Cross-collection Dataset of Public Domain Portuguese-language Works

Mariana O. Silva, Clarisse Scofield, Luiza de Melo-Gomes, Mirella M. Moro

Universidade Federal de Minas Gerais, Brazil
{mariana.santos,clarissescofield,luizademelo,mirella}@dcc.ufmg.br

Abstract. Many datasets are published in English to get more engagement, popularity and reach within a research community. Indeed, most sciences are language-agnostic and thrive on publicly available data. However, such a claim is not always valid for Arts, where Literature and Music are two examples of fields that heavily rely on the language of the work. Especially in Literature, combining human expertise with book consumers' data may generate what is needed to sustain constant changes experienced in the book publishing market. Therefore, we introduce *PPORTAL*, the first public domain Portuguese-language literature dataset that is composed of a wide variety of book-related metadata. After introducing its building process and content, we present an exploratory data analysis with a quantitative description of its main features. We also show its usability as a resource on different research domains through examples of real-world applications, as well as pointing out other potential applications.

Categories and Subject Descriptors: H.2 [**Database Management**]: Database Applications; H.3 [**Information Storage and Retrieval**]: Digital Libraries; J.5 [**Arts and Humanities**]: Literature

Keywords: Dataset, Portuguese Literature, Public Domain Works, Feature Engineering.

1. INTRODUCTION

Creative industries have undergone significant changes in the last decade, mainly due to digital transformation. Such evolution interferes with all productive sectors, including the book industry. Indeed, book publishing has become more accessible to everyone in terms of opportunity, marketing freedom and greater consumption flexibility, at the same time as the book itself has moved to the digital format. Still, technical challenges are not limited to moving from physical to digital books. Instead, book digital transformations require putting the physical and the digital editions side by side, making them coexist to strengthen the publishing industry as a whole.

In such a context, combining human expertise with book-consumer digital data is essential to face existing and upcoming challenges [Champagne 2020]. Along with the publishing industry, researchers rely on book-related data to develop tools whose results feed better informed, faster decisions. Such solutions range from best-sellers prediction models [Maity et al. 2019; Sharma et al. 2020; Wang et al. 2019] to natural language processing techniques to classify raw text [de Araujo et al. 2018; Bao et al. 2021; Shahsavari et al. 2020]. Besides requiring Artificial Intelligence (AI) based methods, all of them are essentially data-dependent, i.e., primarily book-related data-dependent.

However, no solution can be properly developed and tested without considering data on literary works, readers and their reading habits [Lebrun and Audet 2020]. In other words, proper solutions require building and publishing datasets that fully comprise the essential elements of the book industry ecosystem. Although there are efforts for English-written books [Ni et al. 2019; Sabri and Weber 2021], little has been done regarding other lesser-spoken languages, such as Portuguese. A naive solution is to

This work was partially funded by CNPq and FAPEMIG, Brazil.

Table I. Automatically translated examples from Portuguese to English. Fragments from Dias Gomes' *Odorico na Cabeça* and Machado de Assis' *Dom Casmurro*. Errors are highlighted and missed *wrong* words are underlined.

Original text in Portuguese	**Automatically translated English**
– Não vou botar amendoim no vatapá da Oposição! Ou tiram da placa inauguratícia o nome desse patifista subversento, ou a inauguratura fica adiada sine die. Mormentemente porque eu não vou ficar com ele no mesmo palanque.	– I'm not going to put peanuts on the Opposition's vatapa! Either the name of this subversive scoundrel is removed from the inaugural plaque, or the inauguration is postponed sine die. Mainly because I won't be with him on the same platform.
OLHOS DE RESSACA Tudo era matéria às curiosidades de Capitu. [...] – Teimo; hoje mesmo ele há de falar. – Você jura? – Juro! Deixe ver os olhos, Capitu. Tinha-me lembrado a definição que José Dias dera deles, "olhos de cigana oblíqua e dissimilada". Eu não sabia o que era oblíqua, mas dissimulada sabia, e queria ver se se podiam se chamar assim.	HANGOVER EYES Everything was a matter for Capitu's curiosities. [...] – Stubborn; today he will speak. – You swear? – Interest! Let me see the eyes, Capitu. He had reminded me of the definition José Dias had given of them, "the eyes of an oblique and dissimulated gypsy." I didn't know what was oblique, but underhanded I knew, and I wanted to see if they could be called that.

translate works from Portuguese to English and then apply any algorithmic method over them. Still, automatic translation is error-prone, mainly when translating from a language as rich as Portuguese.

For example, Table I shows two snippets from Dias Gomes' *Odorico na Cabeça* and Machado de Assis' *Dom Casmurro*. The character *Odorico Paraguaçu*[1] is famous for creating peculiar words by using wrong suffixes such as "inauguratícia"/"inauguratura" for the actual Portuguese words "inaugural"/"inauguração". The translations to "inaugural"/"inauguration" were correct, but the *wrong* language that defines the character is completely lost. Proper neologisms could be "inauguric" and "inaugurament". Likewise, automatic translations may not work on regionalisms used by great Brazilian authors such as Raquel de Queiroz, Guimarães Rosa and Érico Veríssimo, among others. Even simpler, regular Portuguese statements do not automatically translate well. The second example in the table is a dialogue between *Bentinho* and *Capitu*, in which the former says "Teimo;". A proper translation could be "I insist", in the sense of being stubborn about something. However, it is automatically translated to "Stubborn", which does not make sense within the dialogue. Then Capitu asks "You swear?", to which he surely responds "I do"; and not *interest*, which is the translation to the noun "juro" – not the verb "swear" in the first singular person as written in the original dialogue.

To tackle the aforementioned issues, we present *PPORTAL*: a **P**ublic domain **PORT**uguese-l**A**nguage **L**iterature dataset whose contributions are summarized as follows.

– Data integration of digital libraries for public domain works from Brazil and Portugal: Domínio Público, Projecto Adamastor and Biblioteca Digital de Literatura de Países Lusófonos (BLPL);

– Enriched metadata of the book industry ecosystem: works, authors, readers and online reviews, as extracted from Goodreads;

– Feature engineering on the metadata to create meaningful additional features, including literary genres, popularity information and sentiment analysis scores from online reviews; and

– Access to the data available in three separate versions (Preliminary, Goodreads, and Full) and two formats (SQL dump file and compressed `.csv` files).

This article extends a previous paper from the Dataset Showcase Workshop of Brazilian Symposium on Databases 2021 [Silva et al. 2021a]. Specifically, the related work is updated; we handle missing data by also considering a new data source, the *isbntools* Python library; the dataset considers new features generated by sentiment analysis tools based on online reviews; we also introduce three examples of real-world applications (book genre classification, sentiment analysis on book reviews) and two social network analyses.

[1] *Odorico Paraguaçu* first appeared in the play *Odorico, o Bem-Amado ou Os Mistérios do Amor e da Morte* by the Brazilian playwright Dias Gomes in 1962.

2. RELATED WORK

Digital transformation for book publishing and other creative industries relies on *data* at its core. The huge amount of *digital footprints* left every day on social networks and shopping platforms is perhaps the starting point for facing today's publishing market challenges. However, book publishing is a segmented industry with different publishing categories (academic, literary, comics, etc.), different distributions (physical, digital), and different economic models (self-published, state-funded, privately funded), then making data-based applications and solutions very diverse [Lebrun and Audet 2020]. Nonetheless, extracting, processing and making such digital data available are not trivial tasks, as they require numerous pre-processing steps and advanced methods of data collection.

A common source of books reviews and readers' data is Goodreads,[2] the world's largest site for readers and book recommendations. Due to its high quality, several studies have trusted in its data. For example, Thelwall and Kousha (2017) investigate its user base by comparing users' activity and behavior concerning gender. Still on user behavior, Maity et al. (2019) explore the impact of such factor as an amazon best sellers predictor. Other researchers have focused solely on Goodreads reviews and ratings [Lozano and Planells 2020; Shahsavari et al. 2020]. Yet, few researchers have proposed open, enriched datasets that help to advance research in the book publishing context [Lozano and Planells 2020; Rigau and Tienda 2020; Silva et al. 2021d].

We can divide available book-related datasets into (*i*) books' reviews/ratings [Lozano and Planells 2020; Ni et al. 2019]; (*ii*) books' metadata, with information on editor, price, category and others [Rigau and Tienda 2020]; (*iii*) readers' interactions information, such as spoiler annotations [Wan et al. 2019]; and (*iv*) their combination [Sabri and Weber 2021]. While providing valuable data, each dataset focuses on one or two dimensions of book publishing. Specifically, they typically focus on reader and book reviews data, limiting the potential applications of the data presented, which still requires a more comprehensive, complete dataset to take full advantage of data-driven technologies.

In a different perspective, most datasets are also limited to English-written books [Lozano and Planells 2020; Ni et al. 2019; Rigau and Tienda 2020], which results in a huge research gap for lesser-spoken languages, such as Brazilian Portuguese. Existing public datasets in the Portuguese language are also specific for Natural Language Processing (NLP) applications, then being limited to building a corpus of words extracted from documents [de Araujo et al. 2018; Sousa and Fabro 2019], web content [Wagner Filho et al. 2018], and academic publications [Soares et al. 2018]. Still, Silva et al. use a dataset that comprises cultural, geographic, and socioeconomic information for exploring Brazilian cultural identity through reading preference [Silva et al. 2021c; Silva et al. 2021d].

Seeking to fill the existing research gaps, we introduce *PPORTAL*, a large dataset with information from books in Portuguese that enables studies in the book publishing domain. The dataset considers both Portuguese and Brazilian literature, with works written in European and Brazilian Portuguese languages, henceforth called just Portuguese for simplicity. Besides focusing on the Portuguese-written Literature context, its diverse feature collection can be helpful in different NLP and Machine Learning (ML) applications, as further discussed in Section 4.

3. PPORTAL

We now present *PPORTAL*, a cross-collection dataset with metadata related to public domain Portuguese-language works. First, we describe its building process in Section 3.1. Next, we describe it in quantitative terms in Section 3.2 and through an exploratory data analysis in Section 3.3. Finally, we summarize its format and usage in Section 3.4.

[2] Goodreads: `https://www.goodreads.com/`

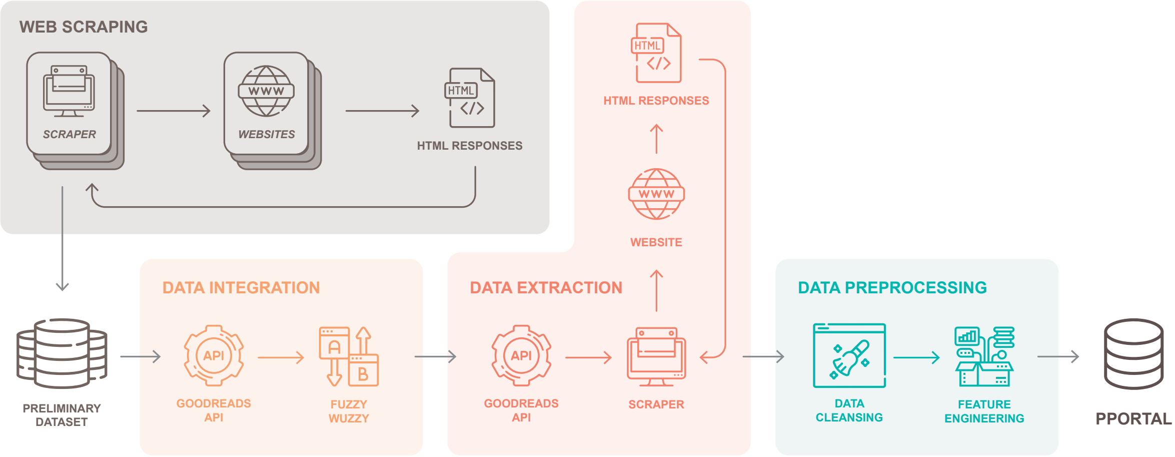


Fig. 1. *PPORTAL* building process graphical summary.

3.1 Dataset Building Process

PPORTAL building process is composed of four main steps: Web Scraping, Data Integration, Data Extraction and Data Preprocessing. Figure 1 shows a schematic diagram of the building process, from web scraping to data preprocessing as explained next.

3.1.1 *Web Scraping.* *PPORTAL* was initially created to support research on NLP and ML tasks over Brazilian and Portuguese literature. Hence, as primary data sources, we consider three well-known digital libraries for public domain works mainly from Brazil and Portugal: Domínio Público,[3] Projecto Adamastor,[4] and Biblioteca Digital de Literatura de Países Lusófonos (BLPL).[5] The first step is coding a web crawler to automatically extract raw data from the platforms (none has API) by using *BeautifulSoup*[6] and *Selenium*,[7] two popular Python libraries for web scraping. We extracted tabular data from the HTML pages using specific web scrapers tailored for each platform's unique structure and formatting between February and August 2021. This step collected download links and metadata from over 80,000 public domain works, as informed in Table II and explained next.

Projecto Adamastor collects more than 1,100 titles in Portuguese from several data sources, including Domínio Público. It provides a Digital Books Database with easy access to public domain works in digital format. All records are presented through a single table, where they can be filtered and/or sorted by some fields. Therefore, we directly extract all records present in this tabular data.

Domínio Público is a digital library maintained by the Brazilian Ministry of Education and includes works duly assigned by the copyright holders. It offers four searchable media types (text, image, sound and video) and several categories and languages for querying. Thus, a pre-filtering is required to extract only texts referring to literature in Portuguese. After the initial filtering, the extraction process is practically the same adopted for the Projecto Adamastor.

BLPL is a large database of Brazilian and Portuguese literature openly available, with more than 80,000 titles. It has an interface based on the alphabet sequence, then requires selecting each letter to advance in the search. Also, as one goal of *PPORTAL* is to mine text from literary works, we create

[3] Domínio Público: `https://www.dominiopublico.gov.br/`
[4] Projecto Adamastor: `https://projectoadamastor.org/`
[5] BLPL: `https://www.literaturabrasileira.ufsc.br`
[6] Beautiful Soup: `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`
[7] Selenium with Python: `https://selenium-python.readthedocs.io/`

Table II. Number of records throughout extraction and integration.

	Domínio Público	Projecto Adamastor	BLPL	Total
Web Scraping Records	2,069	1,036	79,208	**82,313**
Records with Downloads	2,069	1,036	6,480	**9,585**
Integrated Records	1,007	191	1,190	**2,388**

Table III. Collected information present in each digital library.

Data Source	Source	Format	File Size	# Access	Lifetime	Publ. Year	Category	Genre
Domínio Público	✓		✓	✓				
Projecto Adamastor		✓			✓	✓	✓	
BLPL						✓	✓	✓

a binary flag based on the files' availability for download, a distinct feature of *PPORTAL* that allows filtering such documents as well.

3.1.2 *Data Integration.* The preliminary dataset provides information for each data source, such as authors' lifetime (Projecto Adamastor), literary genres (BLPL) and the total number of accesses (Domínio Público). Table III informs the collected metadata from each platform. With such heterogeneous information, we use an additional data source to integrate and centralize the content of the preliminary dataset. We chose Goodreads due to its huge volume of data available and its easy-access API.[8] Through a Python interface API, we searched for all collected works, seeking matches on the Goodreads platform. In particular, records from BLPL were prefiltered to maintain only literary work and with the file available for download, then reducing it from 79,208 to 6,480 records (Table II).

This data integration step also requires a record linkage approach, because each data source has a different book identification system. Such an issue is usually solved through probabilistic or fuzzy matching methods, which apply string similarity functions. Here, we use the Python library *fuzzywuzzy*[9] to map the book records that refer to the same entity in all sources. The library uses Levenshtein Distance to calculate the differences between two strings. With a partial ratio set at 75%, the fuzzy string matching process generates an incomplete result. In total, we were able to map around 25% (i.e., 2,388 records from a total of 9,585) of the initial records collected. Table II presents statistics on the whole process before (*Records with Downloads*) and after integration.

3.1.3 *Data Extraction.* The works' identifiers in the Goodreads integrated dataset enable collecting author information and online reviews. We collected metadata from 966 authors through the same Goodreads API, including name, hometown and fans count. Moreover, we created another web scraper to extract text from each work's first 30 online reviews. Then, this collection ended up with 4,196 reviews from 518 distinct works, plus 1,430 distinct readers.

3.1.4 *Data Preprocessing.* The last two steps of the building process are data cleansing and feature engineering. Although Goodreads remains a valuable source of book information, it is also a source of real-world data. As a result, missing and noisy data are inevitable, which requires cleansing procedures. First, we handled the missing data by dropping irrelevant variables and imputing categorical missing values as an *unknown* category. We also used an additional data source to impute specific missing values by considering the *isbntools*,[10] a Python framework for gathering metadata from ISBN strings. In particular, the *isbntools* framework holds information about books' descriptions, publisher and language, all of which had a considerable rate of missing data in *PPORTAL*. Thus, after data imputation, such a percentage decreased significantly, as shown in Table IV.

[8] Goodreads API: `https://www.goodreads.com/api`
[9] *fuzzywuzzy*: `https://github.com/seatgeek/fuzzywuzzy`
[10] *isbntools*: `https://github.com/xlcnd/isbntools`

Table IV. Improvement ratio (in percentage) after missing data imputation.

Feature	Before	After	Improvement (%)
description	1024	803	21.6%
publisher	1142	1074	5.9%
language_ code	1213	62	94.9%

Table V. Quantitative description of *PPORTAL*.

Data Source	Works	Authors	Genres	Categories	Reviews	Readers
Dominio Publico	2,069	1,767	-	-	-	-
Projecto Adamastor	1,036	390	-	54	-	-
BLPL	79,208	18,289	354	12	-	-
Goodreads	2,388	966	80	-	4,196	1,430

Next, we treat textual data: works description and online reviews. Specifically, we cleaned descriptions and reviews by using regular expressions, removing unnecessary and noisy characters. Also, descriptions were tokenized with Python library *re*.[11] We also parsed and converted structured fields into lists, including *authors*, *popular shelves* and *similar books*. Finally, readers' identification from *GoodreadsReviews* was made anonymous through a hash-based method.

Feature Engineering defined new features from the existing data. In Goodreads, a book can be stored on users' shelves and defined using tags. Following the methodology in [Silva et al. 2021c], we extracted meaningful tags and popularity information (e.g., number of users who labeled the work as a *favorite*, *to-read* and *currently-reading*) from the works' popular shelves. We also created quantitative features related to the total number of authors, popular shelves and similar books; and grouped the work-format categories into *physical*, *digital* and *unknown*. Finally, we used *TextBlob*[12] and *VADER*,[13] popular Python libraries for sentiment analysis, to generate emotional properties from the online reviews of each work (as further discussed in Section 4.1.2).

3.2 Data Content

The storage engine used for *PPORTAL* is a relational database management system (RDBMS), with quantitative information summarized in Table V. Then, Figure 2 depicts its schema with 11 tables divided into three available dataset versions: Preliminary, Goodreads and Full. Such division aims to assist different applications that focus on data at distinct levels of processing. Figure 2 also includes the cardinality of the main tables and versions, briefly described as follows.[14]

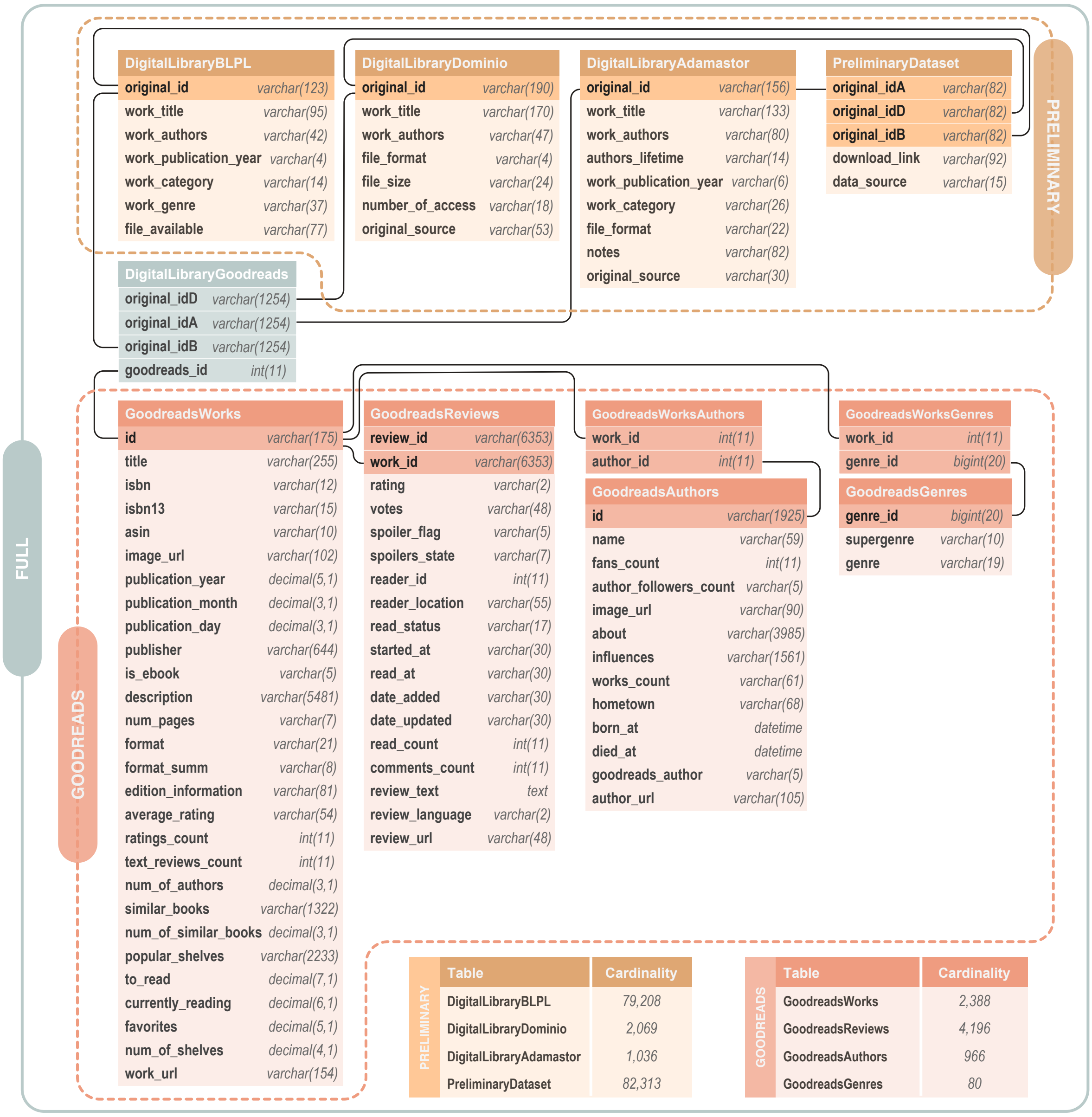
The **Preliminary** version includes four tables referring to the three digital libraries and the preliminary dataset described in Section 3.1. Each digital library presents a set of different features (Table III). Then, *PPORTAL* makes each digital library collection available individually, with *PreliminaryDataset* acting as an auxiliary table that links all records by their ID and includes both source and download link. The **Goodreads** version includes four tables referring to works, authors, online reviews and literary genres. For each of these elements of the book publishing context, there are numerous metadata fields available in Goodreads and additional data generated in the Feature Engineering step (Section 3.1). Furthermore, to represent relationships between such elements, we create two join tables: *WorksAuthors* and *WorksGenres*. The **Full** version combines the first two versions and the *DigitalLibraryGoodreads* table, which stores the data integration result (Section 3.1), making a total of 12 tables.

[11] *re*: `https://docs.python.org/3/library/re.html`

[12] *TextBlob*: `https://textblob.readthedocs.io/`

[13] *VADER*: `https://github.com/cjhutto/vaderSentiment`

[14] Complete descriptions of each table are available in the dataset webpage: `https://bit.ly/PPORTAL`



DigitalLibraryBLPL	
original_id	varchar(123)
work_title	varchar(95)
work_authors	varchar(42)
work_publication_year	varchar(4)
work_category	varchar(14)
work_genre	varchar(37)
file_available	varchar(77)

DigitalLibraryDominio	
original_id	varchar(190)
work_title	varchar(170)
work_authors	varchar(47)
file_format	varchar(4)
file_size	varchar(24)
number_of_access	varchar(18)
original_source	varchar(53)

DigitalLibraryAdamastor	
original_id	varchar(156)
work_title	varchar(133)
work_authors	varchar(80)
authors_lifetime	varchar(14)
work_publication_year	varchar(6)
work_category	varchar(26)
file_format	varchar(22)
notes	varchar(82)
original_source	varchar(30)

PreliminaryDataset	
original_idA	varchar(82)
original_idD	varchar(82)
original_idB	varchar(82)
download_link	varchar(92)
data_source	varchar(15)

DigitalLibraryGoodreads	
original_idD	varchar(1254)
original_idA	varchar(1254)
original_idB	varchar(1254)
goodreads_id	int(11)

GoodreadsWorks	
id	varchar(175)
title	varchar(255)
isbn	varchar(12)
isbn13	varchar(15)
asin	varchar(10)
image_url	varchar(102)
publication_year	decimal(5,1)
publication_month	decimal(3,1)
publication_day	decimal(3,1)
publisher	varchar(644)
is_ebook	varchar(5)
description	varchar(5481)
num_pages	varchar(7)
format	varchar(21)
format_summ	varchar(8)
edition_information	varchar(81)
average_rating	varchar(54)
ratings_count	int(11)
text_reviews_count	int(11)
num_of_authors	decimal(3,1)
similar_books	varchar(1322)
num_of_similar_books	decimal(3,1)
popular_shelves	varchar(2233)
to_read	decimal(7,1)
currently_reading	decimal(6,1)
favorites	decimal(5,1)
num_of_shelves	decimal(4,1)
work_url	varchar(154)

GoodreadsReviews	
review_id	varchar(6353)
work_id	varchar(6353)
rating	varchar(2)
votes	varchar(48)
spoiler_flag	varchar(5)
spoilers_state	varchar(7)
reader_id	int(11)
reader_location	varchar(55)
read_status	varchar(17)
started_at	varchar(30)
read_at	varchar(30)
date_added	varchar(30)
date_updated	varchar(30)
read_count	int(11)
comments_count	int(11)
review_text	text
review_language	varchar(2)
review_url	varchar(48)

GoodreadsWorksAuthors	
work_id	int(11)
author_id	int(11)

GoodreadsAuthors	
id	varchar(1925)
name	varchar(59)
fans_count	int(11)
author_followers_count	varchar(5)
image_url	varchar(90)
about	varchar(3985)
influences	varchar(1561)
works_count	varchar(61)
hometown	varchar(68)
born_at	datetime
died_at	datetime
goodreads_author	varchar(5)
author_url	varchar(105)

GoodreadsWorksGenres	
work_id	int(11)
genre_id	bigint(20)

GoodreadsGenres	
genre_id	bigint(20)
supergenre	varchar(10)
genre	varchar(19)

	Table	Cardinality
PRELIMINARY	DigitalLibraryBLPL	79,208
	DigitalLibraryDominio	2,069
	DigitalLibraryAdamastor	1,036
	PreliminaryDataset	82,313

	Table	Cardinality
GOODREADS	GoodreadsWorks	2,388
	GoodreadsReviews	4,196
	GoodreadsAuthors	966
	GoodreadsGenres	80

Fig. 2. Schema and cardinality for *PPORTAL* divided by versions available.

3.3 Exploratory Data Analysis

We now present an exploratory data analysis over *PPORTAL* and summarize its main characteristics. We start by analyzing the missing values that were not handled in the Data Cleansing step. Mainly, only tables *GoodreadsWorks*, *GoodreadsAuthors* and *GoodreadsReviews* have missing data, and Figure 3 shows their percentage (a–c) and distribution across all variables (d–f). Most missing data refer to dates, which is a complex variable to handle when missing. Moreover, some identification information related to works also has incomplete records, such as ISBN/ISBN-13 and ASIN[15] codes. Overall, the nullity matrix (Figure 3 d–f) shows a correlation between most variables (i.e., if an observation is

[15] ASIN stands for Amazon Standard Identification Number.

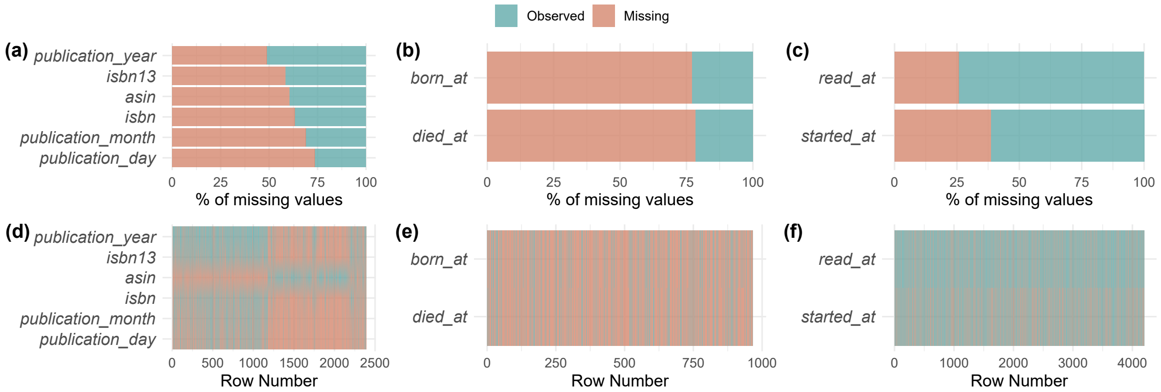


Fig. 3. Missing values of *GoodreadsWorks*, *GoodreadsAuthors* and *GoodreadsReviews* tables, respectively. (a) Percentage of the missing values and (b) the distribution of data across all variables.

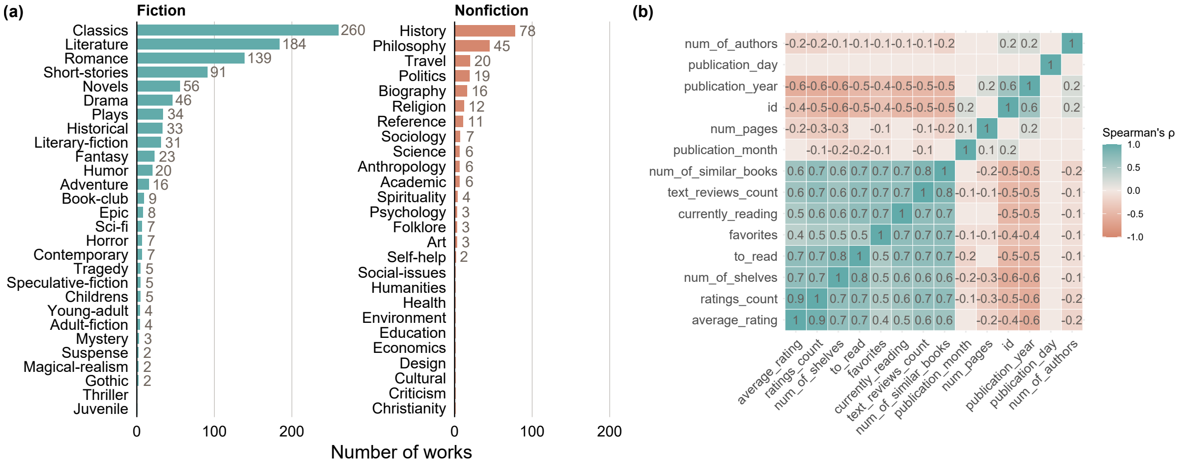


Fig. 4. (a) Breakdown of collected books into genre categories of Fiction and Nonfiction. (b) Spearman's rank correlation matrix of the numeric variables from *GoodreadsWorks* table. Correlations with $p - value \geq 0.05$ are considered as insignificant and are left blank.

missing in a variable, it is also certainly missing in the other one), except for the ASIN code.

With more than two thousand distinct works, the dataset includes 80 different genres classified into *Fiction* or *Nonfiction* categories. Figure 4(a) breaks down all the collected works by category, then genre. Note that about 70% (54) of the available genres are present in our works' collection. Moreover, most of them (24%) fall into *Classics*, *Literature* and *Romance*, all fiction genres. Indeed, about 90% of the works' genres in the dataset are categorized as fiction, whereas the remaining 10% of the Nonfiction genres fall mainly into *History*, *Philosophy*, *Travel* and *Politics* categories.

Regarding *GoodreadsWorks* table, Figure 4(b) presents the Spearman's rank correlation matrix of the numeric variables. There is a clear positive correlation between most of the works' popularity measures, including *average_ rating*, *ratings_ count*, *text_ reviews_ count*, among others. In contrast, such measures are negatively correlated to the works' publication year, indicating a possible preference for classic works. Overall, there are very few perfectly positive or negative numeric attributes, reducing the chances of multicollinearity in future machine learning models using *PPORTAL*.

Finally, Figure 5 displays the distribution of main authors' and online reviews' characteristics. Fig-

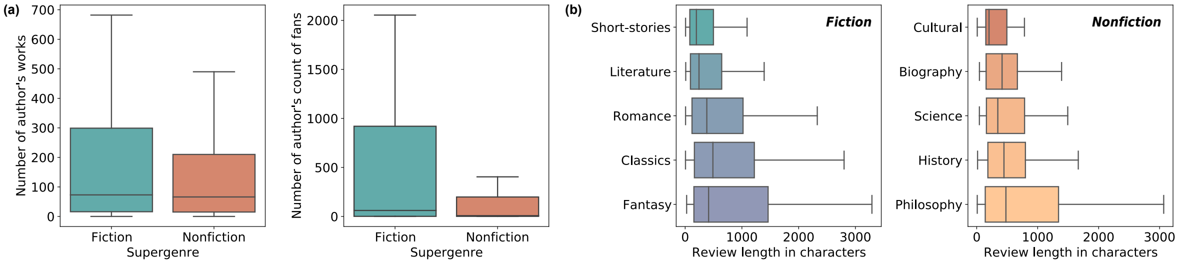


Fig. 5. (a) Distributions of the number of works and fans for Fiction and Nonfiction authors. (b) Distribution of reviews' lenght for the top-5 Fiction and Nonfiction genres, respectively.

ure 5(a) shows a difference between the average number of works by Fiction and Nonfiction authors. Despite the similar number of works, Fiction authors have a more extensive fan base compared to Nonfiction ones. Furthermore, Figure 5(b) shows reviews are lengthier for genres *Fantasy* and *Philosophy*, although they are not very popular within *PPORTAL*. Still, the main genres for Fiction and Nonfiction are *Classics* and *History*, respectively, which also present long reviews on average.

3.4 Format and Usage

PPORTAL dataset is publicly available in an open-access Zenodo repository [Silva et al. 2021b] and can also be downloaded from its project webpage.[14] As aforementioned, all collected and enriched data are available in three separate versions (Preliminary, Goodreads and Full). Hence, we generate a dump file for each version that contains the database structure and content, which can then be imported into any MySQL server. As the dataset is structured in tabular format, we also make all three versions available in `.csv` format, which enables easy process by notebooks, for example.

4. *PPORTAL* APPLICATIONS

PPORTAL can be used to assess different artificial intelligence tasks, feeding a variety of machine learning and natural language processing models. Moreover, its entities can be explored in social network analyses as well. This section shares applications and possible scenarios within such contexts, illustrating the breadth and potential impact of the data available in *PPORTAL*.

4.1 Natural Language Processing

NLP is an essential, valuable branch of Computer Science, allowing machines to understand human language. It spans multiple applications, including automated text classification, entity recognition and sentiment analysis, mostly working over English. Although they can be trained for Portuguese, they still need to be significantly improved to get the nuances and peculiarities of Portuguese. Hence, there is a well-justified necessity for creating tools that operate in Portuguese.

4.1.1 *Text Classification.* It involves automatically understanding, processing and categorizing unstructured text; i.e., assigning a document into predefined categories. Current work usually employ a machine learning approach: a classifier model is built to learn the categories' features from a set of pre-classified documents [Graovac et al. 2015; Sebastiani 2002]. Regardless of methodology, text classifiers automatically structure all types of text in a fast and cost-effective way, saving time, automating business processes and making data-driven business decisions. Next, we show *PPORTAL* resource power in a real-world application of automatic book genre classification.

Genre classification is a relevant task for the publishing industry, as readers can use genre to decide what to read next and editors to choose books to be published and guide top list strategies. Also,

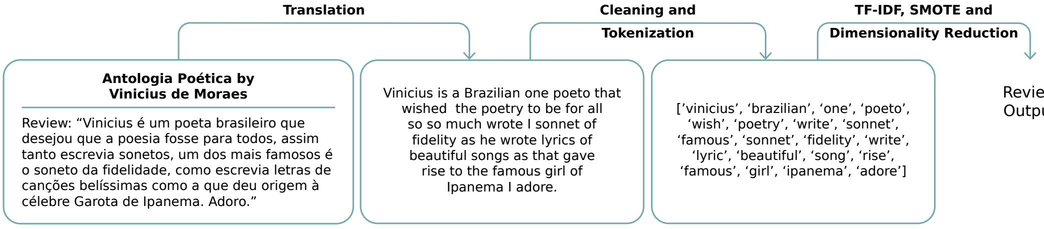

Translation
Cleaning and Tokenization
TF-IDF, SMOTE and Dimensionality Reduction
Antologia Poética by Vinicius de Moraes
Review: "Vinicius é um poeta brasileiro que desejou que a poesia fosse para todos, assim tanto escrevia sonetos, um dos mais famosos é o soneto da fidelidade, como escrevia letras de canções belíssimas como a que deu origem à célebre Garota de Ipanema. Adoro."
Vinicius is a Brazilian one poeto that wished the poetry to be for all so so much wrote I sonnet of fidelity as he wrote lyrics of beautiful songs as that gave rise to the famous girl of Ipanema I adore.
['vinicius', 'brazilian', 'one', 'poeto', 'wish', 'poetry', 'write', 'sonnet', 'famous', 'sonnet', 'fidelity', 'write', 'lyric', 'beautiful', 'song', 'rise', 'famous', 'girl', 'ipanema', 'adore']
Review Output

Fig. 6. Example of review preprocessing from the book "Antologia Poética" (*Poetic Anthology*) by Vinicius de Moraes.

Table VI. Experimental results from classifier algorithms, sorted by F1.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	**0.85**	**0.89**	0.79	**0.81**
K Nearest Neighbor	0.82	0.88	0.79	**0.81**
Gaussian Naive Bayes	0.80	0.81	**0.81**	0.80
Stochastic Gradient Descent	0.81	0.82	0.76	0.78
Decision Tree	0.76	0.76	0.77	0.76
Naïve (baseline)	0.21	0.02	0.08	0.03

book genre analysis is an appealing study for marketing and sales, as it serves as an indicator of interest. Thus, we developed a real-world application using *PPORTAL* to train an automatic book genre classifier based on online reviews. The methodology is composed of: preprocessing, training book genre classifiers, and evaluating results with different metrics.

Preprocessing takes raw reviews and prepares them for the classifiers, as exemplified in Figure 6. It starts by translating all reviews to English using the Google Translator API, unifying the language used and facilitating the data cleaning process. Next, clean-up methods remove symbols, URLs, emojis, stopwords, and any other noise interfering with the classification result. Then, we create a custom dictionary of single words with tokenization (and normalization). Feature representation uses TF-IDF (Term frequency-inverse document frequency) to compute the weight for each term, based on the importance of a term compared to the whole text. As *PPORTAL* is unbalanced – Figure 4(a), we use the SMOTE (Synthetic Minority Oversampling TEchnique) [Bowyer et al. 2011] method to perform an oversampling, such that the number of examples in the minority class better resembles or matches the number of examples in the majority classes. In particular, SMOTE works by selecting close examples in the feature space for a minority class instance and using the k-nearest neighbor algorithm to synthesize new examples from the minority class. Next, we use Latent Semantic Analysis (LSA), typically a dimensionality reduction method, to reduce the number of redundant words.

For **genre classification**, an instance must be placed within one class among all possible ones. Therefore, we chose the multiclass classification approach. Reviews are divided into two sets: 70% of the dataset is used to *train* the model and evaluate the training score, and 30% is used to *test* the model's accuracy (test score). We consider the following classifiers: Naïve Classifier (as a simple baseline), Decision Tree Classifier, Random Forest, Stochastic Gradient Descent, GaussianNaive Bayes and K Nearest Neighbor. All algorithms used default parameters provided by the scikit-learn library.[16] After the model is trained, it classifies genre on the test set and can be evaluated by different metrics.

For **evaluating the performance** of all classifiers, we consider four well-known evaluation metrics: Accuracy, Precision, Recall and F1-score, with results in Table VI. Overall, the results are very good. Random Forest (RF) yields the best performance, except for Recall, reaching 85% accuracy, followed by the K-Nearest Neighbor. In the upcoming analyses, we consider only the best model –RF classifier.

Figure 7(a) shows a confusion matrix to analyze the discrepancies between predicted and true labels.

[16] Scikit Learn: `https://scikit-learn.org/`

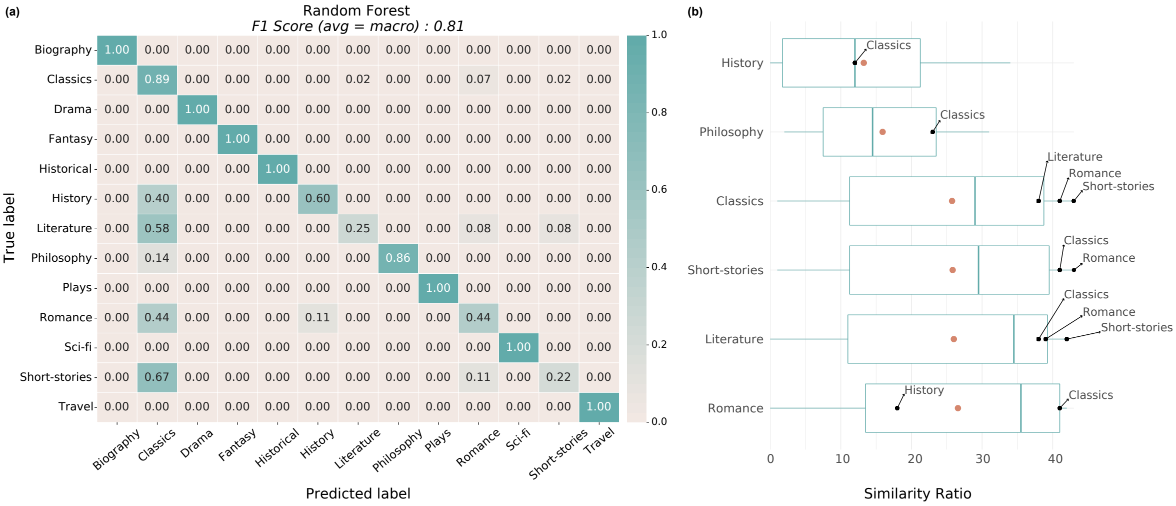


Fig. 7. (a) Normalized confusion matrix of Random Forest classifier results on test set. (b) Similarity ratio distribution for each missclassification, where the erroneously predicted classes are highlighted (red dots represent the mean values).

Although most of the predictions ended up on the diagonal (predicted label = actual label), there are some misclassifications, with *Short-stories* and *Literature* at the top of the misclassification list. As we oversampled the minority classes to balance *PPORTAL* data, such misclassifications should not be related to the lack of samples. Therefore, we also calculate the overall similarity of the reviews for the paired genres. To do so, we applied the same string similarity method used in the Data Integration step (Section 3.1.2). Figure 7(b) shows the similarity ratio distribution for each misclassification, where the erroneously predicted classes are highlighted.

Overall, the similarity ratio between misclassified and predicted classes is high (above average), indicating a possible cause for misclassifications: *Classics* book reviews being similar to *Literature*, *Romance* and *Short-stories* book reviews may have confused the classification model. However, among the 12 existing cases, there are two counterexamples: $\{$*History* and *Classics*$\}$ and $\{$*Romance* and *History*$\}$. For both pairs, the similarity ratio was not very high, especially in the latter, indicating that unknown factors affect the model's performance in both cases. As *History* is a representative class in the dataset, we may rule out class imbalance as a cause; hence, a more robust analysis is needed to understand such unexpected results, which is left as future work.

4.1.2 *Sentiment Analysis.* It is a NLP technique for investigating data opinions, sentiments and emotions, often performed on textual data. Sentiment analysis remains one of the most challenging tasks in NLP since even humans struggle to accurately analyze sentiments [Yadollahi et al. 2017]. However, there are many efforts to improve and advance the state-of-the-art in different contexts [Alves et al. 2016; Harb et al. 2019; Matsuno et al. 2017], even for literature [Maharjan et al. 2018]. As the aforementioned applications, the text of public domain works can be extracted and, consequently, used to feed NLP models. Moreover, table *GoodreadsReviews* may be used to identify and extract subjective information from works' online reviews.

As a practical example, we apply two sentiment analysis libraries (TextBlob and VADER) to online reviews of works available on *PPORTAL*. Specifically, the TextBlob sentiment analyzer returns two properties for a given input sentence: (i) **Polarity**, a float number between $[-1, 1]$, where -1 indicates negative sentiment and $+1$ indicates positive sentiment; and (ii) **Subjectivity**, a float number in the range of $[0, 1]$, which generally refers to opinion, emotion or judgment. VADER (Valence Aware Dictionary and sEntiment Reasoner) is another popular rule-based sentiment tool, which uses a list of lexical features (e.g., words) that are labeled as positive or negative according to their semantic

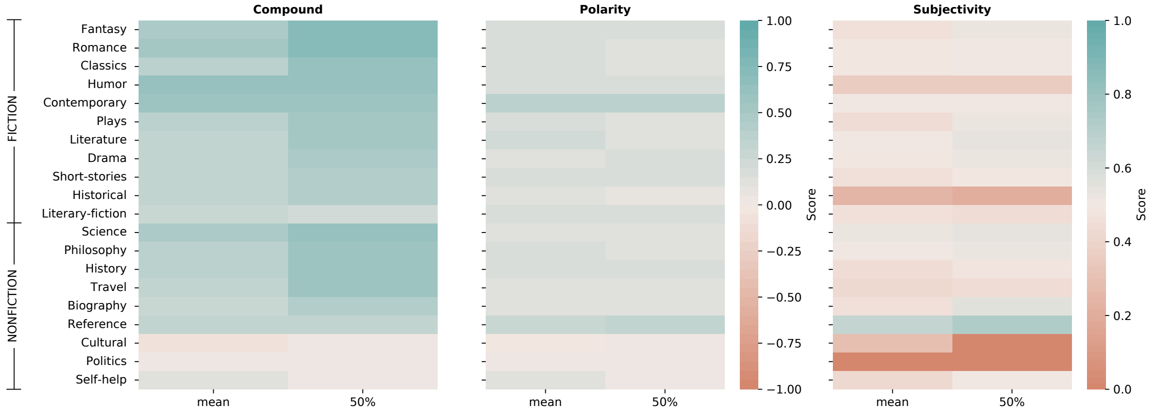


Fig. 8. Sentiment analysis summary on online reviews, categorized by genre and supergenre.

orientation to calculate the text sentiment. VADER sentiment returns the probability of a given input sentence to be positive, negative and neutral.

For sentiment analysis, we first apply the translation process from Section 4.1.1. Then, we run the tools to get each review's sentiment, polarity and subjectivity scores. We group the online reviews into the works' literary genres for a more interesting analysis. Figure 8 shows the mean and median values of Compound (VADER), Polarity and Subjectivity scores. The Compound score is the sum of all lexicon ratings and ranges from -1 (extreme negative) to $+1$ (extreme positive). Overall, all Fiction genres presented positive reviews, mainly *Romance* and *Fantasy*. For Nonfiction, although the majority also received positive reviews on average, genres such as *Cultural*, *Politics* and *Self-help* had more impartial reviews. TextBlob scores show similar results, indicating slightly positive and somewhat subjective reviews, except for *Politics* and *Cultural* genres.

4.1.3 *Named Entity Recognition (NER).* It is an NLP technique that automatically identifies entities in a text and classifies them into predefined categories. Solutions are based on ML methods, using statistical models that need training on a large corpus (labeled data) to achieve good performance. Unfortunately, such datasets are scarce due to costly and time-consuming generation. This reality is worse for Portuguese, with few existing annotated corpora [de Araujo et al. 2018; Soares et al. 2018; Wagner Filho et al. 2018]. Hence, generating benchmark datasets for NER is still an open issue, which may be assisted by *PPORTAL* digital documents plus those available at the download links.

4.2 Social Network Analysis

Social network analysis (SNA) investigates and characterizes social structures using networks and graph theory [Procópio Jr. et al. 2012; Souza et al. 2020]. In the context of book publishing, *PPORTAL* can be used to explore interactions between readers present in the *Reviews* table, the works' professionals (authors, editors, illustrators, translators, etc.), similar works, among others. By building such networks, many SNA studies can be performed, including community detection to identify communities of readers/authors/works; social network-based recommendation (e.g., making personalized recommendations from reader preference information), and user-behavior analysis (e.g., performing a cross-location analysis based on reading preferences as done by our research group in [Silva et al. 2021c]). This section follows with two examples of community detection, which is essential to understanding the structure of complex networks by identifying and extracting groups with similar properties in different contexts, for example.

Here, we first identify communities of book professionals by building an *Interaction network*, where

Table VII. Statistics for the networks of book interactions and similar books (*Avg* is average, *W* is weighted)

Network	**Nodes**	**Edges**	**Avg.Degree**	**Avg.W.Degree**	**Density**	**Diameter**	**Modularity**
Interaction	966	742	1.536	1.588	0.002	6	0.840
Similarity	7,273	70,653	19.429	23.603	0.003	7	0.678

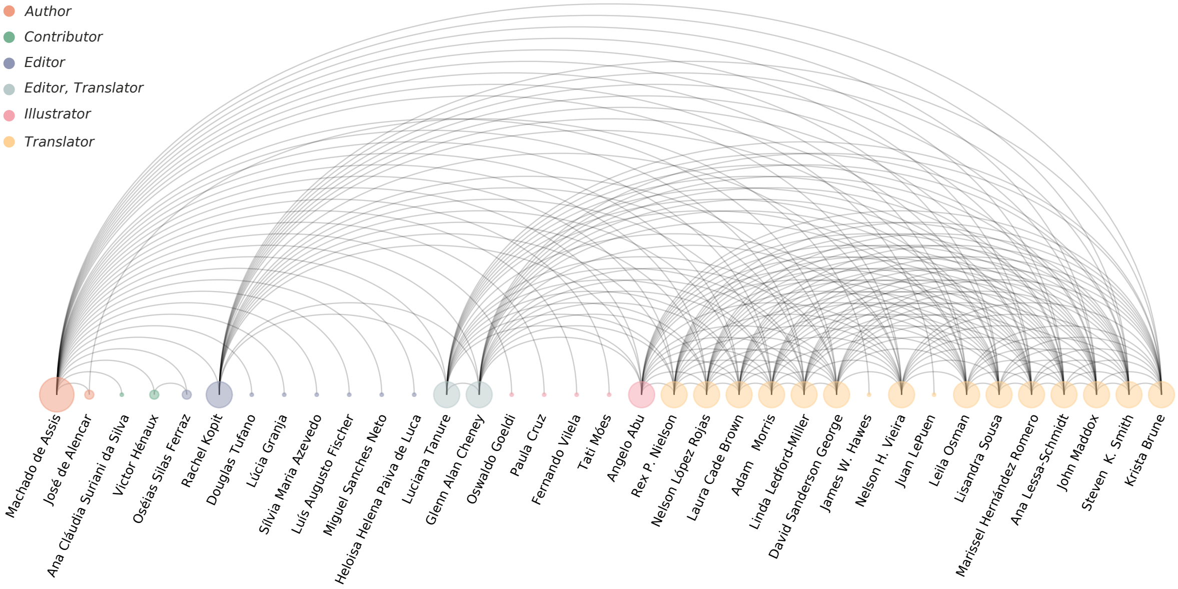


Fig. 9. Largest sub-community in the *Interaction* network, resulting from the Louvain community detection algorithm. Colored nodes inform the role of each book professional. The size of nodes is relative to their degree.

nodes represent people and there are edges between those who have any relationship in making a book, such as *author-illustrator* and *editor-author*. Second, we further explore the *GoodreadsWorks* table and its list of similar books to build another network, called *Similarity*. Table VII presents basic statistics on nodes and edges, as well as structural metrics for both networks. The networks present high modularity, meaning they have dense connections between the nodes within modules but sparse connections between nodes in different modules.

There are different algorithms for network community detection. Here, we apply the Louvain method [Blondel et al. 2008] in the networks, as it is a popular algorithm. Louvain is a heuristic method based on modularity, which maximizes a modularity score for each community. The method returned 715 different communities for the *Interaction* network, where only 13% (91) of these communities have more than one node. The largest community detected accounts for 4.35% of the original network. According to the betweenness centrality score, the community comprises 42 nodes and 199 edges in total, representing the most influential (on average) people of the original network. There are two subcommunities within such a major group, separated by Machado de Assis, who serves as a bridge between them. Figure 9 shows the largest sub-community, where Machado de Assis has the highest betweenness centrality among all nodes, considerably influencing the *Interaction* network's flow. He is also the *subject* of many books, such as "Ex Cathedra: Stories by Machado de Assis" translated by Glenn Alan Cheney and other translators, and his books have school-oriented versions with pedagogical guidance and reading notes written by other professionals, such as Douglas Tufano.

Regarding the book similarity network, Louvain's algorithm detected 1,888 communities, where only 2.6% (50) have more than one node. The largest and most influential community detected accounts for 9.03% of the original network, with 657 nodes and 8,507 edges in total. As the lists of similar books present in the *GoodreadsWorks* table may have books out of *PPORTAL* (i.e., which were not available

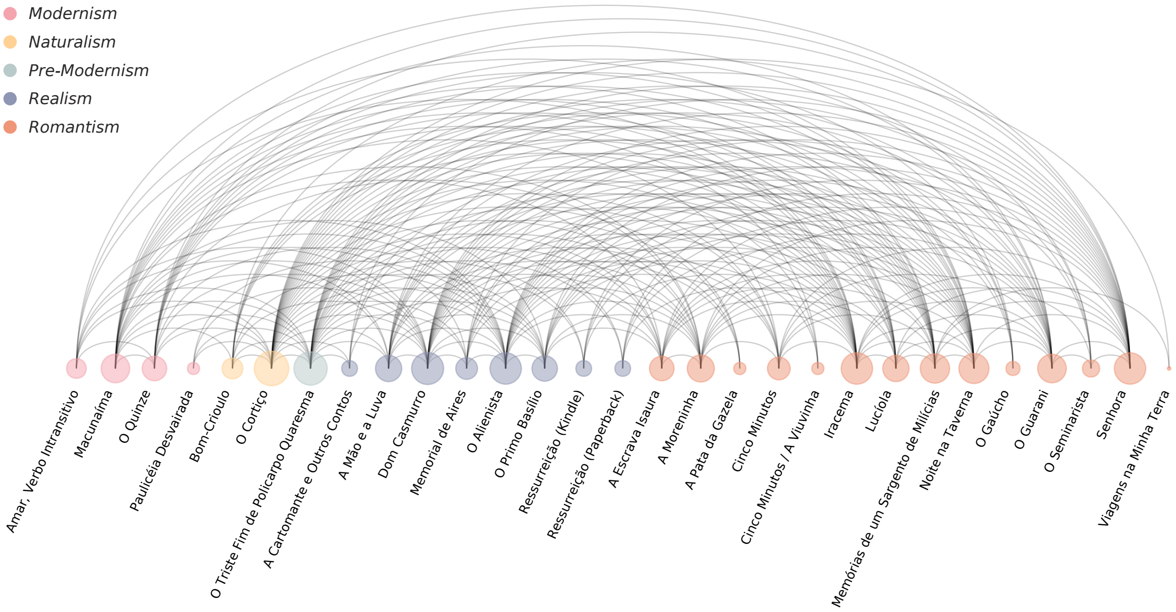


Fig. 10. Largest community in the book *Similarity* network, resulting from the Louvain community detection algorithm. Colored nodes inform the Literary Movement or Period of each work. The size of nodes is relative to their degree.

in any of the digital libraries considered in the collection), their nodes are set to "unknown". Hence, to improve the visualization, such "unknown" nodes are filtered out; i.e., we keep only those books that are within *PPORTAL*. Figure 10 shows the final community with colored nodes representing the Literary Movement or Period of each book. Note that such a community is composed of works with the highest number of online reviews and ratings on average, which may indicate a high preference for works that are mainly part of the Romanticism and Realism literary movements.

4.3 Other Scenarios

AI-powered technology and data-oriented applications (such as many within Data Science) have accelerated over recent years in creative industries, such as music, cinema and literature. With so much digital information available, ML-based solutions have been developed to predict success and recommend items in such industries as well. Both applications may directly use *PPORTAL*, and we refer to the original work for more discussion on the subject [Silva et al. 2021a].

5. CONCLUDING REMARKS

Although digital libraries are excellent sources of literary data, they are frequently limited to language-agnostic science work or English written works when language matters. Especially for Literature, most datasets are composed of English-only works. To tackle such challenges, we introduce *PPORTAL*, an open dataset with metadata related to public domain Portuguese-language literature. Initially, we built a cross-collection preliminary dataset by integrating public domain works from three digital libraries, comprising download links and valuable metadata. Next, we used the Goodreads API to collect additional information from essential elements of the book industry ecosystem: works, authors, readers and reviews.

In summary, we believe that *PPORTAL*'s centralized collection is a valuable resource for Natural Language Processing tasks, including (but not limited to) named entity recognition, text classification, and sentiment analysis. To illustrate the latter two, we also presented applications for book genre classification and book reviews-based sentiment analysis. We also built two complex networks based on book professionals' interaction and book similarity. Thus, we expect *PPORTAL* to be suitable for other Machine Learning applications, such as book recommendation and success prediction models.

PPORTAL is mainly composed of data extracted from the web and social media (i.e., Goodreads). As a result, it still has challenges and limitations. Data integration is one primary problem because joining data from different sources requires a record linkage method. We applied a fuzzy matching approach, where record pairs with probabilities above a certain threshold were considered the same entity. However, such a method is subject to misspellings and formatting errors. In such a case, there is no other solution to works not in Goodreads other than keeping them only in the preliminary set or manually searching on the website for those incorrectly mapped.

Regarding data quality, most content on Goodreads is added by its users, therefore subject to imprecision and lack of information. As further discussed in the article (Section 3.3), some dataset variables have considerable portions of missing values, which could be improved by having an additional data source to impute the incomplete content. Despite the missing values, such integration provides valuable information regarding not only the literary genre of the works, but also regarding success/popularity metrics, online reviews, and additional information about the authors. Finally, another problem resulting from data integration is the genre distinction among data sources, where only two have literary genres (Projecto Adamastor and BLPL). A valid solution is to consider fuzzy matching approaches to finding similar genres.

As future work, we plan to consider more data sources for handling missing data and apply fuzzy matching methods to alleviate the issue of the distinct genre. In particular, we are currently exploring the *isbntools* library to handle other missing data, such as ISBN codes. Given the continued growth of data, we also plan to implement an update-oriented collecting phase. Finally, we are also working on integrating the socioeconomic and cultural information from [Silva et al. 2021d] with *PPORTAL*.

REFERENCES

Alves, A. L. F., Baptista, C. d. S., Firmino, A. A., de Oliveira, M. G., and de Paiva, A. C. (2016). A spatial and temporal sentiment analysis approach applied to twitter microtexts. *Journal of Information and Data Management*, 6(2):118. doi:10.5753/jidm.2015.1563.

Bao, H., He, K., Yin, X., Li, X., Bao, X., Zhang, H., Wu, J., and Gao, Z. (2021). Bert-based meta-learning approach with looking back for sentiment analysis of literary book reviews. In Wang, L., Feng, Y., Hong, Y., and He, R., editors, *Natural Language Processing and Chinese Computing - 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13-17, 2021, Proceedings, Part II*, volume 13029 of *Lecture Notes in Computer Science*, pages 235–247. Springer. doi:10.1007/978-3-030-88483-3_18.

Blondel, V. D. et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008. doi:10.1088/1742-5468/2008/10/p10008.

Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813. URL `http://arxiv.org/abs/1106.1813`.

Champagne, A. (2020). What Is A Reader? How Readers on Goodreads are Changing the Canon in the Twenty-First Century. In *Annual Int. Conf. of the Alliance of Digital Humanities Organizations, Conference Abstracts*.

de Araujo, P. H. L., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). Lener-br: a dataset for named entity recognition in brazilian legal text. In *Int'l Conf. on Computational Processing of the Portuguese Language*, pages 313–323. Springer.

Graovac, J., Kovačević, J., and Pavlović-Lažetić, G. (2015). Language independent n-gram-based text categorization with weighting factors: A case study. *Journal of Information and Data Management*, 6(1):4. doi:10.5753/jidm.2015.1552.

Harb, J. G. D., Ebeling, R., and Becker, K. (2019). Exploring deep learning for the analysis of emotional reactions to terrorist events on twitter. *Journal of Information and Data Management*, 10(2):97–115. doi:10.5753/jidm.2019.2039.

Lebrun, T. and Audet, R. (2020). Artificial Intelligence and the Book Industry. White Paper. *Zenodo*. doi:10.5281/zenodo.4036258.

Lozano, L. C. and Planells, S. C. (2020). Best books ever dataset. *Zenodo*. doi:10.5281/zenodo.4265096.

Maharjan, S., Kar, S., Montes, M., González, F. A., and Solorio, T. (2018). Letting emotions flow: Success prediction by modeling the flow of emotions in books. In *Procs. Conf. of the North American Chapter of the Asso-*

ciation for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 259–265. doi:10.18653/v1/N18-2042.

Maity, S. K., Panigrahi, A., and Mukherjee, A. (2019). Analyzing social book reading behavior on goodreads and how it predicts amazon best sellers. In *Influence and Behavior Analysis in Social Networks and Social Media*, pages 211–235. Springer, Cham.

Matsuno, I. P., Rossi, R. G., Marcacini, R. M., and Rezende, S. O. (2017). Aspect-based sentiment analysis using semi-supervised learning in bipartite heterogeneous networks. *Journal of Information and Data Management*, 7(2):141. doi:10.5753/jidm.2016.1584.

Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Procs. Conf. on Empirical Methods in Natural Language Processing and Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Procópio Jr., P. S., Gonçalves, M. A., Laender, A. H. F., Salles, T., and Figueiredo, D. (2012). Time-aware ranking in sport social networks. *Journal of Information and Data Management*, 3(3):195. doi:10.5753/jidm.2012.1448.

Rigau, P. and Tienda, A. (2020). 100 bestselller books during covid-19 in spain. *Zenodo.* doi:10.5281/zenodo.3820050.

Sabri, N. and Weber, I. (2021). A global book reading dataset. *Data*, 6(8):83. doi:10.3390/data6080083.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47. doi:10.1145/505282.505283.

Shahsavari, S., Ebrahimzadeh, E., Shahbazi, B., Falahi, M., Holur, P., Bandari, R., R. Tangherlini, T., and Roychowdhury, V. (2020). An automated pipeline for character and relationship extraction from readers literary book reviews on goodreads.com. In *12th ACM Conference on Web Science*, WebSci '20, page 277–286, New York, NY, USA. Association for Computing Machinery. doi:10.1145/3394231.3397918.

Sharma, A., Liu, H., and Liu, H. (2020). Best seller rank (bsr) to sales: An empirical look at amazon.com. In *2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 609–615. doi:10.1109/QRS-C51114.2020.00104.

Silva, M. O., Scofield, C., and Moro, M. M. (2021a). PPORTAL: Public Domain Portuguese-language Literature Dataset. In *Anais do III Dataset Showcase Workshop*, pages 77–88, Porto Alegre, RS, Brasil. SBC. doi:10.5753/dsw.2021.17416.

Silva, M. O., Scofield, C., and Moro, M. M. (2021b). PPORTAL: Public domain Portuguese-language literature Dataset. *Zenodo.* doi:10.5281/zenodo.5178063.

Silva, M. O., Scofield, C., Oliveira, G. P., Seufitelli, D., and Moro, M. M. (2021c). Exploring Brazilian Cultural Identity Through Reading Preferences. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 115–126. SBC. doi:10.5753/brasnam.2021.16130.

Silva, M. O., Scofield, C., Oliveira, G. P., Seufitelli, D. B., and Moro, M. M. (2021d). BraCID: Brazilian Cultural Identity Information Through Reading Preferences. *Zenodo.* doi:10.5281/zenodo.4890048.

Soares, F., Yamashita, G. H., and Anzanello, M. J. (2018). A parallel corpus of theses and dissertations abstracts. In *International Conference on Computational Processing of the Portuguese Language*, pages 345–352. Springer.

Sousa, A. W. and Fabro, M. D. D. (2019). Iudicium textum dataset uma base de textos jurídicos para nlp. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2019 Companion.* SBC.

Souza, V., Nobre, J., and Becker, K. (2020). Characterization of anxiety, depression, and their comorbidity from texts of social networks. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 121–132, Porto Alegre, RS, Brasil. SBC. doi:10.5753/sbbd.2020.13630.

Thelwall, M. and Kousha, K. (2017). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, 68(4):972–983.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Wan, M., Misra, R., Nakashole, N., and McAuley, J. J. (2019). Fine-grained spoiler detection from large-scale review corpora. In *Procs. Conf. of the Association for Computational Linguistics (ACL)*, pages 2605–2610. doi:10.18653/v1/p19-1248.

Wang, X., Yucesoy, B., Varol, O., Eliassi-Rad, T., and Barabasi, A.-L. (2019). Success in books: predicting book sales before publication. *EPJ Data Science*, 8(31). doi:10.1140/epjds/s13688-019-0208-6.

Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput. Surv.*, 50(2). doi:10.1145/3057270.