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.

Copyright © 2022 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

Journal of Information and Data Management, Vol. 13, No. 1, August 2022, Pages 95–110.
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.

<table>
<thead>
<tr>
<th>Original text in Portuguese</th>
<th>Automatically translated English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Não vou botar amendoim no vatapá da Oposição! Ou tiram da placa inauguratícia o nome desse patifista subversento, ou a inauguratícia fica adiada sine die. Momentemente porque eu não vou ficar com ele no mesmo palanque.</td>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OLHOS DE RESSACA</th>
<th>HANGOVER EYES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tudo era matéria às curiosidades de Capitu. [...]</td>
<td>Everything was a matter for Capitu’s curiosities. [...]</td>
</tr>
<tr>
<td>Teimo; hoje mesmo ele há de falar.</td>
<td>Stubborn; today he will speak.</td>
</tr>
<tr>
<td>Você jura?</td>
<td>You swear?</td>
</tr>
<tr>
<td>Juro! Deixe ver os olhos, Capitu.</td>
<td>Interest!* Let me see the eyes, Capitu.</td>
</tr>
<tr>
<td>Tinham-me lembrado a definição que José Dias dera deles, “olhos de cigana oblíqua e dissimulada”. Eu não sabia o que era oblíqua, mas dissimulada sabia, e queria ver se se podiam se chamar assim.</td>
<td>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 <em>underhanded</em> I knew, and I wanted to see if they could be called that.</td>
</tr>
</tbody>
</table>

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 Paraguaçu1 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 “inaurigic” 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 PPORITAL: a Public domain PORTuguese-_LANGUAGE Literature 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, 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.

---

[Goodreads: https://www.goodreads.com/]
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, Projecto Adamastor, and Biblioteca Digital de Literatura de Países Lusófonos (BLPL). The first step is coding a web crawler to automatically extract raw data from the platforms (none has API) by using BeautifulSoup and Selenium, 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

---

3Domínio Público: https://www.dominiopublico.gov.br/
4Projecto Adamastor: https://projectoadamastor.org/
5BLPL: https://www.literaturabrasileira.ufsc.br
6BeautifulSoup: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
7Selenium with Python: https://selenium-python.readthedocs.io/
Table II. Number of records throughout extraction and integration.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Domínio Público</th>
<th>Projecto Adamastor</th>
<th>BLPL</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Scraping Records</td>
<td>2,069</td>
<td>1,036</td>
<td>79,208</td>
<td>82,313</td>
</tr>
<tr>
<td>Records with Downloads</td>
<td>2,069</td>
<td>1,036</td>
<td></td>
<td>9,585</td>
</tr>
<tr>
<td>Integrated Records</td>
<td>1,007</td>
<td>191</td>
<td>1,190</td>
<td>2,388</td>
</tr>
</tbody>
</table>

Table III. Collected information present in each digital library.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Source Format</th>
<th>File Size</th>
<th># Access</th>
<th>Lifetime</th>
<th>Publ. Year</th>
<th>Category</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domínio Público</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td></td>
</tr>
<tr>
<td>Projecto Adamastor</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td>✔️ ✔️ ✔️</td>
<td></td>
</tr>
<tr>
<td>BLPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

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 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, 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.

---

8Goodreads API: https://www.goodreads.com/api
9fuzzywuzzy: https://github.com/seatgeek/fuzzywuzzy
10isbn tools: https://github.com/slcad/isbntools
Table IV. Improvement ratio (in percentage) after missing data imputation.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Before</th>
<th>After</th>
<th>Improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>description</td>
<td>1024</td>
<td>803</td>
<td>21.6%</td>
</tr>
<tr>
<td>publisher</td>
<td>1142</td>
<td>1074</td>
<td>5.9%</td>
</tr>
<tr>
<td>language_code</td>
<td>1213</td>
<td>62</td>
<td>94.9%</td>
</tr>
</tbody>
</table>

Table V. Quantitative description of PPORTAL.

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Works</th>
<th>Authors</th>
<th>Genres</th>
<th>Categories</th>
<th>Reviews</th>
<th>Readers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominio Publico</td>
<td>2,069</td>
<td>1,767</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Proyecto Adamastor</td>
<td>1,036</td>
<td>390</td>
<td>-</td>
<td>54</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BLPL</td>
<td>79,208</td>
<td>18,289</td>
<td>354</td>
<td>12</td>
<td>4,196</td>
<td>1,430</td>
</tr>
<tr>
<td>Goodreads</td>
<td>2,388</td>
<td>966</td>
<td>80</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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
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\textsuperscript{15} codes. Overall, the nullity matrix (Figure 3 d–f) shows a correlation between most variables (i.e., if an observation is

\textsuperscript{15}ASIN stands for Amazon Standard Identification Number.
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. 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. 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.
Fig. 5. (a) Distributions of the number of works and fans for Fiction and Nonfiction authors. (b) Distribution of reviews’ length for the top-5 Fiction and Nonfiction genres, respectively.

Figure 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. 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,
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, Gaussian Naive 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.

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.85</td>
<td>0.89</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td>K Nearest Neighbor</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>Gaussian Naive Bayes</td>
<td>0.80</td>
<td>0.81</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>Stochastic Gradient Descent</td>
<td>0.81</td>
<td>0.82</td>
<td>0.77</td>
<td>0.78</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
<td>0.76</td>
</tr>
<tr>
<td>Naïve (baseline)</td>
<td>0.21</td>
<td>0.02</td>
<td>0.08</td>
<td>0.03</td>
</tr>
</tbody>
</table>

16Scikit 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 misclassification, 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 PPORAL 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 PPORAL. 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
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)

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>Avg.Degree</th>
<th>Avg.W.Degree</th>
<th>Density</th>
<th>Diameter</th>
<th>Modularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction</td>
<td>966</td>
<td>742</td>
<td>1.536</td>
<td>1.588</td>
<td>0.002</td>
<td>6</td>
<td>0.840</td>
</tr>
<tr>
<td>Similarity</td>
<td>7,273</td>
<td>70,653</td>
<td>19.429</td>
<td>23.603</td>
<td>0.003</td>
<td>7</td>
<td>0.678</td>
</tr>
</tbody>
</table>

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 Goodreads Works 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 Goodreads Works table may have books out of PPORTAL (i.e., which were not available
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


