BEATRIZ AGUIAR JOÃO MARINHO MARGARIDA VIEIRA

GRUPO 96

NATURAL LANGUAGE PROCESSING

AMAZON REVIEWS FOR SA FINE-GRAINED 5 CLASES CSV

INTELIGÊNCIA ARTIFICIAL

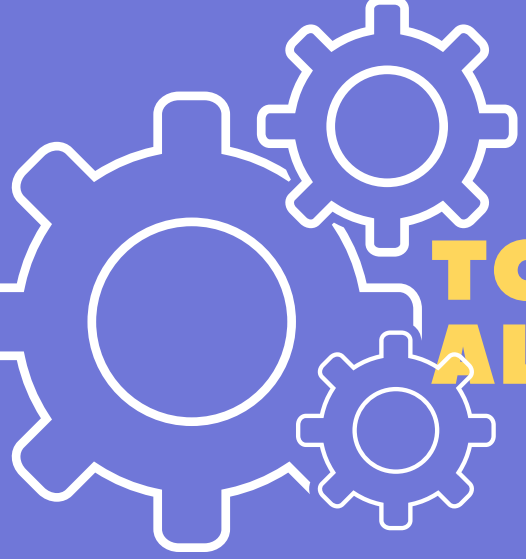
SPECIFICATION OF THE PROBLEM

The aim is to process the amazon reviews' textual data, employing diverse techniques to transform them into appropriate datasets that can then be addressed using supervised learning algorithms to assign each review to one of the 5 available classes.

An initial data preprocessing and exploratory analysis was carried out - class distribution, word distribution per class, data cleaning and so on.

When fully analysed and processed, the new dataset is used in 6 different machine learning algorithms, combined with different data set generating processes.

The results of the different models will be further presented and analysed.



GRUPO 96

TOOLS AND ALGORITHMS USED

Working environment

python, jupyter notebook

Displaying and getting data

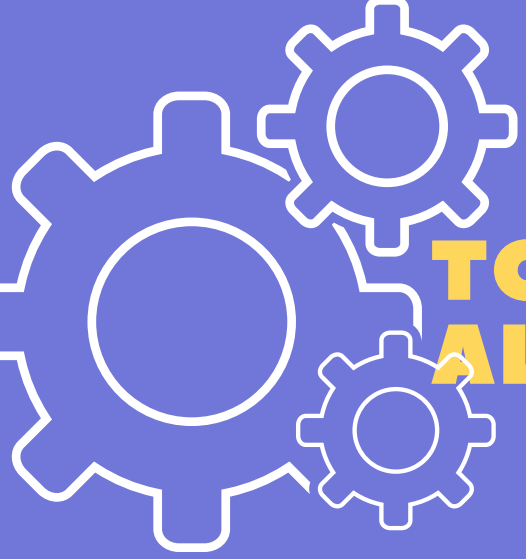
python modules: pandas, wordcloud

Preprocessing dataset

python modules: nltk, pandas, re

Generating & Training dataset

python modules: sklearn



TOOLS AND ALGORITHMS USED

Naive Bayes

Decision Trees

Neural networks

Logistic Regression

K-nearest Neighbors

Support Vector Machine

IMPLEMENTATION WORK

1 Notebook and virtual environment set up

2 Data load

3 Exploratory analysis

4 Data cleaning

5 Model and Training

6 Results and analysis

GRUPO 96

DATA PRE-PROCESSING

Based on the exploratory analysis present in the source code, the following steps were applied to the full review, i.e. review title concatenated with the review text.

1 **Lower case text**

2 **Remove ponctuation, numbers and invalid chars**

3 **Stopwords**

4 **Fix contractions**

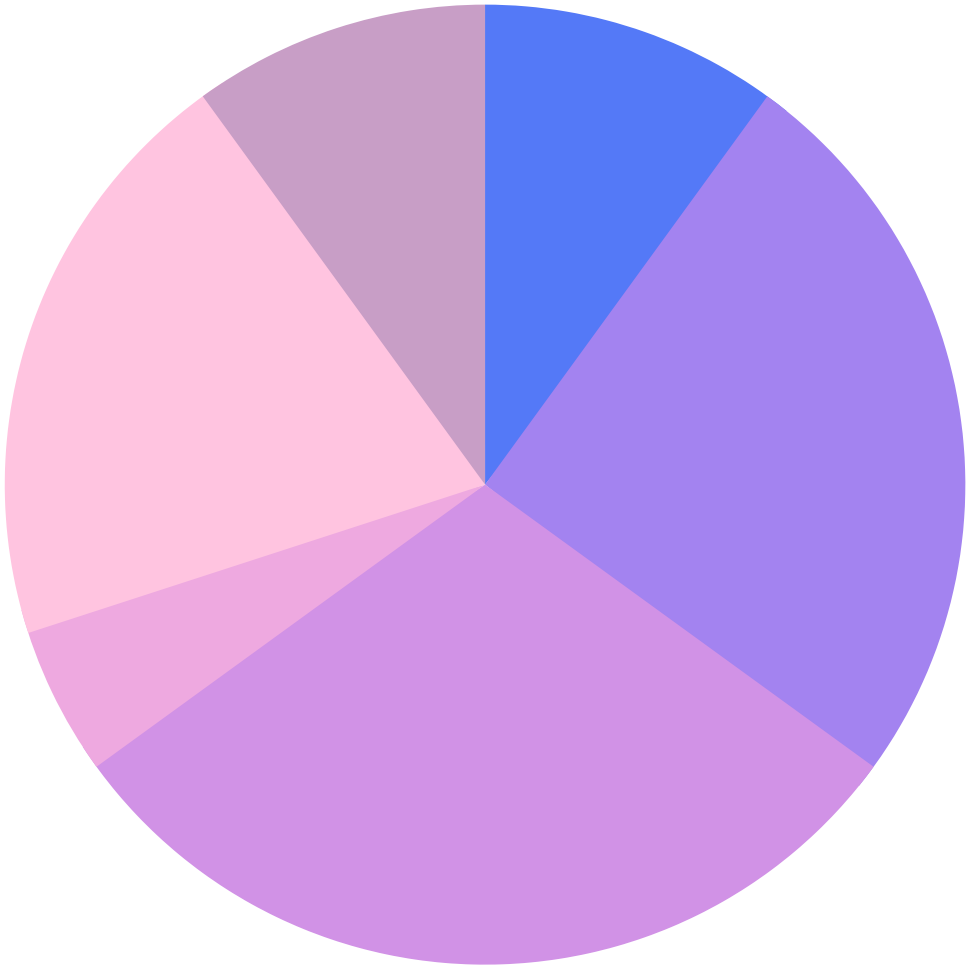
aren't → are not

5 **Stem**

disappointment → disappoint

6 **Handle negation**

not like → NOT_like



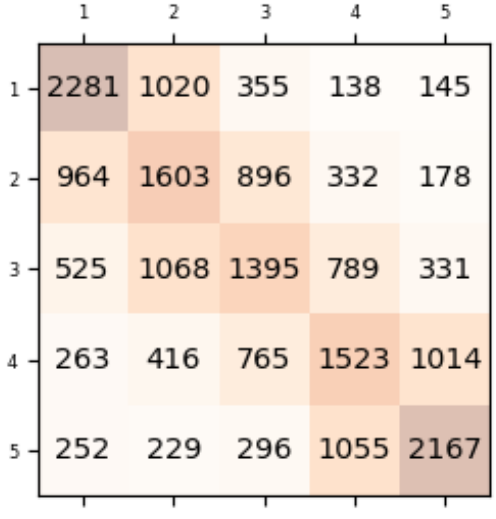
RESULTS & ANALYSIS

Matrix confusion
Accuracy
Precision
F1
Recall

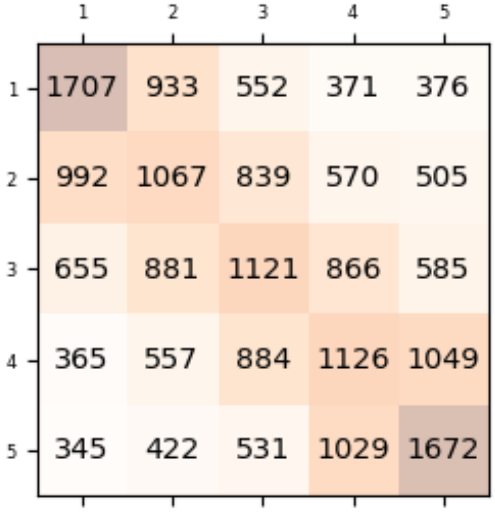
CONFUSION MATRIX

GRUPO 96

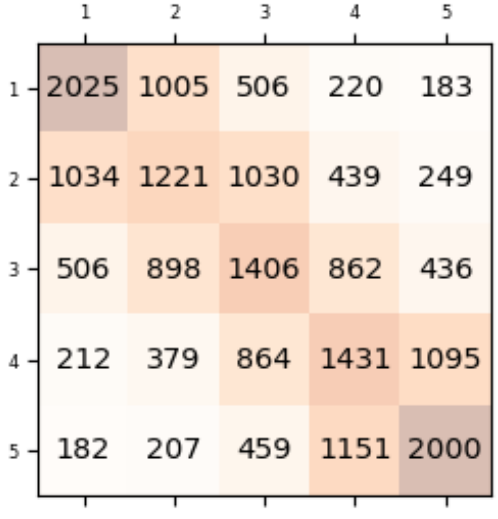
Naive Bayes



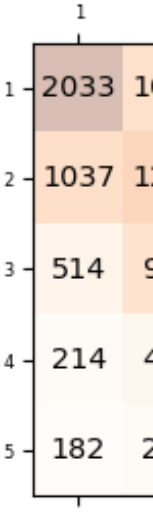
Decision Tree



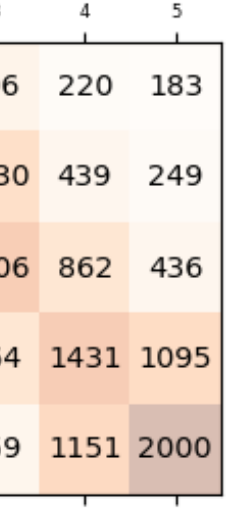
Neural Networks



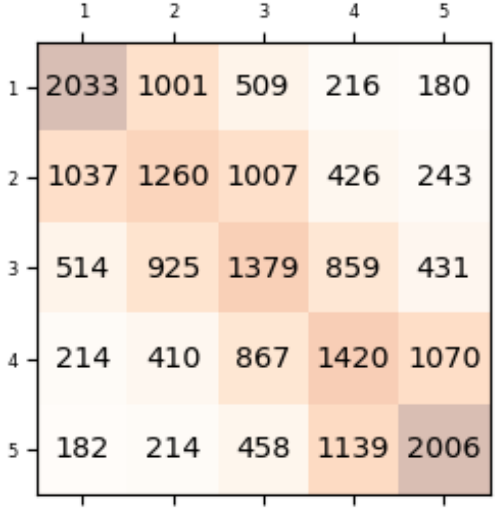
Supo



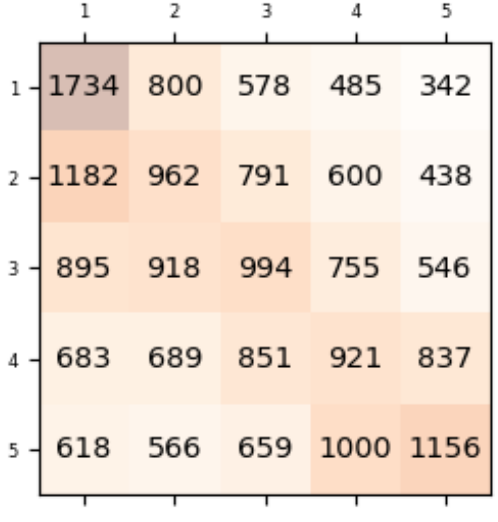
etworks



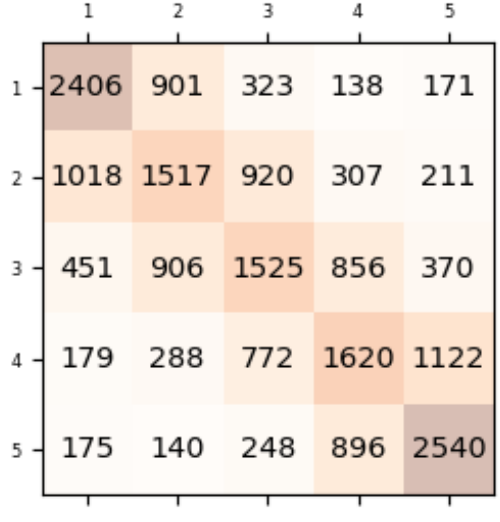
Suport Vector Machine



K-nearest Neighbors

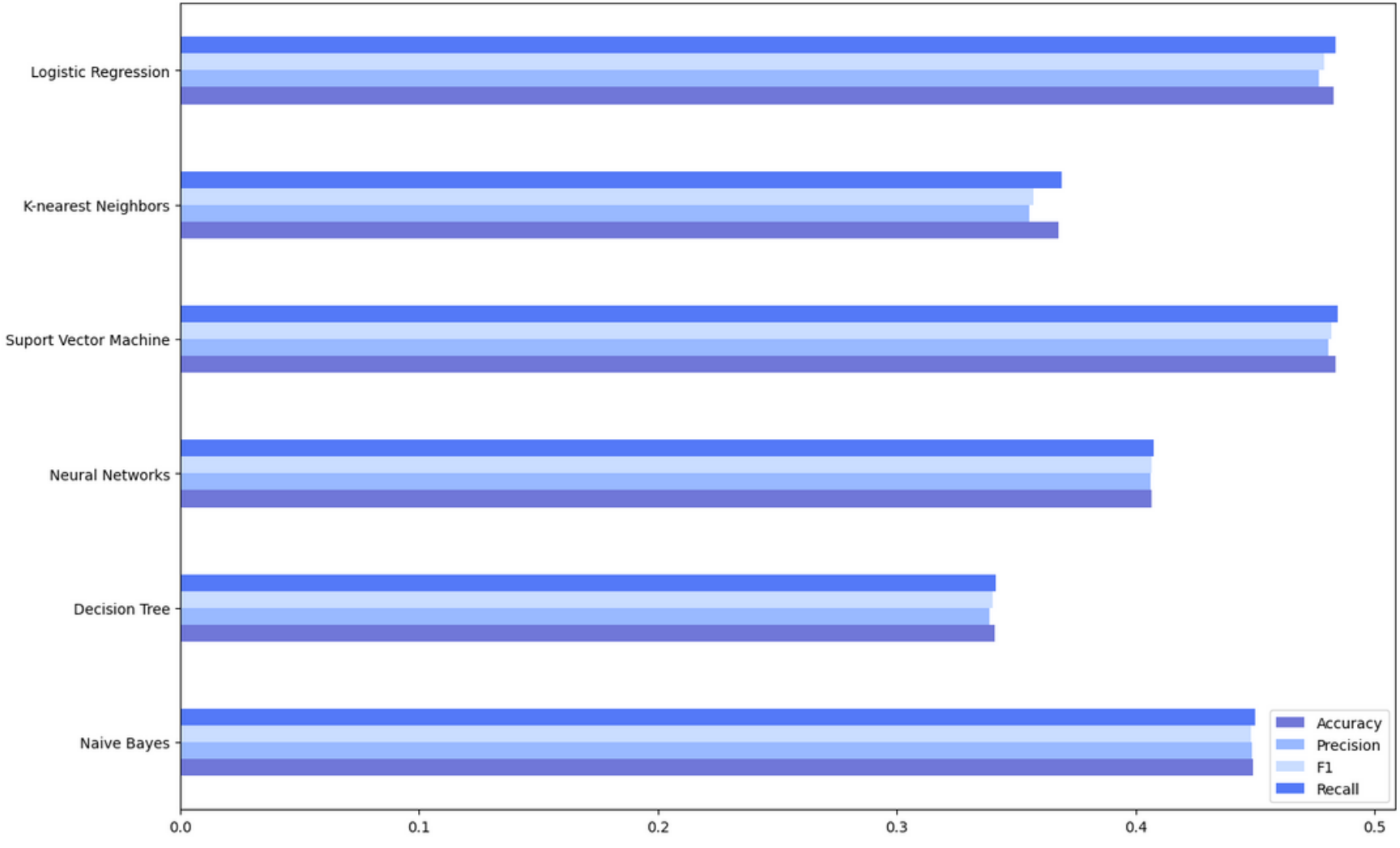


Logistic Regression



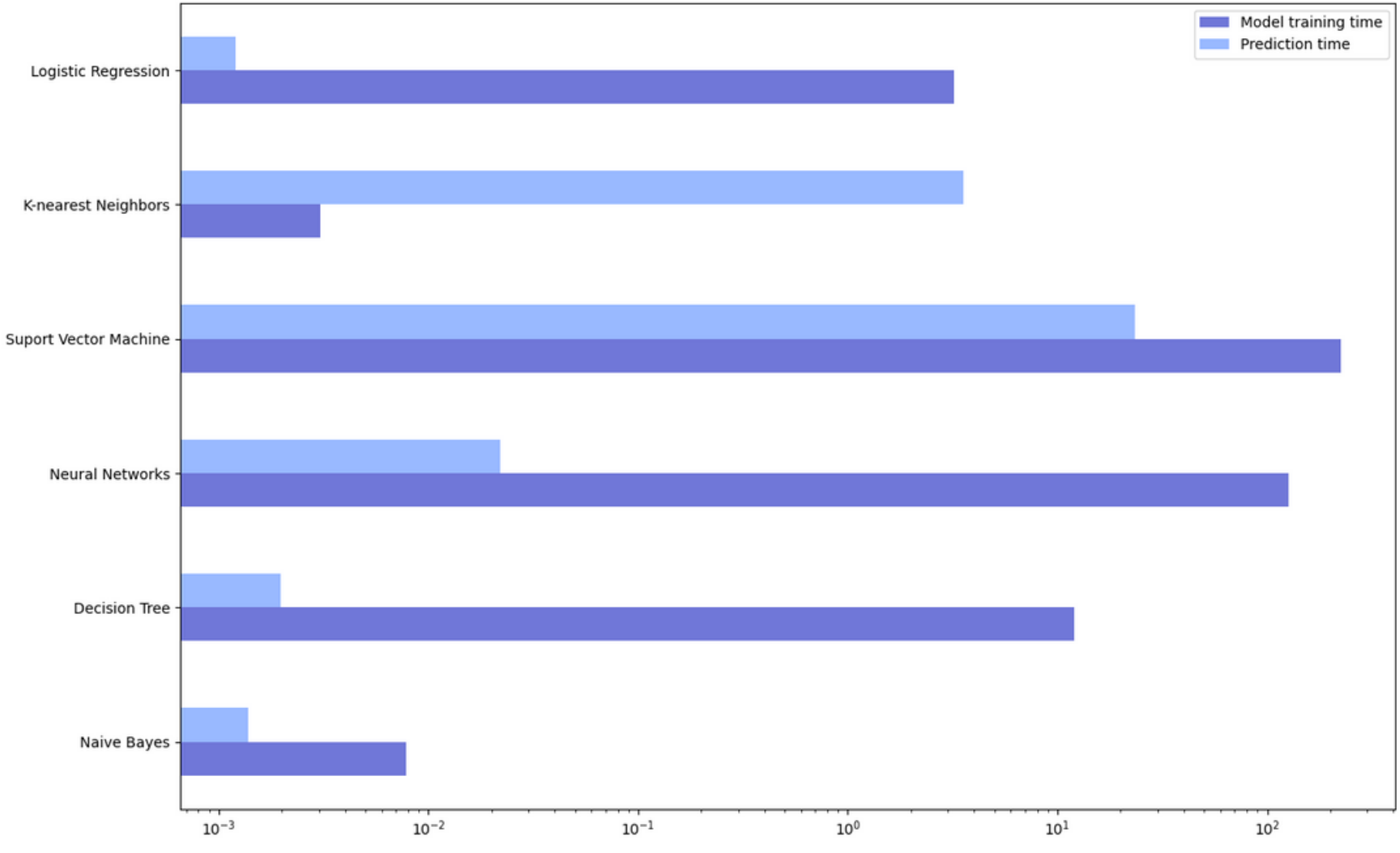
OTHER METRICS

5 TRIAL AVG



OTHER METRICS

5 TRIAL AVG



BEATRIZ AGUIAR JOÃO MARINHO MARGARIDA VIEIRA

CONCLUSIONS

Failed attempts

Using fit sample weights with the difference between the class_index and the sentiment analysis result.

Sentiment analysis

The sentiment analysis returns an average of results that meets the expected rating of the review, however there are samples that do not match. A possible explanation is the absence of sentiment key words and linguist expressions such as irony, which will also impact the training model.

Models

Naive Bayes stands out for its time efficiency. When comparing performance, the top three models are **Logistic Regression, Support Vector Matrix** and **Naive Bayes.** Overall, the Logistic Regression appeared to obtain the best results within the best time window..

Multiclass classification vs Binary classification

The classification is more subjective the wider the range of possible values, in this case 5, given that the reviewer himself will certainly not be able to strictly define the difference between a review with score 4 and 5, for example. This may explain the relatively low accuracy obtained with the models studied.

Performance evolution

The most impactful approach towards performance evolution was the change in the preprocessing step. Having started with the usual cleaning tasks, and refining them along the way, we were able to go from a maximum accuracy of 20% up to 45%.