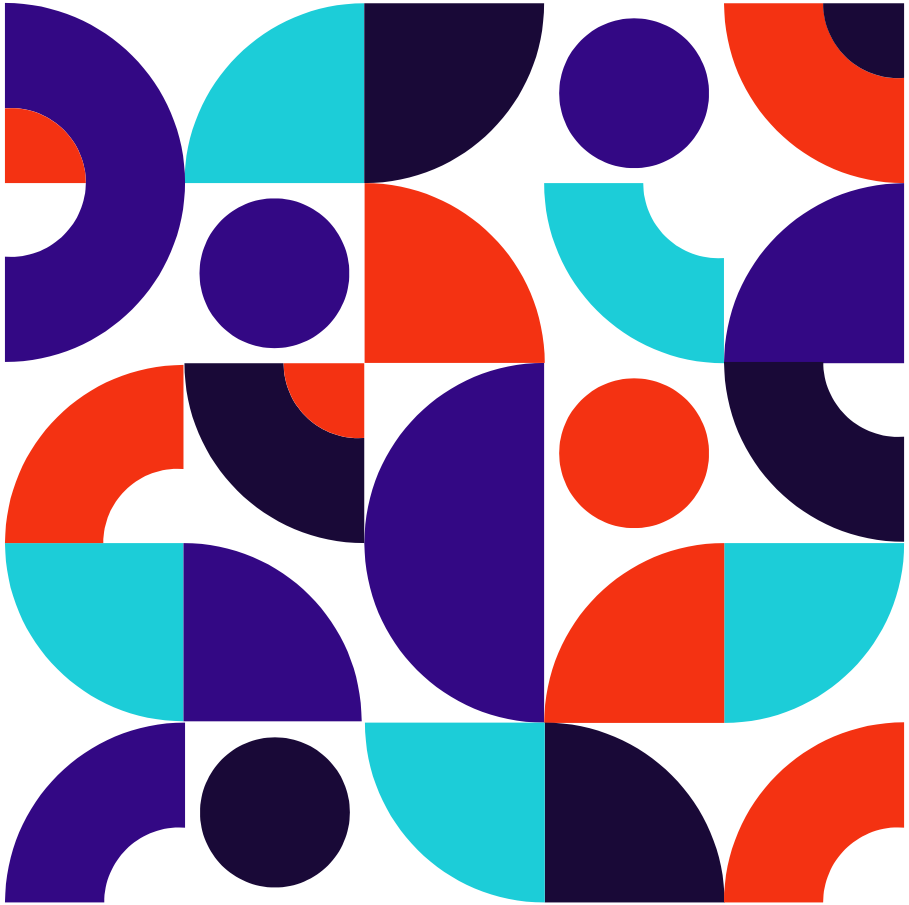
Rating discovery and genre classification from book reviews

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

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

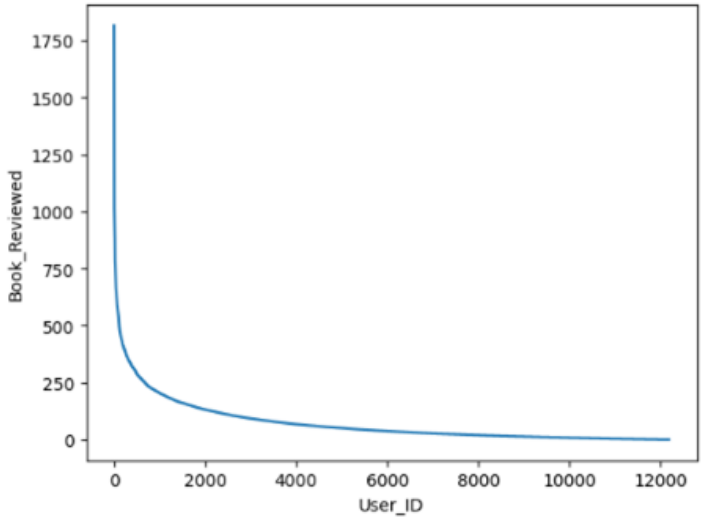
Workflow

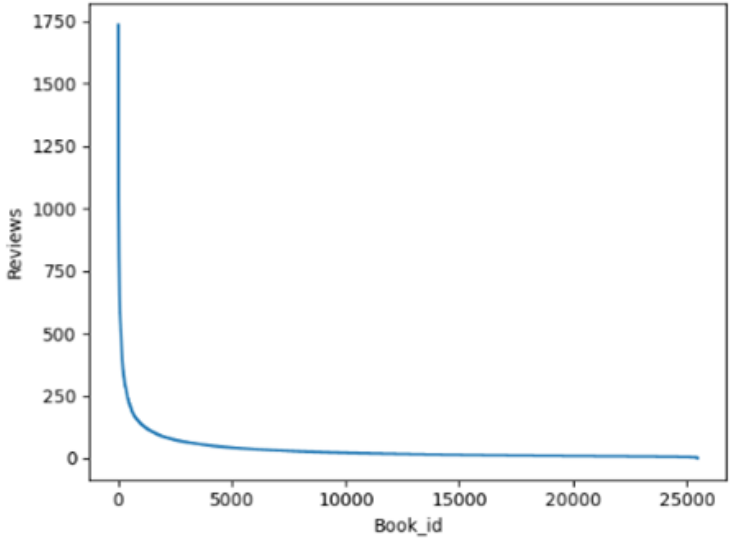
1. **Data understanding** and **preparation**
2. **Web scraping** using XPath to extract the new variable “Genre”
3. **Classification** task using *“Genre”* as target variable
4. **Classification** task using *“Ratings”* as target variable
5. **Explainability** with *SHAP* to explain the prediction of a black box classifier (as Neural network)

EDA – Preprocessing I

Power law of a **few users with lots of reviews** and **too many reviews for some books**....

... So we balanced reviews per book!





EDA – Preprocessing I

- *Cleaning*, Tokenizing, *PoS* Tagging of review texts
- *Removing* stop words, *filtering out* short words, *Lemmatizing*

Size: **8.474 words**!

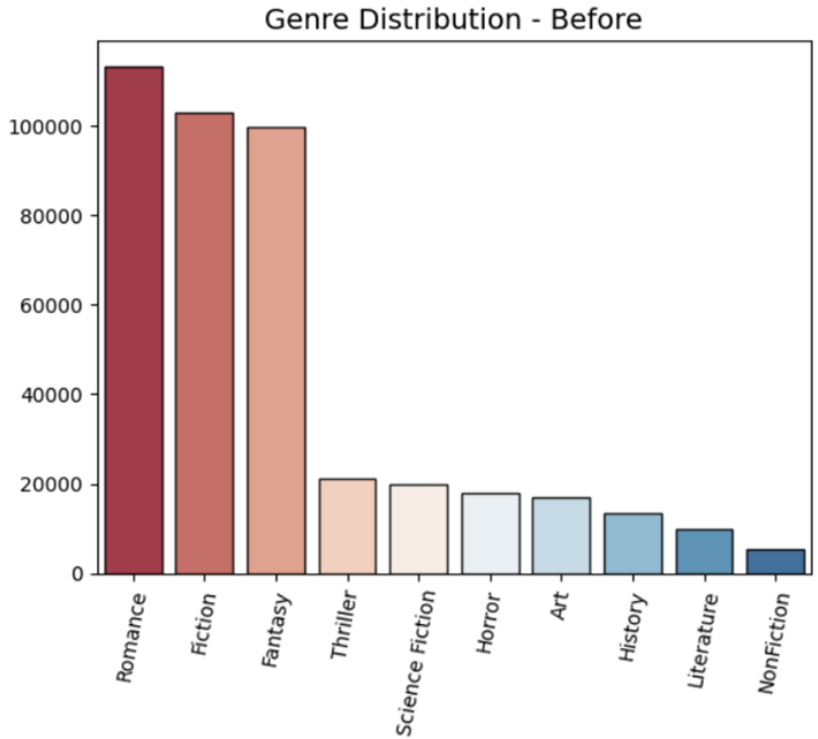
New dataset obtained:

review_text	genre	rating	book_id	tokenized_ text	postagged_ text	lemmatized_ text
i like that i hear all the character voices as...	Art	3	133765	[like, hear, character, voices, read, honestly...	[(like, IN), (hear, VBP), (character, NN), (vo...	[like, hear, character, voice, read, honestly,...

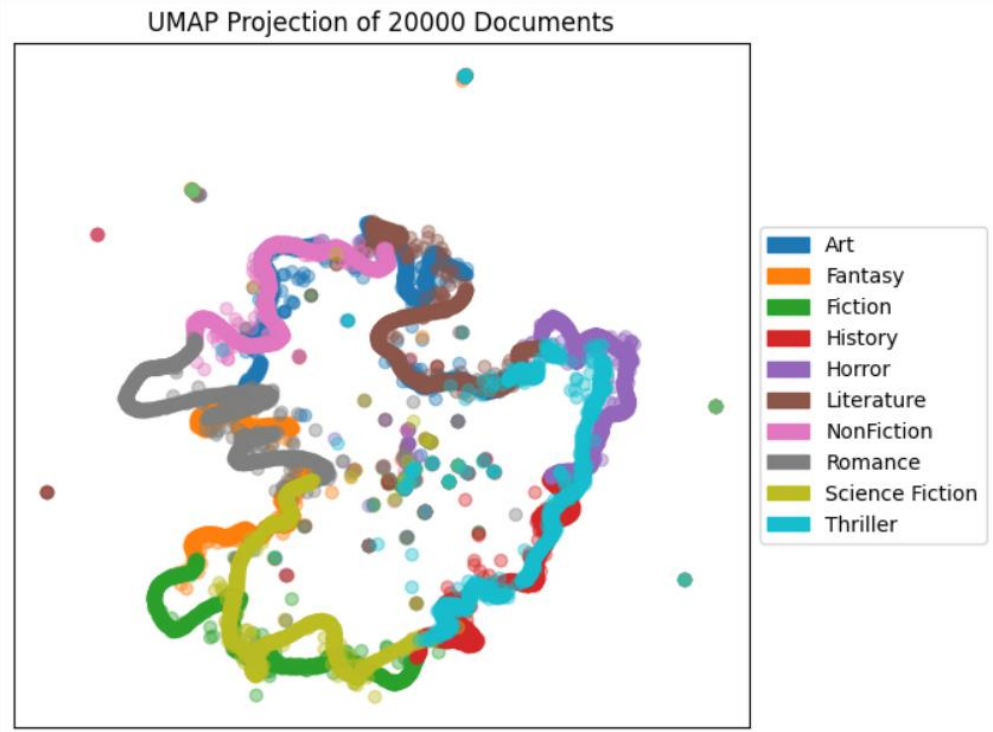
EDA – Preprocessing I

The initial 76 genres has been grouped obtaining, in total, **10 genres**.

Furthermore, the distribution has been **balanced using with an Undersampling**



EDA – Visualizazing genres with UMAP

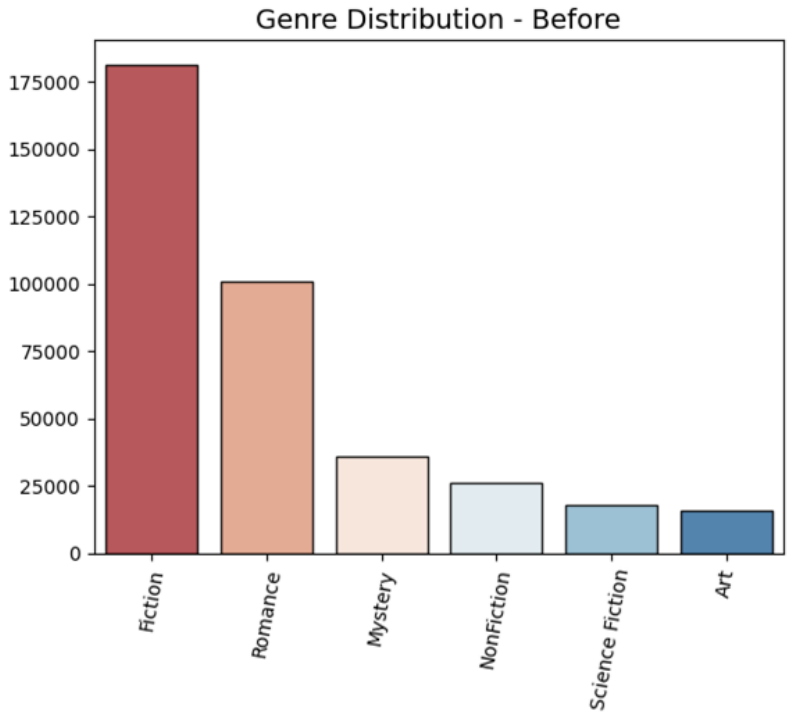


... It can be seen a **similarity** with some genres

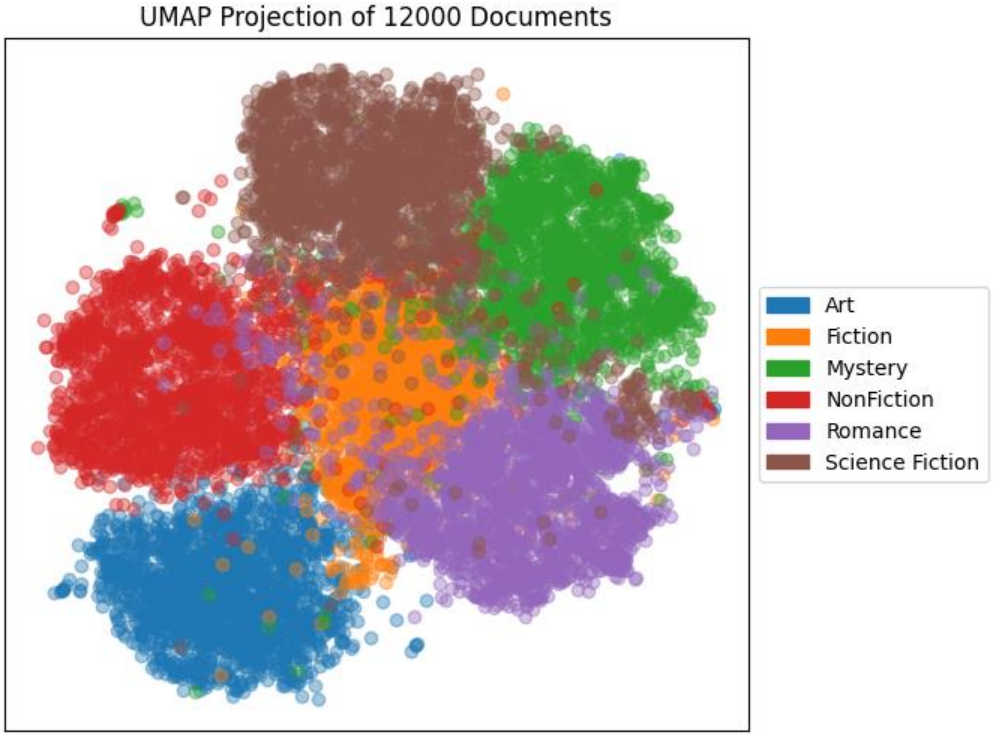
EDA – Preprocessing II

After testing some classifiers we further merged from 10 to **6 genres** according to the visual closeness we saw in the UMAP plot

Then, the distribution has been balanced using a stratified **Undersampling** technique **(2.000 records per genre)**



EDA – Visualizazing genres with UMAP



... After **merging the genres** based on the previous Umap, we got a *new clearer visualization*

Topic Modeling

	genre	freq	cluster	Topic words
0	Art	36	4	ghost
1	Fiction	64	5	narrator
2	Non Fiction	85	5	narrator
3	Romance	211	2	sexy
4	Science Fiction	254	1	alien
5	Mystery	98	4	ghost

Steps:

- Extracting the top n° Topic words (**equal to n° of genres)** treating them as clusters
- Extracting the **maximum frequency for each genre** of the Topic words and looking at their semantics

Sentiment Analysis – Rating/Sentiment metric and correlation

	Precision	Recall	F-1 measure
1	0.12	0.59	0.20
2	0.20	0.21	0.21
3	0.35	0.23	0.28
4	0.46	0.29	0.35
5	0.47	0.44	0.45

***Accuracy**: 0.32*

***Correlation** between rating and sentiment:* 0.29

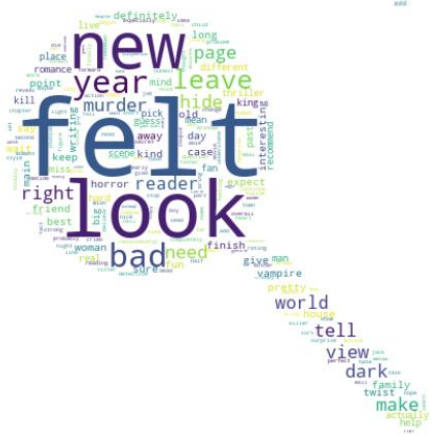
...Conclusions:
Sentiment Analysis is **not** a good predictor of the ratings

Some WordClouds on genres

Romance

Fiction

Mystery



Art

From Classification with 6 genres...

	Validation Accuracy
BERT	***62%***
Naïve	***17%***
LSTM	***... Work in progress!***