# Cross-Cultural Classics: Preliminary Findings from Goodreads Based in the U.S. and Douban Based in China

### Hu, Yuerong

yuerong2@illinois.edu University of Illinois at Urbana-Champaign, United States of America

### **Underwood, Ted**

tunder@illinois.edu

University of Illinois at Urbana-Champaign, United States of America

# Layne-Worthey, Glen

gworthey@illinois.edu

University of Illinois at Urbana-Champaign, United States of America

## Downie, J. Stephen

jdownie@illinois.edu

University of Illinois at Urbana-Champaign, United States of America

Due to the lack of historical records on reading (Milligan, 2016), many aspects of readership studies remained theoretical and/or anecdotal until the 21st century. In the last two decades, the abundance of user-generated book reviews online has filled this gap and created unprecedented opportunities for empirical research on readers, books, and reading activities. Empirical findings based on user-generated book reviews have enriched and advanced our understanding of literary genres, cultural products, readerships, and so on (Bartley, 2009; Bourrier / Thelwall, 2020; Pianzola et al., 2022). For instance, with user-generated reviews collected from Goodreads, a U.S.-based social reading and book reviewing website, Walsh and Antoniak identified that "amateur critics" have curated a narrower sense of "classics", under the impacts of schooling, industrial stakeholders of books, and algorithmic moderation of e-commerce (Walsh / Antoniak, 2020).

However, limitations and gaps also emerge from existing research. Most frequently studied books and readerships remain those with distinguished popularity, commercial success, social impacts, and scholarly prestige in the Anglophone world, which are subject to historical and social-cultural biases such as classism, sexism, racism, colonialism (Antoniak / Walsh, 2020; So / Wezerek, 2020). Within this limited research scope, reciprocal and multiplier effects have emerged, where various book review platforms, booksellers, and book impact indicators (e.g., the MLA International Bibliography, the Open Syllabus Project) echo each other's opinions and endorse the same groups of books (Bao / Chang, 2014; Bourrier / Thelwall, 2020; Maity et al., 2017). To expand the current research landscape and enhance cultural inclusiveness and diversity, as Pianzola et al. advocated in their DH2022

panel, multi-lingual and cross-cultural datasets remain much needed (Pianzola et al., 2022).

Following such advocations, we examined the classics curated by online book reviewers in cross-cultural scenarios. Here cross-cultural scenarios feature linguistic and geological differences in user bases across platforms (data sources). Leveraging existing datasets (Hu et al., 2022; Walsh / Antoniak, 2021; So / Weber, 2021), we conducted a comparative analysis of the "classics" curated on Goodreads based in the U.S., and Douban based in China as an Anglophone vs. non-Anglophone book review case study. Table 1 provides a comparison of the two platforms selected based on our prior work (Hu et al., 2022; Kiu-Chor, 2007; So / Weber, 2021).

| Platform  | Founding<br>Year | User Base  | Coverage of Items                                      | Commercial Dependency |
|-----------|------------------|--|--|-----------------------|
| Goodreads | 2007             | - About 90 million registered members as of 2019;<br>- Users based in the United States, United<br>Kingdom, Canada and Australia make up 49%,<br>while users based in Mainland China, Hong Kong,<br>Macao and Taiwan make up around 0.8% (Sabri &<br>Weber, 2021). | Primarily books  | Acquired by<br>Amazon |
| Douban    | 2005             | - About 220 million registered members as of 2020;<br>- Primarily based in "the national boundaries of<br>China as well as the Greater China region" (Kiu-<br>Chor, 2007).   | Heterogenous<br>cultural<br>products and<br>activities | Independent           |

Table 1. Comparison of Goodreads and Douban

For Goodreads classics, we leveraged a sample of 144 classics identified by Walsh and Antoniak based on users' shelving and tagging (Walsh / Antoniak, 2021). For Douban, we selected the 141 books that had on the Douban Top 250 Books list from 2011 to 2021 and were tagged as "classics" and/or "canons" (Hu et al., 2022). Figure 1 visualizes the distribution of the two sets of "classics" curated on Goodreads and Douban. As we can tell from the overlap in the middle of Figure 1, only 21 works are simultaneously Goodreads and Douban classics, while the others are curated as classics on either Goodreads or Douban. Furthermore, the word clouds in Figure 1 show the most frequent user-generated tags associated with these classics: the blue words on the left are the most frequent tags assigned to the Goodreads classics, while the orange words on the right are those on Douban. The larger the text, the more frequently the tags have been used. For example, we can tell from Figure 1 that "literature" is frequently used on both platforms for tagging classics, while certain genre/theme tags are exclusively popular on either Douban (e.g., "martial arts", "manga") or Goodreads (e.g., "gothic", "dystopia"). Such differences in book lists and tags indicate high-level divergences of collective understanding of classics across Goodreads and Dou-

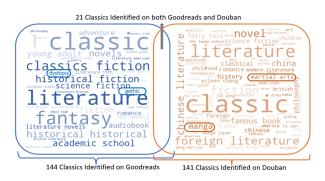


Figure 1. Frequent User Tags for Classics Curated on Goodreads and Douban

| Item | Differences in Ratings (Diff = Douban rating - Goodreads rating, out of 5) | Differences in Numbers of<br>Ratings Received (Diff =<br>Douban num of rating -<br>Goodreads num of ratings) | Divergences of the<br>Distributions of the<br>1-to-5-Star Ratings<br>(Measured by K-L<br>Divergence) |  |
|------|--|--|--|--|
| Min  | -0.290   | -4,908,789   | 0.010  |  |
| Max  | 0.850  | 620,027  | 6.837  |  |
| Mean | 0.262  | -196,121   | 0.252  |  |
| Std  | 0.218  | 683,542  | 0.635  |  |

Table 2. Goodreads and Douban's Differences in (1) Overall Ratings; (2) Numbers of Ratings; and (3) Distributions of Ratings

In addition, we collected and analysed the user-generated ratings and numbers of ratings on both platforms. We measured the differences in parallel ratings, numbers of ratings received per book, and the distributions of the 1-to-5-star ratings for the same books (only the most rated edition/copy) across platforms. Table 2 provides our preliminary findings, as is shown in Table 2. For instance, Table 2 Column 3 shows that on average, each classic receives 196,121 more ratings on Goodreads than on Douban.

Currently, we are further comparing the quantitative and qualitative features of these Goodreads and Douban classics to reveal more differences across the platforms and further understand their implications for ongoing research. While this case study based on Goodreads and Douban makes a unique contribution to the existing cross-cultural analysis of user-generated book reviews, we look forward to expanding our cross-cultural comparisons to other readerships and platforms through collaboration.

# Bibliography

**Antoniak, M.** / **Walsh, M.** (2020): "The Crowdsourced 'Classics' and the Revealing Limits of Goodreads Data", Retrieved January 9, 2022, from https://hcommons.org/deposits/item/hc:31897/.

**Bao, Tong / Chang, Tung-lung Steven.** (2014): "Why Amazon uses both the New York Times Best Seller List and customer reviews: An empirical study of multiplier effects on product sales from multiple earned media." Decision Support Systems, 1-8, DOI: https://doi.org/10.1016/j.dss.2014.07.004.

**Bartley, P.** (2009): "Book tagging on LibraryThing: How, why, and what are in the tags?", Proceedings of the American Society for Information Science and Technology, 46(1), 1–22. https://doi.org/10.1002/meet.2009.1450460228.

**Bourrier, Karen** / **Thelwall, Mike.** (2020): "The Social Lives of Books: Reading Victorian Literature on Goodreads?" Journal of Cultural Analytics, DOI: https://doi.org/10.22148/001c.12049.

Hu, Yuerong / LeBlanc, Zoe / Diesner, Jana / Underwood, Ted / Layne-Worthey, Glen / Downie, J. Stephen. (2022): "Complexities associated with user-generated book reviews in digital libraries: temporal, cultural, and political case studies." In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries, 1-12, DOI: https://doi.org/10.1145/3529372.3530930.

**Kiu-Chor, Ho.** (2007): "A case study of Douban: Social network communities and Web 2.0 in China", Masaryk University Journal of Law and Technology, 2: 43-56.

Maity, Suman Kalyan / Panigrahi, Abhishek / Mukherjee, Animesh. (2017): "Book reading behavior on goodreads can predict the amazon best sellers." In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 451-454, DOI: https://doi.org/10.1145/3110025.3110138.

**Milligan, I.** (2016): "The Problem of History in the Age of Abundance". 5.

Pianzola, Frederico / Marco, Viviano / Alessandro, Fossati / Boot, Peter / Fialho, Olivia / Koolen, Marijn / Neugarten, Julia (2022): "Books' impact in digital social reading: towards a conceptual and methodological framework", Digital Humanities 2022 Tokyo, 94-98.

**So, N.** / **Weber, I.** (2021): "A Global Book Reading Dataset", Data, 6(8), Article 8. https://doi.org/10.3390/data6080083.

**So, R. J.** / **Wezerek, G.** (2020): "Opinion | Just How White Is the Book Industry?", The New York Times. https://www.nytimes.com/interactive/2020/12/11/opinion/culture/diversity-publishing-industry.html

Walsh, Melanie / Antoniak, Maria. (2021): "The Goodreads 'Classics': A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism." Journal of Cultural Analytics, DOI: https://doi.org/10.22148/001c.22221.