What Makes a Book Successful? A Study on Portuguese-language Literature

Clarisse Scofield
clarissescofield@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Mariana O. Silva
mariana.santos@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Mirella M. Moro
mirella@dcc.ufmg.br
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

ABSTRACT

Analyzing the success of books is a matter of interest among publishers, professional book reviewers, expert writers, and even curious readers. Such a task has many influencing factors concerning the intrinsic content and quality of the book (e.g., interest, novelty, writing style, and engaging plot) and others regarding external factors such as social context, author relationships, and luck for publication. Faced with so many variables, recognizing a successful literary work is a challenging endeavor even for specialists in the publishing market. Our objective is: to explore a dataset of books in the Portuguese language created to obtain more knowledge about the different variables of success in the literary context; to understand and evaluate the metrics collected that indicate new perceptions about books using graphical views.

KEYWORDS

Portuguese literature, success analysis, web data, data visualization

1 INTRODUCTION

The book industry has undergone significant changes in the last decade, and experts ensure that this transformation process is just a foretaste of the many changes to come. Specifically, at the 2019 London Book Fair, the north-American Margot Atwell, director of publications at startup Kickstarter, presented five predictions for the book industry in 2025.[1] According to her, using book-consuming-related information will be a crucial tool for decision-making assertiveness, enabling to trace more objective actions from such data and extracting meaningful insights for better business prospects.

If such predictions actually come true, whoever manages to combine human expertise with information drawn from complex digital data will have what it takes to face the upcoming changes in such a fundamental market. It is no surprise that driving forces influencing book success continue to be the main subject of different studies, such as writing styles [1], online reviews [2, 17] and book critics [3]. However, understanding how such factors shape the success of books have received much less attention [15, 16].

The gap is even wider when it comes to Portuguese-language literature studies. Although Portuguese is ranked as the sixth most globally spoken language, little or nothing is known about the factors influencing the success of books in such a context. In fact, existing studies regarding the Portuguese language focus on improving Natural Language Processing (NLP) tasks [7, 8, 11]. Therefore, studies involving the Portuguese language still need to advance to take proper advantage of the language's peculiarities.

Goals. In such a relevant context, this work aims to fill the aforementioned gaps by analyzing what makes a book successful in the Portuguese Literature scenario. Specifically, we introduce an online data-oriented approach to exploring several factors that may contribute to a book's success from different perspectives.

Contributions. Our main contributions consist of: (i) the three perspectives of success - as considering only one aspect to summarize the success of a book can lead to loss of information, we propose a more flexible definition based on three distinct perspectives according to the features available in the created dataset (ii) the different insights related to the success of the books with the methodology proposed (iii) data preparation for future predictions of success (iv) an online data-driven analysis of the success of Portuguese literature, which is still slightly studied.

This work presents the initial part of our research. We develop further studies using the dataset and methodology presented here, such as book genre classification with online reviews [10].

2 RELATED WORK

Currently, limited information is available to assist publishers' decision-making in the book industry [15]. As a result, different exploratory studies investigate factors influencing book success and likability, including literature genres [5], online reviews [2, 17], writing style of the author [1], book critics [3], authors' and publishers' reputations [4]. Whether in the form of exploratory studies or book success predictions, these efforts provide insights to uncover the driving forces behind book success and may act as instrumental in supporting decision-makers.

Although book success analysis remains of interest to many researchers [9, 16], there is still little systematic research involving Portuguese-language literature. As Portuguese has its own literary peculiarities, the evaluation of specific data is of fundamental importance for literature in Brazil. However, existing studies have been limited to improving Natural Language Processing (NLP) tasks [7, 8, 11]. In this sense, to the best of our knowledge, this is the first exploratory study on the success driving factors of Portuguese-language literature books.

3 METHODOLOGY

We now describe the dataset (Section 3.1), the clustering and pillars of success (Section 3.2), and the generation of metadata for analysis (Section 3.3). Figure 1 summarizes the methodology followed.

[1]Publishing 2025: a vision: https://www.thebookseller.com/blogs/publishing-2025-vision-975401

In: I Concurso de Trabalhos de Iniciação Científica (CTIC 2021), Minas Gerais, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2021.

ISSN 2596-1683

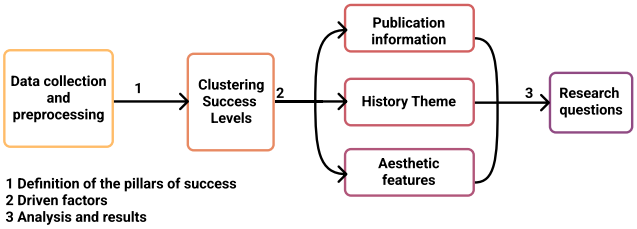

Figure 1: Overview of the methodology for this work.

3.1 Dataset

In this work, we use PPORTAL (Public Domain Portuguese-language Literature Dataset)[13], a cross-collection dataset of public domain Portuguese-language books. PPORTAL is primarily composed of 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] all integrated with additional data obtained from Goodreads[5] platform. The database has information about books and authors and metrics related to book reviews. PPORTAL contains 12 tables divided into three available dataset versions: Preliminary, Goodreads, and Full. With more than two thousand distinct works, the dataset includes 80 different genres classified into *Fiction* or *Nonfiction* categories[6]. PPORTAL dataset is publicly available in an open-access Zenodo repository [14]. Here, to study success in the literary field, we mainly use the metadata related to book ecosystem elements, i.e., books, authors, and readers.

3.2 Book Success Definition

Book success may be defined from different viewpoints, including official bestseller lists [16], the number of online reviews [2], download counts [1] and representative sales data [9]. Such diversity is a result of the success' subjective and abstract nature. Indeed, what may seem successful to some people may not be successful to others. Therefore, proposing a single, objective definition is a challenging task.

In addition to the challenges, considering only one aspect to summarize the success of a book can lead to loss of information about its comprehensive nature. Hence, to fully incorporate book success criteria, we propose a more flexible definition based on three distinct perspectives: *Recognition*, *Popularity* and *Interest*. As a result, we can look deeper into the intrinsic aspects of what makes a book successful. We briefly describe the success measures of each proposed perspective as follows.

Recognition. Recognition refers to how important the book is to be rated and considered by readers. The metric is the count of ratings made for a book and the average ratings.

Popularity. A book's popularity is how much that book has been read and rated by readers in text descriptively. Thus, the metric is the count of review texts.

[2]Domínio Público: https://www.dominiopublico.gov.br/
[3]Projecto Adamastor: https://projectoadamastor.org/
[4]BLPL: https://www.literaturabrasileira.ufsc.br
[5]https://https://www.goodreads.com/
[6]Complete descriptions of each table are available in: https://bit.ly/PPORTAL

Interest. The success of an interest-based book is about how much readers consider that book as a favorite, whether they are reading it at the time or intend to read it.

3.3 Success Driving Factors

The success of a given book may be associated with a collection of factors related to the writing and publishing scenario. In addition to the primary and explicit features such as genre information and publication details, recent research also considers multiple factors such as the writing styles, social networks, critics, and book reviews, expanding research on book success to another level [15].

Such perspectives use a large set of features (e.g., author visibility, previous sales, genre information, etc.) in the publishing market context, making them the basis of the prediction models. However, there is no unique set of features for a successful model. This section proposes a novel taxonomy for the most frequently used features in studying a book's success. Furthermore, it shows how obtaining new, more descriptive data - metadata - is possible from raw data. We can divide such descriptors into three main groups according to their relation to the book itself.

Publication Information. It comes from the book's publication, including year, publisher, sales formats, and the number of pages. This data can provide valuable insights into predicting success related to the publication's marketing, the influence of a specific publisher on a sales market, and book size and success analysis. Publisher features and publication months are often used to analyze book success, mainly from a sales perspective, as in [15].

History Theme. The data directly associated with the book's content includes its description, genre, similar books, and text. It is possible to analyze the gender compared to others or groupings to draw ideas from success, as in the study by Maharjan et al.[6], in which they use a flow of emotions to analyze the book's success. Investigating both the description and quality of the text using Natural Language Processing (NLP) can also bring new data for the success of a book, according to the topic covered and its form [1].

Aesthetic features. Image processing methods can generate new data from the book's cover, such as scores for brightness, contrast, primary colors, etc. In this sense, it is possible to assess the overall style of the cover and produce an assessment to verify comparisons of data related to the cover and the success of a book, creating a new way of investigating success.

3.4 Fuzzy Clustering Method

To determine the book's success, we propose using fuzzy clustering to assess the level of success of a book based on more than one perspective. We describe the method for evaluation in the following section. Cluster analysis is based on partitioning a collection of data points into several subgroups, where the objects inside a cluster (a subgroup) show a certain degree of closeness or similarity. Hard clustering assigns each data point (feature vector) to one and only one of the clusters, with a degree of membership equal to one, assuming well-defined boundaries between the clusters.

In soft or fuzzy clustering, a probability of that point being in that cluster is assigned instead of putting each data point into separate clusters. Each data point can belong to multiple clusters along with its probability score or likelihood. One of the widely used

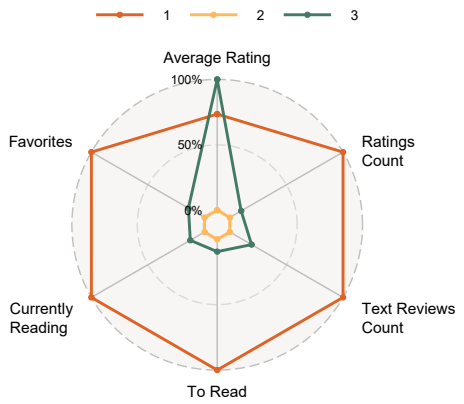

Figure 2: Clustering results

soft clustering algorithms is the Fuzzy C-means clustering (FCM) Algorithm. Fuzzy C-Means clustering is a soft clustering approach, where each data point is assigned a likelihood or probability score to belong to that cluster. In this sense, we can more thoroughly investigate the potential of a book.

Figure 2 shows that three clusters were identified, using six success measures as the clustering algorithm's features. Cluster 1 indicates the success of books with all high feature values, except for *average_rating* with medium value. Cluster 2 indicates the non-success profile, with all features with low values. Finally, Cluster 3 defines the profile of the books with only the *average_ratings* revealing success and the other attributes with low value.

4 RESULTS

This section presents a characterization of the literature in the Brazilian context according to each success perspective.

4.1 Success Level Analysis

After grouping the works concerning the level of success, we explored some of the factors that were most relevant in the preliminary analyses. Next, we present the analyses for the different levels of success according to the factors. The various visualizations revealed some insights about the factors, bringing us closer to the answers to the research questions.

Genres. Initially, using the degree of relevance returned by the clustering algorithm, we analyzed the most frequent literary genres on average for each cluster. For visualization, we use a heatmap widespread to represent multivalued data. In Figure 3, the rows represent the genres, and each column represents a cluster. The color variation indicates the average degree of membership to each cluster. Dark green cells indicate a higher degree of membership and lighter cells a lower degree.

For the high level of success represented by Cluster 1, Religion, Academic, Sci-Fi, and Self-help genres stand out the most; i.e., works of such genres are, on average, more recognizable, popular, and attractive. About the low level, represented by Cluster 2, genres Childrens, Politics, and Travels are the most frequent. Finally, for

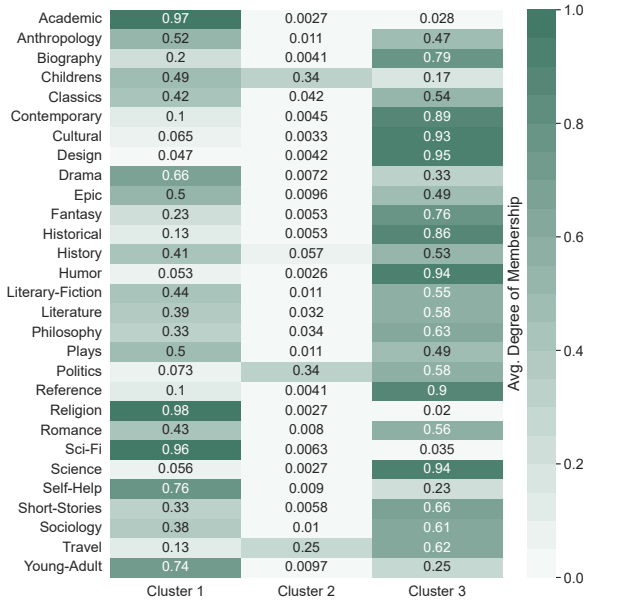

Figure 3: Relationship between genres and levels of success

Figure 4: Relationship between book size and success levels

the medium level, represented by Cluster 3, Design, Science, Humor, and Cultural genres stand out the most.

Size. In addition to literary genres, we also analyzed the average size of works concerning the level of success. In Figure 4, the rows represent the size of the book, and each column represents a cluster (i.e., level of success). Again, the color variation indicates each cluster's average degree of membership. Dark orange cells indicate a higher degree of membership and lighter cells a lower degree.

For the medium and high levels of success, represented by Clusters 3 and 1, the results did not indicate a predominance in the size of the books, on average. That is, such a factor may not influence the reach of literary success. However, at the low level, represented by Cluster 2, there is an average predominance of long/medium books. It may indicate that very long books are not very recognizable, popular, and engaging.

Publication Year. The analyses so far considered only categorical variables. Among our numerical variables, through a preliminary analysis, we identified that the year of publication of the works could be an essential factor in the definition of success. Therefore, we also analyzed this factor, considering the three levels of success

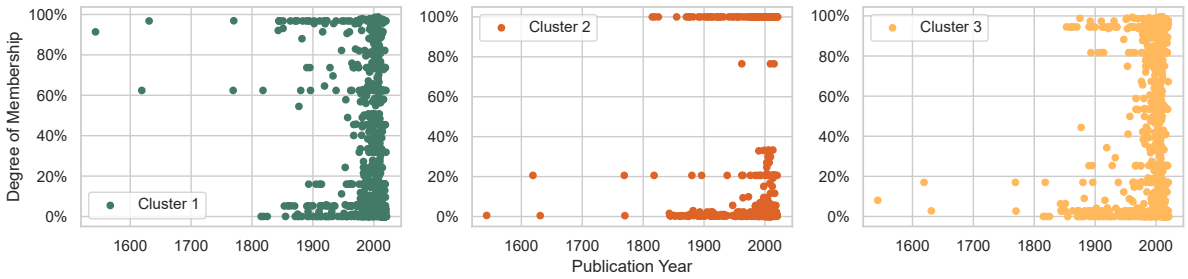

Figure 5: Relationship between publication year and success levels

identified in the clustering. For visualization, we used scatter plots (Figure 5), investigating whether the variation in the years of publication is correlated with the degree of relevance of the works for each cluster (i.e., level of success).

For the medium and high levels of success, represented by Clusters 3 and 1, the results indicate a negative correlation, where the higher the degree of relevance, the lower the year of publication of the books. This result demonstrates that older works receive greater recognition, interest, and popularity. Conversely, at the low success level, there is a positive correlation. That is, the greater the degree of membership of the cluster, the greater the year of publication, indicating that newer books are not as popular. Overall, the results were as expected, as newer books have a shorter time for readers to rate them. Thus, the measures of success we have considered penalize such later-released books.

5 CONCLUSION

Despite being an area still not well explored in the Portuguese language, the study of literature data is highly relevant from the perspective of authors, publishers, writers, or even readers. This work was developed into two main parts: analyzing successful books and classifying genres from their reviews. As products and by-products of this work, we have the following submissions:

Products:. [10] Clarisse Scofield et al., 2022. *Book Genre Classification Based on Reviews of Portuguese-Language Literature.*

By-Products:. [13] Mariana O. Silva et al., 2021. *PPORTAL: Public Domain Portuguese-language Literature Dataset.*
[12] Mariana O. Silva et al., 2021. *Exploring Brazilian Cultural Identity Through Reading Preferences.*

Future Work. We intend to develop classification models for predicting the success of books, as has already been done in some other studies in the area, explicitly applied to Portuguese literature. Finally, we plan to assess the effect of language on rating performance by comparing book success scores for different languages.

ACKNOWLEDGMENTS

This work was partially supported by CNPq, Brazil.

REFERENCES

[1] Vikas Ganjigunte Ashok et al. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1753–1764.
[2] Judith A. Chevalier and Dina Mayzlin. 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* 43, 3 (2006), 345–354. https://doi.org/10.1509/jmkr.43.3.345
[3] Michel Clement, Dennis Proppe, and Armin Rott. 2007. Do critics make bestsellers? Opinion leaders and the success of books. *Journal of Media Economics* 20, 2 (2007), 77–105. https://doi.org/10.1080/08997760701193720
[4] Alain d'Astous et al. 2006. Factors influencing readers' interest in new book releases: An experimental study. *Poetics* 34, 2 (2006), 134–147. https://doi.org/10.1016/j.poetic.2005.12.001
[5] Suraj Maharjan et al. 2018. A Genre-Aware Attention Model to Improve the Likability Prediction of Books. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL, 3381–3391. https://doi.org/10.18653/v1/d18-1375
[6] Suraj Maharjan et al. 2018. Letting Emotions Flow: Success Prediction by Modeling the Flow of Emotions in Books. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 259–265. https://doi.org/10.18653/v1/N18-2042
[7] Miguel V. Oliveira and Tiago de Melo. 2020. Investigating sets of linguistic features for two sentiment analysis tasks in Brazilian Portuguese web reviews. In *Anais Estendidos do XXVI Simpósio Brasileiro de Sistemas Multimídia e Web*. SBC, 45–48. https://doi.org/10.5753/webmedia_estendido.2020.13060
[8] Denilson Alves Pereira. 2021. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review* 54, 2 (01 Feb 2021), 1087–1115. https://doi.org/10.1007/s10462-020-09870-1
[9] Christina Schmidt-Stölting, Eva Blömeke, and Michel Clement. 2011. Success drivers of fiction books: An empirical analysis of hardcover and paperback editions in Germany. *Journal of Media Economics* 24, 1 (2011), 24–47.
[10] Clarisse Scofield, Mariana O. Silva, Luiza de Melo-Gomes, and Mirella M. Moro. 2022. Book Genre Classification Based on Reviews of Portuguese-Language Literature. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR* (Fortaleza, Brazil). 188–197. https://doi.org/10.1007/978-3-030-98305-5_18
[11] Silva et al. 2021. An Improved NER Methodology to the Portuguese Language. *Mobile Networks and Applications* 26, 1 (01 Feb 2021), 319–325. https://doi.org/10.1007/s11036-020-01644-x
[12] Mariana Silva, Clarisse Scofield, Gabriel Oliveira, Danilo Seufitelli, and Mirella Moro. 2021. Exploring Brazilian Cultural Identity Through Reading Preferences. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining* (Evento Online). SBC, Porto Alegre, RS, Brasil, 115–126. https://doi.org/10.5753/brasnam.2021.16130
[13] Mariana O. Silva, Clarisse Scofield, and Mirella Moro. 2021. PPORTAL: Public Domain Portuguese-language Literature Dataset. In *DSW*. SBC, Porto Alegre, RS, Brasil, 77–88. https://doi.org/10.5753/dsw.2021.17416
[14] Mariana O. Silva, Clarisse Scofield, and Mirella M. Moro. 2021. PPORTAL: Public domain Portuguese-language literature Dataset. https://doi.org/10.5281/zenodo.5178063
[15] Xindi Wang et al. 2019. Success in books: predicting book sales before publication. *EPJ Data Science* 8, 1 (17 Oct 2019), 31. https://doi.org/10.1140/epjds/s13688-019-0208-6
[16] Burcu Yucesoy et al. 2018. Success in books: a big data approach to bestsellers. *EPJ Data Science* 7, 1 (06 Apr 2018), 7. https://doi.org/10.1140/epjds/s13688-018-0135-y
[17] Zhou et al. [n. d.]. Measuring book impact based on the multi-granularity online review mining. *Scientometrics* 107, 3 ([n. d.]), 1435–1455. https://doi.org/10.1007/s11192-016-1930-5