Università degli Studi di Milano-Bicocca
AA 2022-2023

Potertì Daniele
Sanvito Alessio

Text Classification and Text Clustering on Steam Reviews Dataset

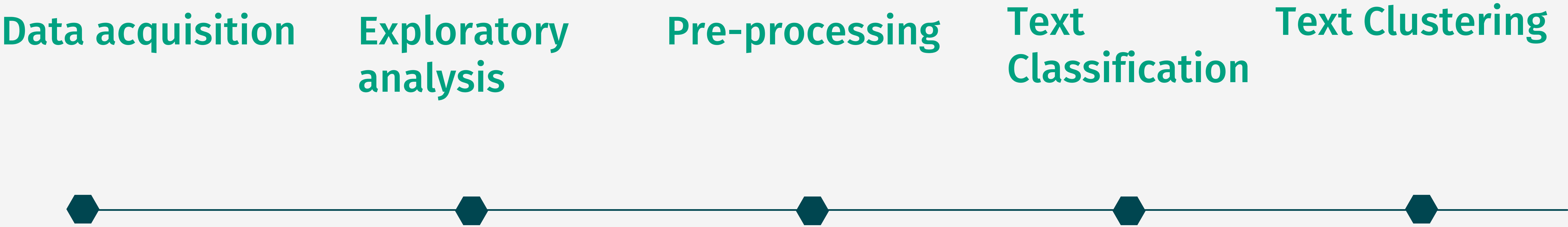
TM&S Project

Dataset

This dataset is a collection of reviews for a game on the Steam platform. Each row represents a single review, and the columns provide various details about that review.
Some of the columns in the dataset are:

- **review: the text of the review**
- **voted_up: whether the review received upvotes**
- **num_words: the number of words in the review**

Workflow



Data acquisition

Exploratory analysis

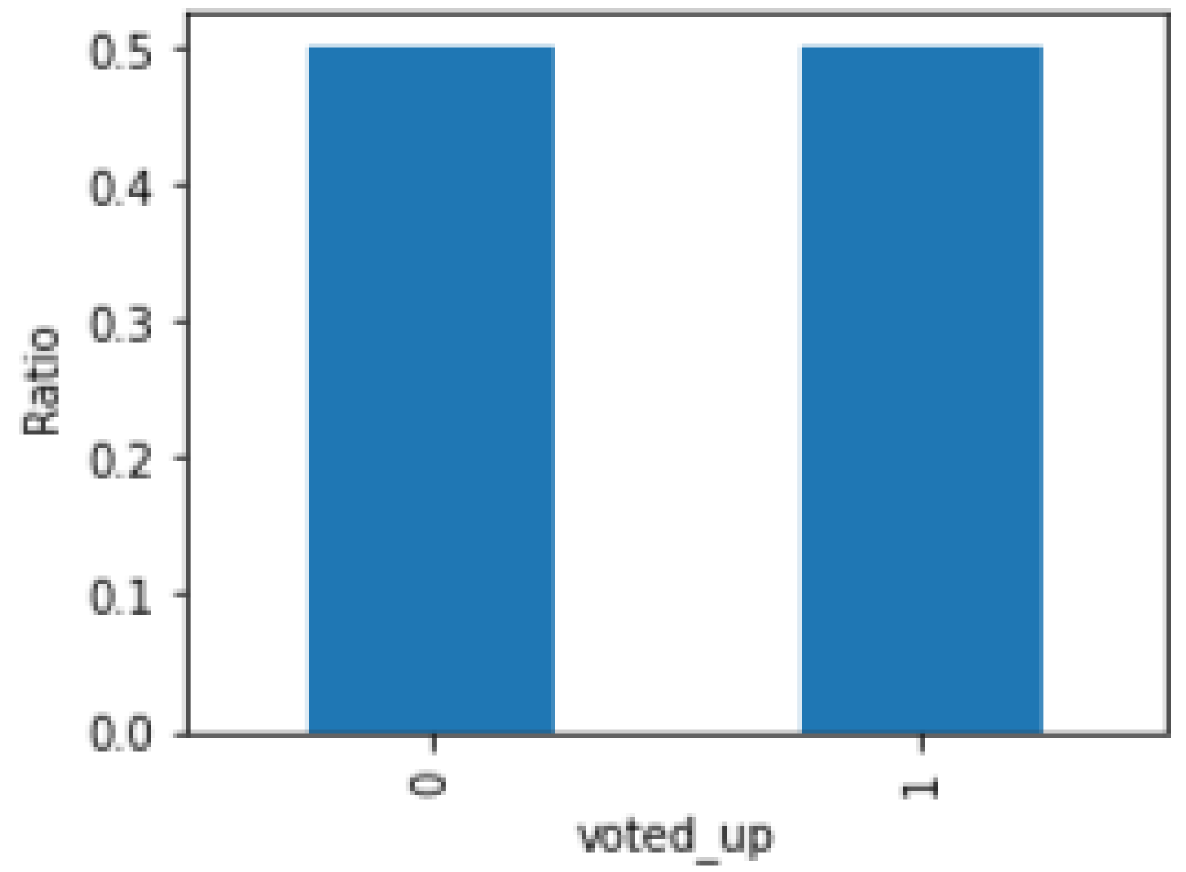
Pre-processing

Text Classification

Text Clustering

Torna al Programma

Exploratory Analysis



The values on the training dataset are balanced

Pre-processing

- Remove special characters and digits
- Transform all letters into lowercase
- Remove stop words
- Get the stem of the words (in order to reduce their size)
- Remove links using regular expressions
- Remove the most frequently and infrequently appearing words (by setting thresholds that were determined through cross-validation)
- Remove short reviews that contained fewer than a certain number of words (in order to minimize the noise in the dataset)

Sample

A sample of the data has been used in order to reduce the time needed to perform the algorithms and the computational complexity needed.

Text Representation

- Bag of Words
- TF-IDF (using trigrams)

Dimensionality Reduction (SVD)

Calculated an optimal number of components to perform SVD using the co-occurrence matrix:

- Bag of Words obtained an explained variance ratio of 0.9789503414025237
- TF-IDF obtained an explained variance ratio of 0.6141355453052243

Applying both of them to the clustering data, the results are similar, so just the tf-idf version was maintained.

Text Classification

Data Preparation

- Remove unnecessary columns from the DataFrame
- Check for any missing values or 'nan' values in the data, which could potentially skew or compromise the results of any subsequent analysis.
- These steps ensure the quality and integrity of the data being used for the analysis.

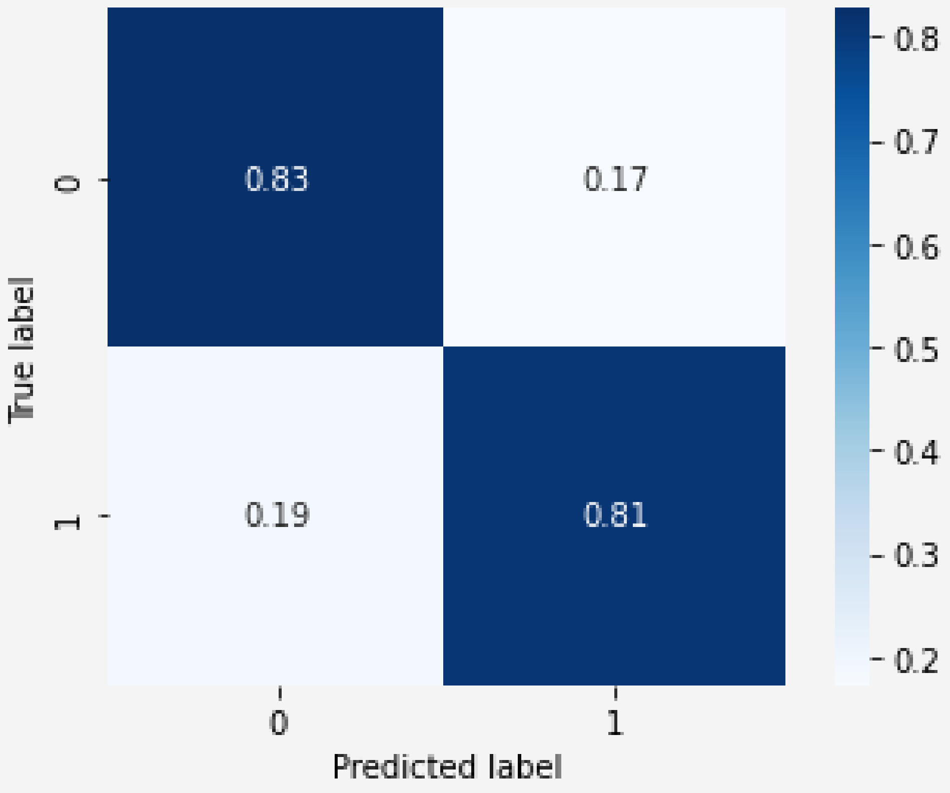
Data Distribution and Split

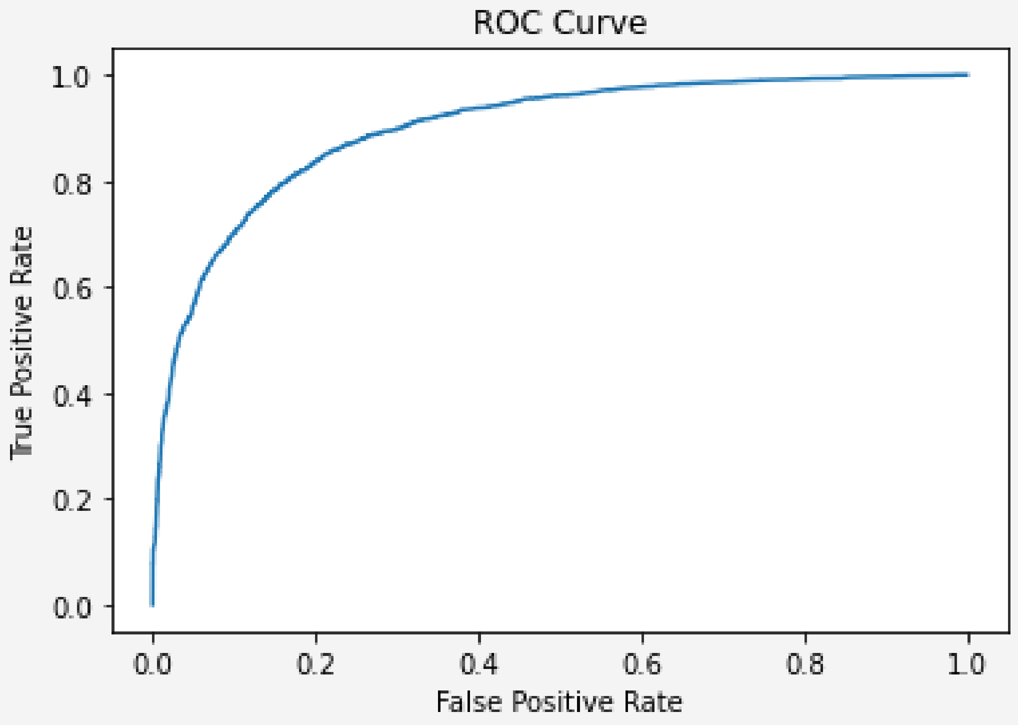
- 'train_test_split' function to randomly split the DataFrame into three new DataFrames: 'df_train', 'df_dev', and 'df_test'
- Use a downsampling technique to balance the class distribution of the 'df_train' and ‘df_test’ data frames (important for building a good-performing model).

Tokenization

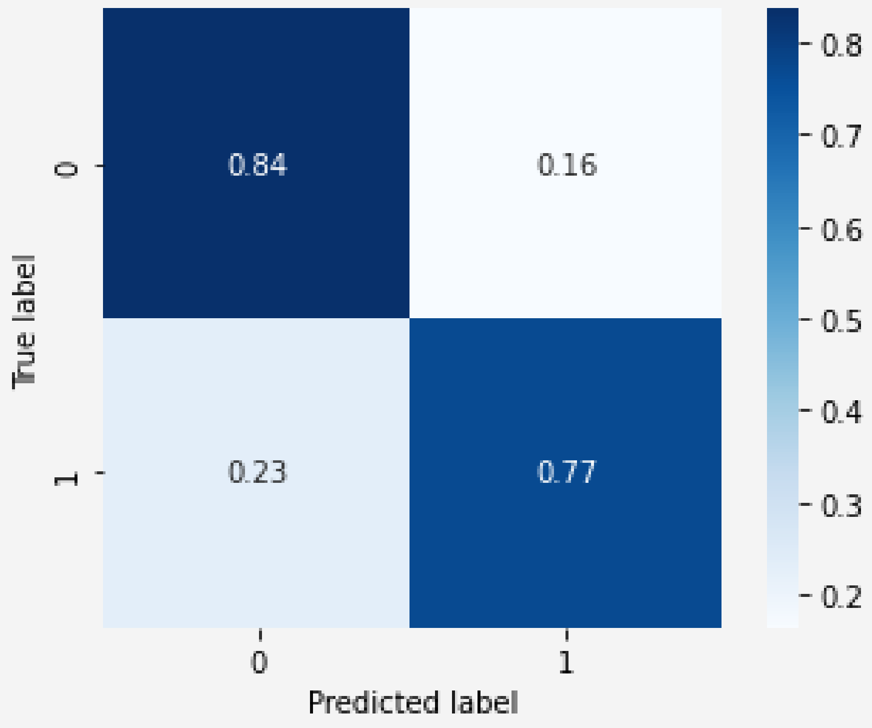
- tokenize the text, count the number of tokens, and visualize the distribution of the token count.
- calculate the ratio of reviews that have less than 128, 256, and 384 tokens (beneficial in determining the maximum input length for a language model such as BERT).

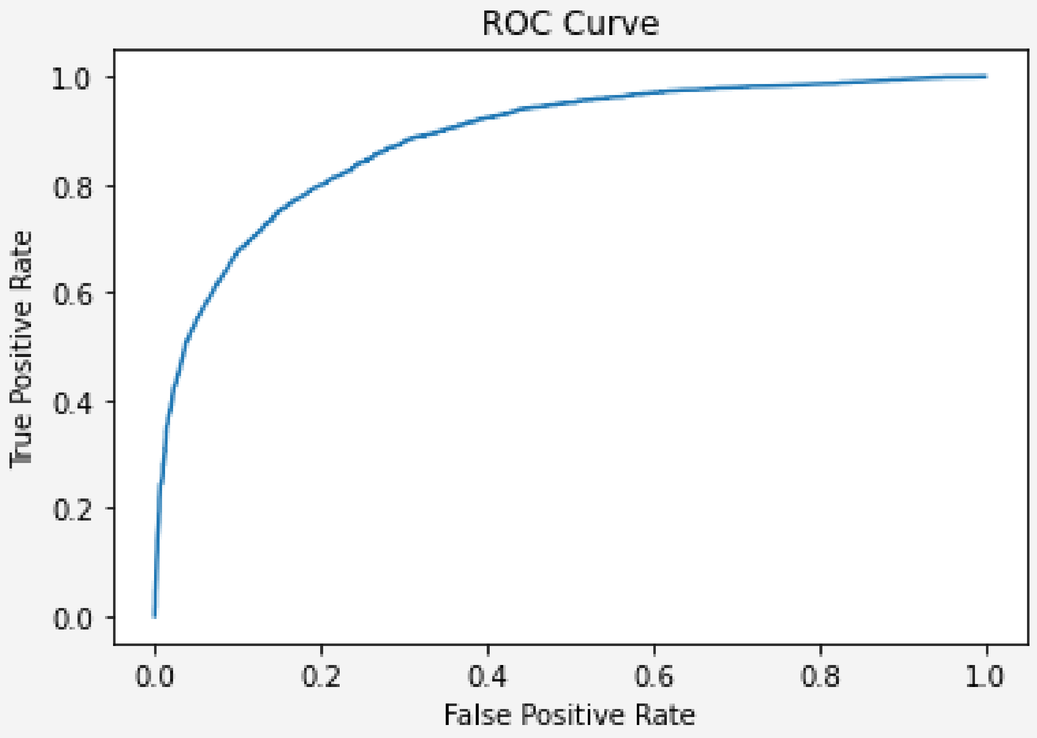
BERT



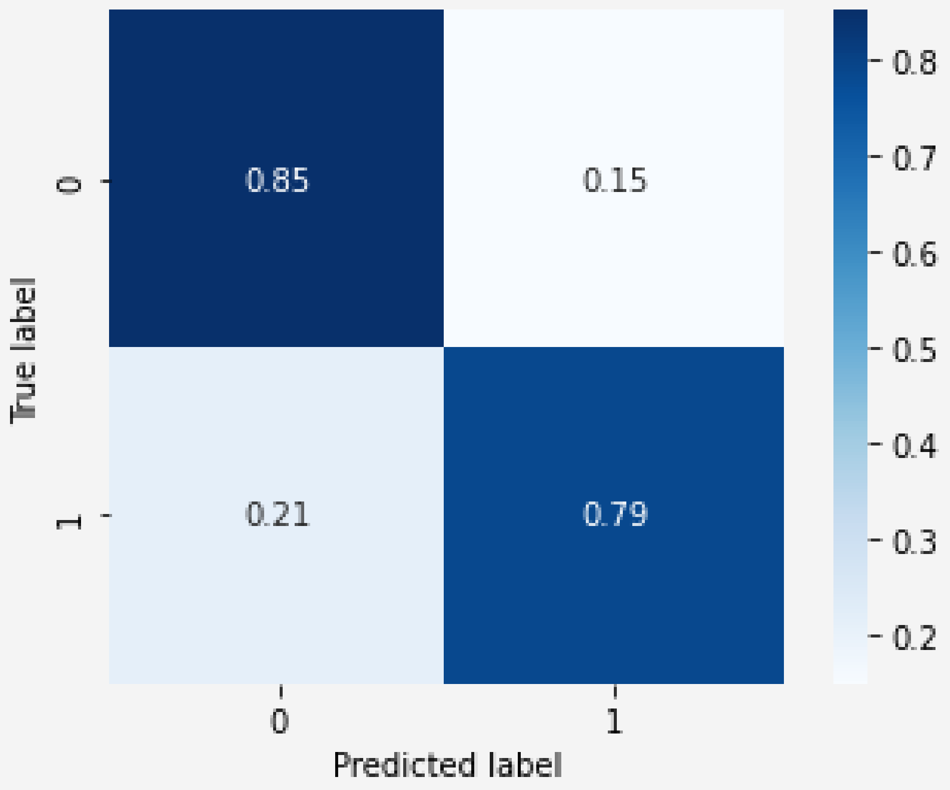


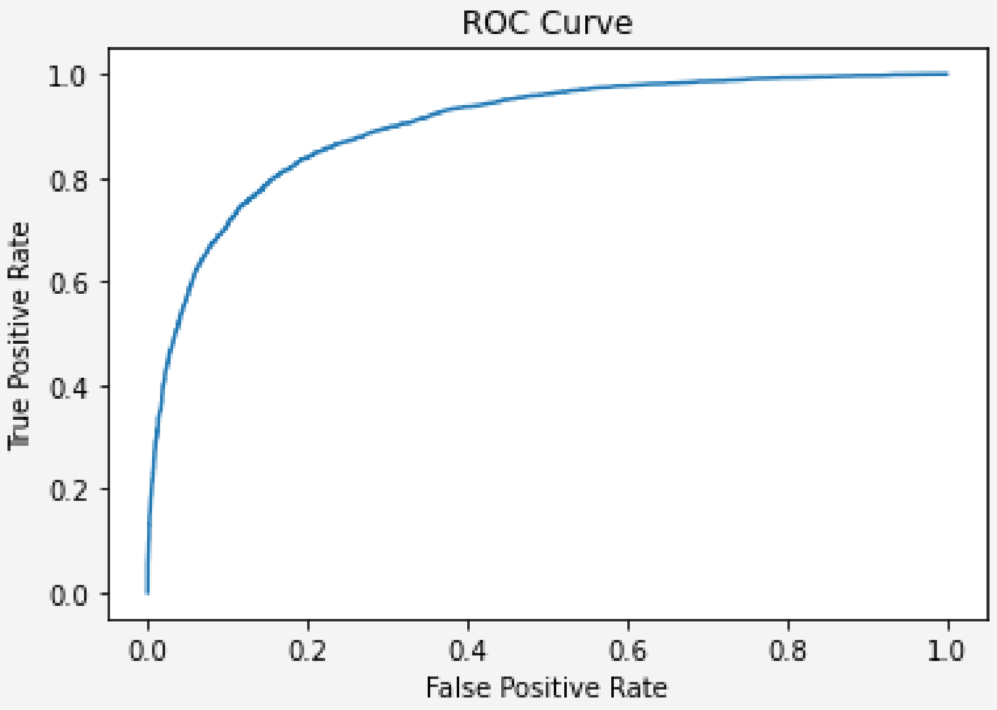
BERT-LARGE



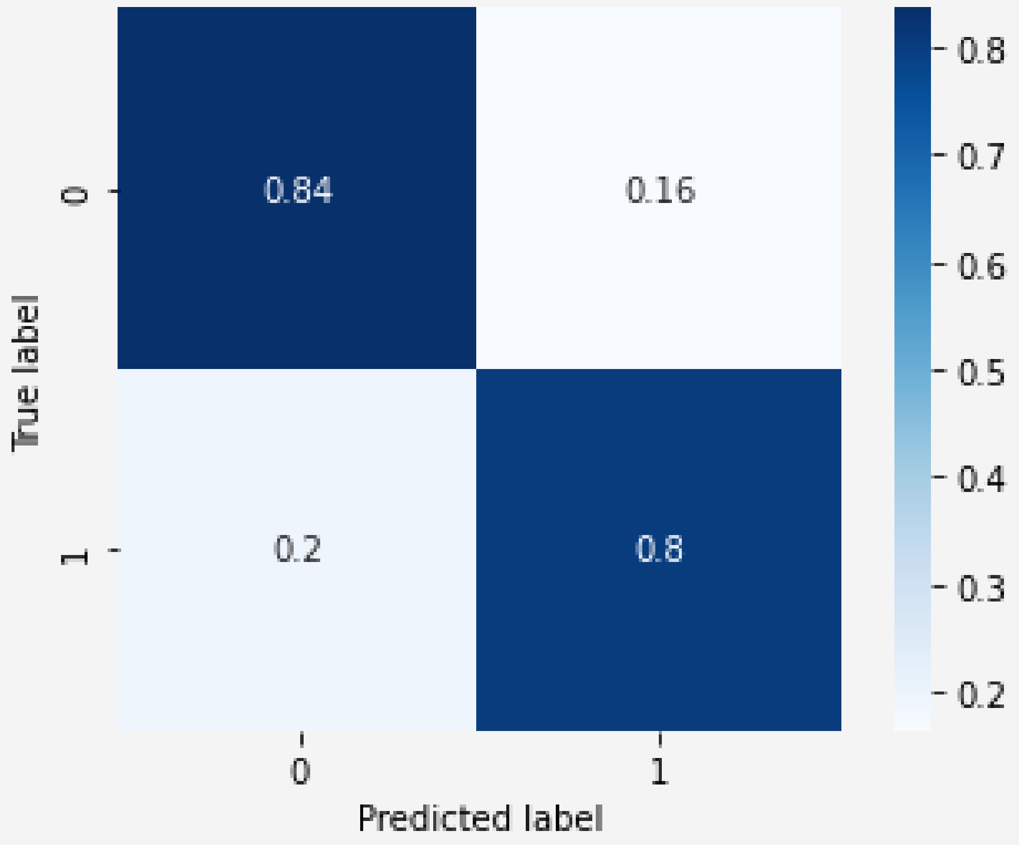


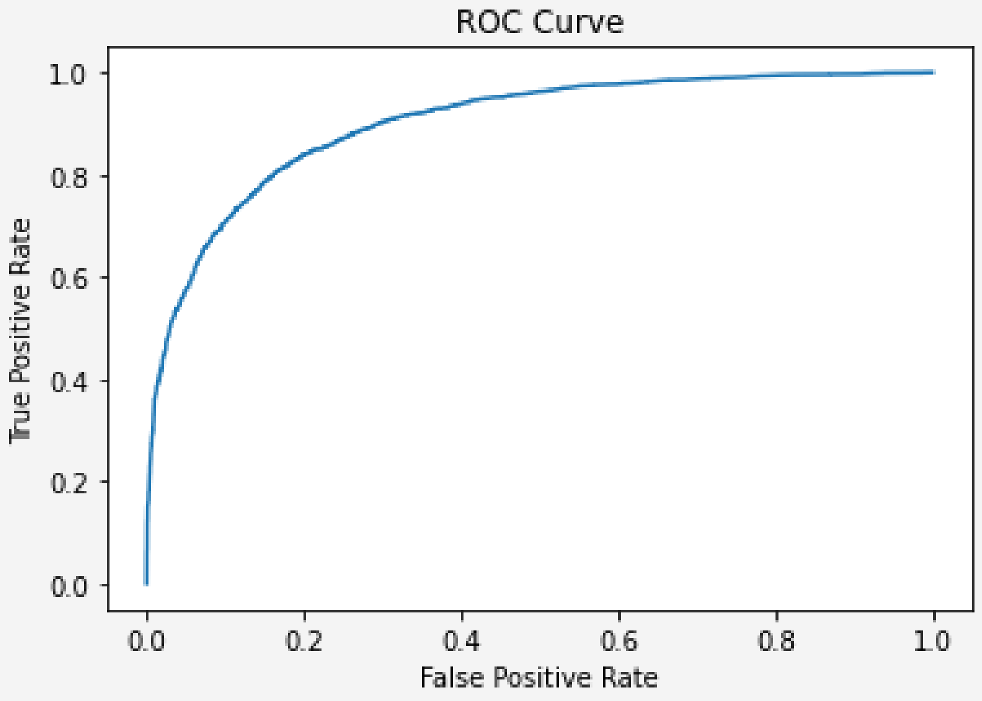
Distil-BERT





RoBERTa





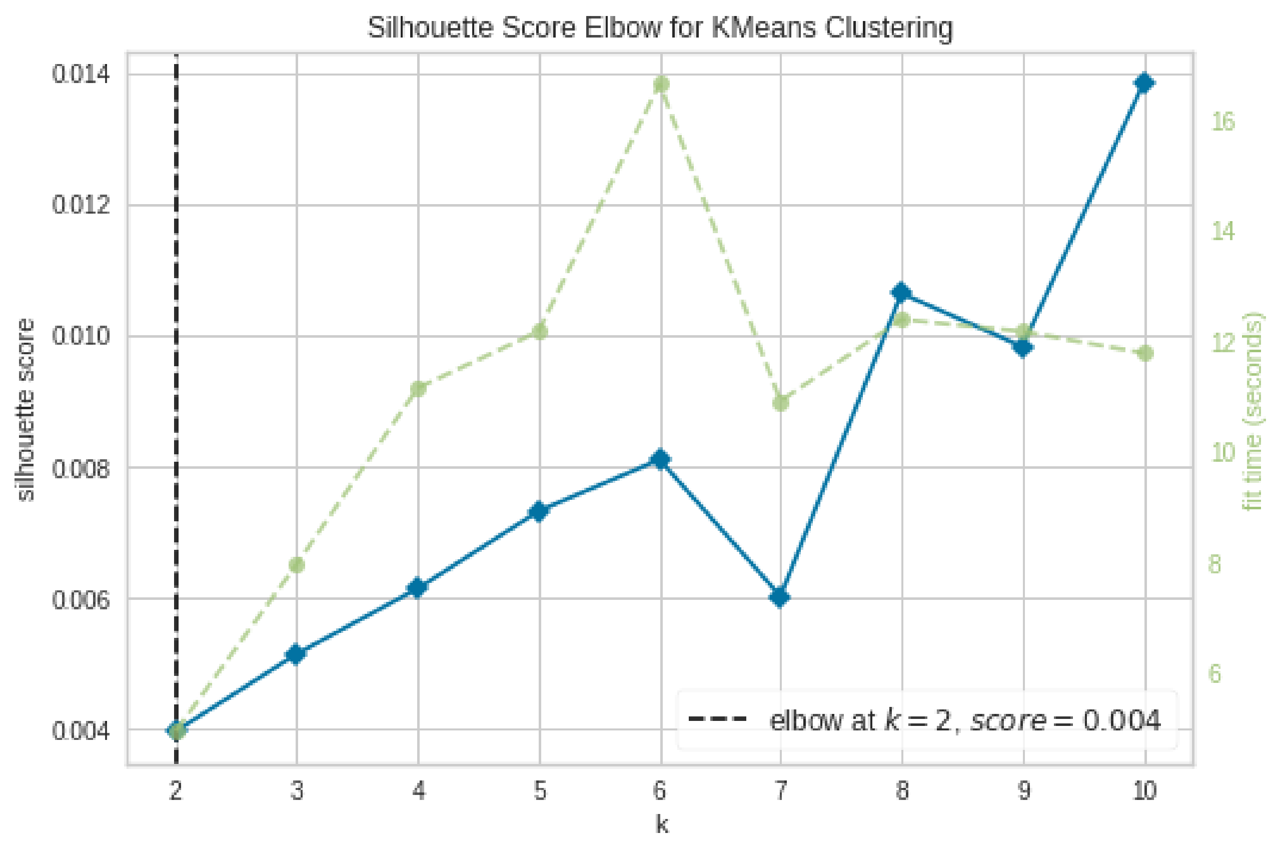
Text Clustering

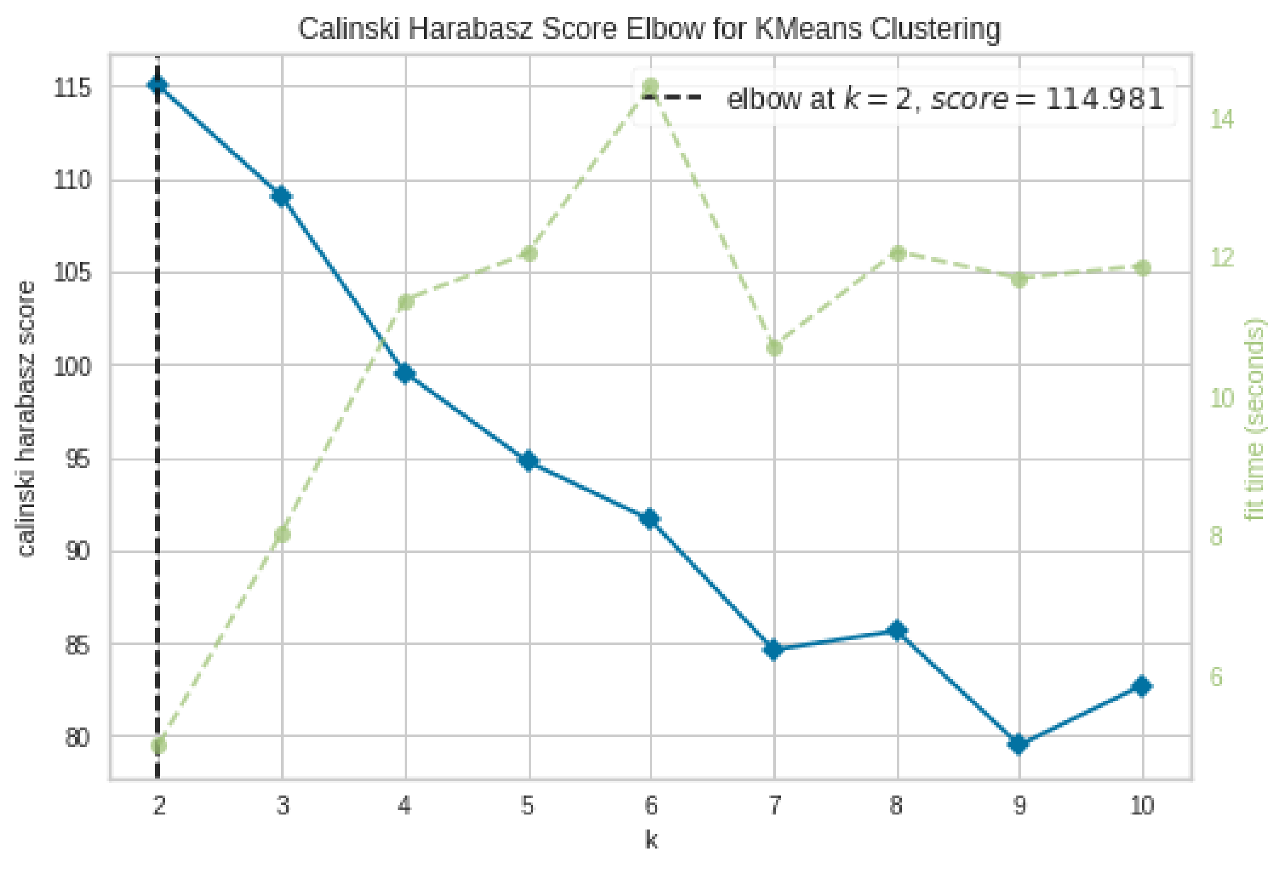
Various clustering techniques have been implemented (K-Means, Hierarchical and DBSCAN with different affinity measures and linkage techniques, Birch, OPTICS)

The results reported will be of the three algorithms with the best performance and with the possibility of selecting the number of clusters:

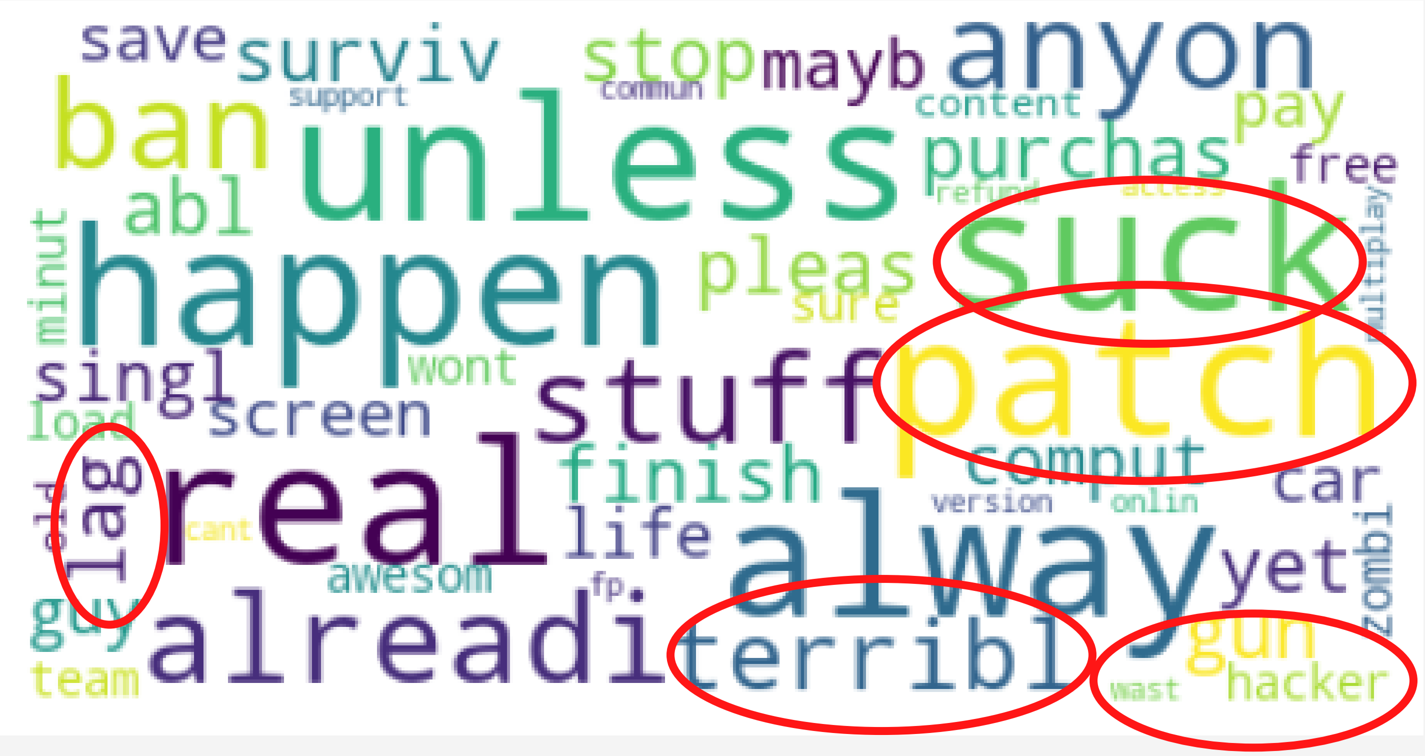
- K-Means
- Hierarchical (with Euclidean Affinity e Ward Linkage)
- Birch

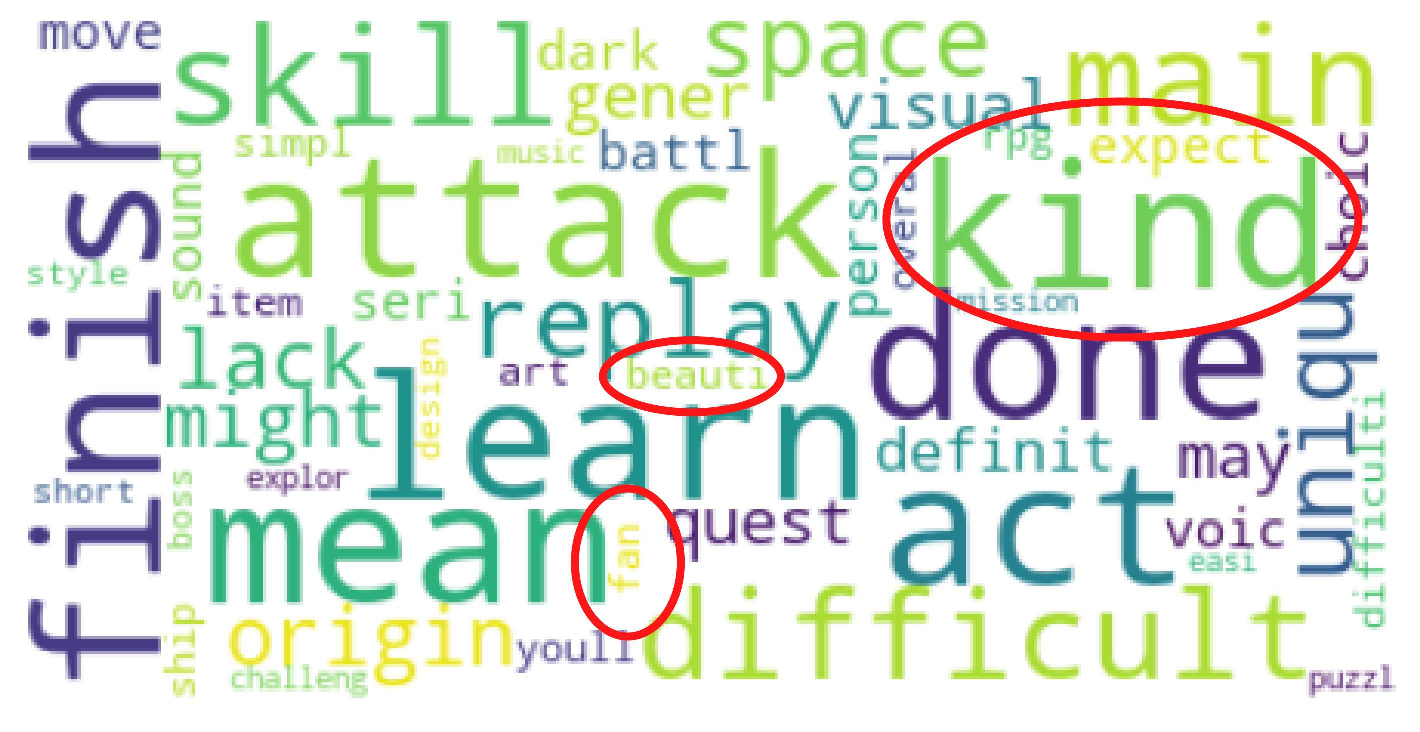
The optimal number of clusters was calculated using the k-elbow method. With Silhouette and Calinski Harabasz metrics, the recommended optimal number of clusters was 2; so, since there are only two instances of the voted_up class, 2 has been used as the number of clusters

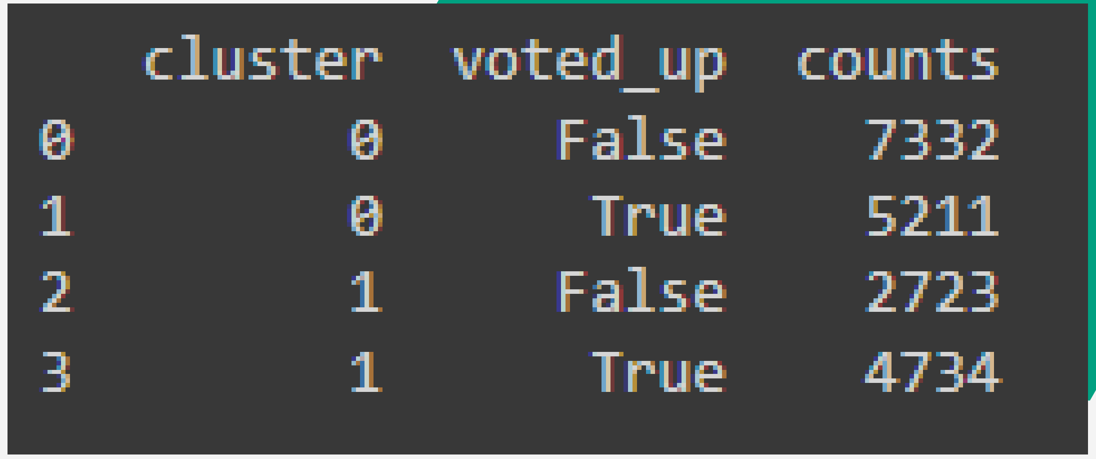




Text Clustering
K-Means

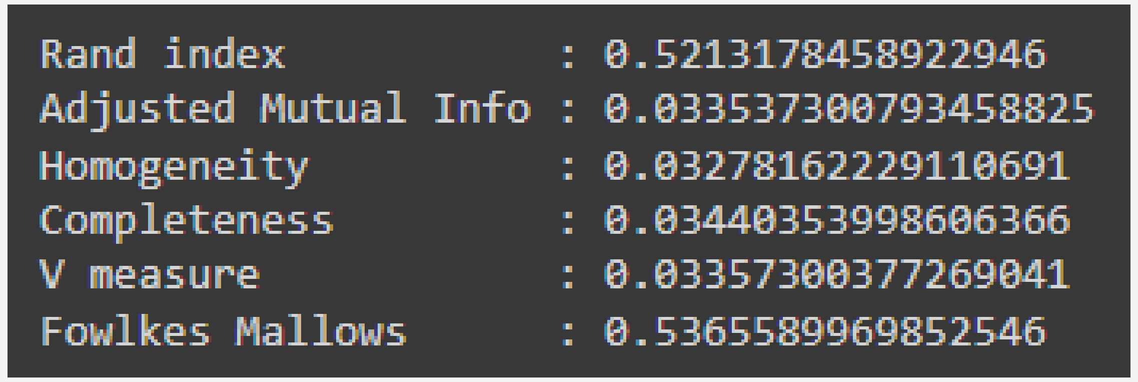






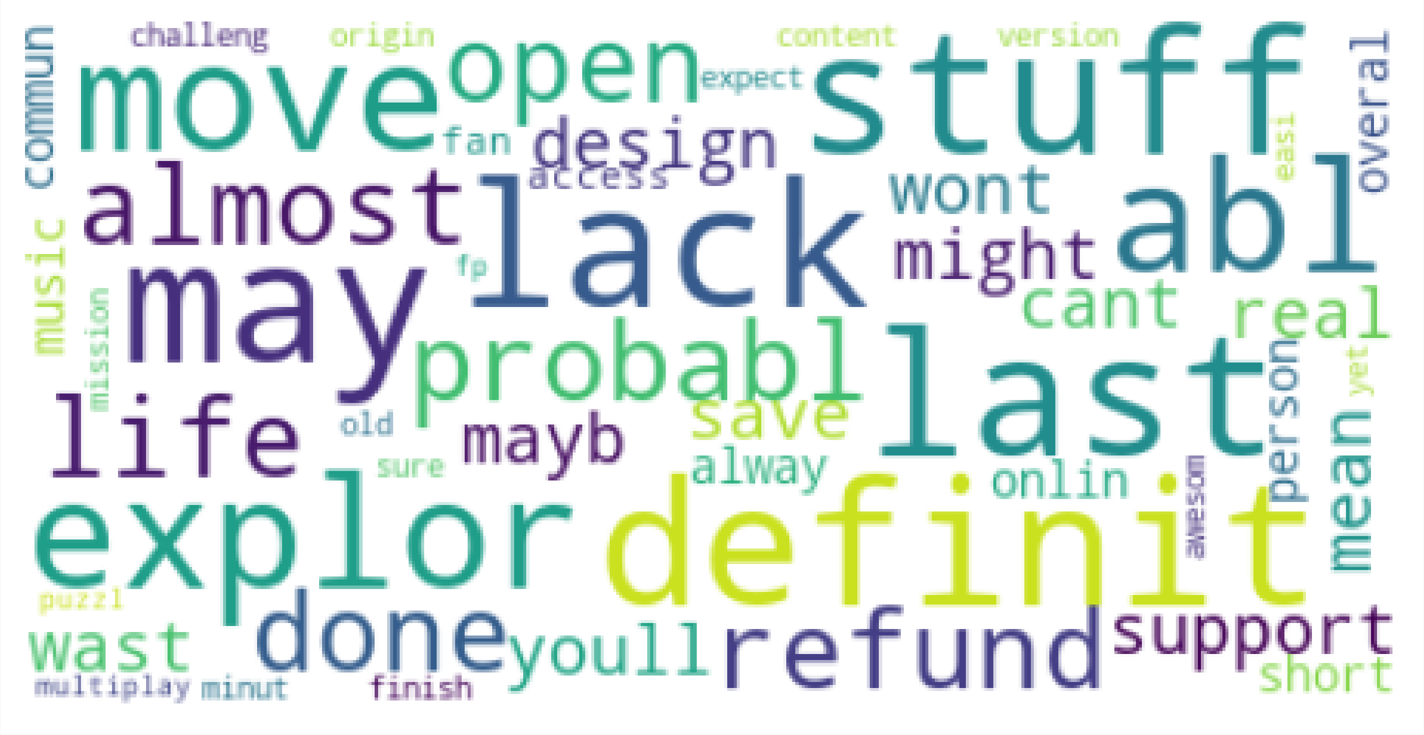
```
   cluster  voted_up  counts
0        0     False    7332
1        0      True    5211
2        1     False    2723
3        1      True    4734
```

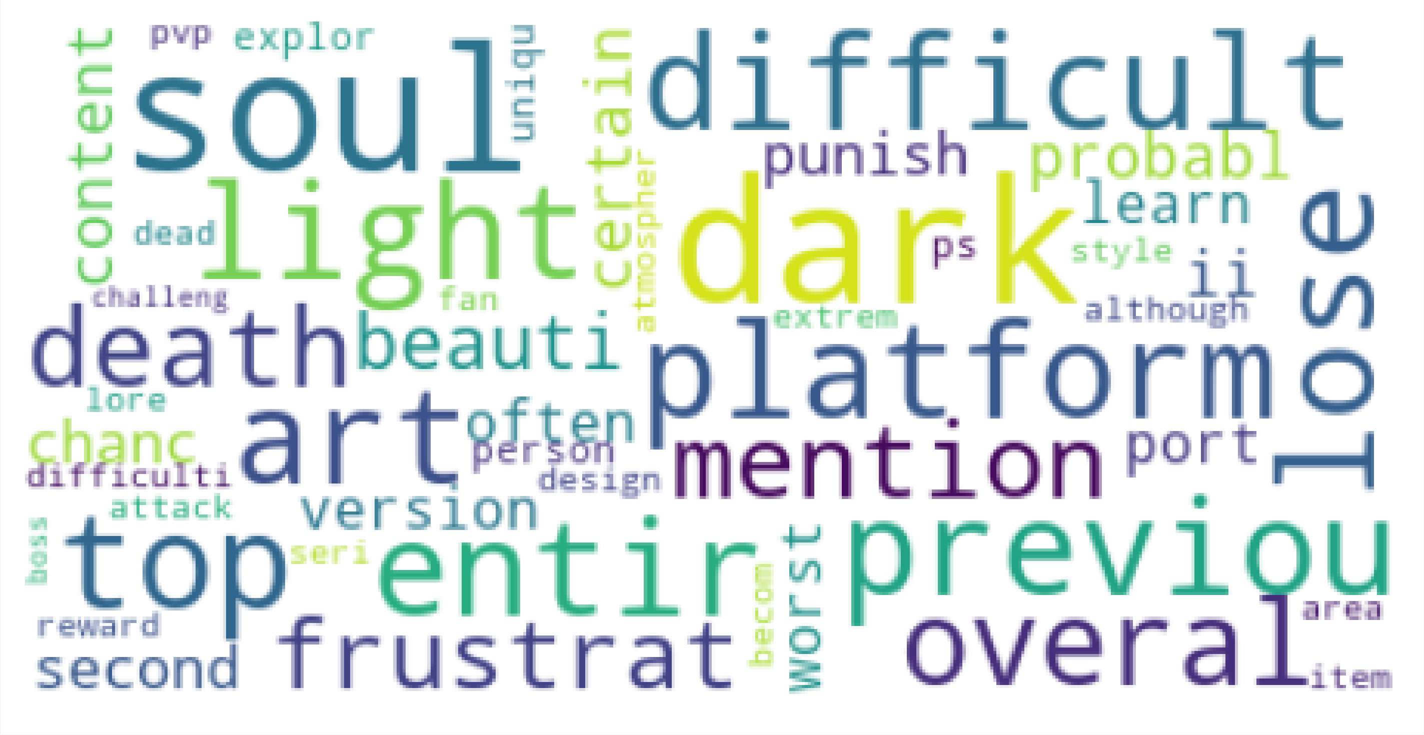
- Clustering with 2 clusters was performed
- The first cluster (cluster 0) is associated with the rows with voted_up = -1, and the second with the other rows.
- Quality of clustering is good (Rand Index and Fowlkes Mallows have values above 0.5)
- Good ability to separate clusters but poor consistency within clusters.

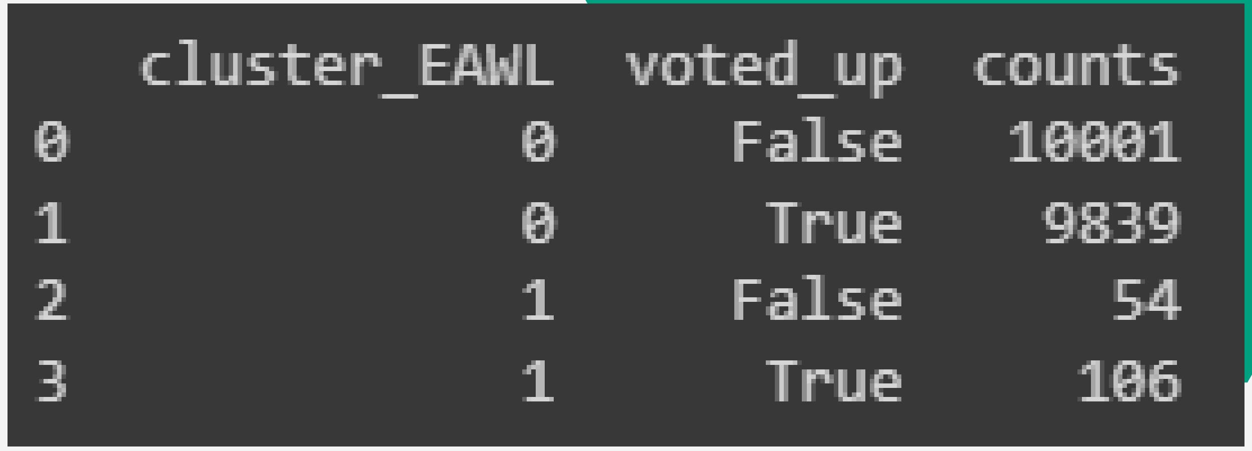


```
Rand index                 : 0.5213178458922946
Adjusted Mutual Info       : 0.033537300793458825
Homogeneity                : 0.03278162229110691
Completeness               : 0.03440353998606366
V measure                  : 0.03357300377269041
Fowlkes Mallows            : 0.5365589969852546
```

Text Clustering Hierarchical

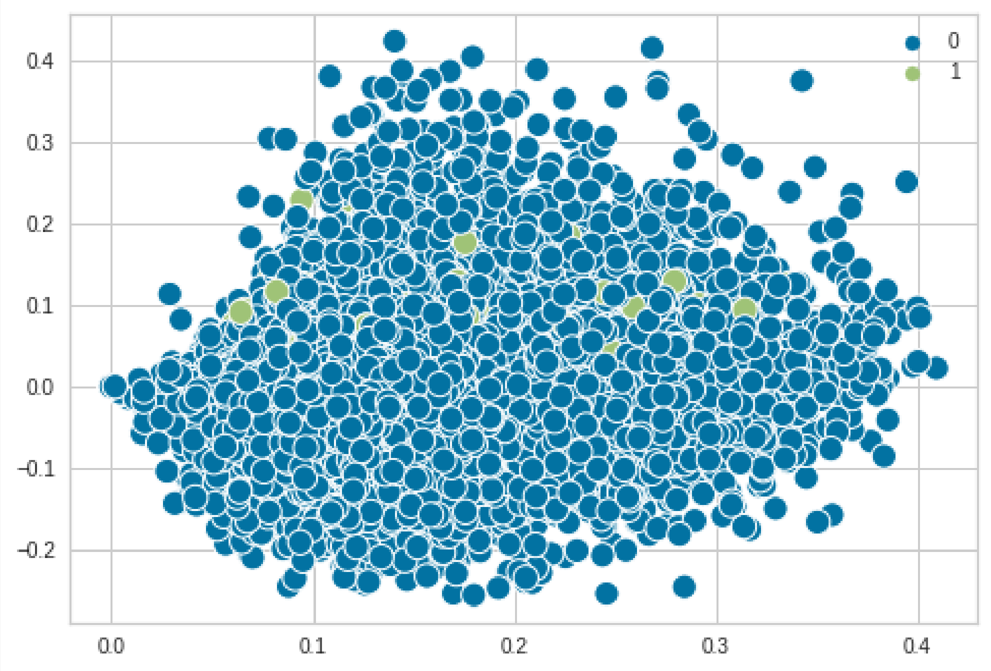




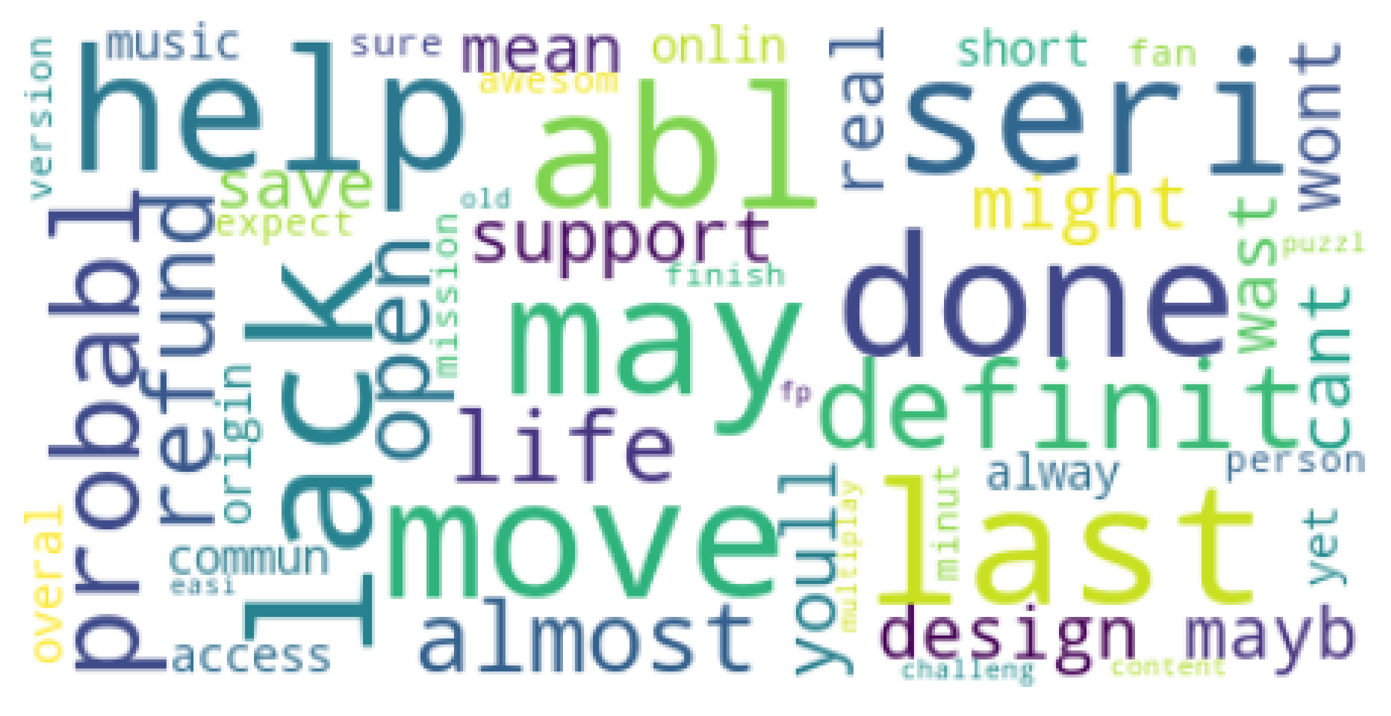


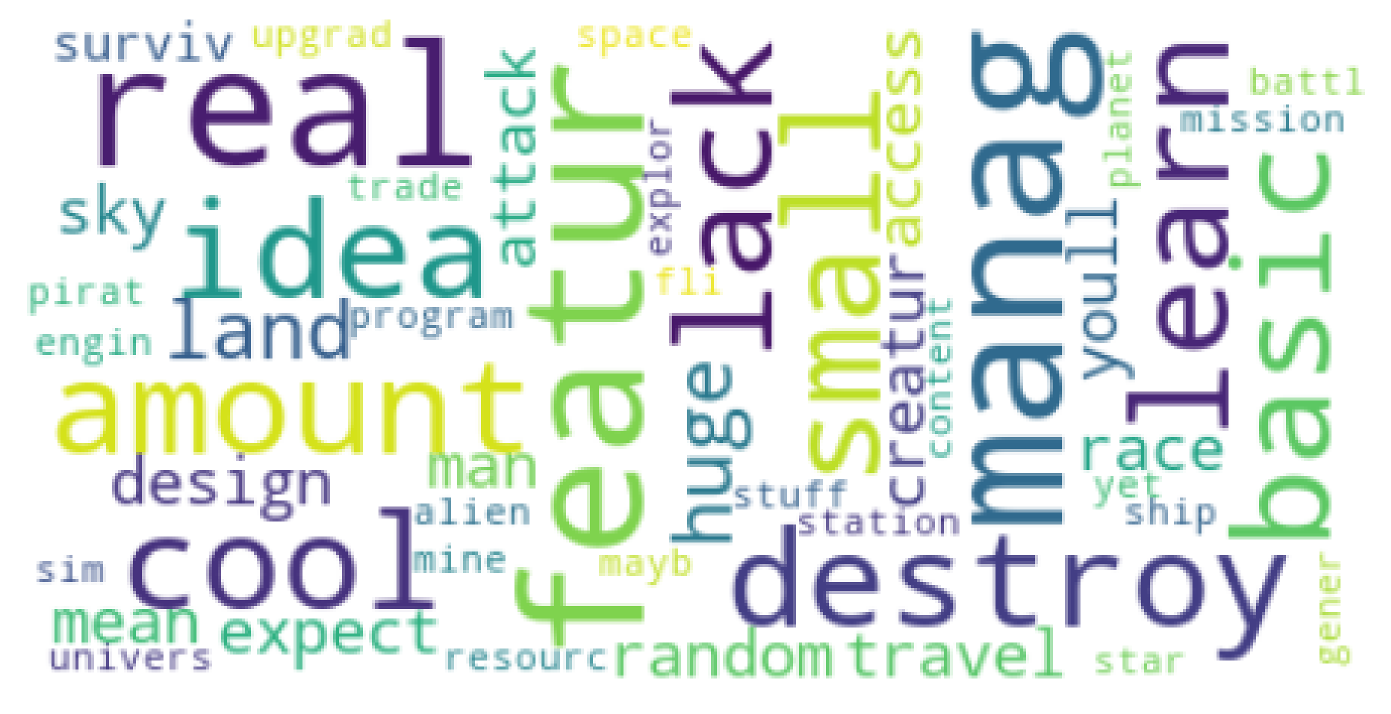
```
   cluster_EAWL  voted_up  counts
0             0     False   10001
1             0      True    9839
2             1     False      54
3             1      True     106
```

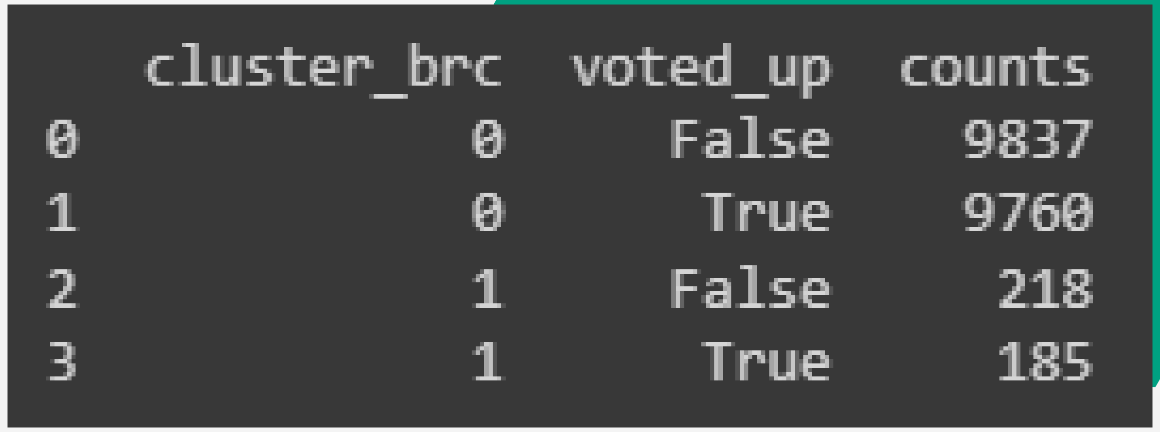
- Clustering with 2 clusters was performed
- The distinction between the two clusters is not good in this case, indeed almost all the reviews are in the first cluster
- This is shown both in the word clouds and in the scatterplot



Text Clustering
Birch

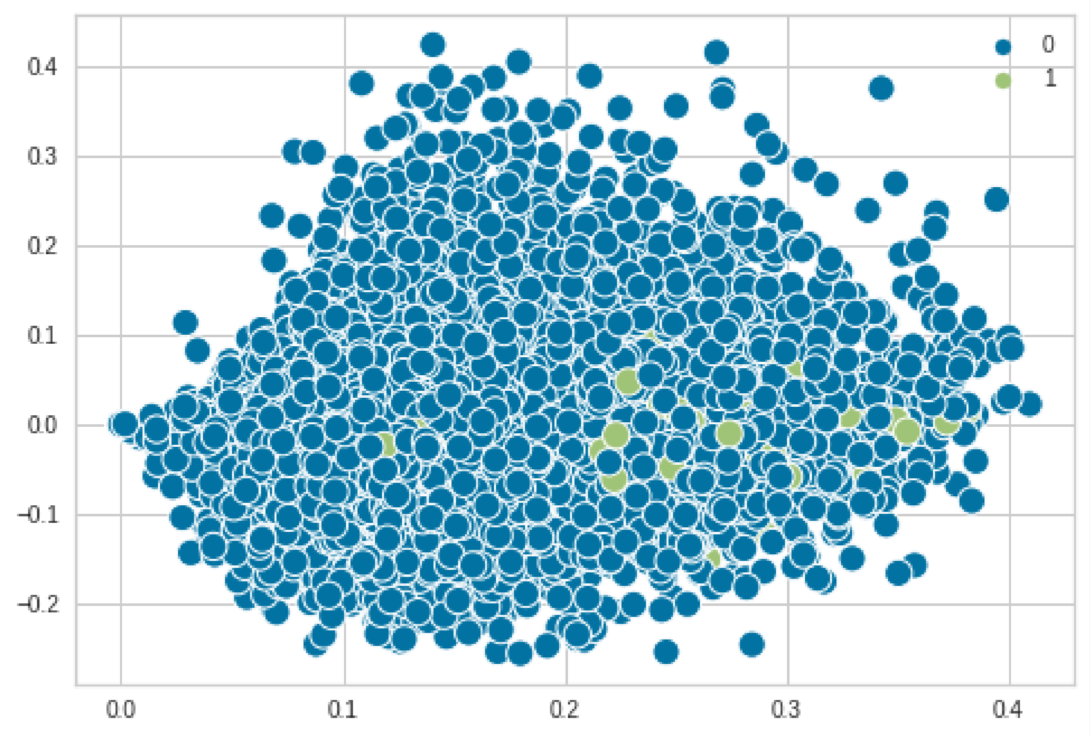






```
   cluster_brc  voted_up  counts
0            0     False    9837
1            0      True    9760
2            1     False     218
3            1      True     185
```

Same results as the previous one both in the performance and in the partition of the clustering.



With both DBSCAN and OPTICS, no number of clusters can be set; the results are in any case similar to the ones obtained by Birch, so they are not much useful.

Conclusions

Text Classification

- models have relatively high accuracy, precision, and recall, indicating that they are performing well on the testing set. However, there is still room for improvement.
- Looking at the results some models are misclassifying samples from class 1. Since we are using BERT, which is a pre-trained transformer model, it is possible that the model has not been fine-tuned enough on this specific dataset and task. Fine-tuning BERT on a large dataset can be computationally expensive
- However, our BERT models performed better than the Decision Tree model, as they achieved an accuracy of around 80%, while the Decision Tree model achieved the highest accuracy of 75% according to Zhen Zuo's paper.

Text Clustering

- poor quality of clustering using word frequency (tf-idf)
- The task of text clustering applied to this dataset did not bring a great added value for a better understanding and interpretation of the data, as all approaches identify a main group that contains almost all observations.
- Hierarchical methods and density-based models do not seem to be suitable for addressing this type of problem

Thank you