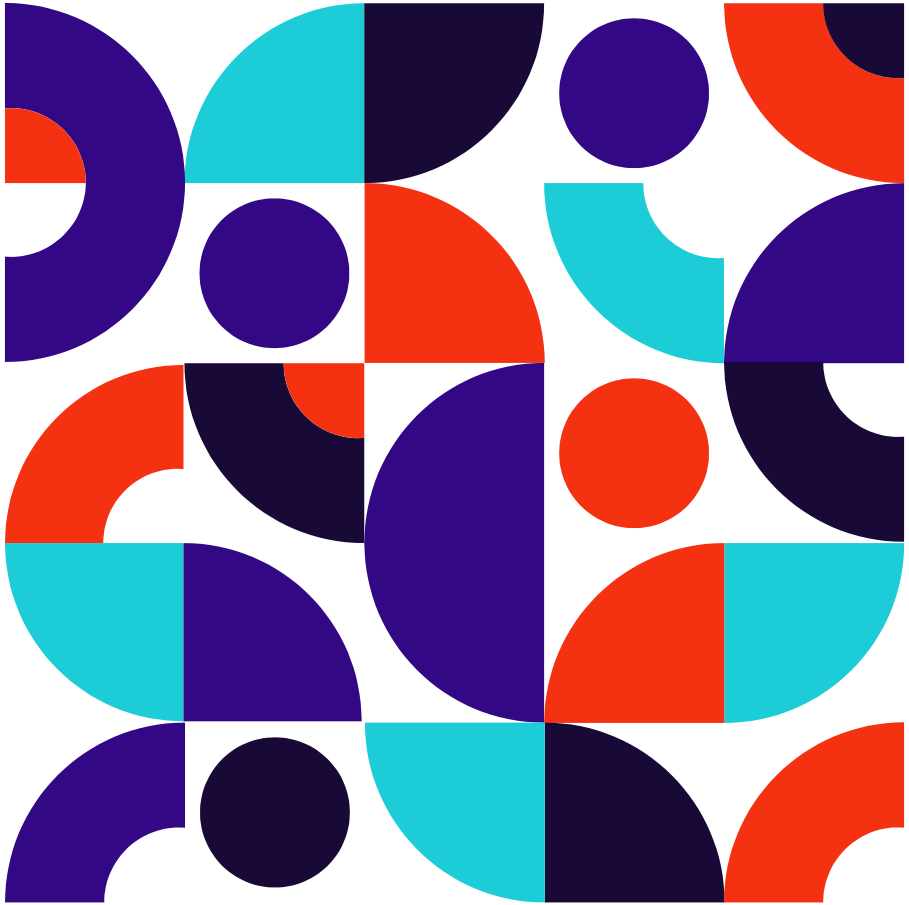
Rating discovery and genre classification from book reviews

Text Analytics

Group 7



UNIVERSITÀ DI PISA

Our Team

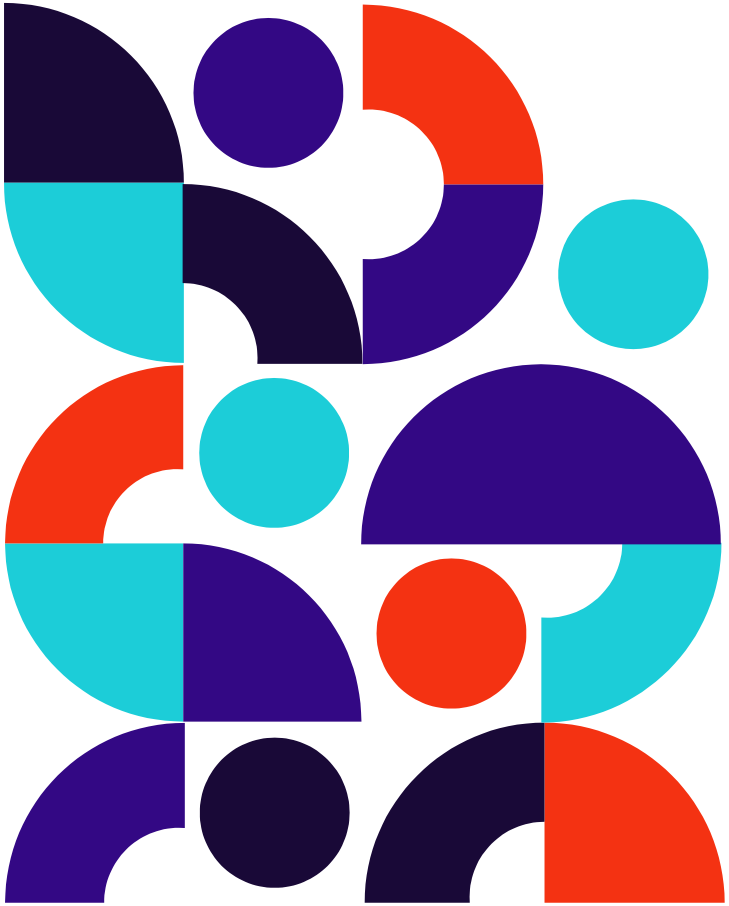
Lia Trapanese
Data Science and
Business Informatics

Davide Innocenti
Data Science and
Business Informatics

Ludovico Lemma
Data Science and
Business Informatics

Maria Grazia Antico
Data Science and
Business Informatics

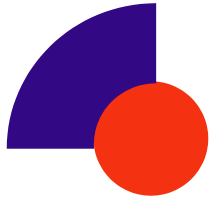
Chiara Germelli
Digital Humanities



Why this proposal?

Reviews allow **optimal purchasing choices** for book buyers

New platforms give the opportunity to users to express their **opinions** regarding their reading experience



The Goal

- *Developing an automatic system for **predictive product evaluation** based on reviews*

- *Retrieving **customer experiences for the classification** of the book genre*

… How?

Analyzing the structure of the reviews to get insights regarding a book

The dataset

It has been taken from Kaggle and contains more than **1.3M book reviews** (rows), about **25,475 books** and **18,892 users** from the *Goodreads* website

From...

- User id
- Book id
- Review id
- Rating
- Review text
- Date added
- Date updated
- Read at
- Started at
- N° votes
- N° comments

To...

- User id *(int)*
- Book id (int)
- Review id (int)
- **Rating (from 0 to 5)** → *Target*
- Review text (string)
- **Genre (string)** → *Target*

Source:
https://www.kaggle.com/competitions/goodreads-books-reviews-290312/overview

How we reach the “Genre” variable?

… Web Scraping!

https://www.goodreads.com/book/show/ + { **book_id** }

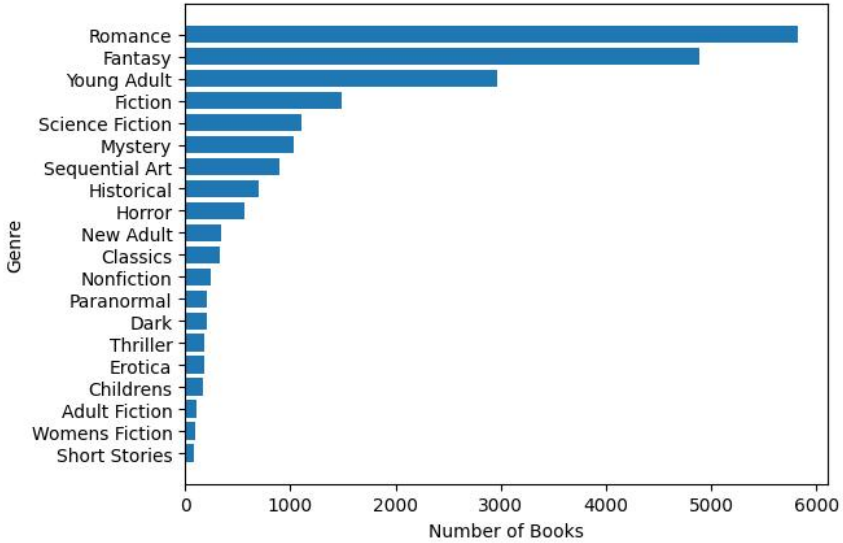
XPath on the page

In total we found **76** genres

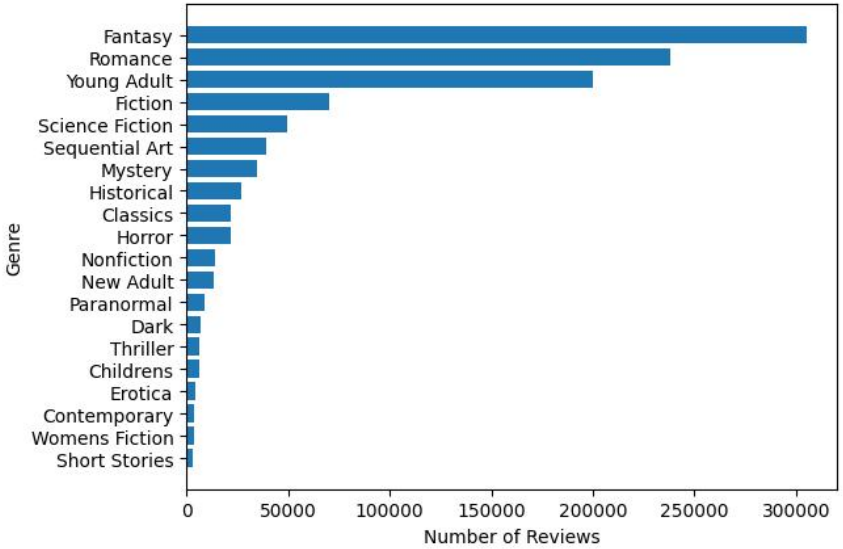
More on “Genre” variable

Since there is a strong genre imbalance, we will **group similar genres** and, if necessary, apply other techniques

In the following bar plots are represented only the first 20 genres



3.3K / 25K **Books** with **NULL** values



280K / 1M **Reviews** with **NULL** values

Classification

Target variable: **Genre**

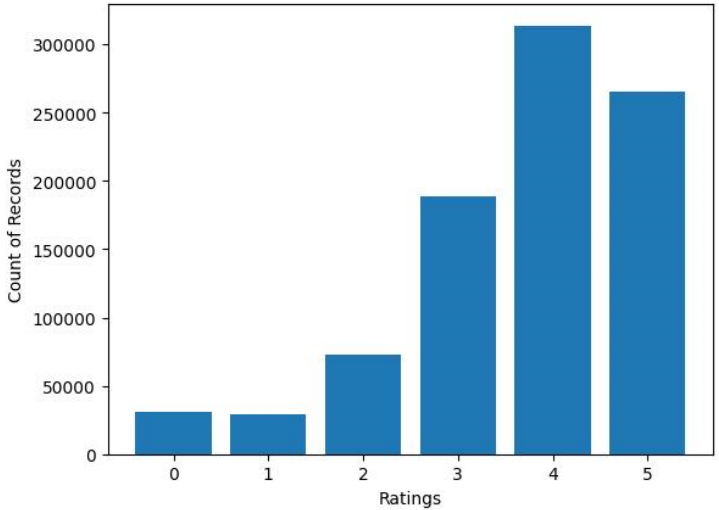
Aim: *Identifying the book genre from the "shelves" of users on the platform*

- SVM
- Neural Network (like BERT)
- Naïve Bayes

Regression

Target variable: **Rating**

Aim: *Identifying the rating of the books from the user emotions expressed by book reviews*



Workflow

1. **Web scraping** using XPath to extract the new variable “Genre”
2. **EDA** (Exploratory Data Analysis)
3. **Regression** task using *“Ratings”* as target variable
4. **Classification** task using *“Genre”* as target variable
5. **Feature Extraction** and **Explainability** (if meaningful) with *SHAP* to explain the prediction of a black box classifier

Thank you!