TEXT MINING & SEARCH

SUPERVISED AND UNSUPERVISED LEARNING ON AMAZON FINE FOOD REVIEWS

Data Science Master's Degree 23/24

Sara Campolattano – 906453
Induni Sandapiumi Nawarathna Pitiyage – 906451

TABLE OF CONTENTS

INTRODUCTION

From the earliest days of online shopping, reviews have been crucial in shaping customer decisions, directly influencing a product’s success or failure. As Amazon continues to evolve, reviews have become more than just feedback: they are now a critical factor in the overall success of the platform. While positive reviews can significantly boost a product's ranking and sales, negative reviews can quickly lead to its downfall, reducing its attractiveness and the visibility of the platform itself.

The goal of this project is therefore to gain valuable insights into consumer feedback to possibly enhance its understanding and future management. We do so by employing:

- Binary Classification Models
- Clusterning

2. DATASET INSPECTION

The Amazon Fine Food Reviews dataset is publicly available on Kaggle and, as the name says, it consists of reviews written by consumers about fine foods.

It comprises of a total of 568.4549 reviews and 9 features; however, in this project only the following ones were employed:

- Time -> reviews time frame (1999-2012)
- Score -> reviews rating (1-5)
- Text -> reviews corpus

2. DATASET INSPECTION

A first glimpse at the raw data allowed to identify some critical issues, such as the presence of duplicated reviews:

- written by the same user with identical ratings, timestamps, and text, but associated with different product codes;
- written by the same user about the same product with almost identical text, but different timestamps.

As they were found not to provide additional informative value to our purpose, they were eliminated. After removing such duplicates, the dataset then comprised of 392.969 reviews.

3. PREPROCESSING

As human-generated content, like reviews, presents with various criticalities preprocessing becomes an essential task to ensure the accuracy and effectiveness of the analysis that are going to be carried out.

1. NORMALIZATION

Lowercase Conversion.

Abbreviation Replacement.

Accent Standardization.

Emojis, URLs, Tags.

2. STOP-WORDS REMOVAL

Common list, keeping “not”.

Amazon Context Adjustments:

“amazon-product-purchase-cart”.

3. TOKENIZATION

word_tokenize

[“love”, “coffee”]

4. STEMMING

Porter Stemmer

[“love”, “coffe”]

4. TEXT REPRESENTATION
& Dimensionality Reducation

As the reviews span a period of 13 years, they were analyzed for length differences over time. Despite initial concerns, the analysis has shown no significant difference in the number of words used across the period.

Bag of Words

Words Frequencies

TF-IDF

Prioritizing Relevant Words

Truncated SVD

Streamlining text representation

BoW Explained Variance

78%

TF-IDF Explained Variance

74%

5. BINARY CLASSIFICATION

Our initial task is to classify reviews as positive or negative based on their scores, converted into classes (0 for negative, 1 for positive). The reviews dataset was therefore split into train (70%) and test set (30%). We implemented three classification models:

- Light GBM
- Logistic Regression
- N-Grams Logistic Regression

For comparison, both BoW and TF-IDF representations were used in training the first two models.

Note: during the analysis it was discovered that the feature “Score” is unbalanced. Therefore, before proceeding with this task the classes were balanced.

5. BINARY CLASSIFICATION: LightGBM

The LightGBM classifier is a highly efficient and scalable gradient boosting framework. For this task, it was trained on both BoW and TF-IDF features.

Results:

BoW

Training

	Precision	*Recall*	*F1-Score*	*Accuracy*
Negative	0.84	0.86	0.85	0.85
Positive	0.85	0.84	0.85	

Test

Negative	0.43	0.82	0.57	0.80
Positive	0.96	0.80	0.87	

TF-IDF

	Precision	*Recall*	*F1-Score*	*Accuracy*
Negative	0.86	0.88	0.87	0.87
Positive	0.88	0.85	0.86	

Negative	0.47	0.85	0.61	0.83
Positive	0.97	0.82	0.89	

5. BINARY CLASSIFICATION: Logistic Regression

As second approach, the focus was shift to Logistic Regression, a simpler yet effective classification model. As it was done with the LightGBM classifier, also in this case the logistic regression classifier was trained using both BoW and TF-IDF representations.

Results:

BoW

Training

	Precision	*Recall*	*F1-Score*	*Accuracy*
Negative	0.87	0.87	0.87	0.87
Positive	0.87	0.88	0.87	

Test

Negative	0.55	0.86	0.67	0.87
Positive	0.97	0.87	0.92	

TF-IDF

	Precision	*Recall*	*F1-Score*	*Accuracy*
Negative	0.88	0.89	0.89	0.89
Positive	0.89	0.88	0.89	

Negative	0.57	0.88	0.70	0.88
Positive	0.98	0.88	0.92	

5. BINARY CLASSIFICATION: N-Grams Logistic Regression

To experiment a solution that could better capture the data nuances, the N-Grams technique was implemented with a logistic regression model and the TF-IDF representation.

Specifically, the focus was on bigrams.

Results:

BoW

Training

	Precision	*Recall*	*F1-Score*	*Accuracy*
Negative	0.87	0.87	0.87	0.87
Positive	0.87	0.88	0.87	

Test

Negative	0.55	0.86	0.67	0.87
Positive	0.97	0.87	0.92	

TF-IDF

	Precision	*Recall*	*F1-Score*	*Accuracy*
Negative	0.88	0.89	0.89	0.89
Positive	0.89	0.88	0.89	

Negative	0.57	0.88	0.70	0.88
Positive	0.98	0.88	0.92	

6. CLUSTERING

The second goal consists in determining whether reviews naturally form distinct clusters based on their scores or if there are hidden groupings related to the content topics, independent of the given scores. To achieve this, two methodologies were employed:

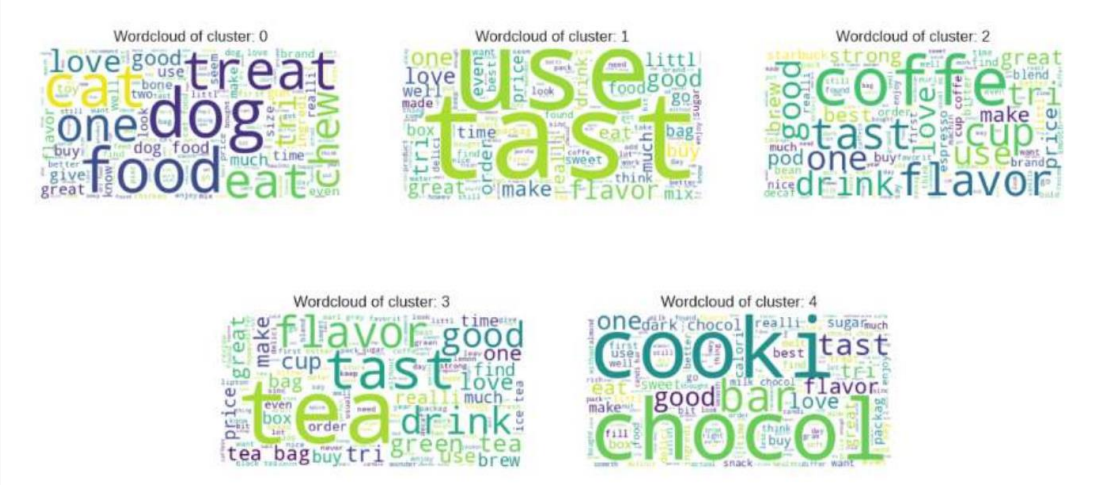
- Agglomerative Hierarchical Clustering
- K-Means

Both clustering methods utilized TF-IDF to leverage the optimal text representation of the reviews and standardization was performed to reduce the influence of value distribution on clustering accuracy.

6. CLUSTERING: Agglomerative Hierarchical Clustering

Despite the algorithm not requiring to specify the number of clusters in advance, here we defined the exact number of clusters that we expect based on the reviews scores, i.e., 5 clusters.

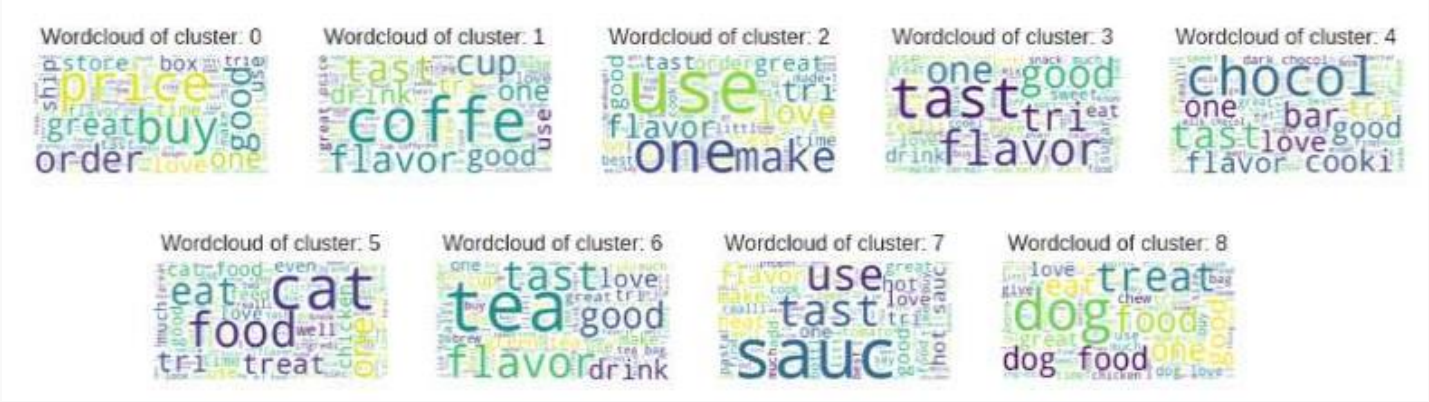
However, the clusters obtained do not seem to represent the scores.



6. CLUSTERING: K-Means

A second trial was made to identify 5 clusters based on score. However, the K-Means algorithm run with $k = 5$ gave similar results to the agglomerative hierarchical clustering algorithm.

The optimal number of clusters was therefore identified using the Elbow method, maximizing the Silhouette coefficient. The K-Means was then run with $k = 9$.



CONCLUSIONS

The analyses carried out allowed to conclude that:

- despite the challenges encountered when classifying negative reviews, the logistic regression model with TF-IDF, which turned out to be the best representation, and N-Grams provided the best performance;
- the application of clustering algorithms revealed that reviews are more likely to group by content themes rather than by scores, suggesting that the text of review might containes valuable information about product categories, independent of user ratings.

THANK YOU!