JOSE
The Journal of Open Source Education

Goodreader: An Open-Source R Package for Teaching Computational Text Analysis with Goodreads Reviews

Chao Liu [1¶]

1 Cedarville University ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Submitted: 04 March 2025
Published: unpublished

Summary

The *Goodreader* R package provides an accessible and structured approach to analyzing crowdsourced book reviews from Goodreads, with the goal of making computational text analysis more approachable for educators and students. By streamlining web scraping, sentiment analysis, and topic modeling, *Goodreader* enables users to engage with large-scale literary data without requiring advanced programming skills.

In educational settings, *Goodreader* supports digital humanities, computational social science, and marketing research where students can gauge reader sentiment, thematic trends, and the impact of book awards. Instructors can design hands-on assignments for students to analyze book reception, track genre trends, or investigate the relationship between literary themes and public perception.

By lowering the technical barriers to computational text analysis, *Goodreader* facilitates active learning, critical thinking, and digital literacy. As an open-source tool, *Goodreader* encourages reproducibility and collaborative research that inspires students to explore the intersection of literature, data, and technology.

Statement of Need

Computational approaches to text analysis are increasingly important in humanities and social science education, yet many students and educators face significant barriers to entry. Traditional literary analysis relies on close reading (Brooks, 1947; Byron, 2021), surveys (Busselle & Bilandzic, 2009; Miall & Kuiken, 1995), and small-scale content analysis (Filipović, 2018; Suico et al., 2023), while large-scale computational methods often demand extensive programming skills. This gap between theoretical understanding and practical application can hinder broader adoption of computational techniques in related fields. The *Goodreader* R package fills this gap by providing an accessible tool for retrieving and analyzing book reviews from Goodreads (**?**), the world's largest site for readers and book recommendations. By eliminating the need for advanced technical expertise, *Goodreader* enables educators and students to engage with real-world textual data in meaningful ways.

Additionally, existing text analysis tools primarily focus on general sentiment analysis or topic modeling, often without domain-specific applications relevant to literature and the social sciences. While platforms such as Twitter and news archives are commonly used for text mining exercises, few educational tools are tailored specifically for studying literary reception and public engagement with books. *Goodreader* addresses this limitation by offering structured functions for collecting, processing, and analyzing crowdsourced book reviews. Students could use these reviews to better understand reading preferences across different demographics (Thelwall & Kousha, 2017), assess the impact of book awards on reader reception (Peters, 2023), or explore the relationship between literary style and

JOSE
The Journal of Open Source Education

reader engagement (Koolen et al., 2022). *Goodreader* is particularly designed to support courses in digital humanities, data-driven literary studies, marketing research, and social science methodologies.

In an educational setting, *Goodreader* can serve multiple purposes:

- **Enhancing computational literacy**: Students learn practical skills in web scraping, data processing, and text analysis within a structured R environment.
- **Facilitating interdisciplinary learning**: The package supports integration between literature, linguistics, psychology, and data science, helping students apply quantitative methods to qualitative research questions.
- **Supporting active learning**: By engaging with real-world book review data, students develop hands-on experience in analyzing sentiment, identifying thematic trends, and visualizing findings.
- **Enabling research-oriented assignments**: Instructors can design coursework where students explore questions such as *How do readers respond to award-winning books differently from non-award winners?* or *What themes emerge in reader reviews of books on artificial intelligence?*

Goodreader is developed to provide a ready-to-use tool for instructors seeking to integrate data-driven approaches into their teaching that can make computational methods more accessible and engaging for students across disciplines.

Uses and Functionality

The *Goodreader* package collects book-related information from Goodreads without the need for API access. The package uses the `rvest` package (Wickham, 2024) to scrape data directly from Goodreads web pages, eliminating the need for API access. When a user provides search input, the package scans relevant book pages on Goodreads for targeted information. The collected data is then processed and returned as a user-friendly R data frame, which researchers can easily manipulate to suit their specific needs.

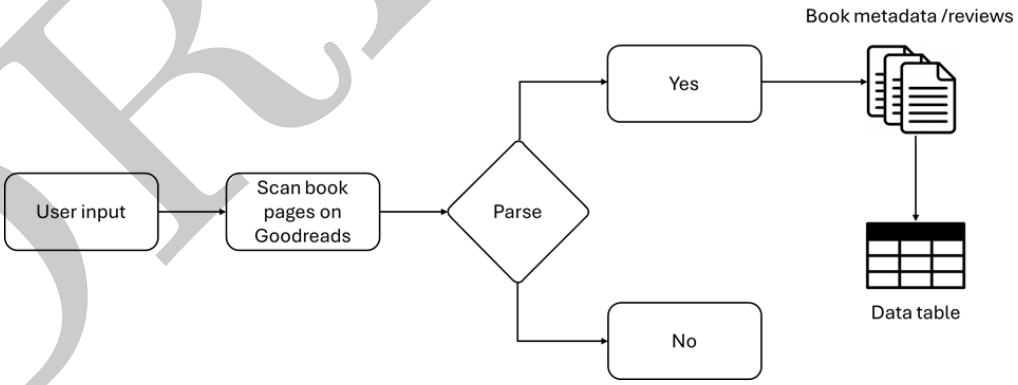

Figure 1: Figure 1: Goodreader package workflow

The package streamlines the process of accessing and analyzing Goodreads data through the following steps:

1. **Book Search**: Users initiate the process with the `search_goodreads()` function, which returns a list of matching titles (or author) and their corresponding book IDs.
2. **Book Information Retrieval**: Using the book IDs obtained from the search, users can then run the `scrape_books()` function to gather detailed information about each book, including genres, publication details, and rating distributions.

3. **Review Collection**: For a more comprehensive analysis of reader opinions, users can apply the `scrape_reviews()` function to collect individual reviews for specific books of interest.

Table 1: *Functions of the Goodreader package*

Function	Returned objects	Description
Search and scrape functions		
`search_goodreads()`	Data frame	Search books on Goodreads based on user's supplied search criteria
`scrape_books()`	Data frame	Scrape book related information (e.g., title, author, summary, genre, average rating)
`scrape_reviews()`	Data frame	Scrape book reviews
Sentiment analysis functions		
`analyze_sentiment()`	Data frame	Perform sentiment analysis on collected reviews
`average_book_sentiment()`	Data frame	Calculate average sentiment score per book
`sentiment_histogram()`	Plot	Create a histogram of sentiment scores for collected reviews
`sentiment_trend()`	Plot	Plots the average sentiment score for collected reviews over time
Topic modeling functions		
`preprocess_reviews()`	List	Preprocess the review text (e.g., remove stopwords, punctuation, non-English, etc.)
`fit_lda()`	LDA model	Fit LDA model on the preprocessed reviews
`top_terms()`	List	Extract and print the top terms for each topic in the LDA model
`model_topics()`	List	Perform topic modeling and print the results
`plot_topic_terms()`	Plot	Create a bar plot of the top terms for each topic

JOSE
The Journal of Open Source Education

Function	Returned objects	Description
`plot_topic_heatmap()`	Plot	Create a heatmap of the topic distribution across documents
`plot_topic_prevalence()`	Plot	Create a bar plot of the overall prevalence of each topic
`gen_topic_clouds()`	Plot	Create a word cloud for each topic
Utility functions		
`get_book_ids()`	Text file	Retrieve the book IDs from the input data and save to a text file
`get_book_summary()`	List	Retrieve the summary for each book
`get_author_info()`	List	Retrieve the author information for each book
`get_genres()`	List	Extract the genres for each book
`get_published_time()`	List	Retrieve the published time for each book
`get_num_pages()`	List	Retrieve the number of pages for each book
`get_format_info()`	List	Retrieve the format information for each book
`get_rating_distribution()`	List	Retrieve the rating distribution for each book

To enhance performance when handling large volumes of books or reviews, both the `scrape_books()` and `scrape_reviews()` functions include options for parallel processing. Additionally, the package implements appropriate delays between requests to respect Goodreads' server resources and avoid overwhelming their system.

The package also offers a suite of functions for performing sentiment analysis and topic modeling on the review data. These functions generate visualizations that depict the emotional tone of the reviews and identify key themes within the collection of text.

For detailed guidance, the user manual is available here, and a step-by-step tutorial demonstrating the package's functionality can be accessed here.

Example Teaching Applications:

- **Digital Humanities & Literary Studies**: Students can analyze how literary themes evolve across different reader demographics, explore reader engagement with award-winning books, or examine sentiment trends in classic and contemporary literature.

- **Data Science & Computational Text Analysis**: The package provides an entry point for students learning natural language processing, allowing them to work with structured textual data without needing extensive programming experience.
- **Social Sciences & Psychology**: Instructors can use *Goodreader* to explore questions related to public opinion, cultural trends, and media reception, making it an ideal tool for courses in media studies, consumer psychology, and communication research.
- **Business & Marketing Research**: Students studying book marketing or consumer behavior can analyze how reviews impact book sales, author branding, and genre preferences, providing insights into audience reception and market trends.

References

Brooks, C. (1947). *The well wrought urn: Studies in the structure of poetry*. Harcourt Brace.

Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology*, *12*(4), 321–347.

Byron, M. (2021). Close reading. In *Oxford research encyclopedia of literature*. https://oxfordre.com/literature/display/10.1093/acrefore/9780190201098.001.0001/acrefore-9780190201098-e-1014

Filipović, K. (2018). Gender representation in children's books: Case of an early childhood setting. *Journal of Research in Childhood Education*, *32*(3), 310–325. https://doi.org/10.1080/02568543.2018.1464086

Koolen, M., Neugarten, J., & Boot, P. (2022). 'This book makes me happy and sad and i love it'. A rule-based model for extracting reading impact from english book reviews. *Journal of Computational Literary Studies*, *1*(1). https://doi.org/10.48694/jcls.104

Miall, D. S., & Kuiken, D. (1995). Aspects of literary response: A new questionnaire. *Research in the Teaching of English*, *29*(1), 37–58.

Peters, B. (2023). The impact of literary awards on reader perception and book sales. *European Journal of Literature Studies*, *1*(1), 49–60.

Suico, T., Donovan, S. J., Boyd, A., Hill, C., Bickmore, S., & Unsicker-Durham, S. (2023). Research: Exploring trends in a growing field: A content analysis of young adult literature scholarly book publications 2000–2020. *English Education*, *55*(2), 116–135. https://doi.org/10.58680/ee202332216

Thelwall, M., & Kousha, K. (2017). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, *68*(4), 972–983. https://doi.org/10.1002/asi.23733

Wickham, H. (2024). *Rvest: Easily harvest (scrape) web pages*. https://github.com/tidyverse/rvest, https://rvest.tidyverse.org/