Università degli Studi di Milano – Bicocca
Master Degree in Data Science
Academic Year 2022/2023

CLASSIFICATION & CLUSTERING OF AMAZON FINE FOOD REVIEWS DATASET

Gianluca CAVALLARO - n. 826049
Remo MARCONZINI - n. 883256

/ Outline

- ❑ Introduction
- ❑ Data Cleaning & Data Exploration
- ❑ Preprocessing
- ❑ Text Representation
- ❑ Text Classification
- ❑ Text Clustering
- ❑ Conclusions

INTRODUCTION

/ Introduction

- Online shopping is an increasingly common practice
- When making an online purchase, one of the **most important aspects are the reviews** left by previous customers
- The leader in the sector is undoubtedly **Amazon**
- In this project, we want to study the **Amazon Fine Foods dataset**
- The main purpose of this study is to apply **text classification** and **text clustering techniques**

TEXT CLASSIFICATION

TEXT CLUSTERING

/ Dataset

- The dataset contains **568,454 reviews**
- They cover the time period between **October 1999** and **October 2012**
- 256,059 **unique** users and 74,258 **different** products
- The dataset contains 10 features

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

/ Research questions

- **Text classification**
  1. *Can a review be classified as good or bad from its text?*
  2. *Is it possible to predict the user's rating starting from the text of the review?*

- **Text clustering**
  1. *Is it possible to group similar reviews?*

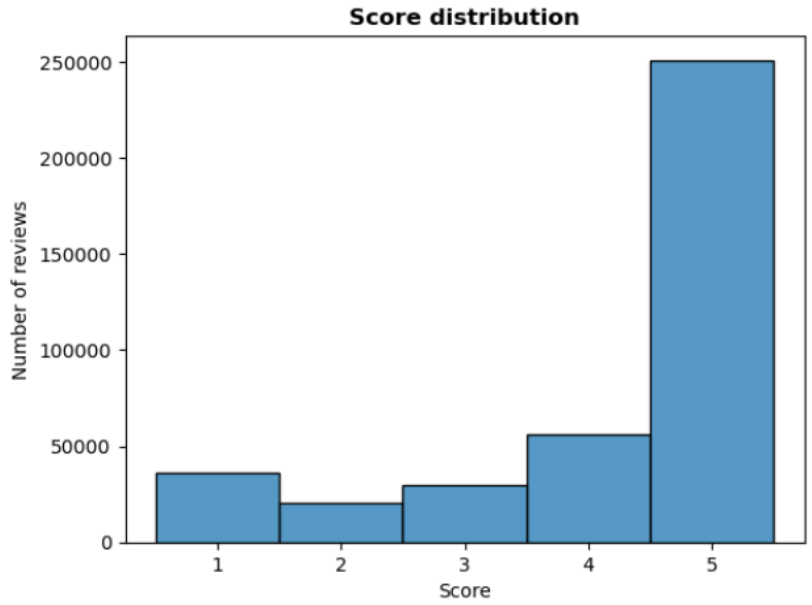
DATA CLEANING & DATA EXPLORATION

/ Data Cleaning

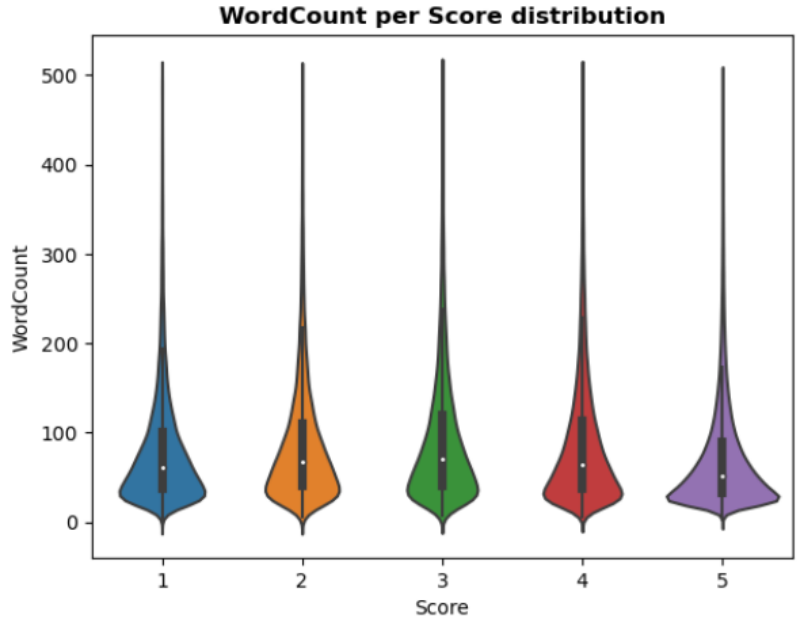
- Deleting rows with inconsistent values for **HelpfulnessNumerator** and **HelpfulnessDenominator**
- Missing values
  - Missing values for **ProfileName** and **Summary** only
- Removing duplicates
  - We have removed rows with identical values for **UserId**, **Time**, and **Text**
- The **dataset is reduced to 393,931 reviews**

/ Data Exploration

- The main focus is on the attributes **Score** and **Text**:



Number of reviews per Score value



Number of words per review per Score value

PREPROCESSING

/ Normalization & Tokenization

❑ NORMALIZATION

- Text is transformed into a single canonical form
  - Conversion to lower case
  - Removing alphanumeric characters
  - Removing emoji, URLs and HTML tags
  - Removal of duplicate characters and spelling correction
  - Removal of punctuation and special characters

❑ TOKENIZATION

- Texts are splitted into tokens
- We decided not to consider n-grams

/ Stopwords removal & Lemmatization

❑ STOPWORDS REMOVAL

- We remove very frequent words (articles, pronouns, …) without semantic meaning

❑ LEMMATIZATION

- All the inflected forms of a word are grouped into a single token
- More expensive than Stemming, but preserves the meaning of the words

```
'I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like a stew than a processed meat and it smells better.
My Labrador is finicky and she appreciates this product better than  most.'
```

```
"['bought', 'several', 'vitality', 'canned', 'dog', 'food', 'product', 'found', 'good', 'quality', 'product', 'look', 'like', 'stew', 'process', 'meat', 'smell', 'better', 'labrador',
'finicky', 'appreciated', 'product', 'better']"
```

TEXT REPRESENTATION

/ Text Representation

- 3 possible text representation for both **binary** and **multi-class** classification:
  - **Bag-of-Words**
  - **TF-ID**
  - **Word2Vec**
    - 300 components
    - CBOW

Representation	Binary	Multi-class
BOW	80999	72807
TF-IDF	80999	79518

Original no. of **features**

- **Dimensionality reduction** for both BOW and TF-IDF:
  - **Singular Value Decomposition (SVD)**
  - Cumulative variance analysis:
    - 85% of the total variance in the dataset (**threshold**)

Representation	Binary	Multi-class
BOW	1466	1413
TF-IDF	4776	4719

Optimal no. of **components**

CLASSIFICATION

/ Objective

- Answer the following two questions:
  - *Can a review be classified as **good** or **bad** from its text?*
  - *Is it possible to predict the **user's rating** starting from the text of the review?*

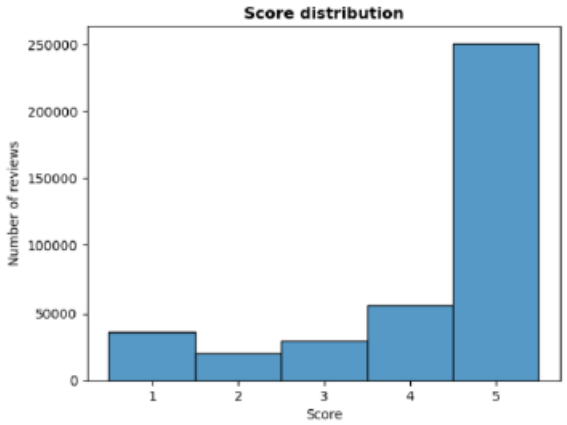
- With the following techniques:
  - **Binary** Classification
  - **Single-Label-Multi-Class** Classification

/ Preprocessing

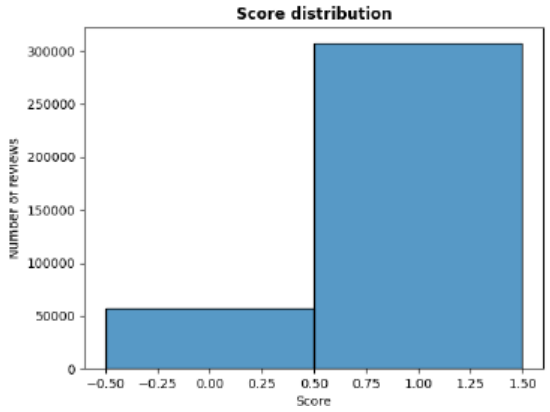
- Building the **Binary** datasets:
  - **Negative** class: merging class “1” and “2”
  - **Positive** class: merging class “4” and “5”
  - Excluding the **intermediate** class 3.

- **Class imbalance** problem:
  - For both binary and multi-class datasets
  - **Downsampling** of the datasets
    - rebalancing with respect to the least populated class

- **Final** datasets, divided into **training** (70%) and **test** (30%) set
  - Binary datasets: **114216** reviews
  - Multi-class datasets: **104010** reviews



Multi-Class score distribution



Binary score distribution

/ Algorithm Implementation

- We used four different **algorithms**:
  - Logistic Regression,
  - XGBoost
  - K-NN
  - Random Forest

- These classifiers were estimated using all three **text representations**:
  - BOW,
  - F-IDF
  - W2V

- We performed **fine-tuning** with **cross-validation**
  - Random Search
  - K = 2

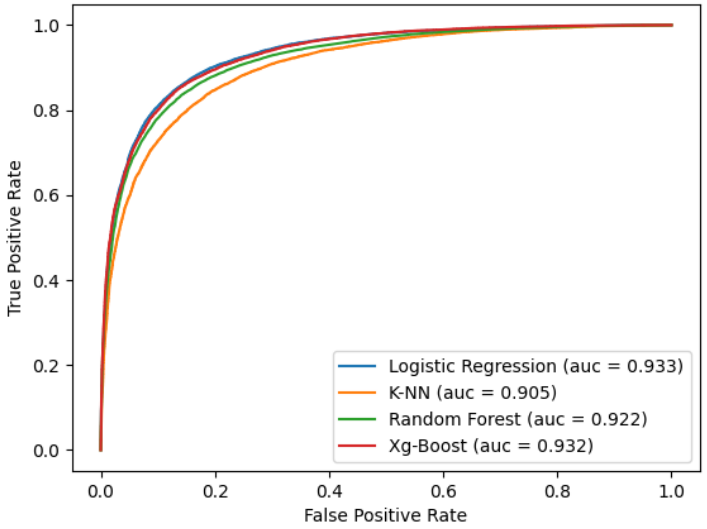
- **Performance evaluation**
  - Accuracy
  - AUC-ROC curve (binary classification)
  - Confusion matrix (multi-class classification)

/ Binary Classification - Analysis Of The Results

- Can a review be classified as good or bad from its text?

Model	Bag-Of-Word		TF-IDF		Word2Vec	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Logistic Regression	**0.87**	0.94	**0.88**	0.95	**0.86**	0.93
K-NN	0.55	0.58	0.64	0.70	0.82	0.90
Random Forest	0.74	0.83	0.79	0.87	0.85	0.92
XG-Boost	0.81	0.89	0.83	0.91	**0.86**	0.93

- In term of accuracy and AUC:
  - **Logistic regression** and **XGBoost** have a better performance
    - Across all the text representation
- **AUC** values for the **Word2Vec** representation are relatively high across all models
  - Word2Vec captures **semantic information**



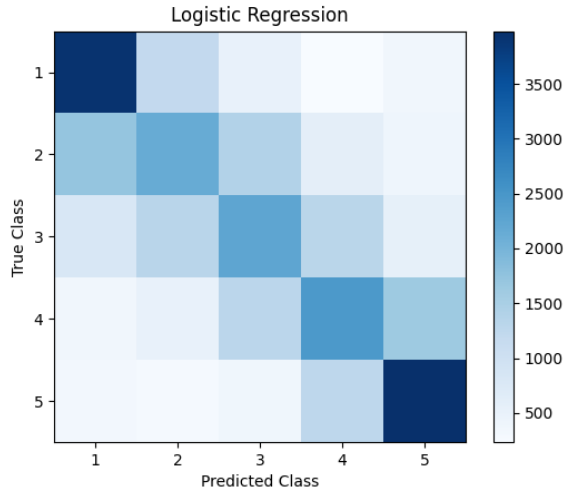
AUC-ROC curve, Word2Vec

/ Multi-Class Classification - Analysis Of The Results

- Is it possible to predict the user's rating starting from the text of the review?

Model	**Bag-Of-Word** Accuracy	**TF-IDF** Accuracy	**Word2Vec** Accuracy
Logistic Regression	0.22	0.31	**0.47**
K-NN	0.22	0.22	0.38
Random Forest	**0.28**	0.29	0.45
XG-Boost	0.27	**0.43**	0.45

- In term of **accuracy**:
  - **Lower** with respect to the binary classification task
  - Word2Vec representation has the **best performance** across all models
    - capture semantic information
    - is a **dense vector** representation
- The models performed well in classifying the **extreme classes**
  - **Intermediate** ratings may contain **mixed sentiments**

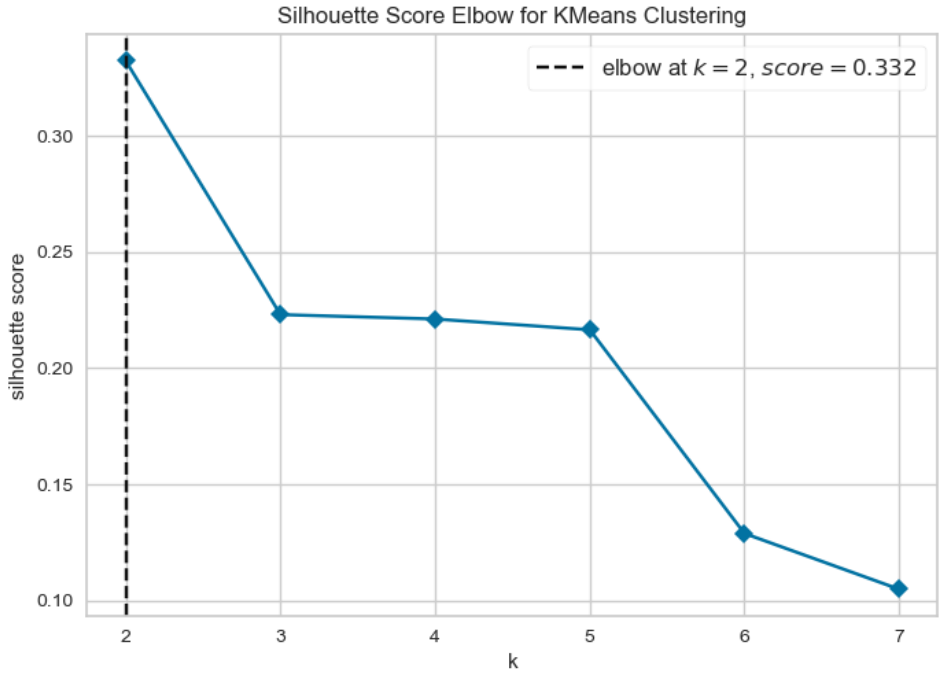


Confusion Matrix, Word2Vec, Logistic Regression

CLUSTERING

/ Clustering: choosing the optimal number of clusters

- **Is it possible to group similar reviews?**
- Two different clustering algorithms: **K-Means and Agglomerative clustering**
- The optimal number of cluster is chosen considering the **Silhouette coefficients**
  - The optimal number corresponds to the highest Silhouette coefficient



/ Clustering: K-Means

- Iterative process
  - Random selection of centroids
  - Assignment of each record to the closest centroid
  - Iterate until stop criterion
- **Optimal number of clusters:**
  - BOW = 2
  - TF-IDF = 3
  - W2V = 2
- We use Silhouette coefficients and Davies-Bouldin index as validation metrics
- **Both metrics are generally bad**
- **No common concept for the clusters**

Text representation	Silhouette	Davies-Bouldin
BOW	0.336	5.376
TF-IDF	0.013	6.257
W2V	0.086	3.035

Evaluation metrics

Text representation					
BOW - cluster 1	taste	like	product	good	one
BOW - cluster 2	like	taste	one	coffee	product
TF-IDF - cluster 1	tea	taste	like	flavour	good
TF-IDF - cluster 2	coffee	like	taste	cup	flavour
TF-IDF - cluster 3	like	taste	product	good	one
W2V - cluster 1	taste	like	coffee	flavour	good
W2V - cluster 2	product	one	like	food	would

Most frequent words per cluster

/ Clustering: Agglomerative Clustering

- Hierarchical clustering algorithm
- Observations starts in their own cluster and then are progressively merged until a single cluster is obtained
- **Optimal number of clusters is 2**
  - For all text representations
- **Metrics are still bad**
  - Slight improvement for W2V
- **Still difficult to a common concept for the clusters**

Text representation	Silhouette	Davies-Bouldin
BOW	0.401	5.642
TF-IDF	0.002	8.022
W2V	0.178	1.825

Evaluation metrics

Text representation					
BOW - cluster 1	taste	like	product	good	one
BOW - cluster 2	like	coffee	taste	one	food
TF-IDF - cluster 1	like	taste	product	good	one
TF-IDF - cluster 2	coffee	like	taste	cup	flavour
W2V - cluster 1	taste	like	coffee	flavour	good
W2V - cluster 2	product	like	one	food	would

Most frequent words per cluster

CONCLUSIONS

/ Conclusions

- CLASSIFICATION
  - Binary Classification
    - **Logistic Regression** and **XG-Boost** performed the best across all text representations
    - **TF-IDF** or **Word2Vec** representation with these models may yield the best results
  - MultiClass Classification
    - **Word2Vec** representation gives the best results in combination with **Logistic Regression**

- CLUSTERING
  - **No satisfactory results**
  - There are probably **no well-defined clusters in the data**
    - Possible reasons:
      - Presence of many semantically useless words
      - The proposed text representations generate sparse vectors when applied on short texts on short texts tend to generate sparse vectors
  - Ad-hoc preprocessing could lead to better results
    - Example: a rigorous application of the Zipf's law

References

- Stanford Network Analysis Project. Amazon Fine Food Reviews. URL: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews
- Chen H-H Cheng C-H. "Sentimental text mining based on an additional features method for text classification". In: (2019). DOI: https://doi.org/10.1371/journal.pone.0217591
- Dilip Valeti. Classification using Word2vec. URL: https://medium.com/@dilip.voleti/classification-using-word2vec-b1d79d375381
- G. Pasi and M. Viviani. Text Mining and search course lecture notes and slides. 2022-23
- Tune Hyperparameters for Classification Machine Learning Algorithms. URL: https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/
- Scikit-learn. URL: https://scikit-learn.org/stable/index.html