AMAZON FINE FOOD REVIEWS

CONFALONIERI RICCARDO (830404) | RANIERI SILVIA (878067)

CONTENTS

INTRODUCTION and GOALS

Amazon is one of the most popular e-commerce sites in the world. One of its strengths is the reviews system that allows all customers to express opinions on the products purchased. Over the years this review system has become more and more organized and is a strong influence during purchases. Some studies report that over half of customers rely on reviews to decide which product to buy.

The aim of the project is therefore to create a system that allows, automatically, to verify the score assigned to the reviews. By doing this, it is possible to prevent vendors using automatic systems to obtain more positive reviews and consequently greater visibility. For this reason, two different techniques have been developed:

CLASSIFICATION

CLUSTERING

1 DATASET

The data used for the project are available on Kaggle. The dataset consists of reviews of fine foods from Amazon and contains different features including those of interest to the project:

- *ProductId*. Unique identifier for the product.
- *UserId*. Unique identifier for the user.
- *Score*. Rating assigned to the review between 1 and 5.
- *Time*. Review date in UNIX format.
- *Text*. Text of the review.

The dataset contains 568.454 reviews made by over 200k users about 74.258 products.

1.3 DUPLICATE REMOVAL

Reviews containing the *same pair of (UserId, ProductId)* are considered to be duplicate. In fact, on Amazon each user can only review a product once, which is why the *most recent* reviews were kept.

Also lines with *same (UserId, Score, Time, Text)* have been deleted. In fact, it seems that the review is automatically duplicated on multiple similar products, but obviously it is unusual for a user to buy equivalent products on the same day and give the same reviews.

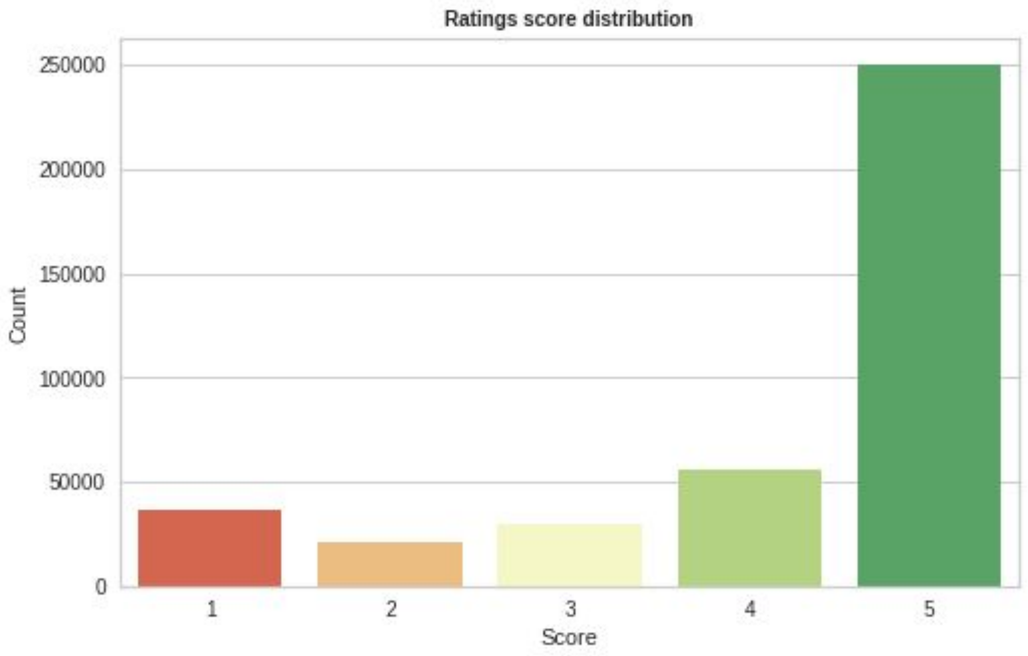
ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
B000PMJLJO	AZYMD9P9F9UZ6	W. Coombe	0	0	5	1239148800	Good Jerky	I like the peppered flavor a lot better than t...
B000GW46D4	AZYMD9P9F9UZ6	W. Coombe	0	0	5	1239148800	Good Jerky	I like the peppered flavor a lot better than t...
B000GW6786	AZYMD9P9F9UZ6	W. Coombe	0	0	5	1239148800	Good Jerky	I like the peppered flavor a lot better than t...

In total, 30% of the data was deleted, thus making 392,969 reviews available.

1.4 TEMPORAL SHIFT and SCORE

From the time shifting analysis it emerged that the reviews refer to the time period 1999-2012. However, most refer to the years 2006 onwards and there was *no substantial textual difference* between the reviews of the early years. As regards the variable score, it was instead noted that:

1. It is heavily *unbalanced*
2. From 2006 onwards there is an increasing trend in the number of reviews, however 5-star reviews have an *anomalous trend*. This can be due to unverified or fake reviews.



2. PRE PROCESSING

The preprocessing step made it possible to standardize the representation of the text indexes, in particular the following steps were carried out.

Normalization

- Lowercase
- Handling of abbreviations (not)
- Accents
- Particular cases (html, emoji)

Stopwords removal

- Predefined set
- Contextual words (Amazon/Order)
- Preserve 'not'

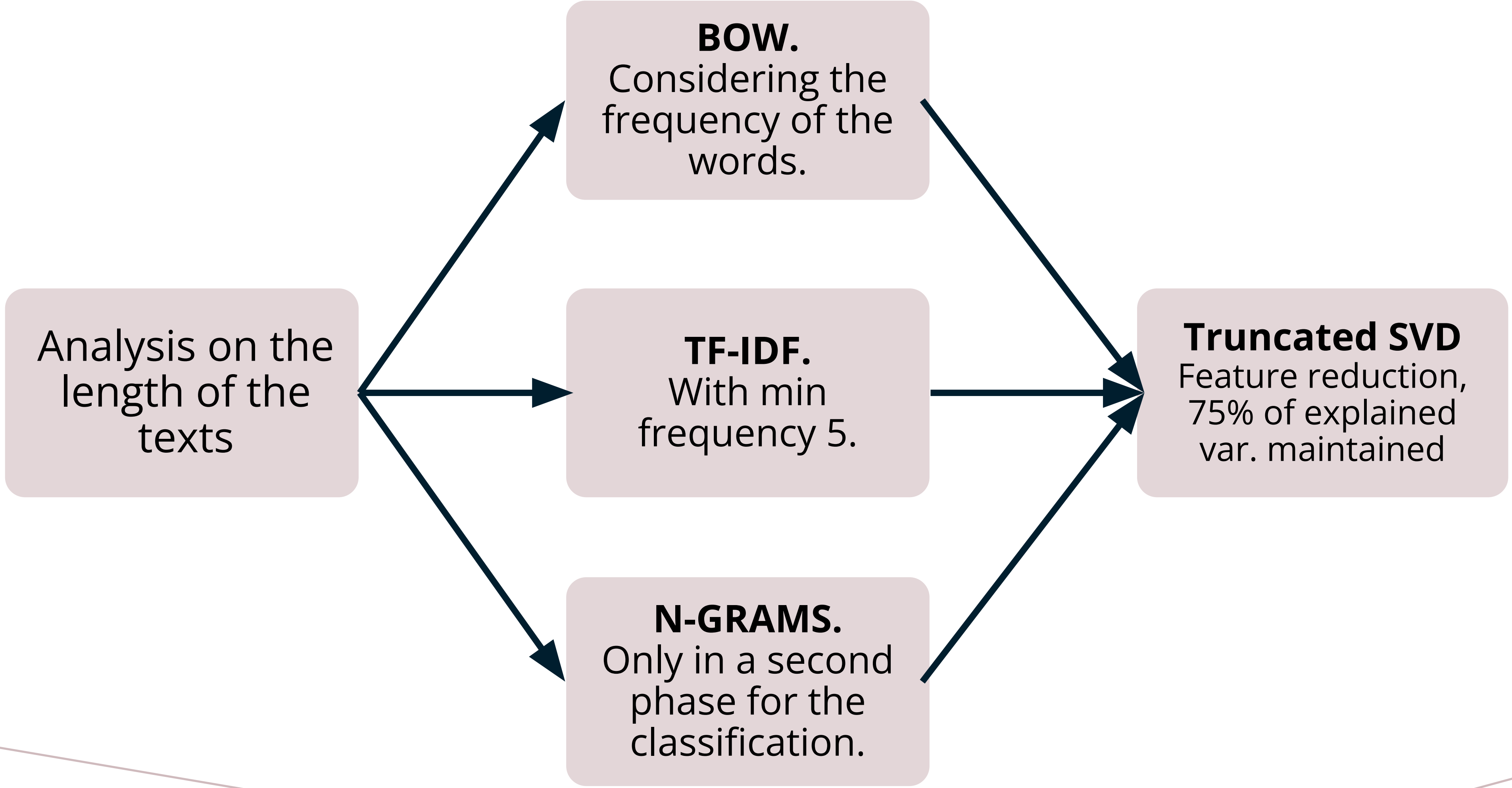
Tokenization e stemming

- Porter stemmer
- More than 90.000 token

```
<-------- Before remove stopwords -------->
Example1:  I do not like sour taste and this has a sour kind of taste which i don't like. The smell isn't that great either
Example2:  I just love it, and I am Not a major Indian cooking fan--just enough. Really, it mixes with anything you are doing like <a href="http://www.amazon.com/gp/product/B000FIXT2I"> ...

<-------- After remove stopwords -------->
Example1:  not like sour taste sour kind taste not like smell not great either
Example2:  love not major indian cooking fan enough really mixes anything like steamed brown rice bowl organic microwaveable ounce bowls pack use convenience not ...
```

3 TEXT REPRESENTATION



4. CLASSIFICATION

The first NLP task considered was *classification*, specifically the goal is to classify the reviews into two macro categories: *positive* (score ≥ 4) or *negative* (score ≤ 2). To do this, the data was further manipulated to:

1. **Score conversion** to binary. All neutral reviews have been eliminated (Score = 3) and the remaining ones have been binarized.
2. The dataset was **balanced** by eliminating reviews from the majority class (Positive)

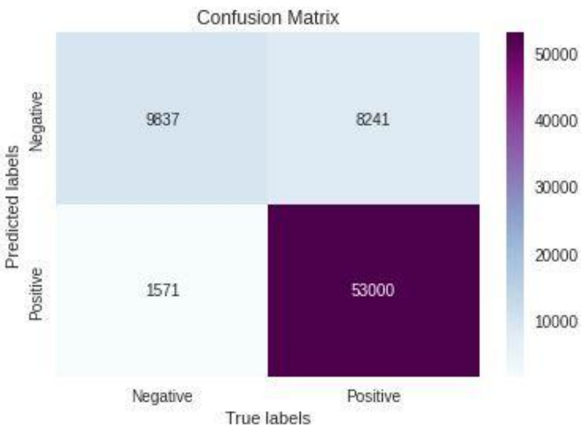
The binary classification task was therefore carried out by applying different models:

- Logistic regression.
- Light SVM.
- LGBM.

4.1 LOGISTIC REGRESSION

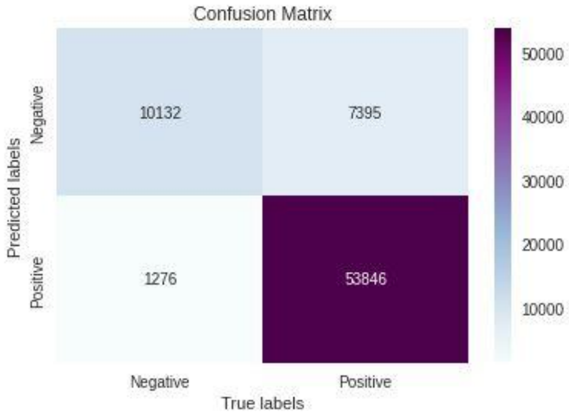
Linear classifier that assigns a probability between 0 and 1 for each class, with the sum of one. The default threshold value, which was used in this project, is ≥ 0.5. It *only takes 13s* in the case of TF-IDF.

	precision	recall	f1-score	support
Negative	0.86	0.54	0.67	18078
Positive	0.87	0.97	0.92	54571
accuracy			0.86	72649
macro avg	0.86	0.76	0.79	72649
weighted avg	0.86	0.86	0.85	72649



Results with BOWs

	precision	recall	f1-score	support
Negative	0.89	0.58	0.70	17527
Positive	0.88	0.98	0.93	55122
accuracy			0.88	72649
macro avg	0.88	0.78	0.81	72649
weighted avg	0.88	0.88	0.87	72649

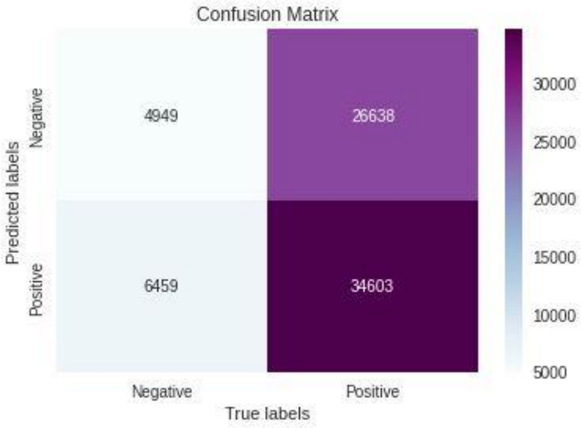


Results with TF-IDF

4.2 SVM

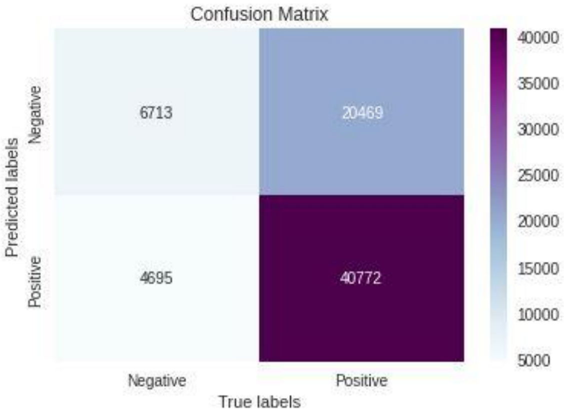
The classic SVM classifier cannot be applied due to the complexity of the data, so an *approximate version* was used which allows for very fast execution times. In fact, it only takes 3s for TF-IDF.

	precision	recall	f1-score	support
Negative	0.43	0.16	0.23	31587
Positive	0.57	0.84	0.68	41062
accuracy			0.54	72649
macro avg	0.50	0.50	0.45	72649
weighted avg	0.51	0.54	0.48	72649



Results with BOWs

	precision	recall	f1-score	support
Negative	0.59	0.25	0.35	27182
Positive	0.67	0.90	0.76	45467
accuracy			0.65	72649
macro avg	0.63	0.57	0.56	72649
weighted avg	0.64	0.65	0.61	72649

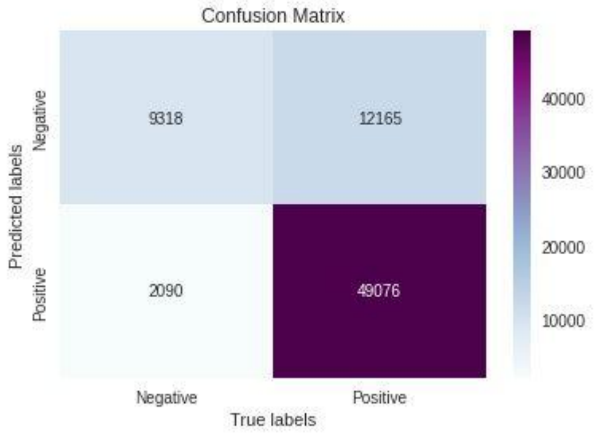


Results with TF-IDF

4.3 LGBM

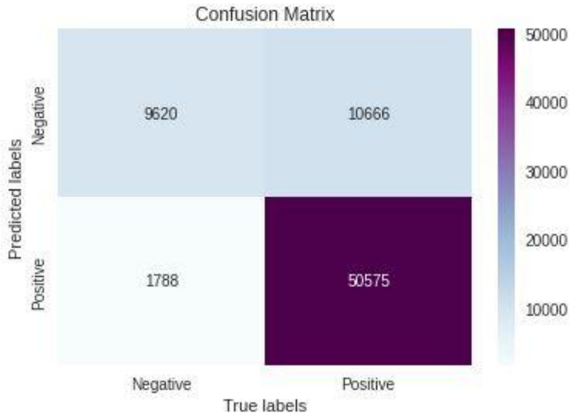
This is a very fast, distributed, high-performance gradient boosting framework based on decision tree algorithms. Results are similar to logistic regression but *takes 5min* on TF-IDF.

	precision	recall	f1-score	support
Negative	0.82	0.43	0.57	21483
Positive	0.80	0.96	0.87	51166
accuracy			0.80	72649
macro avg	0.81	0.70	0.72	72649
weighted avg	0.81	0.80	0.78	72649



Results with BOWs

	precision	recall	f1-score	support
Negative	0.84	0.47	0.61	20286
Positive	0.83	0.97	0.89	52363
accuracy			0.83	72649
macro avg	0.83	0.72	0.75	72649
weighted avg	0.83	0.83	0.81	72649



Results with TF-IDF

4.4 INSPECTION OF RESULTS

By analyzing the incorrect classifications of the various previous classifiers, it emerged that many sentences *contained the word 'not'*. Furthermore, the most *significant words* to disambiguate the reviews seem to be random and *not very useful.*

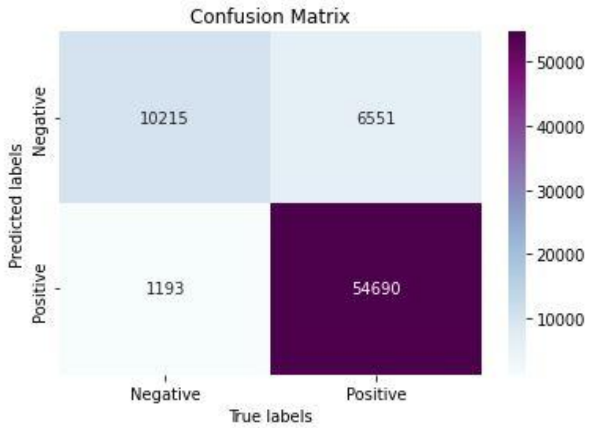
-1.8676	abl	1.2743	antisept
-1.5903	advertis	1.2472	acut
-1.4381	abhor	1.1295	aesthet
-1.2205	abund	1.1274	achiev
-1.0414	apiec	1.0715	aafco
-0.9961	apex	1.0427	aggrav
-0.9757	arginin	1.0349	abid
-0.9682	acidophilu	1.0152	adventuresom
-0.9650	accessori	1.0073	anti
-0.9465	apart	0.9997	alik
-0.9286	appet	0.9401	aerogarden
-0.9209	ambul	0.9266	aback
-0.9126	antidot	0.9168	absinth
-0.8698	asterisk	0.9008	adren

Therefore the TF-IDF with 2-grams has been considered.

4.5 LOGISTIC REGRESSION (2-GRAMS)

The results are more encouraging, all scores increase and the most significant words significantly improve.

	precision	recall	f1-score	support
Negative	0.90	0.61	0.73	16766
Positive	0.89	0.98	0.93	55883
accuracy			0.89	72649
macro avg	0.89	0.79	0.83	72649
weighted avg	0.89	0.89	0.89	72649



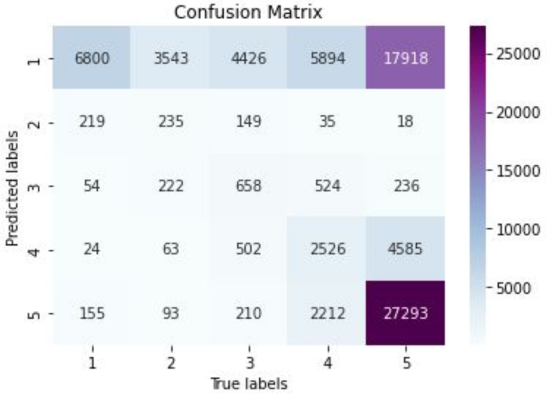
Results with TF-IDF

-13.8783	disappoint	13.5828	great
-10.3715	not	11.4268	delici
-9.7754	not recommend	11.3074	best
-8.8912	worst	10.6514	love
-8.6430	not good	9.8481	perfect
-8.5169	not buy	9.2003	good
-8.0209	not worth	8.8540	not disappoint
-7.6340	terribl	8.1196	excel
-7.4375	unfortun	7.2719	favorit
-7.3566	aw	7.1035	nice
-6.7330	horribl	6.9527	amaz
-6.5910	return	6.6659	happi

4.6 MULTICLASS

Then the problem of multiclass classification as been approached, thus keeping the *real score* in the interval [1,5]. A modern approach was attempted using a *recurrent neural network*, in particular LSTM with an embedding layer that maps to vectors of size 100.

	precision	recall	f1-score	support
1	0.18	0.94	0.30	7252
2	0.36	0.06	0.10	4156
3	0.39	0.11	0.17	5945
4	0.33	0.23	0.27	11191
5	0.91	0.55	0.68	50050
accuracy			0.48	78594
macro avg	0.43	0.38	0.30	78594
weighted avg	0.69	0.48	0.52	78594



Results with LSTM

Results are not good especially for class 1 classified almost randomly!

5. CLUSTERING

The second task considered was *clustering*, specifically the initial goal is to group reviews into *5 different groups*. The idea is therefore to find representative clusters for the different scores.
Subsequently, by semantically analyzing the results, we moved to clustering in a *greater number of clusters* trying to maximize certain metrics.

The algorithms considered for clustering are:
- K-means.
- Agglomerative hierarchical.

5.1 RESULTS

By analyzing the results of the two different clusters, valid performance is not obtained

```
No. of reviews in Cluster-0: 4841
No. of reviews in Cluster-1: 6215
No. of reviews in Cluster-2: 40903
No. of reviews in Cluster-3: 23890
No. of reviews in Cluster-4: 7266

Rand index           : 0.5981333172465243
Adjusted Mutual Info : 0.008051656398375396
Homogeneity          : 0.007286547678001394
Completeness         : 0.009163022529309783
V measure            : 0.00811775623228642
Fowlkes Mallows      : 0.2670176345603583
Silhouette           : 0.011880372905302599
```

Results k-means (k=5)

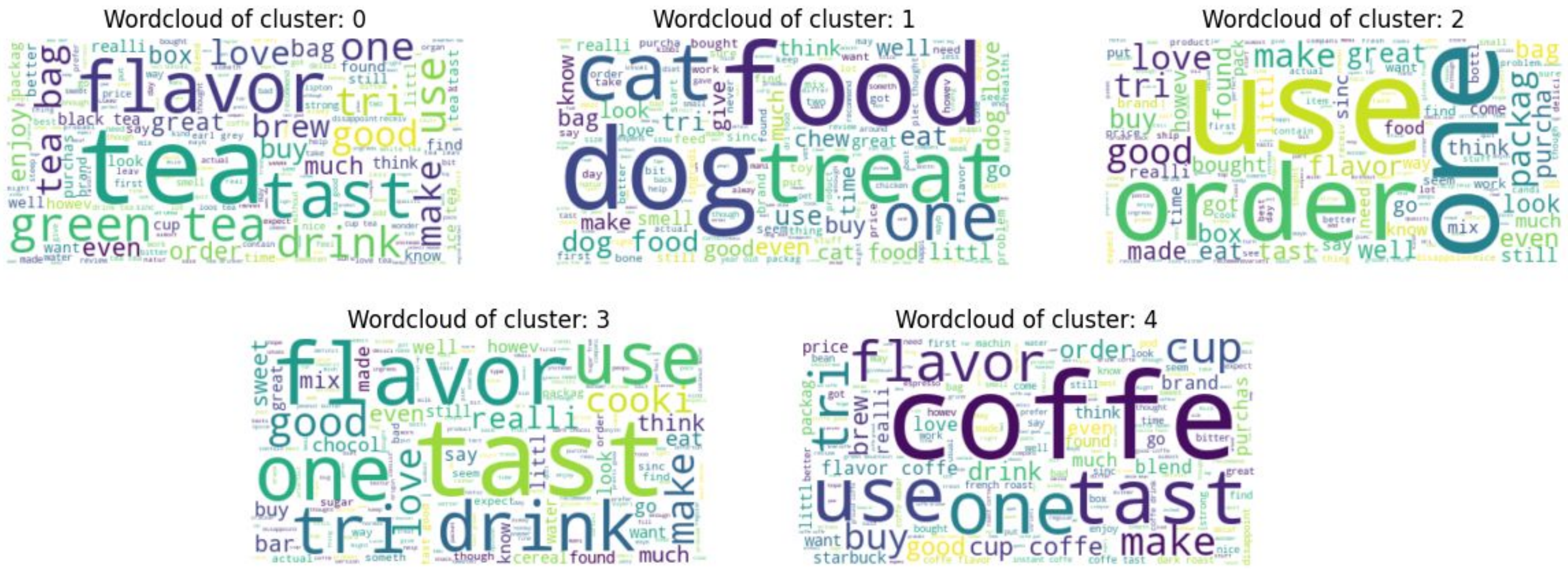
```
No. of reviews in Cluster-0: 11747
No. of reviews in Cluster-1: 641
No. of reviews in Cluster-2: 1534
No. of reviews in Cluster-3: 798
No. of reviews in Cluster-4: 280

Rand index           : 0.4230088583683357
Adjusted Mutual Info : 0.0011221649188717187
Homogeneity          : 0.0011680343902362516
Completeness         : 0.0023802389639944535
V measure            : 0.0015670725952446239
Fowlkes Mallows      : 0.3549546141524873
Silhouette           : 0.005335375457491175
```

Results hierarchical (k=5)

5.2 SEMANTIC EVALUATION

Analyzing the semantics of the clusters it has been noticed that there is a possible subdivision based on the topics but also in this case there are several repetitions and unique clusters are not obtained.



Wordcloud k-means (k=5)

5.3 K-MEANS (K=9)

An attempt was therefore made to maximize the silhouette metric to find the optimal number of clusters which turned out to be 10. The goal is to form clusters that divide the content of the reviews.

Wordcloud k-means (k=10)

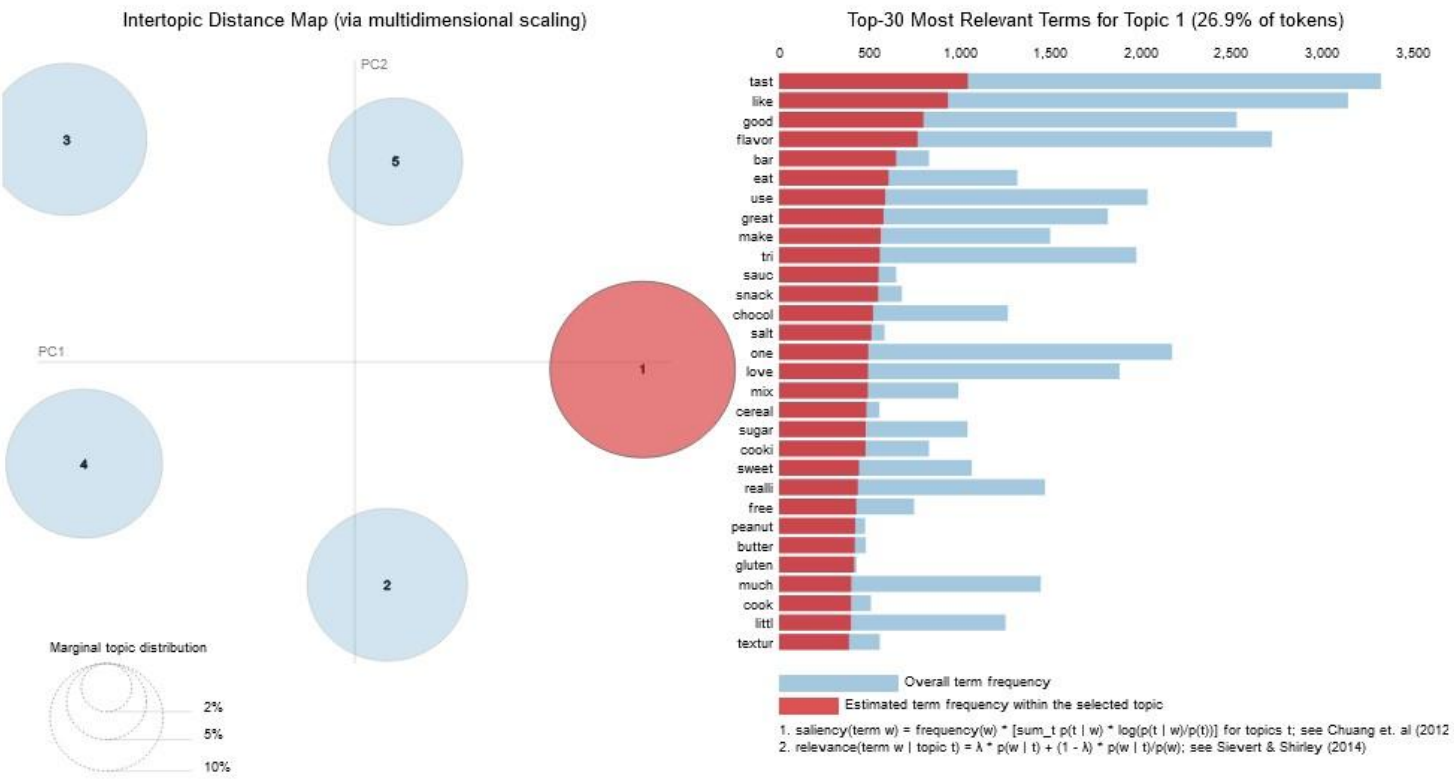
6. TOPIC MODELING

Given the partially encouraging results obtained at the semantic level with k-means, an attempt was made to approach the problem of extracting the contents (topic) present in the various reviews.
The previous analysis highlighted the presence of at least *5 topics* (animals, coffee, tea, orders, chocolate / snacks). We therefore searched for a number of topics ≥ 5 that minimize the *perplexity* metric, the optimal choice for extracting the topics turned out to be precisely that of extracting 5 different topics.

The LDA technique was used to extract the topics.

6.1 RESULTS

The results obtained with this technique, although basic, are encouraging and it seems that it is actually possible to extract the topics present in the reviews.



LDA topic extraction

Specifically:

- topic1 = generic, sweets
- topic2 = order, shipment
- topic3 = animals
- topic4 = the
- topic5 = coffee

7. CONCLUSION

In conclusion, it can be stated that:

- The TF-IDF representation turns out to be more performing than BOW.
- The classification gives good results but there are problems with the recall of the negative class.
- Multiclass classification does not give ideal results, highlighting the limitations of the model. Perhaps also due to the fact that there is no clear textual distinction for the different score.
- Clustering did not give the desired results and proved to be more complex than the classification task. Also because of the difficult evaluation.
- The topic modeling, even if approached quickly, highlights how there are several topics that can be extracted from the reviews with satisfactory results.

Overall it is possible to say that the classification models are able to predict the binary class with some accuracy but it is difficult to create a complete model that, given a new review, automatically returns the score of the same.

QUESTION?

THANKS!

7. REFERENCES

- G. Pasi and M. Viviani, “lecture notes and slides of text mining and search course” 2021.
- J. McAuley and J. Leskovec, “From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews”
- S. N. A. Project, “Amazon fine food reviews” 2017. https://www.kaggle.com/snap/amazon-fine-food-reviews