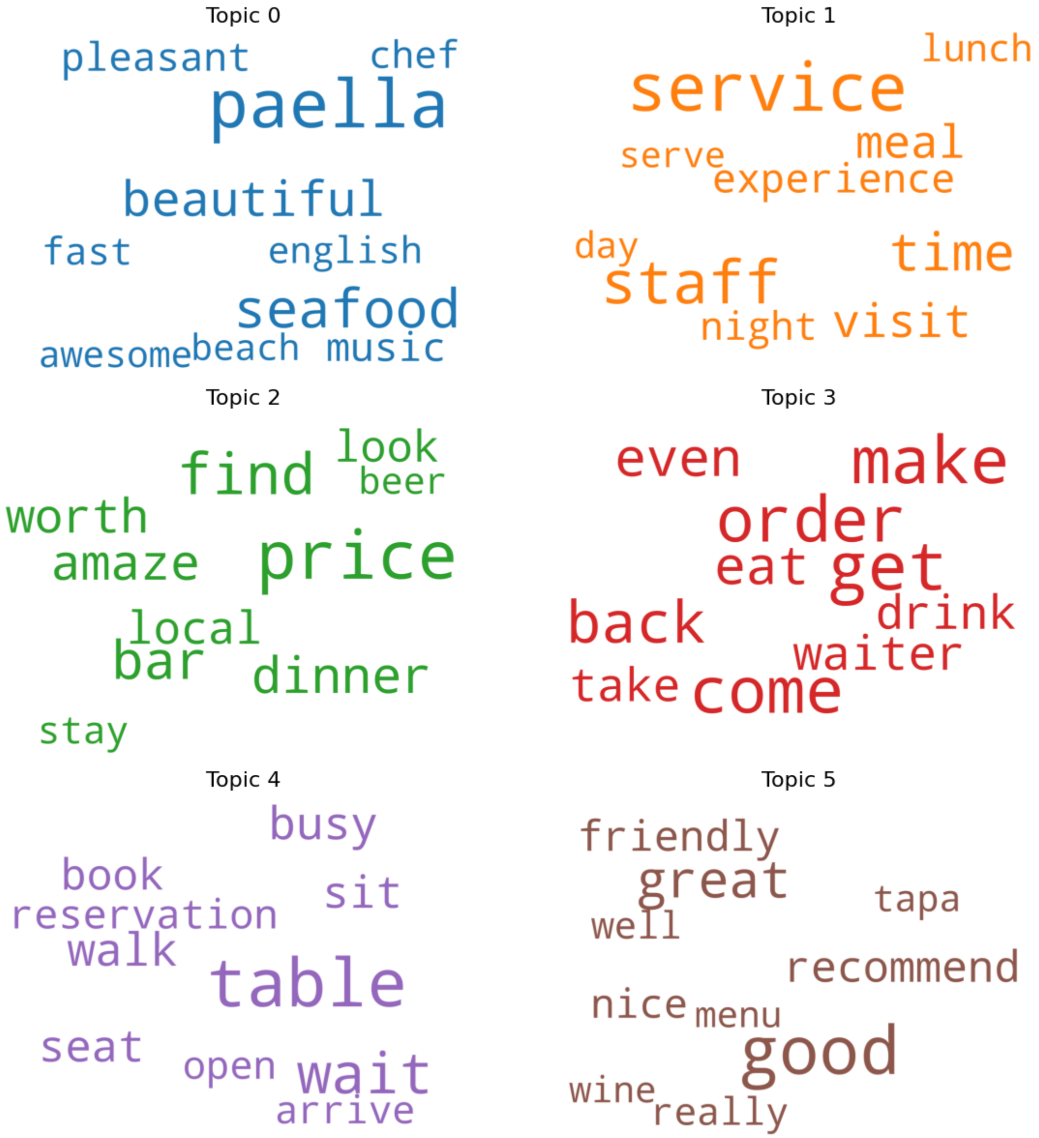
Topic 2: *price, find, bar, dinner, amaze, worth, local, look, stay, beer, spanish, reasonable, area, hotel, attentive, location, much, perfect, stop, close, drink, visit, fun, tourist, many, home, city, view, expensive, include, list, relax, street, twice, high, pretty, sure, especially, sangria, reasonably*

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*

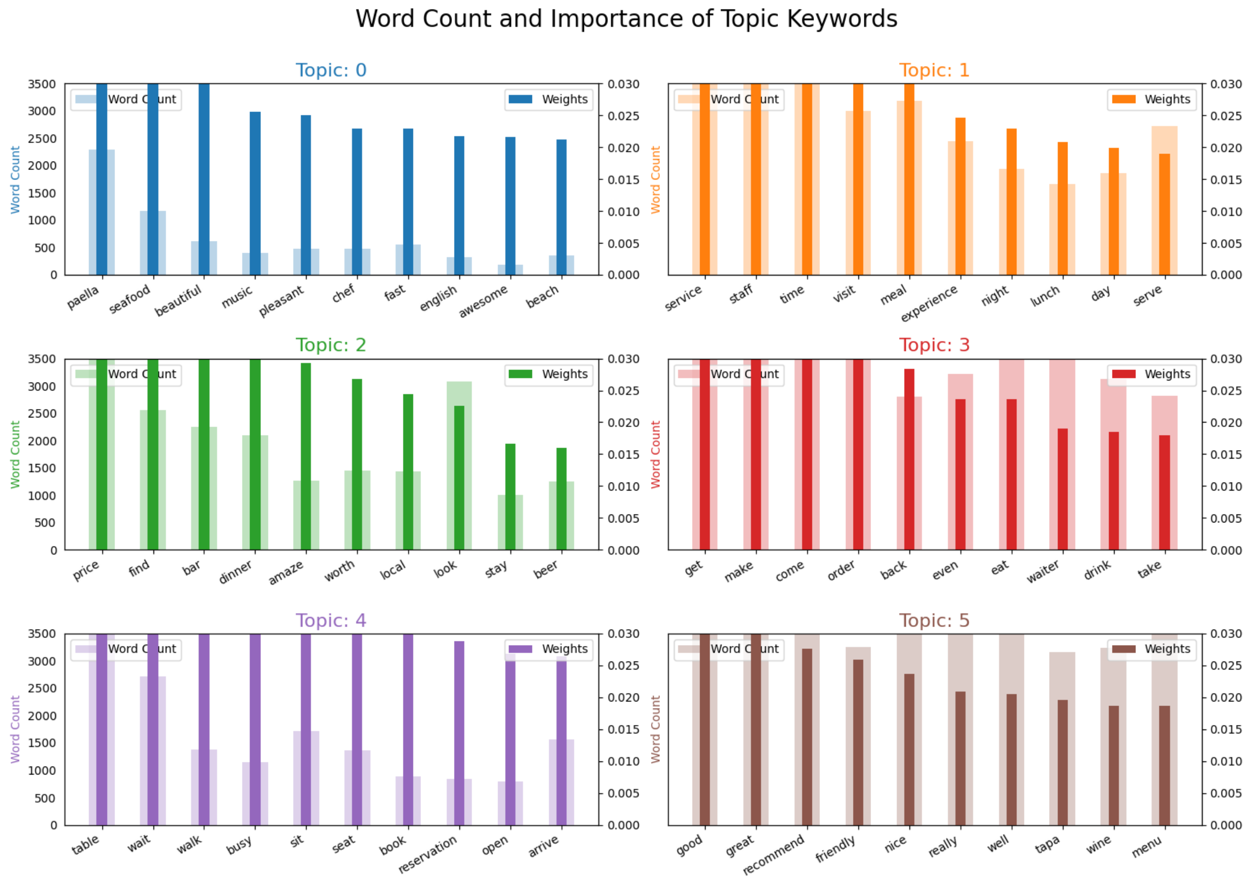
Topic 4: *table, wait, walk, busy, sit, seat, book, reservation, open, arrive, long, breakfast, minute, coffee, early, hour, pm, incredible, able, ambiance, door, later, space, easy, vegan, brunch, line, accommodate, front, head, enjoyable, host, gorgeous, finally, square, empty, tender, language, manage, greet*

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*

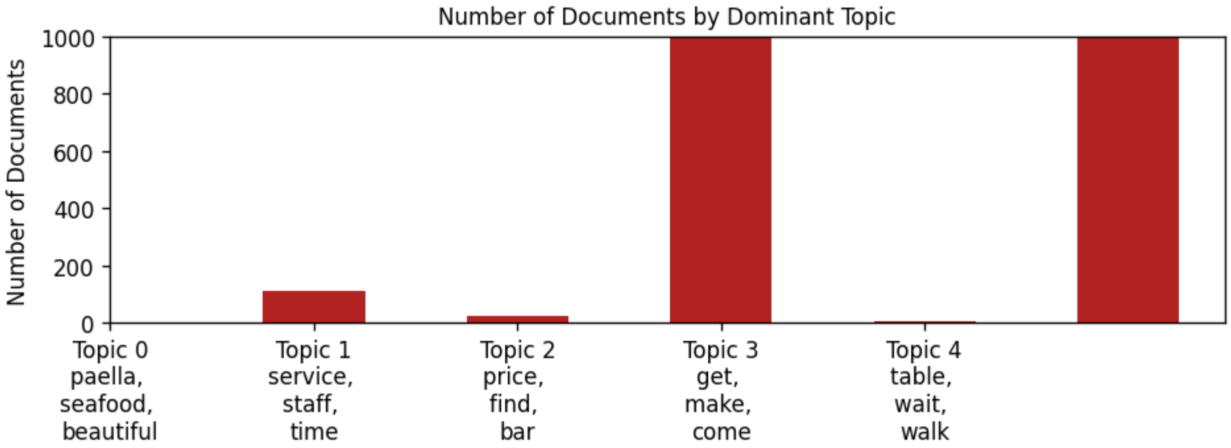
The word cloud for the 6 identified Topics are as:

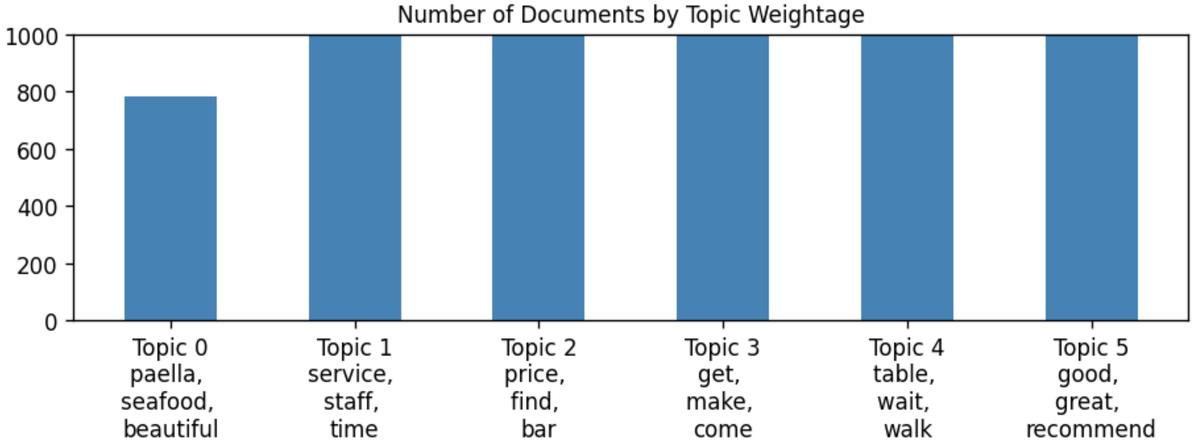


From the Graph below, we see that Topic 5 and Topic 3 has the most weighted keywords considering the bag-of-words in the corpus.



From the graph below, we see that the most discussed Topic is Topic 5 and Topic 3 (Dominant Topic).





The final list of Topics identified by importance and most important keywords are:

- ***Topic 5:*** *'good, great, recommend, friendly, nice, really, well, tapa, wine, menu',*
- ***Topic 3:*** *'get, make, come, order, back, even, eat, waiter, drink, take',*
- ***Topic 1:*** *'service, staff, time, visit, meal, experience, night, lunch, day, serve',*
- ***Topic 2:*** *'price, find, bar, dinner, amaze, worth, local, look, stay, beer',*
- ***Topic 4:*** *'table, wait, walk, busy, sit, seat, book, reservation, open, arrive'.*

8. Train and Test Split-

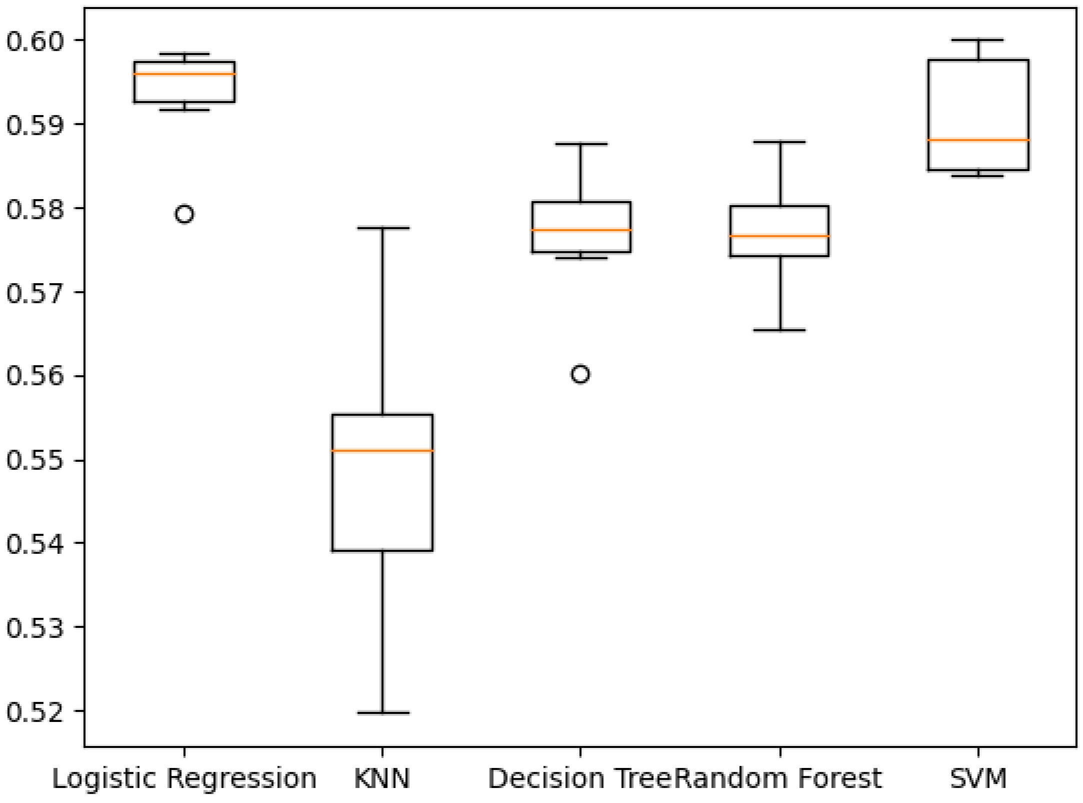
For this Classification problem, we will consider the wordcount and 6 Topics as Regressor to classify the reviews. The training and test split of the Dataset was 80%-20% after upsampling the dataset.

Modelling:

As our Classifiers, we considered 5 different Algorithms-

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- SVM

Below is the box plot of the distribution of the k-fold=6 cross validation score for different classifiers-



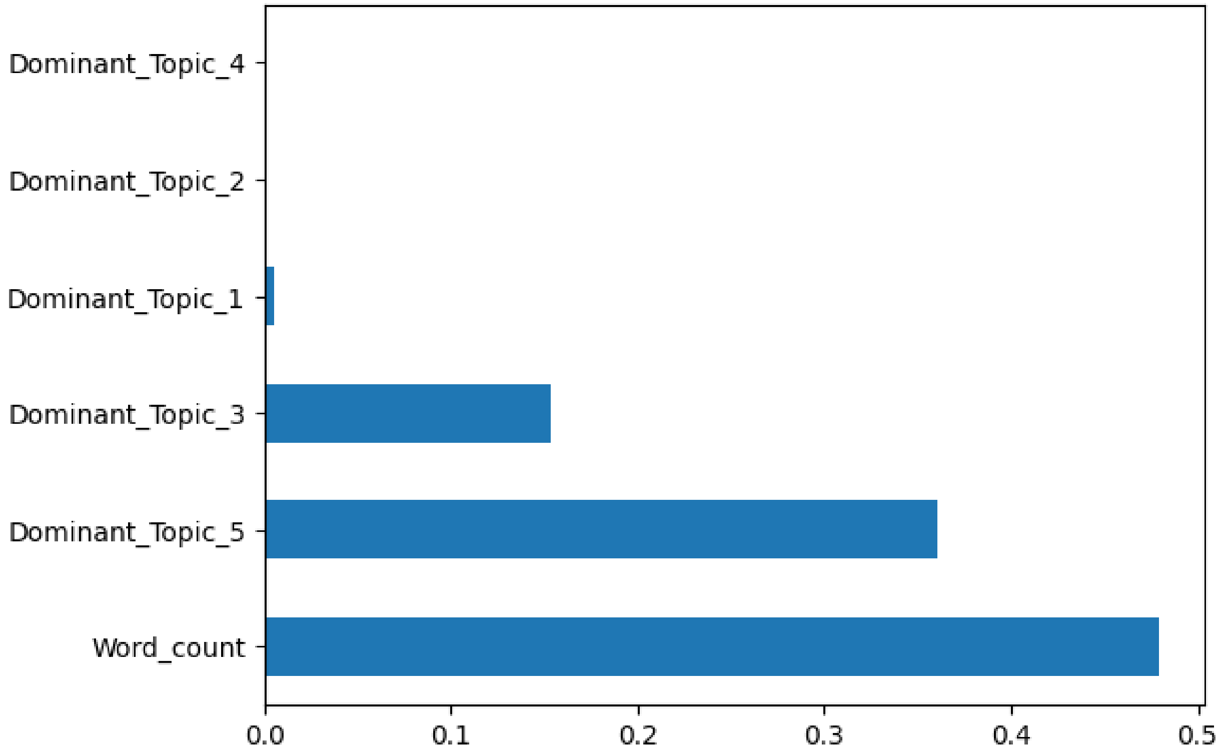
Our Parameters of consideration on Model Performance would be-

- Accuracy
- F1 Score (for Precision and Recall).

The Random Forest and Logistic Regression models performed well in terms of accuracy (RF 58.25% accurate and Logistic 59.05%).

But, we would also emphasize on the F1 score since our count of false positives and false negatives. For business policy planning and design, it's the management's call to decide on which performance metrics should be prioritized (Precision or Recall); in other words- which segment of customers should we target to churn. This will help to contribute towards developing effective customer retention policies for the company.

While identifying the important features for modeling, I found that wordcount, Topic 5 and Topic 3 are the most important factors that contributed to the performance.



Summary:

There might be varieties of classifiers that might classify the Sentiment accurately/ precisely, but what I found fascinating was- the idea of Topics as Regressors for classifying Sentiment. I found this to be a more modern method for Dimensionality Reduction and Categorizing Important Topics. Up to now, Topic Modelling has been implemented as Unsupervised Learning. In this analysis, identification of the Topics are not only used for text processing and categorizing keywords, but also for predictive modeling.

- Considering the important features, the **Word Count of reviews showed up to be the most contributing factor**, from which we can interpret that- people tend to use more words while leaving negative reviews than positive.
- **Topic 5 and Topic 3 has overall proven to be the Dominant topic** with important keywords.

This categorization of keywords can lead us to a whole new dimension of exploring not only for the overall sentiment, but also exploring more business prospects in terms of-

  - demography (e.g. these keywords are for the Barcelona dataset, the Topics finding for other 5 cities might be interesting and comparable),
  - business opportunity (findings on the associated Topic keywords for restaurant specific records and customer interest e.g. service, food, ambiance, menu etc.

The scope for further analysis lies on different factors according to the findings, majorly due to high performance computational support. (This analysis was run on Google Colab).

The further analysis has prospect for:

- Building a Pipeline.
- Fit a Neural Network Model for Topic Modelling.
- Cross Validation over sample selection and Topic Identification.
- Automation in Topic Labeling based on high frequency keywords.

Acknowledgement:

- Dataset: The details on the dataset can be found here at Kaggle.
- Analysis: The Github Repo of this Analysis is here.
- Tools and Software: Python 3.9 (Most Commonly libraries listed below):
    - Numpy,
    - Pandas,
    - Matplotlib,
    - Seaborn,
    - NLTK,
    - Scikitlearn
    - Genism
    - Spacy.