Università degli studi di Milano - Bicocca
A.Y. 24/25

CALIFORNIA STAYS

Explainable Sentiment Analysis on TripAdvisor Data

Gamberetti: Alice Brunazzi, Daniele Lepre, Pedrazzini Ernesto

DATA ANALYSIS & CLEANING

1- Cleaning the price field:
Identified issues: prices with non numeric symbols, text expressions and price ranges, not suitable for further analysis

2 - Converting the ratings field:
Identified not-numeric values (strings), inconsistent formats (or missing values).

3 - Extracting and cleaning cities
The address field in the database contains HTML tags, and have extra information. The extracted city is then normalized by capitalizing the first letter of each word.

DATA QUERYING

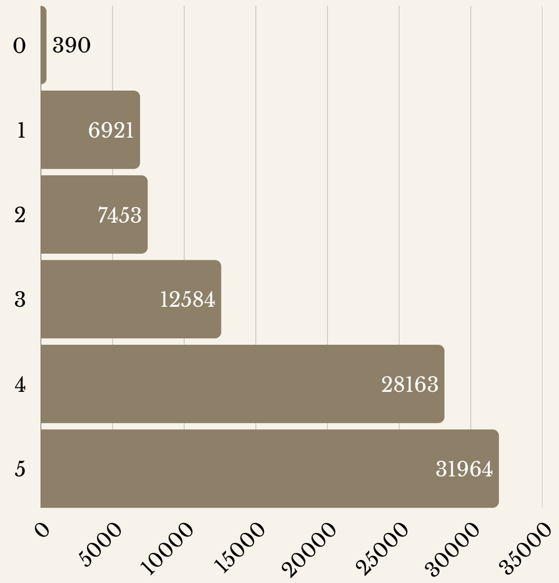
Number of autors: 72 595

Number of reviews: 87 475

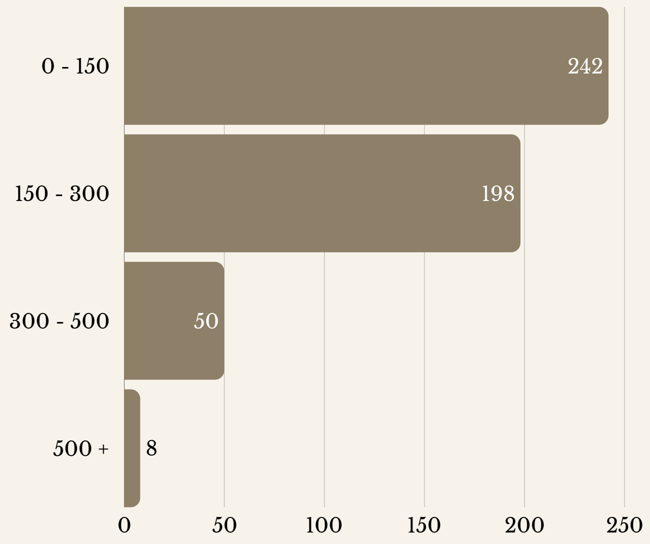
Number of reviews per author: 1.2

Average overall rating: 3.7959 stars

Reviews distribution per overall category:



Hotel distribution per average price:



Average overall rating for each price range:

0 - 150: 3.08 stars

150 - 300: 3.73 stars

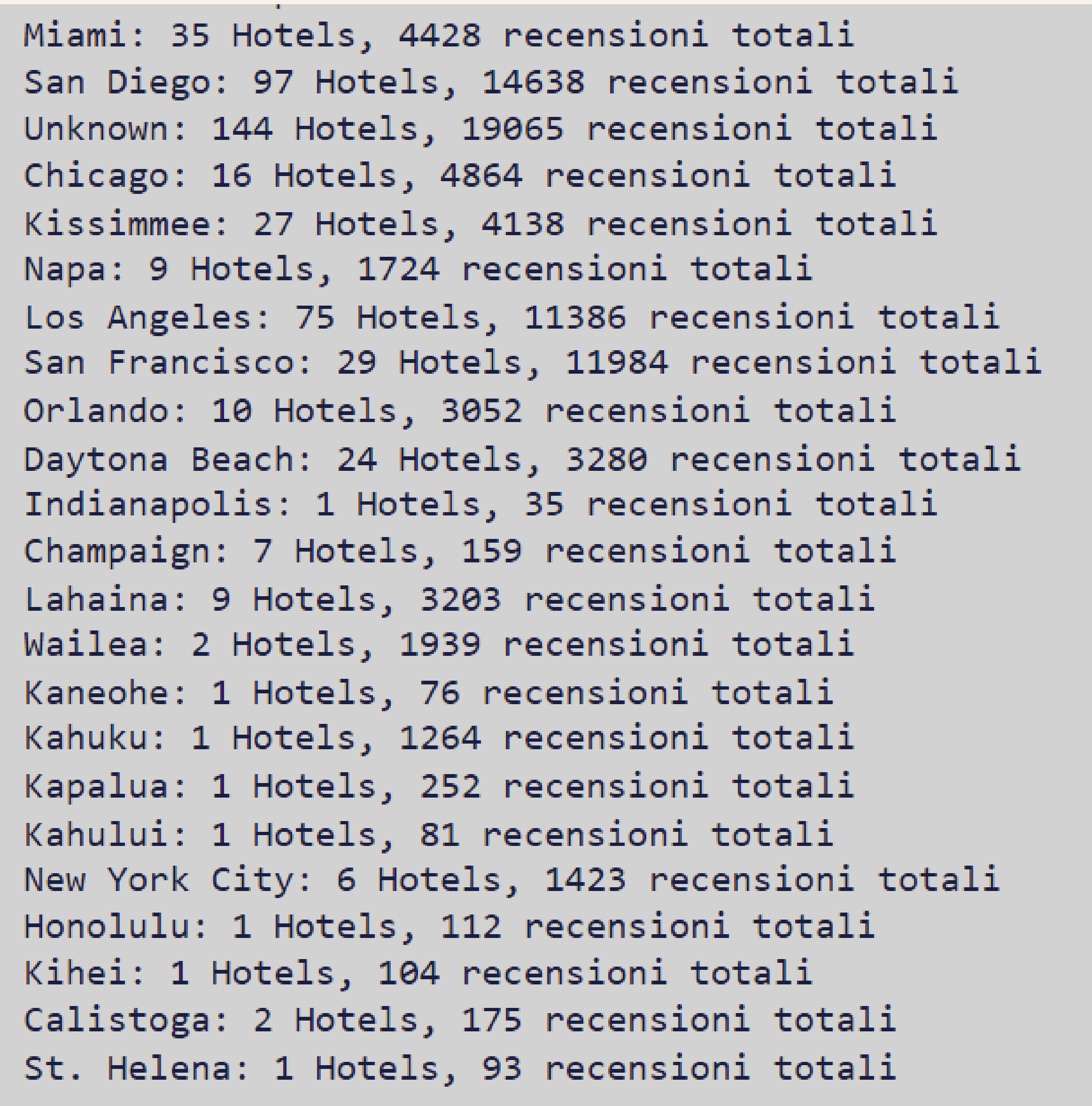
300 - 500: 4.02 stars

500 + : 4.40 stars

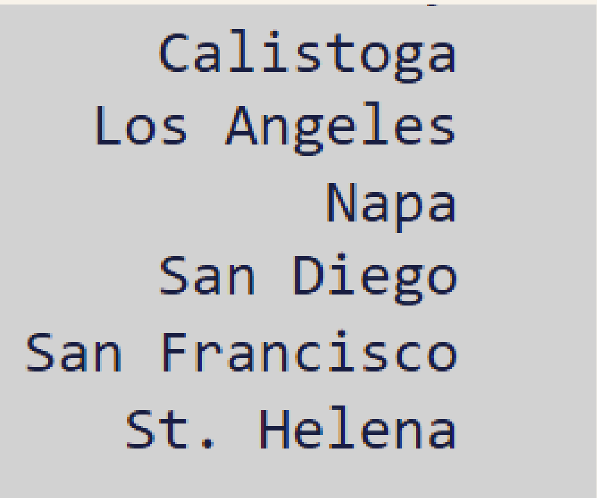
Number of reviews and average price in California

City	Positive	Negartive	Average price
Calistoge	50	11	152.25
Los Angeles	563	80	170.96
Napa	1003	75	222.78
San Diego	3812	580	164.74
San Francisco	2920	177	328.59

DATA FILTERING

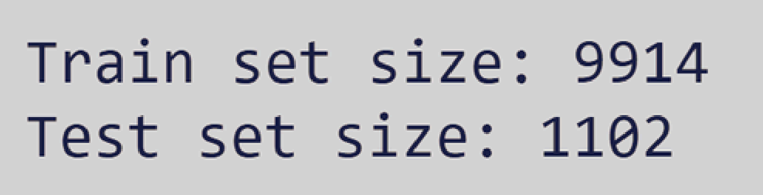


```
Miami: 35 Hotels, 4428 recensioni totali
San Diego: 97 Hotels, 14638 recensioni totali
Unknown: 144 Hotels, 19065 recensioni totali
Chicago: 16 Hotels, 4864 recensioni totali
Kissimmee: 27 Hotels, 4138 recensioni totali
Napa: 9 Hotels, 1724 recensioni totali
Los Angeles: 75 Hotels, 11386 recensioni totali
San Francisco: 29 Hotels, 11984 recensioni totali
Orlando: 10 Hotels, 3052 recensioni totali
Daytona Beach: 24 Hotels, 3280 recensioni totali
Indianapolis: 1 Hotels, 35 recensioni totali
Champaign: 7 Hotels, 159 recensioni totali
Lahaina: 9 Hotels, 3203 recensioni totali
Wailea: 2 Hotels, 1939 recensioni totali
Kaneohe: 1 Hotels, 76 recensioni totali
Kahuku: 1 Hotels, 1264 recensioni totali
Kapalua: 1 Hotels, 252 recensioni totali
Kahului: 1 Hotels, 81 recensioni totali
New York City: 6 Hotels, 1423 recensioni totali
Honolulu: 1 Hotels, 112 recensioni totali
Kihei: 1 Hotels, 104 recensioni totali
Calistoga: 2 Hotels, 175 recensioni totali
St. Helena: 1 Hotels, 93 recensioni totali
```



```
    Calistoga
  Los Angeles
         Napa
    San Diego
San Francisco
   St. Helena
```

	POSITIVE	NEGATIVE
PRE BALANCING	28819	5508
POST BALANCING	5508	5508



```
Train set size: 9914
Test set size: 1102
```

SENTIMENT & CLASSIFICATION

01 - Preprocess

1. Data Filtering : city-based
2. Sentiment Division: each review is labeled as **Positive** if equal or greater then 4, **Negative** if smaller or equal to 2 and greater than 0
3. Dataset Balancing and shuffling (stratified sampling)
4. Train - Test splitting, with **10%** of the created dataset being used as a test set
5. Text Vectorizatio: TF-IDF Vectorizer **OR** BERT tokenizer

02 - Classification

1. Model selection (Random Forest, Support Vector Machine and Feed Forward Neural Network)
2. Pipeline Creation
3. Model training
4. Prediction

03 - Results

MODELS - TD-IDF

01 - RANDOM FOREST

WHY:
Computationally efficient, good generalization

ACCURACY:
89.8%

499	52
60	491

02 - SVM

WHY:
Excels in text classification, better generalization

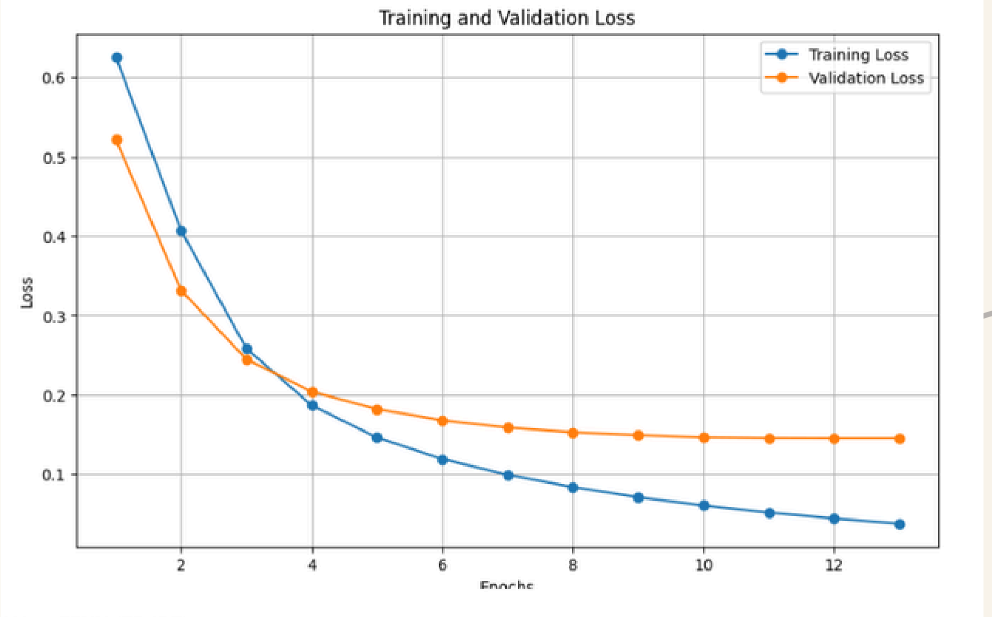
ACCURACY:
92.8%

520	31
37	514

03 - FFNN

WHY:
Computationally expensive, tries to capture non-linear patterns

ACCURACY:
94.2%

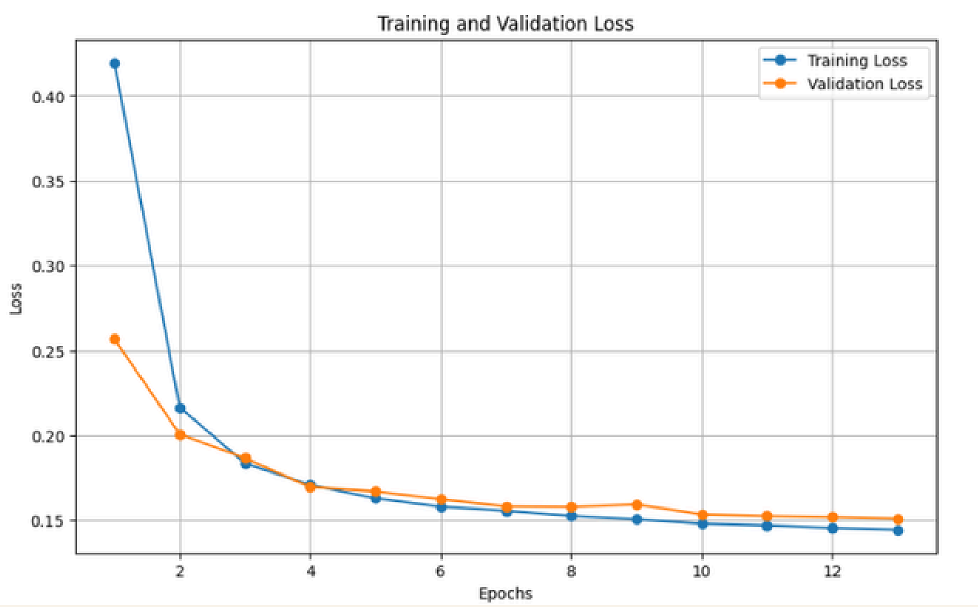


MODELS - BERT

TD-IDF : Every word is an isolated token, the more representative a word, the more weigth it will have. Stop-words are removed (149)
BERT : The embedding is contextual, the embedding is dense. No stop-words removed.

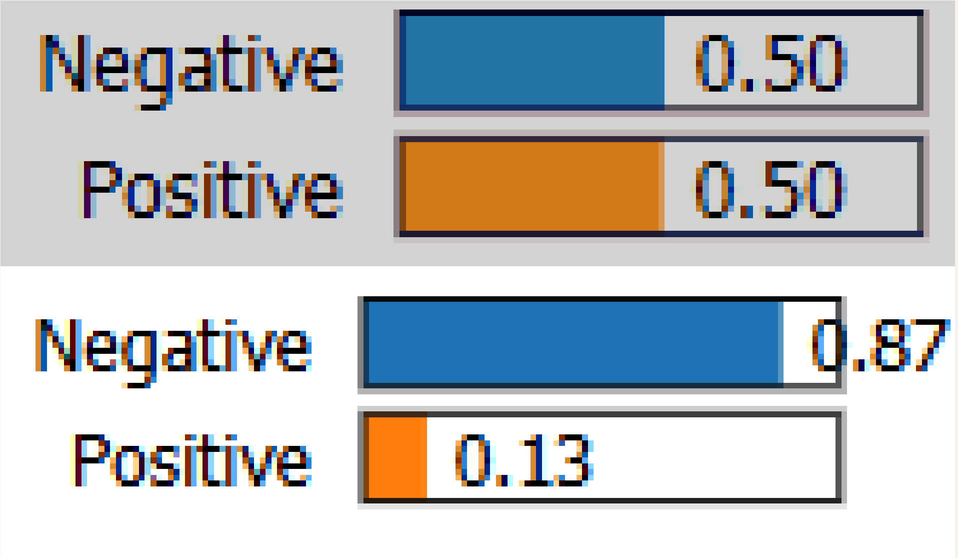
04 - FFNN

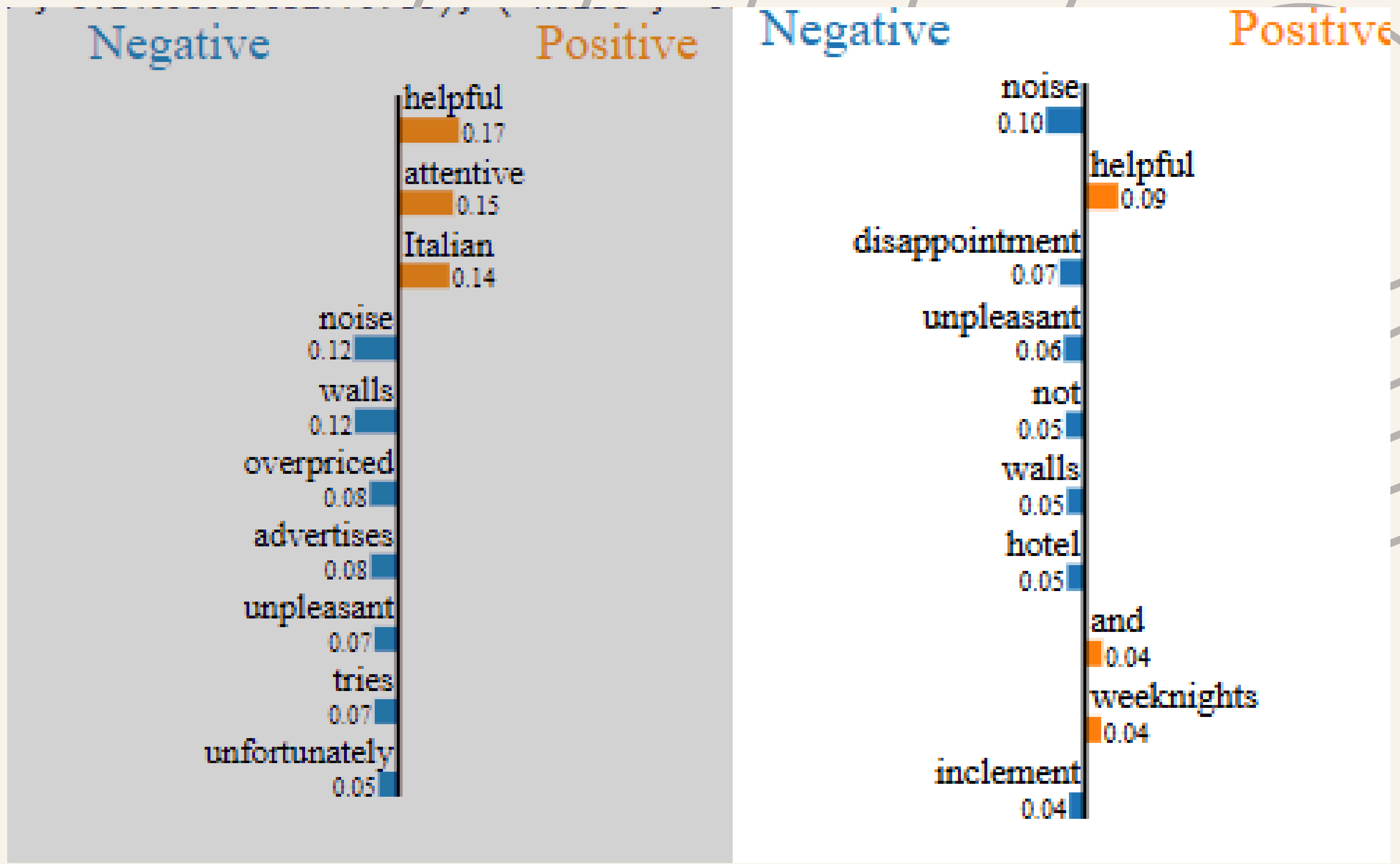
ACCURACY:
94.2%

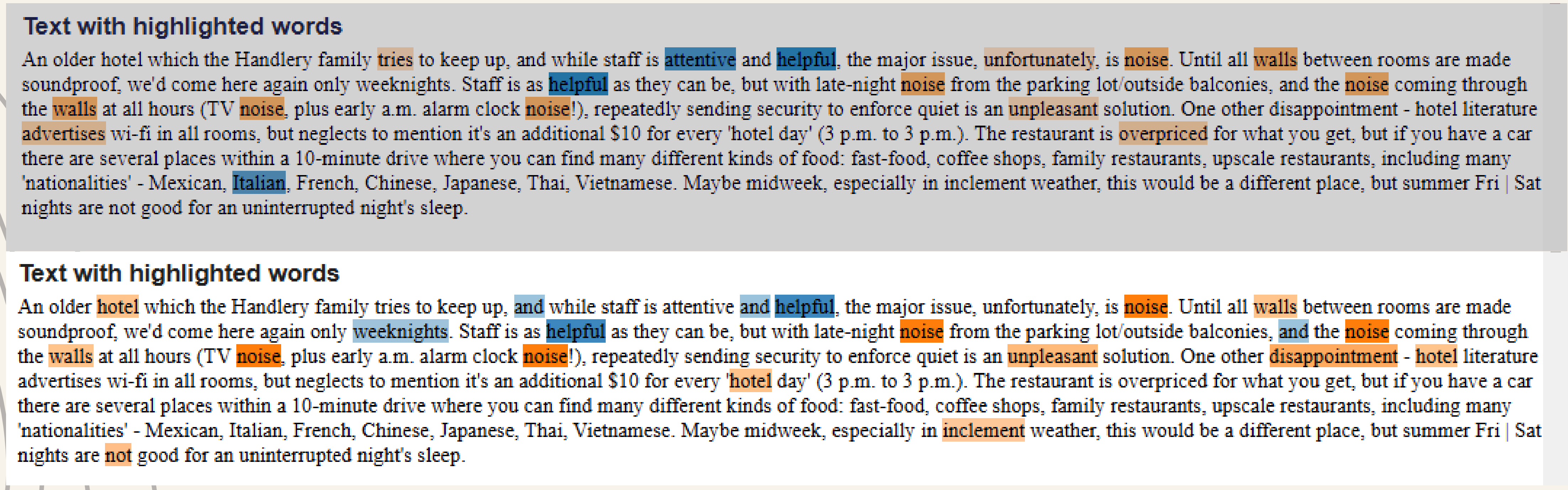


EXPLAINABLE AI

MODEL : FFNN, XAI: LIME, IDX: 52







Text with highlighted words

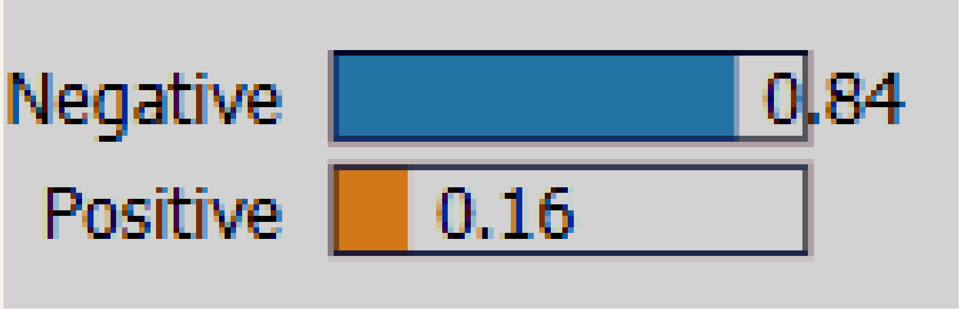
An older hotel which the Handlery family tries to keep up, and while staff is attentive and helpful, the major issue, unfortunately, is noise. Until all walls between rooms are made soundproof, we'd come here again only weeknights. Staff is as helpful as they can be, but with late-night noise from the parking lot/outside balconies, and the noise coming through the walls at all hours (TV noise, plus early a.m. alarm clock noise!), repeatedly sending security to enforce quiet is an unpleasant solution. One other disappointment - hotel literature advertises wi-fi in all rooms, but neglects to mention it's an additional $10 for every 'hotel day' (3 p.m. to 3 p.m.). The restaurant is overpriced for what you get, but if you have a car there are several places within a 10-minute drive where you can find many different kinds of food: fast-food, coffee shops, family restaurants, upscale restaurants, including many 'nationalities' - Mexican, Italian, French, Chinese, Japanese, Thai, Vietnamese. Maybe midweek, especially in inclement weather, this would be a different place, but summer Fri | Sat nights are not good for an uninterrupted night's sleep.

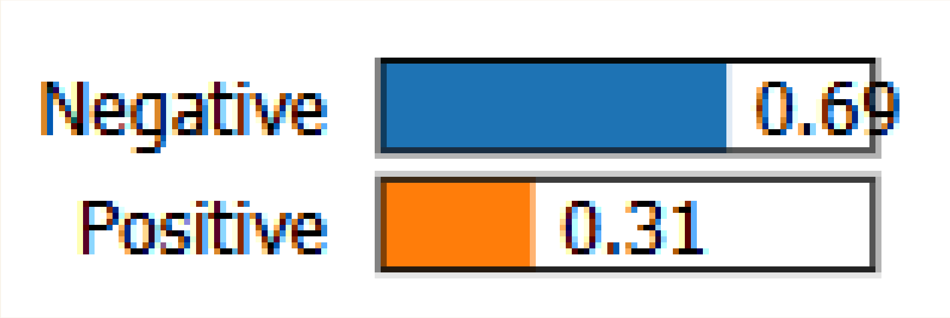
Text with highlighted words

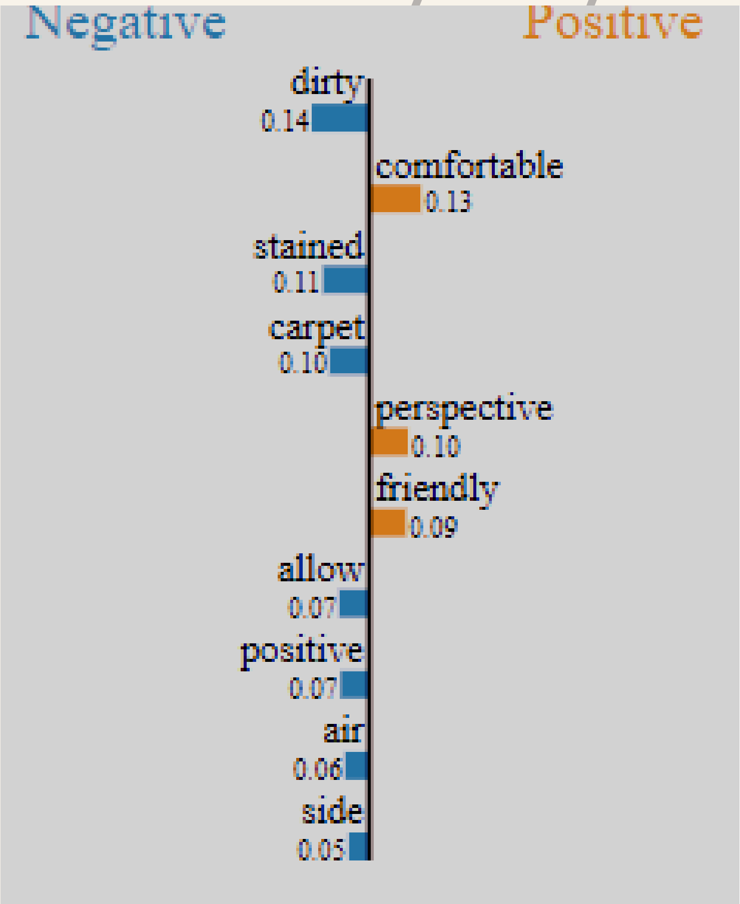
An older hotel which the Handlery family tries to keep up, and while staff is attentive and helpful, the major issue, unfortunately, is noise. Until all walls between rooms are made soundproof, we'd come here again only weeknights. Staff is as helpful as they can be, but with late-night noise from the parking lot/outside balconies, and the noise coming through the walls at all hours (TV noise, plus early a.m. alarm clock noise!), repeatedly sending security to enforce quiet is an unpleasant solution. One other disappointment - hotel literature advertises wi-fi in all rooms, but neglects to mention it's an additional $10 for every 'hotel day' (3 p.m. to 3 p.m.). The restaurant is overpriced for what you get, but if you have a car there are several places within a 10-minute drive where you can find many different kinds of food: fast-food, coffee shops, family restaurants, upscale restaurants, including many 'nationalities' - Mexican, Italian, French, Chinese, Japanese, Thai, Vietnamese. Maybe midweek, especially in inclement weather, this would be a different place, but summer Fri | Sat nights are not good for an uninterrupted night's sleep.

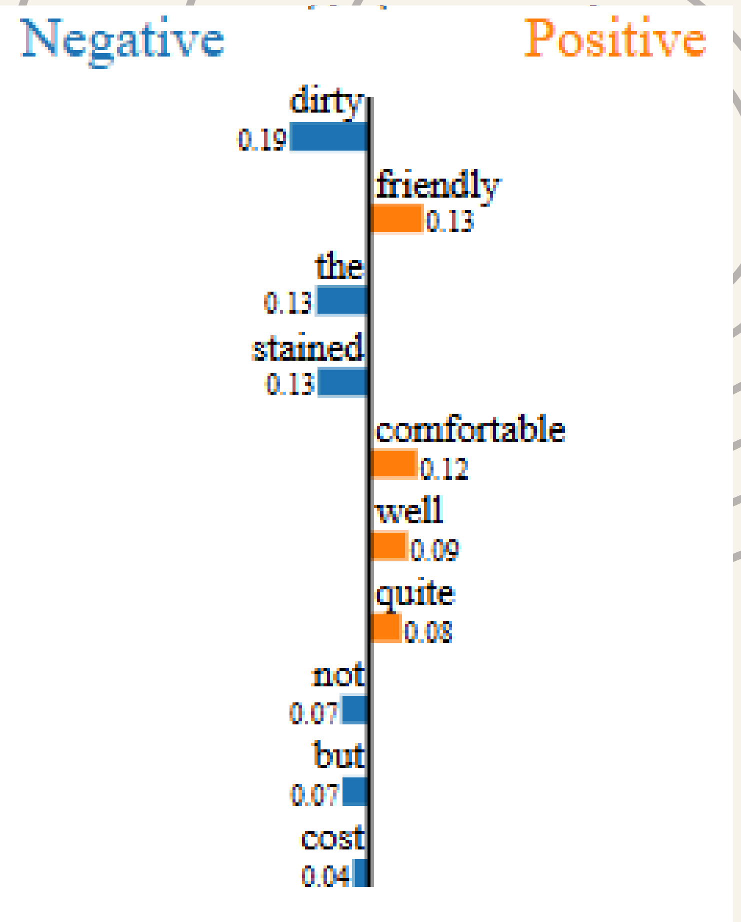
EXPLAINABLE AI

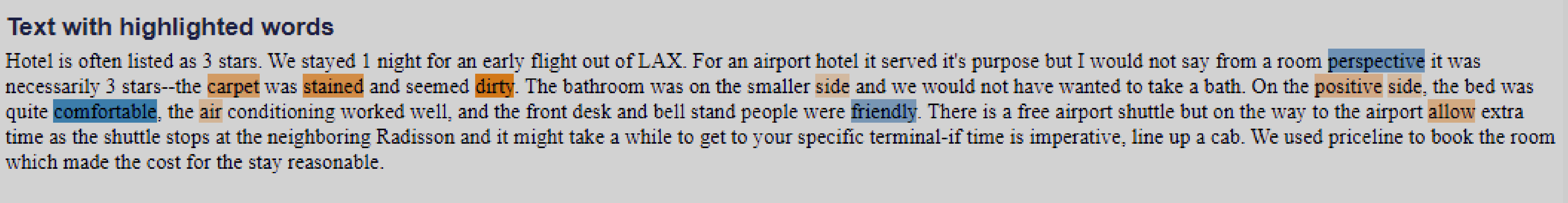
MODEL : FFNN, XAI: LIME, IDX: 43





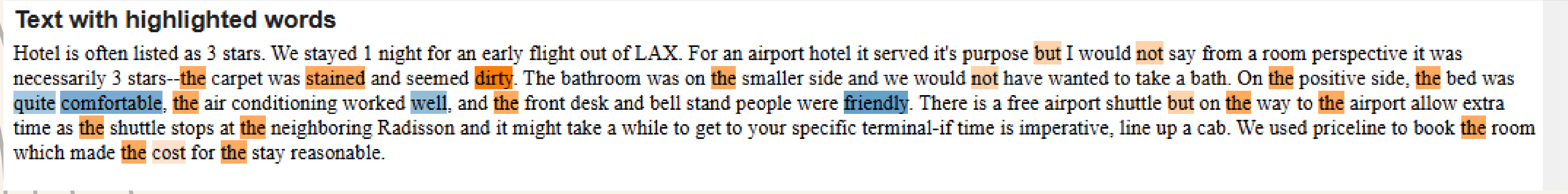






Text with highlighted words

Hotel is often listed as 3 stars. We stayed 1 night for an early flight out of LAX. For an airport hotel it served it's purpose but I would not say from a room perspective it was necessarily 3 stars--the carpet was stained and seemed dirty. The bathroom was on the smaller side and we would not have wanted to take a bath. On the positive side, the bed was quite comfortable, the air conditioning worked well, and the front desk and bell stand people were friendly. There is a free airport shuttle but on the way to the airport allow extra time as the shuttle stops at the neighboring Radisson and it might take a while to get to your specific terminal-if time is imperative, line up a cab. We used priceline to book the room which made the cost for the stay reasonable.



Text with highlighted words

Hotel is often listed as 3 stars. We stayed 1 night for an early flight out of LAX. For an airport hotel it served it's purpose but I would not say from a room perspective it was necessarily 3 stars--the carpet was stained and seemed dirty. The bathroom was on the smaller side and we would not have wanted to take a bath. On the positive side, the bed was quite comfortable, the air conditioning worked well, and the front desk and bell stand people were friendly. There is a free airport shuttle but on the way to the airport allow extra time as the shuttle stops at the neighboring Radisson and it might take a while to get to your specific terminal-if time is imperative, line up a cab. We used priceline to book the room which made the cost for the stay reasonable.

XAI

LIME

WHY? Computational efficiency and semplicity. Local.

CONSIDERATIONS:

1. The terms "great", "perfect", and "clean" are consistently identified across all models.
2. The term "dump", identified by two models, carries a strong negative impact.
3. Explanations from BERT are less intuitive compared to other models, as they are not tied to single words but rather to broader semantic concepts.

Thank You