PPORTAL: Public Domain Portuguese-language Literature Dataset

Mariana O. Silva, Clarisse Scofield, Mirella M. Moro

Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{mariana.santos,clarissescofield,mirella}@dcc.ufmg.br

Abstract. *Combining human expertise with book-consumers data may generate what is needed to sustain constant changes experienced in the book publishing market. Then, building and making available datasets that entirely comprise the essential elements of the book industry ecosystem is essential. However, little has been done in such a context for non-English languages, such as Portuguese. Hence, we introduce* PPORTAL*, a public domain Portuguese-language literature dataset composed of books-related metadata. After an overview of its building process and content, we discuss a brief exploratory data analysis to summarize its main characteristics. We also highlight potential applications, showing how* PPORTAL *is useful as a resource on different research domains.*

1. Introduction

Digital transformation has changed how we produce, consume and relate to ourselves and others. Such evolution has also affected all productive sectors, and the book industry is no exception. Indeed, it is still changing the way we read and, more generally, the book itself. However, the challenge is by no means limited to moving from physical to digital books. Instead, it is a matter of 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, Wang et al. 2019] to natural language processing techniques to classify raw text [de Araujo et al. 2018, Soares et al. 2018, Wagner Filho et al. 2018]. Besides requiring Artificial Intelligence (AI) methods, all of them are essentially data-dependent; i.e., mostly book-related data-dependent.

However, no solution can be properly developed and tested without a preliminary collection of 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 some efforts have been made for English-written books [Ni et al. 2019, Sabri and Weber 2021], little has been done regarding other lesser-spoken languages, such as Portuguese. Hence, we present *PPORTAL*: a **P**ublic domain **PORT**uguese-l**A**nguage **L**iterature dataset whose contributions are summarized as follows:

- Data integration of numerous public domain works from three digital libraries;
- Enriched metadata for works, authors and online reviews extracted from Goodreads;
- Feature engineering on the metadata to create meaningful additional features; and
- Unrestricted access in two formats (SQL database and compressed `.csv` files).

2. Related Work

The real engine of digital transformation for book publishing and other creative industries comes down to having *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, 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. Indeed, few researchers have proposed open and enriched datasets that face such challenges and help to advance research in the book publishing context [Lozano and Planells 2020, Rigau and Tienda 2020, Silva et al. 2021c].

Book publishing is a very segmented industry regarding different publishing categories (academic, literary, etc.), different distributions (physical, digital), and different economic models (self-published, state-funded, privately funded), then making data-based applications and solutions very diverse. We can divide some available book-related datasets into: (*i*) books' reviews/ratings [Lozano and Planells 2020, Ni et al. 2019]; (*ii*) books' metadata [Rigau and Tienda 2020]; and (*iii*) readers' interactions information [Sabri and Weber 2021, Wan et al. 2019]. Although providing valuable data, each focuses on one dimension of the problem, which still requires a more comprehensive, complete dataset to take full advantage of data-driven technologies.

In current book publishing research, most datasets are also limited to English-written books. Hence, there is a considerable research gap for lesser-spoken languages, such as Portuguese and Brazilian Portuguese. Existing public datasets in such languages 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]. From a different perspective, Silva et al. use a dataset that comprises cultural, geographic, and socioeconomic information for exploring Brazilian cultural identity through reading preference [Silva et al. 2021a, Silva et al. 2021c].

Seeking to fill the existing research gaps, we introduce *PPORTAL*, a large dataset with information from books in Portuguese[1] that facilitates studies in the book publishing domain. Besides focusing on the Portuguese-written Literature context, its diverse feature collection can be useful for different Natural Language Processing (NLP) and Machine Learning (ML) applications, as further discussed in Section 4.

3. PPORTAL

We now present *PPORTAL*, a cross-collection dataset of public domain Portuguese-language books. 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 and 3.3. Finally, we summarize its format and usage in Section 3.4.

3.1. Dataset Building Process

Figure 1 shows a compiled schematic diagram of the building process, from web scraping to data pre-processing as explained next.

[1]The dataset considers both Portuguese and Brazilian literature, with works written in both Portuguese and Brazilian Portuguese languages, henceforth called just Portuguese for simplicity.

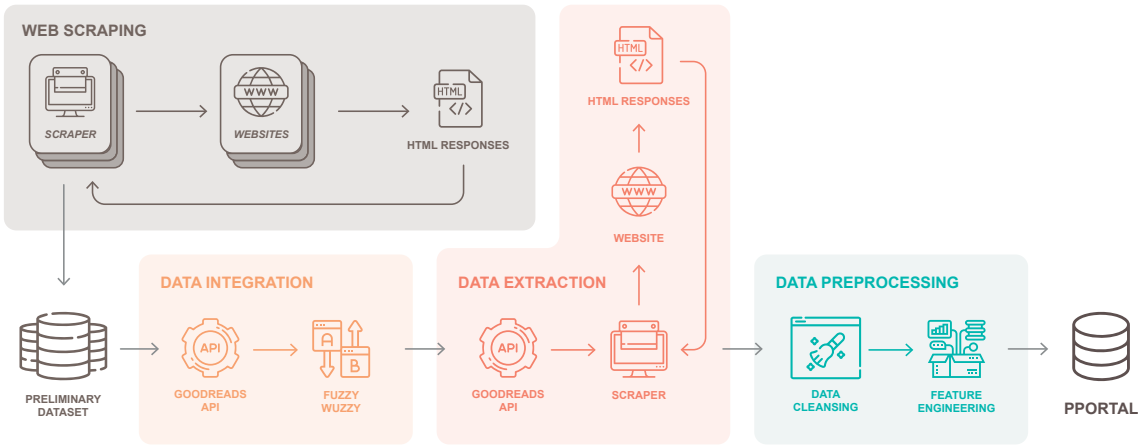

Figure 1. *PPORTAL* building process graphical summary.

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,[2] Projecto Adamastor,[3] and Biblioteca Digital de Literatura de Países Lusófonos (BLPL).[4] Domínio Público is a digital library maintained by the Brazilian Ministry of Education and includes works duly assigned by the copyright holders. Projecto Adamastor collects more than 1,100 titles in Portuguese from several data sources, including Domínio Público. BLPL is a large database of Brazilian and Portuguese literature openly available, with more than 80,000 titles.

The first step is to implement a web crawler to automatically extract raw data from all three platforms, as none of them provides an API. To do so, we used *BeautifulSoup*[5] and *Selenium*,[6] two popular Python libraries for web scraping. We extracted tabular data from the HTML pages by using specific web scrapers, as each platform has unique structure and formatting. This process happened between February and August 2021.

For Projecto Adamastor, this phase was simply extracting all records from its database. Domínio Público offers four searchable media types (text, image, sound, video), and many categories and languages for querying, which required filtering to extract only texts referring to literature in Portuguese. BLPL has an interface based on the sequence of the alphabet, then requiring to select each letter to advance in the search. Also, as one goal of *PPORTAL* is to mine text from literary works, we create a binary flag based on the files' availability for download, a distinct feature of *PPORTAL* that allows filtering such documents as well. As a result, this step collected download links and metadata from over 80,000 public domain works, as informed in Table 1.

Data Integration. The preliminary dataset provides different and unique information for each data source, such as authors' lifetime (Projecto Adamastor), literary genres (BLPL), and the total number of accesses (Domínio Público). Table 2 informs the collected meta-

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

Table 1. 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 2. Collected information present in each digital library.

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

data 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 – the world's largest site for readers and book recommendations, due to its huge volume of data available and its easy-access API.[7] 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 1).

This data integration stage also requires a record linkage approach, as each data source has a different book identification system. Such a problem is usually solved through probabilistic or fuzzy matching methods, which apply string similarity functions. Here, we use the Python library *fuzzywuzzy*[8] 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 1 presents statistics on the whole process before (*Records with Downloads*) and after integration.

Data Extraction. The works' IDs in the Goodreads integrated dataset enabled to collect author and online review information. Through the same Goodreads API, we collected metadata from 966 authors, including name, hometown, and fans count. We also created another web scraper to extract text from the first 30 online reviews of each work. At the end, 4,240 reviews were collected from 518 distinct works, plus 1,430 distinct readers.

Data Preprocessing. The last two steps of the 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. Then, we cleaned and tokenized the work description field using Python library *re*.[9] We also parsed and converted some structured fields into lists, including *authors*, *popular shelves*, and *similar books*. Finally, all readers' identification data from *GoodreadsReviews* was protected by using a hash-based anonymizing method, converting it to a numeric code.

[7] Goodreads API: `https://www.goodreads.com/api`

[8] *fuzzywuzzy*: `https://github.com/seatgeek/fuzzywuzzy`

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

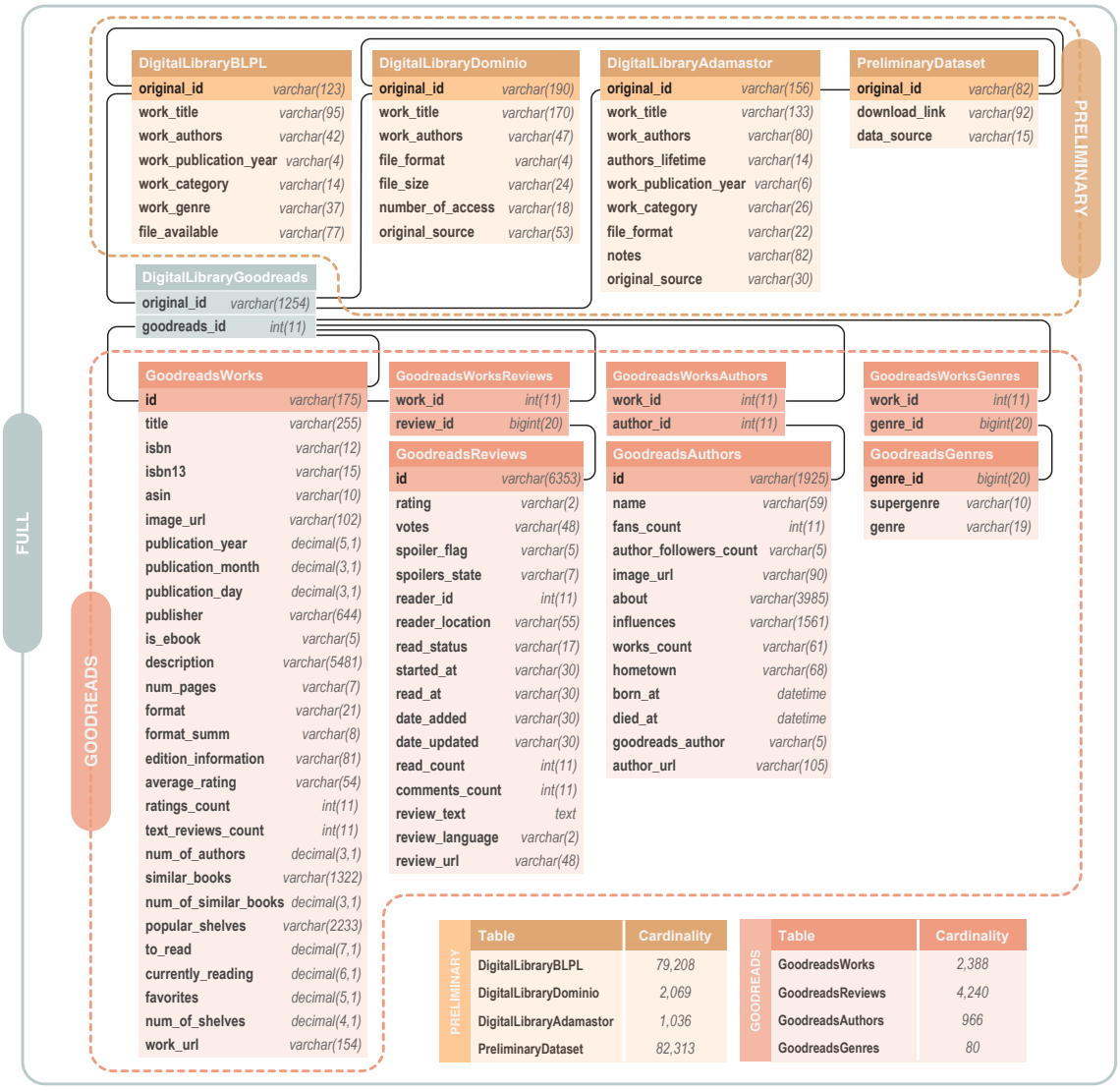

Figure 2. Schema and cardinality for *PPORTAL* divided by versions available.

In the Feature Engineering step, we created new features from the existing data collected. In Goodreads, a book can be stored on users' shelves and defined by using fixed and custom tags. Therefore, following the same methodology applied in [Silva et al. 2021a], we extracted meaningful tags from the works' popular shelves related to the literary genre and popularity information (e.g., number of users who labeled the work as *favorite*, *to-read* and *currently-reading*). We also created some quantitative features related to total number of authors, popular shelves and similar books; and grouped the work-format categories into three: *physical*, *digital*, and *unknown*.

3.2. Data Content

The storage engine used for *PPORTAL* is a relational database management system (RDBMS). Figure 2 depicts its *schema* with 12 tables divided into three available dataset versions: Preliminary, Goodreads and Full. Such division aims to assist different applications of end-users, facilitating access to data that is actually relevant to them. Figure 2 also includes the cardinality of the main tables, whereas Table 3 presents quantitative

Table 3. Quantitative description of *PPORTAL*.

Data Source		# Works	# Authors	# Genres	# Categories	# Reviews	# Readers
Preliminary	*Dominio Publico*	2,069	1,767	-	-	-	-
	Projecto Adamastor	1,036	390	-	54	-	-
	BLPL	79,208	18,289	354	12	-	-
Goodreads		2,388	966	80	-	4,240	1,430

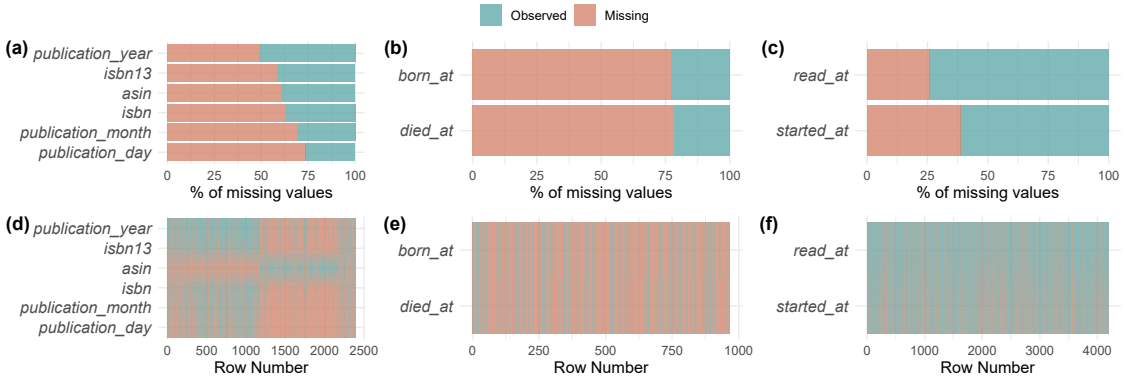

Figure 3. Missing values of *GoodreadsWorks*, *GoodreadsAuthors* and *GoodreadsReviews* tables, respectively. (a) Percentage of the missing values and (b) the distribution of data across all variables.

information about the versions, briefly described as follows.[10]

Preliminary. The Preliminary version includes four tables referring to the three digital libraries and the preliminary dataset described in Section 3.1. As previously mentioned, each digital library presents a set of different features (Table 2). Therefore, we make each one available separately, with the *PreliminaryDataset* acting as a compiled table concatenating all records by their ID, the digital library source, and the link to download.

Goodreads. The Goodreads version includes seven tables referring to works, authors, online reviews, and literary genres. For each of these elements of the book publishing context, there are numerous metadata available in the Goodreads API and the additional features generated in the Feature Engineering step (see Section 3.1). Furthermore, to represent relationships between such elements, we create three join tables: *WorksAuthors*, *WorksReviews* and *WorksGenres*.

Full. This version combines the first two versions and the *DigitalLibraryGoodreads* table, which stores the data integration result (see Section 3.1), resulting in 12 tables.

3.3. Exploratory Data Analysis

We now present a brief exploratory data analysis to investigate the curated and enhanced dataset *PPORTAL* and summarize its main characteristics. We start by analyzing the missing values that were not handled in the Data Cleansing step. In particular, 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. In addition, some identification information related to works has also incomplete records, such

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

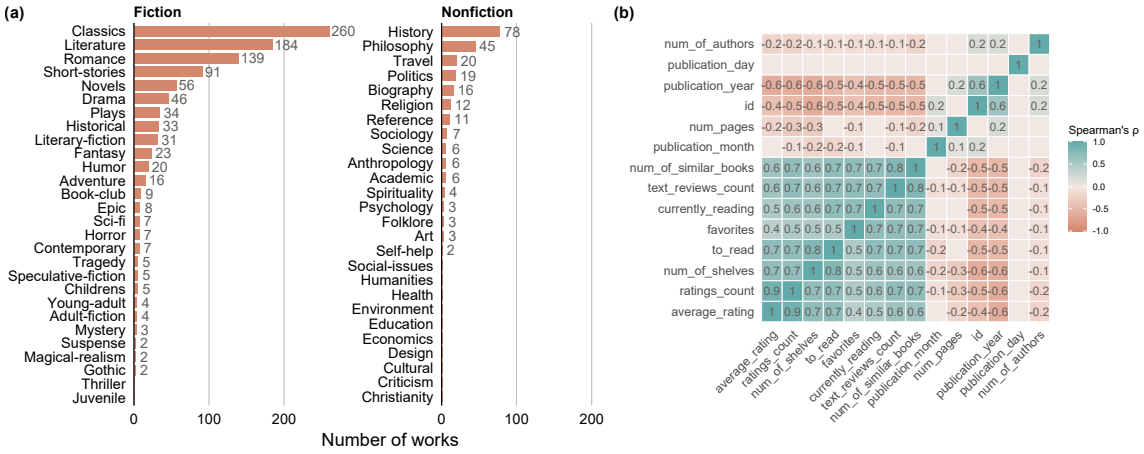

Figure 4. (a) Breakdown of collected books into genre categories of Fiction and Nonfiction. (b) Spearman's rank correlation matrix of the numeric variables from *GoodreadsWorks* table. Correlations with $p - value \geq 0.05$ are considered as insignificant and are left blank.

as ISBN/ISBN-13 and ASIN[11] codes. Overall, the nullity matrix (Figure 3 d–f) shows a correlation between most variables (i.e., if an observation is 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 4a 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.

Finally, Figure 4b presents the Spearman's rank correlation matrix of the numeric variables from *GoodreadsWorks* table. 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*.

3.4. Format and Usage

PPORTAL dataset is publicly available in an open-access Zenodo repository [Silva et al. 2021b], but can also be downloaded from its project webpage.[10] 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.

[11] ASIN stands for Amazon Standard Identification Number.

4. Applicability

PPORTAL's collection of metadata and valuable information can be used to feed a wide variety of machine learning and natural language processing models. Furthermore, the entities in the dataset can be extensively explored in social network analyses. This section shares scenarios and possible applications within such contexts, illustrating the breadth and potential impact of the data available in *PPORTAL*.

4.1. Natural Language Processing (NLP)

Natural Language Processing 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. Currently, most of these applications operate primarily in the English language. Although they can be trained in the Portuguese language, they still need to be significantly improved to understand the nuances and peculiarities of Portuguese. Hence, there is a well justified necessity for creating tools that operate in Portuguese. Next, we briefly describe how our dataset is useful for such applications.

Text Classification. It involves automatically understanding, processing and categorizing unstructured text. Such a task consists of assigning a document into predefined categories. Previous works usually tackle this problem through a machine learning approach: a classifier model is built for learning the characteristics of the categories from a set of pre-classified documents [Sebastiani 2002]. Therefore, from the download links available in our preliminary dataset, the end-user can download the full text of over 80,000 works and extract text structure features by using text mining methods. As labeled data (i.e., with the predefined categories settled), it allows to use the documents from the BLPL and Projecto Adamastor digital libraries (as indicated in Table 2); whereas all documents from the Domínio Público can be used as a test set.

Named Entity Recognition (NER). It is an NLP technique that automatically identifies named entities in a text and classifies them into predefined categories, such as persons, locations, and organizations. State of the art entity recognition systems are based on machine learning methods, employing statistical models that need to be trained on a large amount of labeled data (i.e., corpus) 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 the digital documents at *PPORTAL* plus those available from the download links.

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, even in the literary context [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.

4.2. Machine Learning (ML)

AI-powered technology has accelerated dramatically over the past few years in creative industries, such as music, movies, and books. With so much digital information available, machine learning solutions have been developed to predict the next hit song [Martín-Gutiérrez et al. 2020], the next blockbuster [Ahmad et al. 2017], or even the next bestseller [Wang et al. 2019]. Moreover, book recommendation systems have also been a hot application in ML. Despite the many possibilities, little has been done on Portuguese-language literature regarding the publication scenario due to the scarce availability of benchmark datasets. Therefore, we describe how *PPORTAL* is a valuable data source to enrich and advance research in these different ML applications, as follows.

Recommender System. It is essential to some creative industries as a powerful tool for making sound decisions. The fundamental purpose of a recommender system is to suggest relevant items to users. Traditional solutions are either collaborative or content-based, or a combination of both. While the former is based solely on the past interactions between users and items, the latter uses additional information about users and/or items to produce new recommendations. In particular, *PPORTAL* can be a great resource to boost research and development in the book publishing context. For collaborative filtering techniques, the *GoodreadsReviews* table can be used to create the items and users sets, containing direct (rating on a scale of 1 to 5) and implicit (read count) interactions. *PPORTAL* also provides meaningful metadata for content-based recommendations, such as literary genre, description, format, publisher, authors, among others.

Success Prediction. The increasing volume of consumer data is driving growth and innovation in the publishing industry. Book publishers are increasingly investing in AI-powered and data-driven strategies to offer a holistic view of the books' performance, extracting meaningful insights for better business prospects. Consequently, potential success drivers of books have been of great interest to many researchers [Maity et al. 2019, Wang et al. 2019]. However, understanding how such factors shape the success of books written in languages other than English (e.g., Brazilian literature) has received much less attention. Furthermore, most of the datasets used in such studies are not openly available. Therefore, *PPORTAL* is an essential resource for Portuguese-language book success prediction, providing numerous popularity-based information and an enriched metadata collection.

4.3. Social Network Analysis (SNA)

Social network analysis investigates and characterizes social structures using networks and graph theory. In the context of book publishing, *PPORTAL* can be used to explore interactions between readers present in the Reviews table, between the works' authors (analyzing a co-authorship network), between similar works, among others. By building such networks, many SNA studies can be performed, including, but not limited to: community detection to identify communities of readers/authors/works; social network-based recommender system, e.g., making personalized recommendations from reader preference information; and user-behavior analysis, e.g., perform a cross-location analysis based on reading preferences as done by our research group in [Silva et al. 2021a].

5. Challenges and Limitations

PPORTAL is not free of challenges, which may be tackled in future versions, as follows.

Data Integration. When joining different datasets based on entities that may (or not) share a common identifier, a record linkage method is often necessary, which was also done here. We applied a fuzzy matching approach, where record pairs with probabilities above a certain threshold were considered the same entity. However, although fuzzy logic is typically a better method than deterministic matching, it is prone to misspellings and formatting errors. *Possible solutions:* (*i*) there is not much to do regarding works that are not actually present in Goodreads except keeping them only in the preliminary set; and (*ii*) another option is to manually search on the website for those incorrectly mapped.

Data Quality. Besides being the world's largest site for readers and book recommendations, Goodreads is also a social media platform. Therefore, like any other real-world data source, much of its available content is imputed by its users, then being subject to imprecision and missing information. As presented Section 3.3, some valuable dataset variables have a considerable portion of missing values. *Possible solution:* to consider an additional data source to try imputing the incomplete content.

Distinct Genres. Another problem resulting from the data integration is the genre distinction among data sources, where only two have literary genres: Projecto Adamastor and BLPL. In the former, the *work_category* field comprises 54 genres; whereas in the latter, the *work_genre* field contains 354 genres. When comparing both sets, they share only four literary genres (*Stories*, *Biography*, *Letters*, and *Memories*). *Possible solution:* to consider fuzzy matching approaches to find similar genres.

6. Concluding Remarks

Although digital libraries are excellent sources of literary data, they usually provide documents in tabular structures with their download links and some metadata. Consequently, it is time-consuming to manually collect all available data to be further processed and applied to any task or automated analysis. In order to tackle such a challenge, we introduce *PPORTAL*, an open dataset of public domain Portuguese-language literature. Initially, we built a cross-collection preliminary dataset by integrating public domain works from three different digital libraries, comprising the 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.

PPORTAL is organized in a relational database and available in a Zenodo repository in two formats (SQL dump and compressed `.csv` files). Such a centralized collection is a valuable resource for Natural Language Processing tasks, including (but not limited to) text classification, named entity recognition, and sentiment analysis. We also expect our enhanced dataset to be suitable for various machine learning applications, ranging from recommender systems to book success prediction.

Future Directions. We plan to consider more data sources for handling missing data, and apply fuzzy matching methods for mitigating the distinct genres issue. Given the continued growth of data, we also plan to implement an update-oriented collecting phase. Finally, we are currently working on two tasks: extracting text files from the collected works to create a NER annotated corpus of Portuguese-language literature; and integrating the socioeconomic and cultural information from [Silva et al. 2021c] with *PPORTAL*.

Acknowledgments. The work was partially funded by CNPq and FAPEMIG, Brazil.

References

Ahmad, J., Duraisamy, P., Yousef, A., and Buckles, B. (2017). Movie success prediction using data mining. In *Int'l Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–4. doi:10.1109/ICCCNT.2017.8204173.

Champagne, A. (2020). What Is A Reader? How Readers on Goodreads are Changing the Canon in the Twenty-First Century. In *15th Annual International Conference 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.

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 Association 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*, pages 211–235. Springer International Publishing, Cham.

Martín-Gutiérrez, D., Hernández Peñaloza, G., Belmonte-Hernández, A., and Álvarez García, F. (2020). A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374. doi:10.1109/ACCESS.2020.2976033.

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.

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.

Silva, M., Scofield, C., Oliveira, G., Seufitelli, D., and Moro, M. (2021a). 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., 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. B., and Moro, M. M. (2021c). 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.

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.