# RATING PREDICTIONS FROM REVIEWS GIVEN TO PRODUCTS IN ONLINE MARKETS

### CPSC-6300 Applied Data Science

Spring 2023

### Checkpoint 1

Dmitry Lukyanov

Sejal Bansal

Adithya Ravi

Shareef Shaik

## The dataset peculiarities

1. Almost all reviews contain only evaluation of products with average presence of non-product aspects in reviews approximately equal 1 to 100. It makes it virtually impossible to utilize the dataset for training a model that would be able to distinguish such facets. To be addressed during the next stages of the project.
2. The dataset is highly imbalanced in ratings distribution. To be addressed during the next stages of the project.

## EDA Unit

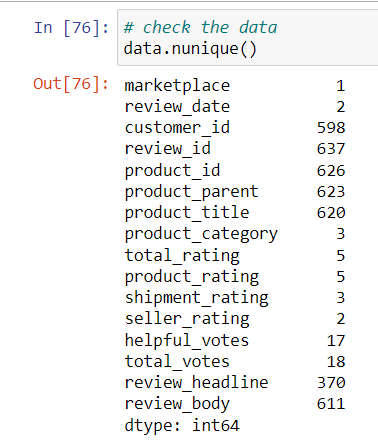
A unit of analysis is a review of a product at Amazon. Each row in the dataset represents a single review, with features such as a product category, a review text, a rating and so on. Therefore, the analysis performed on this dataset focuses on the characteristics and distribution of individual reviews.

## Dataset size

The dataset contains a total of 637 reviews.

## Unique observations

The below table specifies the number of unique observations



## Covered period of time

All reviews within the sample fall on 08/31/2015.

## Data cleaning

The following steps were executed on the dataset.

#### Unnecessary columns were dropped

1. id: has no value to the project
2. marketplace: all values are the same

#### Columns’ format has been unified

1. review\_date: had two different formatting

#### Types were adjusted

| **Column** | **Original** | **Target** |
| --- | --- | --- |
| review\_date | object | datetime |
| customer\_id | int | string |
| review\_id | object | string |
| product\_id | object | string |
| product\_parent | int | string |
| product\_title | object | string |
| product\_category | object | string |
| total\_rating | int | string |
| product\_rating | int | string |
| shipment\_rating | float | string |
| seller\_rating | float | string |
| review\_headline | object | string |
| review\_body | object | string |

#### Missed values

There were no missing values except missing labels shipment\_rating and seller\_rating. Corresponding units were labeled manually with marking lack of relevant aspects in reviews as “NA”.

#### Reviews cleaning

Texts of reviews were cleaned from html-tags.

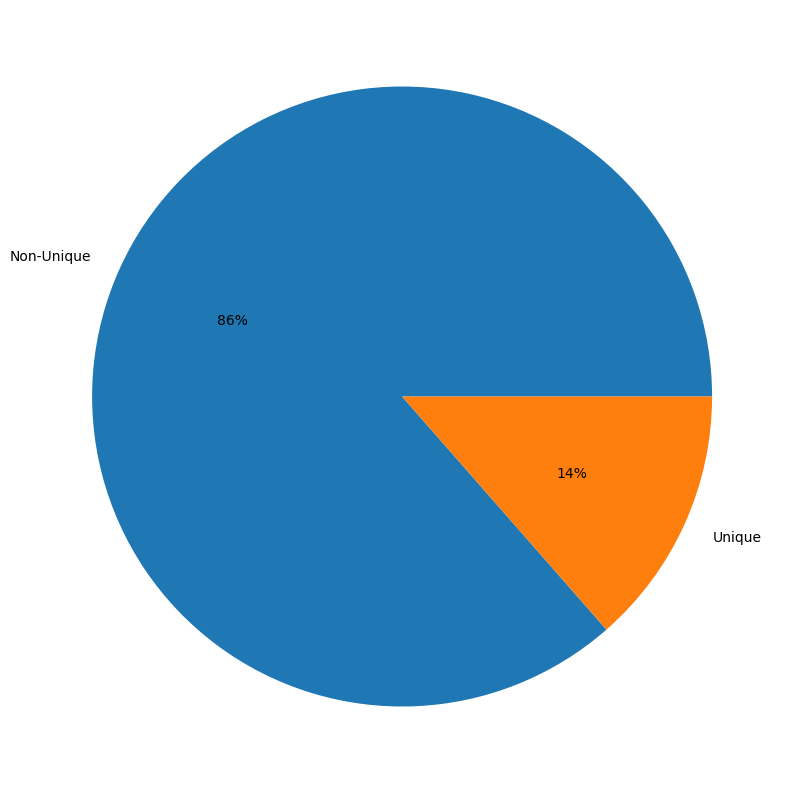
#### Reviews normalizing

* Text was lowercase.
* Punctuation marks were removed.
* Contradictions in text were expanded.
* Words were lemmatized.

## Visualizations and Analysis

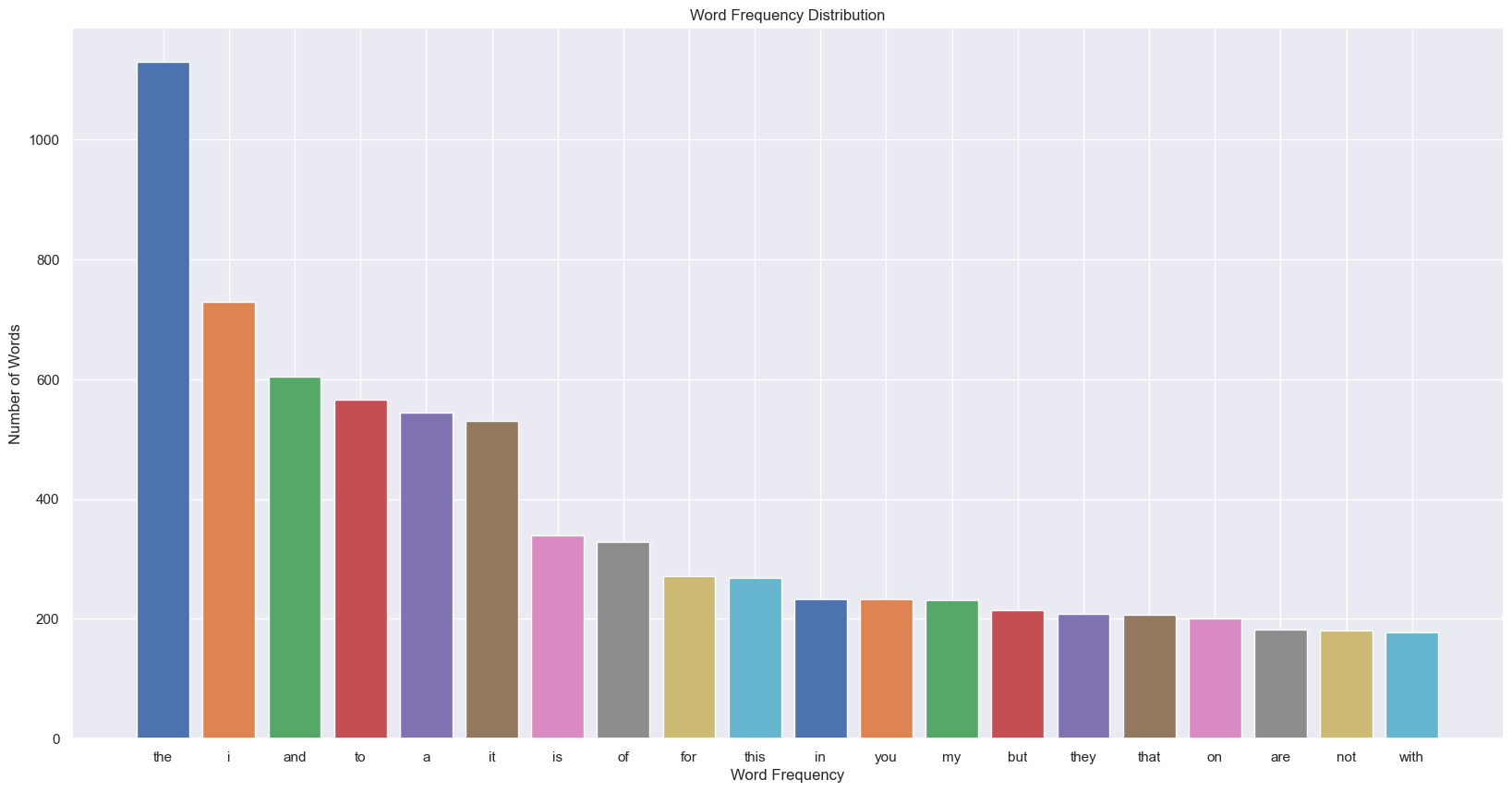
### Words

In total, the dataset contains 22496 words with the distribution on the unique and non-unique words.



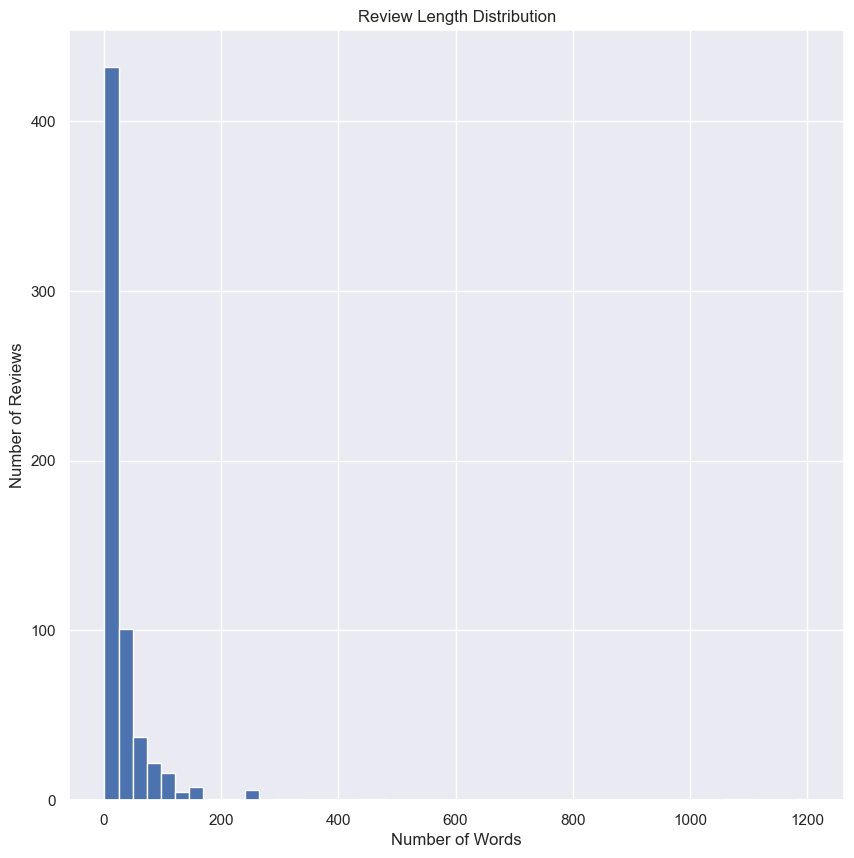
It shows the tendency of usage of approximately the same set of words in most reviews.

The next words are the most frequent in the reviews



The plot revealed that the most frequent words in the reviews are articles, pronouns and prepositions that, probably, should be addressed during the further stages of the project.

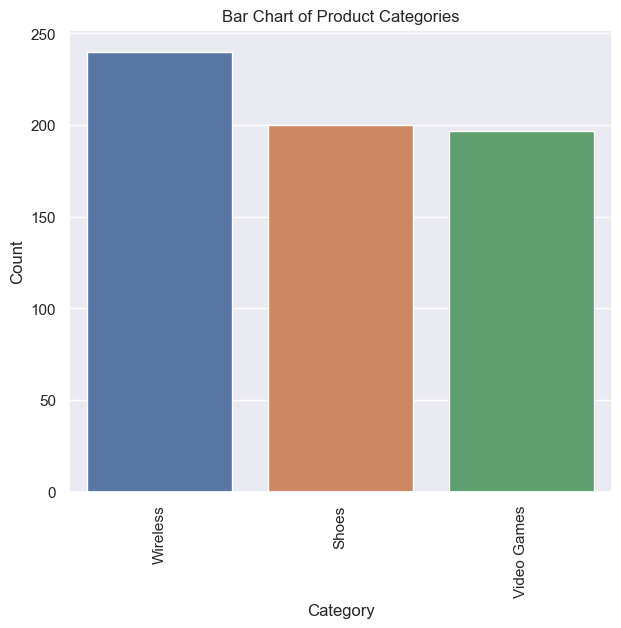
Reviews’ lengths are distributed in the next way



The plot reveals that the size of reviews gravitates to the short side that can affect a model in the aspect of handling reviews that have a size that is different from the majority of the training dataset.

#### Categories

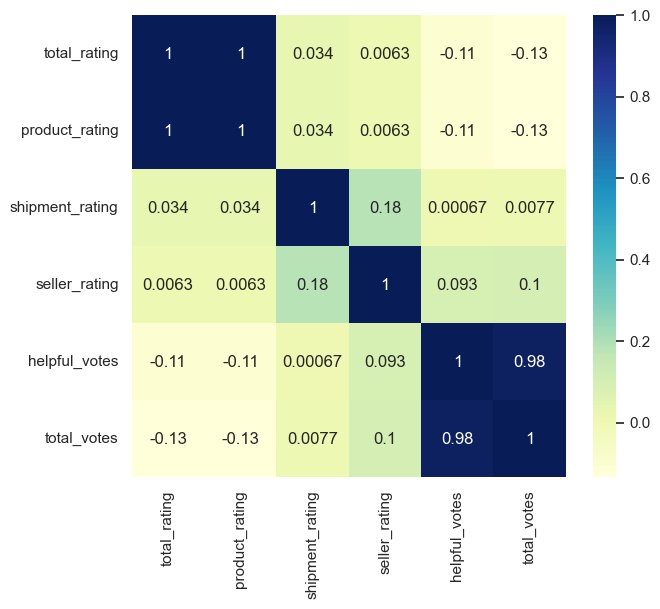
Categories of products in the dataset sample are distributed in the next way.



The plot shows that products are distributed across the dataset almost uniformly, which is a good quality for a training dataset. There is still an insignificant imbalance that can be addressed during the next stages of the project.

### Correlation

The correlation among features if we consider rating-related ones as quantitative is the next.



The correlation heatmap shows that there is no visible positive or negative correlation between ratings and votes. However, there is direct correlation between a total rating and a product rating. As almost all reviews are about products only, it’s expected. As there is almost none of reviews that include ratings for shipments and sellers, the lack of correlation between them and all other variables are understandable.

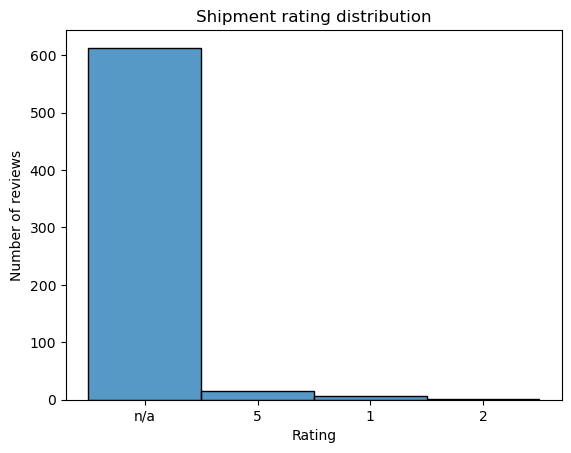
### Rating

The rating distribution among reviews is the next



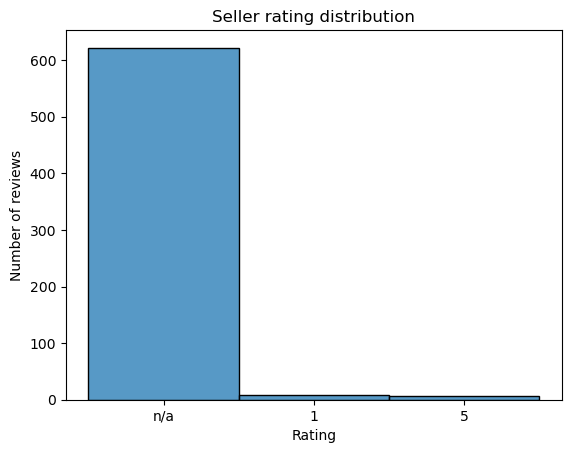
The plot reveals an imbalance in the ratings distribution as it was mentioned in the beginning of the document. As it can affect the model training, it should be addressed during the next stages of the project.

The current shipment rating distribution that to be address on the next stages of the project is the next



As there is almost no mention of shipments in reviews, it’s virtually impossible to train a model to handle reviews with such mentionings properly. It should be addressed during the next stages of the project.

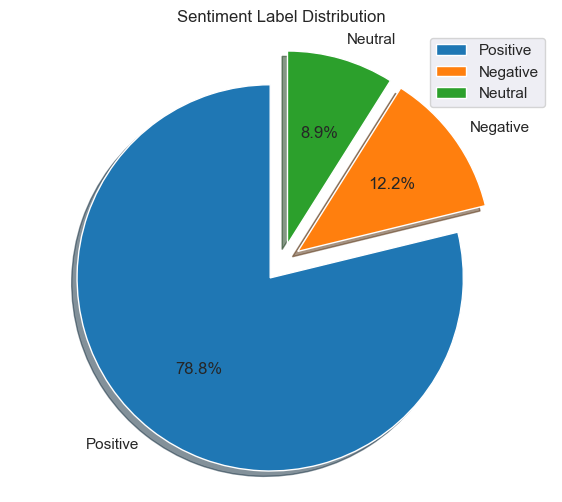
The current seller rating distribution that to be addressed on the next stages of the project is the next



As there are almost no mentions of sellers in reviews, it’s virtually impossible to train a model to handle reviews with such mentionings properly. It should be addressed during the next stages of the project.

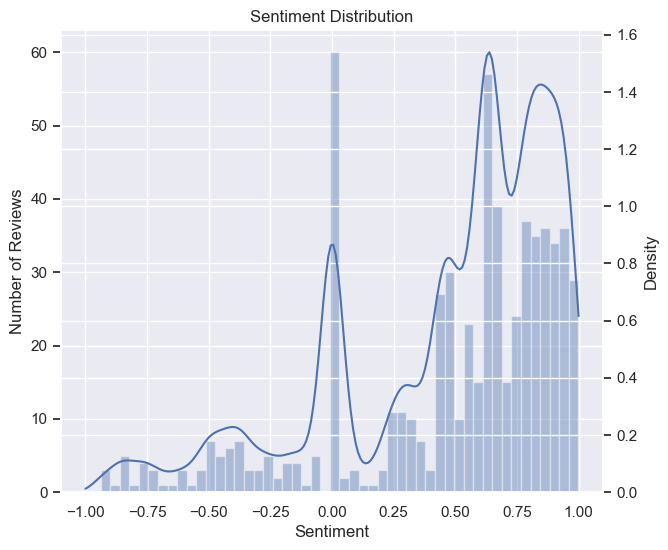
### Sentiment

The sentiment distribution across all reviews with categorization is the next



The chart shows that the majority of reviews in the dataset are Positive, followed by Negative and then Neutral. This creates an imbalance that is to be addressed in the further stages of the project.

The sentiment distribution across all reviews after normalization is the next



The plot provides insight into the sentiment distribution of the Amazon product reviews dataset, showing that the majority of reviews have a positive sentiment score, with a spike of the neutral reviews and a small number of negative reviews. The density plot also indicates that the distribution is skewed towards positive sentiment. As it was already mentioned, this creates an imbalance that is to be addressed in the further stages of the project.

## Key predictors

Considering the nature of the project, the features of the dataset and the executed correlation analysis, the most reasonable choice of predictors will include in order of importance

1. review\_body - a text of a review
2. review\_headline - a title of a review