Information Research - Vol. 30 No. iConf (2025)

Towards nuanced and critical data analysis: an exploratory investigation into complexities of Goodreads reviews for children's books

Yuerong Hu and Chen Ling

DOI: https://doi.org/10.47989/ir30iConf47269

Abstract

Introduction. With the advance of artificial intelligence and big data, digital humanities have been empowered with new research affordances. However, computational modelling on cultural datasets without a critical examination of the datasets risks decontextualized datafication and misinterpretation, which does not necessarily advance our knowledge into humanities and social sciences questions. This paper presents an exploratory investigation into 37,408 Goodreads reviews on children's books to showcase the complexities of cultural datasets online, to inform contextualized and critical cultural data analysis.

Method. We conducted topic modelling and qualitative analysis of content on the reviews retrieved by regular expressions and visualized our findings.

Analysis. Our analysis empirically illuminates (1) participation in reviewing children's literature online, and (2) presentation of topics and opinions in the reviews.

Results. We found that although various family members across generations participate in reading children's books and reviewing them online, female members play particularly active roles, and children's opinions are underrepresented.

Conclusions. In addition to gaining insights into reviews on children's books, this study also demonstrates the importance of exploratory data analysis in gaining a more nuanced understanding of cultural datasets online, which is essential for advancing insightful and critical data analysis.

Introduction

In the last decade, artificial intelligence has become increasingly impactful in both socio-cultural life and humanities research, from generating poetry that competes against '*the work of professional writers*'(Köbis and Mossink, 2021) to deciphering ancient scrolls (Marchant, 2024). However, tensions persist between humanities methodologies and AI technologies (Chun and Elkins 2023). One key question is AI applications' capabilities to depict the complexities and nuances of humanities data and questions (Chun and Elkins 2023; Windhager and Mayr, 2024). This is a particularly pressing problem for digital humanities research on social media data '*produced within the fabric of sociomaterial*' (Wu, 2023). Due to factors ranging from the platform owners' restricted data access to heterogeneous user behaviour, rarely can scholars comprehensively evaluate the limitations, biases, and representativeness of the datasets available for research. This lack of understanding of the datasets remains a bottleneck for critical usage of large humanities datasets, despite advances in AI (Nguyen et al. 2020).

To address this problem, digital humanities and information sciences scholars have proposed multiple theories and methods, such as data archaeology (Lee, 2021), predictive modelling (Underwood 2019), and speculative bibliography (Cordell, 2020). It is increasingly acknowledged that a nuanced understanding of the data as it is (e.g., how the data has been created, curated, and used within social, cultural, and technical systems) is key to computational modelling that adequately reflects the dataset's complexities and contexts.

Digital humanities research on online book review studies provides a modelling research case. Although advanced AI models have been applied to reviews across various platforms and languages, answers to some research questions remain unclear (Herrmann et al., 2024). For example, are all reviewers merely book readers (Landow, 2006)? To what extent are reviews incentivized (Hu et al., 2023), copied (David and Pinch, 2006), or generated by AI? Why do some reviewers rate their favourite books with 1 star and less favoured ones with 5 stars (citation anonymized for blind review)? Without addressing these questions, AI applications can hardly make reliable and insightful interpretations of millions of book reviews available.

Among various book reviews, those for children's books are particularly challenging to interpret computationally as they tend to involve multiple stakeholders and perspectives. For example, many self-identified parent reviewers provide opinions of their own and those of their children in the same review; therefore, opinions and sentiments presented in the same review might diverge drastically. Yet only one numeric rating is associated with each review. Such features pose practical questions to data analysis: to what extent and how are children readers' opinions represented by online book reviews, and how can researchers identify them? When multiple opinions are presented in one review, how can the computer disentangle the discourses or enrich the single-valued ratings?

It is essential to explore these questions for more advanced computational data analysis, which advances humanities understanding of children's literature and yields practical implications. For decades, a number of heated discussions on children's literature remains theoretical or anecdotal, partially due to the lack of reading records (Hu, et al. 2023): Who is in a position to judge children's literature, adults, children, or both (Abate, 2020)? What are the perceived differences between literature for children and young (Gubar, 2011; Trites, 2001)? Unprecedented online book reviews enable scholars to empirically understand the reception of children's literature from different perspectives and with varied granularities, answering existing research questions while raising new ones.

Given such backgrounds, we conducted exploratory data analysis on 37,408 reviews on 17,407 children's books collected from the social reading platform Goodreads. This is designed as an exemplar case study of inspecting characteristics of social media datasets for advancing cultural

data analytics. We leveraged both computational models and qualitative analysis of content to explore three questions: Who engages with reading and reviewing children's literature? How are topics and opinions presented in the reviews? Based on the first two questions, what are the practical implications for AI-infused research on children literature reviewing and datasets alike? Specifically, our research focuses on reviews that explicitly mention family members, as reading and reviewing children's books within families is a prominent setting identified in our dataset. In the remainder of this paper, first, we introduce the research dataset and method. Then we present our data analysis and outcomes. Finally, we summarise our findings and their limitations, along with proposals for future work.

Dataset and methods

We leveraged existing research datasets called Goodreads book graph datasets (Wan et al. 2019; Wan and McAuley 2018). The data was scraped from Goodreads, a social reviewing platform based in the U.S. and one of the most prominent data sources for digital humanities research on online book reviewing (Walsh and Antoniak 2021). Our analysis focused on its subset of reviews for children's books. We pre-processed the reviews by dropping the reviews that are not in English or too short to be semantically meaningful (with less than 50 English characters). This process returned 297,222 reviews. Then, we developed a small dictionary for computationally identifying the reviews that mentioned family members. Based on our preliminary data annotation and testing, the regular expression of 'my' followed by a singular or plural form of a noun for a family member effectively and efficiently retrieves reviews of our research interest. Please see Appendix 1 for the family members included in the dictionary. We include 'friends' as an exceptional 'family member' because our pilot research shows that friends are often referred to with similar intimacy and significance as one's family members. With this dictionary and regular expression matching, we retrieved 37,408 reviews that mention family members. It is to be noted that neither our dictionary nor the regular expression-based method can exhaustively identify all reviews mentioning family members, as our goal was to identify sufficient instead of all reviews in this category for exploratory analysis.

We applied two complementary analyses to the reviews. First, we conducted a computational overview of the review regarding (1) mentions of family members and (2) topics that emerge from the reviews. Second, we conducted a qualitative analysis of content (Wildemuth, 2017) on a stratified sample of 400 reviews. For topic modelling, we used the topic coherence measure implemented in Gensim (Rehurek and Sojka, 2011) to optimise the number of topics: using LDA modelling and searching across a range from 10 to 90 topics, the topic coherence optimised at 40. Therefore, we set the number of topics to 40. Please see the Appendix 2 for details of the topics. For content analysis, we randomly selected an even number of reviews associated with each of the family members appearing in the reviews. After deduplication, we got a sample of 400 reviews which we manually annotated and analysed.

Our research proposal was reviewed by the Institutional Review Board (IRB) of our current institution. A decision was made that the application for non-human subjects' research or IRB application was not required. Nevertheless, our research complies with the guidelines proposed by latest research on the legal and ethical use of online book reviews in research (Hu, Layne-Worthey, et al. 2023). For instance, reviews quoted in this paper were all paraphrased and further anonymized to minimise the risks that might be caused to reviewers (e.g., reidentification).

Analysis and findings

Family participation in reading and reviewing children's books

First, to understand who participates in reading and reviewing children's literature in a family, we checked the occurrences of each cluster of family members in the 37,408 book reviews, as is summarised in Table 1. We grouped the 66 named entities of family members that appear in the

reviews into 19 primary groups for presentation purposes. We can tell from Table 1 that the group of words for '*my child*' (including '*my child*', '*my daughter*', '*my son*', and many others), appear 36,383 times in the 37,408 book reviews. Figure 1, the sunburst plot, visualises the distributions of the occurrences in Table 1 in a hierarchical manner: the inner ring represents numbers of occurrence in groups; the outer ring shows how the groups are broken into more detailed counts, with numbers after the words indicating their total occurrences. According to the counts, a variety of family members were mentioned 56, 625 times in total. The most frequently mentioned members are the reviewers' children (64%, n=36, 383), siblings (6%, n=3,393), and friends (5%, n=2,793). The reviewers' parents also appear many times, including mother (5%, n=2,741), parent (3%, n=1,928), and father (2%, n=1,249). Other frequently mentioned family members include reviewers' grandchild (3%, n=1,756), grandparent (2%, n=1,023), and family (as a whole) (2%, n=958).

In addition, the weight of female family members stands up (based on conventional constructions of family members' gender identities). When referring to family members senior to the reviewers, female family members are more likely to be mentioned. Despite wording differences, the occurrences of a reviewer's mother vs. father are 2,741 vs. 1,249, the occurrences of grandmother vs. grandfather are 485 vs.290, and the occurrences of aunt vs. uncle are 137 vs. 74. In these three cases, mentions of female members approximately doubled mentions of their male counterparts. The unbalance is more apparent when reviewers mention their partners/spouses mentions of husband (n=1,204) triple mentions of wife (n=367) and mentions of boyfriend (n=125) quadruple girlfriend (28). However, mentions of younger generations in the family are more balanced gender-wise. For instance, reviewers' sons are mentioned 6,743 times, while their daughters are mentioned 7,212 times; grandsons are mentioned 496 times, and granddaughters are mentioned 488 times. One possible explanation for these findings is that female readers/reviewers have been more actively involved in reading children's literature at home and reviewing books online: When female readers review books online, they may mention their spouses or partners who are often male, as well as their children of various genders; on the other hand, when non-female family members write book reviews, they often mention their female family members who play a significant role in reading at home.

group	entity	count	total	group	entity	count	total
	child	6,708			grandmother	309	
	daughter	6,632			grandpa	171	
	son	6,402			grandma	160	
	kid	5,521			grandparent	131	
	kids	4,864			grandparents	117	
	children	3,474			grandfather	115	
	boys	842			grandmothers	16	
	girls	681		grandparent	grandfathers	4	1,023
	daughters	580			niece	1,051	
	sons	341		niece	nieces	263	1,314
	baby	280			nephew	546	
	babies	32		nephew	nephews	144	690
child	twins	26	36,383		mother-in-law	32	
	friend	1,782		parent-in-law	father-in-law	13	45
friend	friends	1,011	2,793		sister-in-law	30	
	mom	1,395		sibling-in-law	brother-in-law	12	32
mother	mother	1,346	2,741		aunt	130	
	dad	722		aunt	aunts	7	137
father	father	527	1,249		uncle	68	
	parent	983		uncle	uncles	6	74
parent	parents	945	1,928		sister	1,117	
family	family	958	958		brother	690	
husband	husband	1,204	1,204		younger	638	
wife	wife	367	367		older	428	
	granddaughter	439			sisters	115	
	grandson	433			sibling	115	
	grandchild	390			siblings	110	
	grandchildren	382			brothers	89	
	grandsons	63			twin	62	
grandchild	granddaughters	49	1,756	sibling	elder	29	3,393
	cousin	233			boyfriend	125	
cousin	cousins	61	294		partner	54	
					girlfriend	28	
				partner	spouse	27	234

Table 1. Occurrences of the family members being mentioned in the reviews

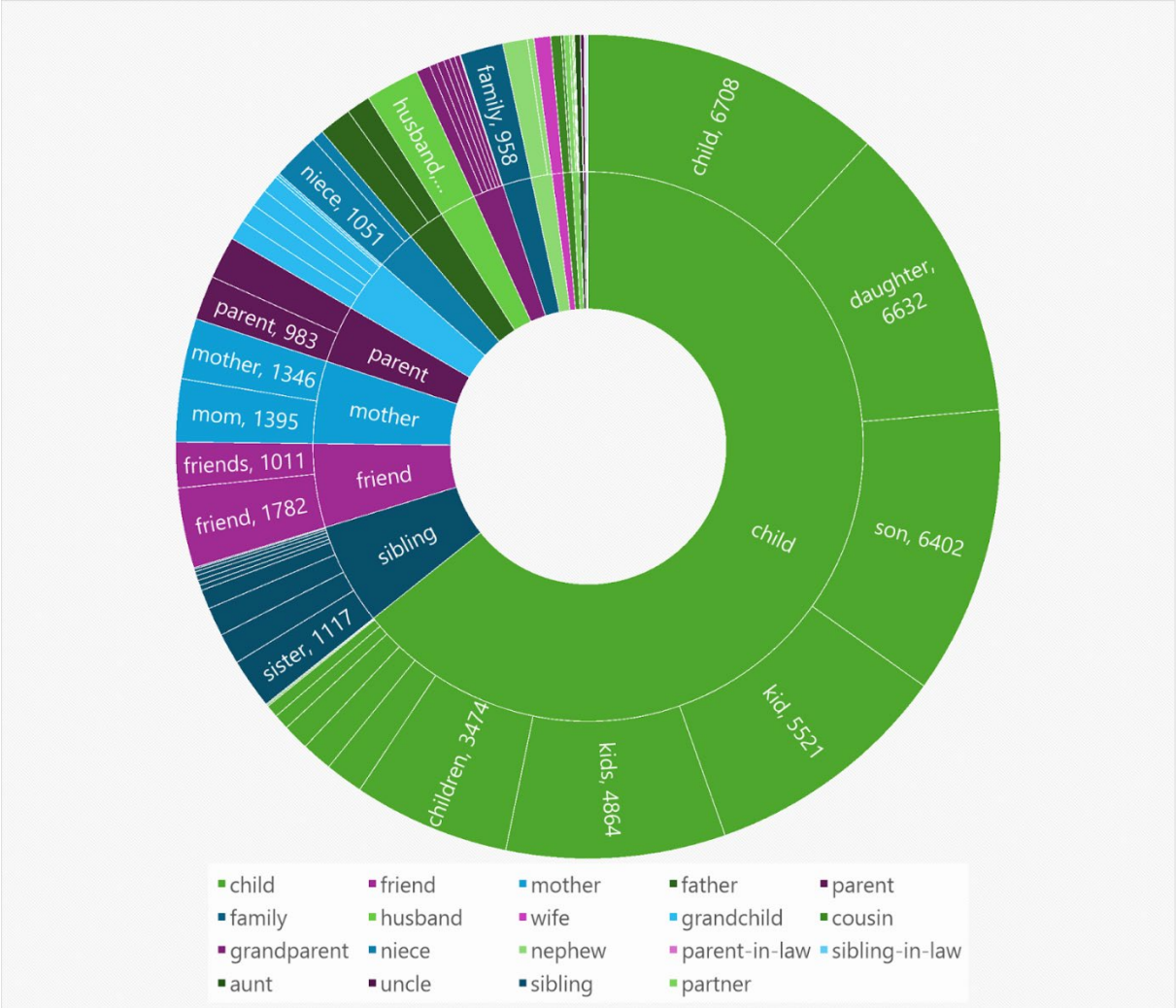

Figure 1. A sunburnt plot depicting distributions of occurrences of family members

Furthermore, we analysed the co-occurrence of different groups of family members in the same reviews. Figure 2 provides an overview: the number in each coloured block indicates the number of times that two different groups of family members appear in the same review; the corresponding textual variables on this coloured block's x-axis (the horizontal line) and y-axis (the vertical line) tell which two groups of members they are. The colour of the block represents the relative frequency of the paired codes' co-occurrence: the darker colours indicate that the paired codes were brought up together more frequently, while the lighter colours suggest fewer co-occurrences. For instance, we can tell from the darkest blue block that reviewers' siblings and children were mentioned simultaneously in 482 reviews, which indicates a stronger proximity or correlation between the two groups. Reviewers' children also co-occur with the reviewers' siblings' mothers (n=452), husbands (n=340), and friends (n=208). In addition, reviewers' mothers and siblings frequently appear in the same reviews. Our close reading reveal that this often happens when reviewers write about their childhood memories of story time with their mothers and siblings, which echoes the hypothesis that female members play a significant role in reading children's books at home.

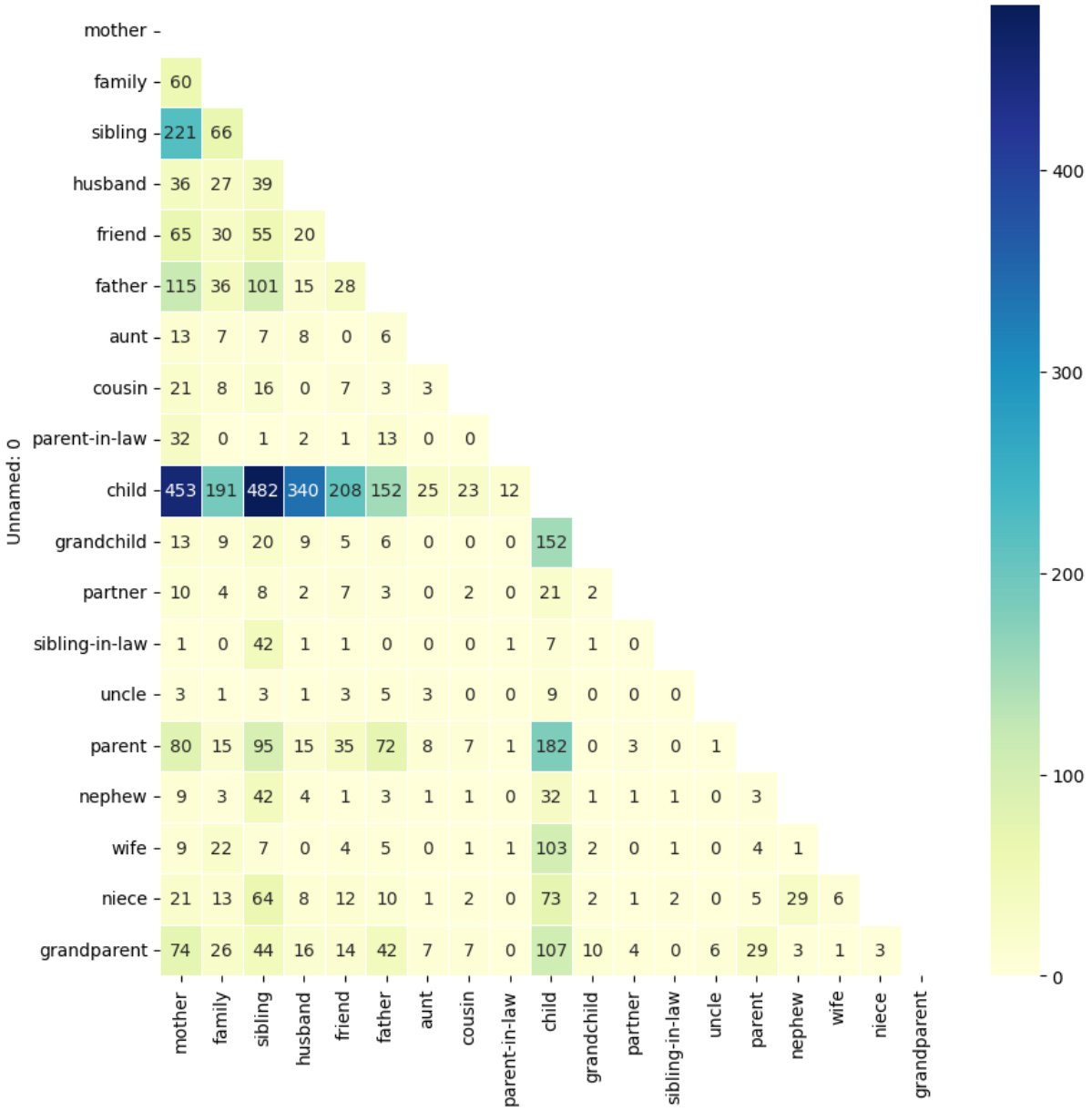

Figure 2. Co-occurrences of family members mentioned in the reviews

Topics and opinions presented in the reviews

First, our topic modeling on the 37,408 reviews provides an overview of topics discussed. In addition to family members, topical words center on book genres and features (e.g., illustration, version, and graphic novel), sentiment and emotions (e.g., amazing, humorous, and thrill), education and schooling (e.g., teacher, school, and study), as well as reading behaviour and scenarios (e.g., reread, bedtime, and listen). For instance, words associated with fairytales and fantasy stories are also commonly seen (e.g., magic, mystery, and enchant), which aligns with genre features of children's literature. Comparing mentions of family members with the topics, we find that reviews involving senior family members, particularly grandparents, are more likely to involve nostalgia emotions, childhood memories, and retrospective reflections. It is also noted that many topic words are conventionally associated with female characteristics, such as actress, witch, ballet dancer, and princess, which might reflect gender stereotypes in children's literature.

Second, we applied close reading and qualitative analysis to the sample of 400 reviews. 70 reviews explicitly mentioned children's reactions to the book being reviewed, while none of the reviewers self-identified themselves as children. Most of the reviews were created (1) to share personal opinions about a children's book as a child supervisor and/or care provider; (2) to express retrospective opinions about a book they had read in their childhood; and (3) to express their thoughts about a book they recently got introduced to or re-read.

These suggest that the reviews might be dominated by adults' opinions. Specifically, according to the reviews written by self-identified parents, parents tend to make decisions for their children, and sometimes even impose their preferences on their children. For instance, one parent wrote:

> *if I cannot bear with a book, I'm not going to read it to my kids by giving it away or hiding it somewhere they won't find or request the book.*

Children's opinions might be underrepresented in numeric book ratings as well. For instance, a parent gave a 3-star rating to a book and commented '*I read this book to my girls who are 5 and 11. They both loved it; but I was bored*'. While acknowledging their children's love for the book, this review did not give it 4-star or 5-star ratings, which are generally regarded as positive. A few parents further explained why they failed to relate to their children's favor for a book, such as '*my children love this book, probably because it contains the word fart, but I do not think this is one of my favourite books on the shelves*'.

Another parent commented:

> *my children love this book because I cannot help crying when reading it aloud to them, then they point at me and laugh, when their father weeps, trying to get the words out without turning into a sobbing mess. sigh. I do love my damn kids.*

Such disagreements and divergences of opinions across reviewers and their children also exist between reviewers and their more senior family members. For instance, a reviewer wrote '*I will never get why my grandmom was so fond of this book. Who needs an upsetting story about cruelty to animals? I am not*'.

Nevertheless, there are also cases where opinions about books converge across generations and books get to be passed on in a family, as the review below shows:

> *my grandma used to read this book to me when I was a kid every time I stayed overnight at her place. I still vividly remember how I sat on her lap in a recliner and listened to her voice. I can still hear it. One day I will read this book to my own kids.*

Out of the 70 reviews, only two offer perspectives from multiple children, as shown in Figure 3. The first review includes the opinions of two child readers but does not reference the adult reviewer's perspective. The second review presents ratings and feedback from each family member, including two adults and three children. For such reviews that feature multiple sets of opinions and ratings from different individuals, the standard approach of treating one review as representative of a single reviewer no longer applies. A more nuanced data analysis, accommodating varying granularities of opinions, is needed for a better interpretation of the data.

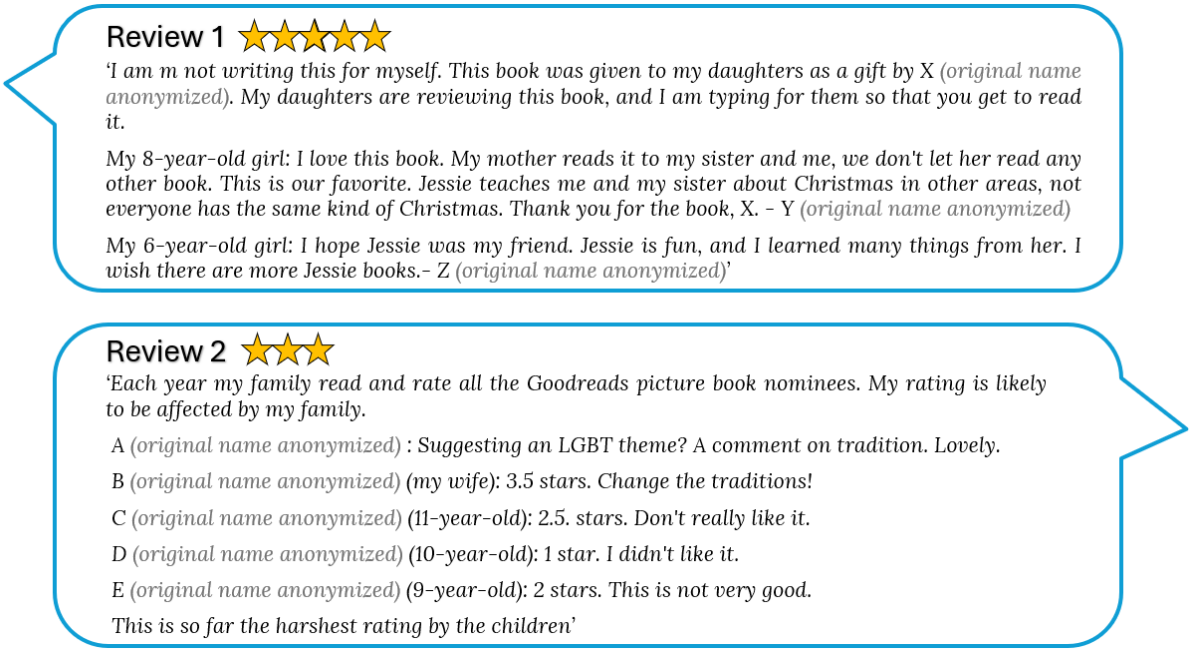

Figure 3. Two reviews that present multiple children's opinions

Conclusions and discussions

This case study of 37,408 Goodreads reviews on children's books has three primary findings. First, a variety of family members were mentioned in online book reviews on children's books, which suggests their broad participation in reading children's literature. Parents and female family members played especially prominent roles in reading such books in households and reviewing them online. Second, while various topics were discussed in the reviews, the reviews were predominantly adults' voices. Children's feedback on books appears to be underrepresented in both review texts and numeric ratings. Third, findings of our study, such as the gender-specific topic words and some parents' irrelevance to children's love for certain books, speak to problems identified by critical scholarship on children's literature, such as sexism and gender issues in children's literature (Varga-Dobai 2013), paternalism (Gubar, 2020) and censorship (Jenkins, 2010; MacLeod, 1983) in children's reading activities. Such connections demonstrate the promise of enriching critical scholarship on children's literature with online book reviews and empirical methods. Questions worthy of future investigation also emerge from our preliminary study, such as (1) how cross-generation readers of the same family negotiate with each other, given divergent opinions about the same books; and (2) how female family members' participation might have shaped online book reviewing.

As an exploratory analysis of a specific dataset, these preliminary findings are subject to further examination and might not apply to other datasets. Therefore, we intend further examine the prominent roles played by female family members through hypothesis testing and analysis on other datasets. Nevertheless, existing findings demonstrate the complexities of online reviews of children's books and have practical implications for cultural data analysis. For instance, this case study demonstrates how mixed-method analysis contributes to a more comprehensive understanding of online cultural datasets; such understanding should help researchers identify suspicious data instances such as fake reviews and AI-generated content, so as to improve data quality in scholarly research. In the future, we plan to deepen our analysis by conducting more qualitative investigations and testing our findings on other datasets. We are currently working on surveys on online book reviews and interviews on reading children literature at home, to advance our understanding of the dissemination and evaluation of children's literature.

Acknowledgements

We thank the Goodreads book reviewers who shared their ratings and reviews. This research would be impossible without their contributions. We also want to thank anonymous reviewers of iConference 2025 for their critical feedback.

About the authors

Yuerong Hu is an Assistant Professor at the Department of Information and Library Science of Luddy school of Informatics, Computing, and Engineering, Indiana University Bloomington, 700 N. Woodlawn Ave, Bloomington, Indiana, 47408, USA. She received her Ph.D. from University of Illinois Urbana-Champaign. Her research combines digital humanities, cultural analytics, and critical data studies to advance contextualized analysis of cultural datasets. She can be contacted at yuerhu@iu.edu.

Chen (Cici) Ling is an Assistant Professor at Informatics Department of Luddy School of Informatics, Computing, and Engineering at Indiana University Bloomington, 700 N. Woodlawn Ave, Bloomington, Indiana, 47408, USA. She received her PhD from Boston University. Her research spans security, privacy, psychology, and computational social science, with a particular focus on social media and niche communities, aiming to promote accessibility and inclusivity in digital applications. She can be contacted at ccling@iu.edu.

References

Bannon, L. (2011). Reimagining HCI: toward a more human-centered perspective. Interactions, 18(4), 50–57. https://doi.org/10.1145/1978822.1978833

Abate, M. (2020). No Kids Allowed: Children's Literature for Adults. Baltimore: Johns Hopkins University Press. https://muse.jhu.edu/pub/1/monograph/book/78274.

Chun, J, & Elkins, K. (2023). The Crisis of Artificial Intelligence: A New Digital Humanities Curriculum for Human-Centred AI. International Journal of Humanities and Arts Computing 17 (2): 147–67. https://doi.org/10.3366/ijhac.2023.0310.

Cordell, R. (2020). Speculative Bibliography. Anglia 138 (3): 519–31. https://doi.org/10.1515/ang-2020-0041.

David, S, & Pinch, T. 2006. "Six Degrees of Reputation: The Use and Abuse of Online Review and Recommendation Systems." First Monday 11 (3).

Gubar, M. (2020). Toothless Pedagogy? Problematizing Paternalism in Children's Literature and Childhood Studies. Children's Literature 48 (1): 153–97. https://doi.org/10.1353/chl.2020.0007.

Herrmann, B., Pianzola, F., van Zundert, J. J., Hu, Y., Nielbo, K., Koolen, M., ... & Boucher, M. C. Unsiloing and Unsilencing Online Reader Reviews - Mapping Reception of Fiction Novels Across Cultures and Languages. In Proceedings of the 11th Edition of the DH Benelux Conference. Belgium. https://2024.dhbenelux.org/wp-content/uploads/2024/05/DHB24_paper_Herrmann_et-al.pdf.

Hu, Y., Layne-Worthey, G., Martaus, A., Downie, J.S., Diesner, J. (2023). Research with User-Generated Book Review Data: Legal and Ethical Pitfalls and Contextualized Mitigations. In: Sserwanga, I., et al. Information for a Better World: Normality, Virtuality, Physicality, Inclusivity. iConference 2023. Lecture Notes in Computer Science, vol 13971. Springer, Cham. https://doi.org/10.1007/978-3-031-28035-1_13

Hu, Y., LeBlanc, Z., Diesner, J., Underwood, T., Layne-Worthey, G., & Downie, J. S. (2024). Complexities of leveraging user-generated book reviews for scholarly research: transiency, power dynamics, and cultural dependency. International Journal on Digital Libraries, 25(2), 317-340. https://doi.org/10.1007/s00799-023-00376-z.

Jenkins, C. A. (2011). Censorship: Book Challenges, Challenging Books, and Young Readers. In Handbook of Research on Children's and Young Adult Literature (pp. 443-454). Routledge. Kidd, Kenneth B. 2007. “Prizing Children’s Literature: The Case of Newbery Gold.” Children’s Literature 35 (1): 166–90.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in human behavior, 114, 106553. https://doi.org/10.1016/j.chb.2020.106553.

Landow, G. P. (2006). Hypertext 3.0: Critical theory and new media in an era of globalization. The Johns Hopkins University Press.

Lee, B. (2021). Compounded mediation: A data archaeology of the newspaper navigator dataset. Digital Humanities Quarterly, 15(4).

MacLeod, A. S. (1983). Censorship and children's literature. The library quarterly, 53(1), 26-38.

Marchant, J. (2024). First passages of rolled-up Herculaneum scroll revealed. Nature 626 (7999): 461–62. https://doi.org/10.1038/d41586-024-00346-8.

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How we do things with words: Analyzing text as social and cultural data. Frontiers in Artificial Intelligence, 3, 62. https://doi.org/10.3389/frai.2020.00062.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre*, *Faculty of Informatics*, *Masaryk University*, *Brno*, *Czech Republic*, *3*(2).

Underwood, T. (2019). Distant horizons: digital evidence and literary change. University of Chicago Press.

Varga-Dobai, K. (2013). Gender Issues in Multicultural Children's Literature—Black and Third-World Feminist Critiques of Appropriation, Essentialism, and Us/Other Binary Oppositions. Multicultural Perspectives, 15(3), 141-147. https://doi.org/10.1080/15210960.2013.809303.

Walsh, M., & Antoniak, M. (2021). The goodreads “classics”: a computational study of readers, Amazon, and crowdsourced amateur criticism. Journal of Cultural Analytics, 6(2), 243-287. https://doi.org/10.22148/001c.22221.

Wan, M., & McAuley, J. (2018). Item recommendation on monotonic behavior chains. In Proceedings of the 12th ACM conference on recommender, 86–94. Vancouver British Columbia Canada: ACM. https://doi.org/10.1145/3240323.3240369.

Wan, M., Misra, R., Nakashole, N., & McAuley, J. (2019). Fine-grained spoiler detection from large-scale review corpora. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2605–10. https://doi.org/10.18653/v1/P19-1248.

Windhager, F., & Mayr, E. (2024). Digital Humanities and Distributed Cognition: From a Lack of Theory to its Visual Augmentation. Journal of Cultural Analytics, 7(4). https://doi.org/10.22148/001c.121866.

Wu, P. F. (2023). Veni, vidi, vici? On the rise of scrape-and-report scholarship in online reviews research. Journal of the Association for Information Science and Technology, 74(2), 145-149.

Appendix

Appendix 1. The list of words used for retrieving reviews that mention family members

'my mother', 'my mom', 'my father', 'my dad', 'my parent', 'my parents', 'my child', 'my children', 'my kid', 'my kids', 'my baby', 'my babies', 'my son', 'my sons', 'my daughter', 'my daughters', 'my boys', 'my girls', 'my sister', 'my sisters', 'my brother', 'my brothers', 'my twin', 'my twins', 'my sibling', 'my siblings', 'my grandmother', 'my grandmothers', 'my grandma', 'my grandpa', 'my grandfather', 'my grandfathers', 'my grandparent', 'my grandparents', 'my grandson', 'my grandsons', 'my granddaughter', 'my granddaughters', 'my grandchild', 'my grandchildren', 'my aunt', 'my aunts', 'my uncle', 'my uncles', 'my niece', 'my nieces', 'my nephew', 'my nephews', 'my cousin', 'my cousins', 'my husband', 'my wife', 'my partner', 'my spouse', 'my sister-in-law', 'my sisters-in-law', 'my brother-in-law', 'my brothers-in-law', 'my mother-in-law', 'my father-in-law', 'my boyfriend', 'my girlfriend', 'my fiancé', 'my fiancée', 'my younger', 'my older', 'my elder', 'my friend', 'my friends', 'my family'.

Appendix 2. The forty topic models (ten words per topic)

(0, '0.189*"reread" + 0.126*"level" + 0.119*"silly" + 0.117*"perhaps" + 0.069*"indeed" + 0.038*"typical" + 0.034*"intend" + 0.030*"behaviour" + 0.028*"happily" + 0.027*"insist"')

(1, '0.635*"friend" + 0.095*"pass" + 0.084*"positive" + 0.065*"discussion" + 0.048*"appropriate" + 0.000*"teorii" + 0.000*"sursa" + 0.000*"toate" + 0.000*"studiat" + 0.000*"urmaresc"')

(2, '0.238*"movie" + 0.206*"hear" + 0.158*"listen" + 0.131*"adorable" + 0.119*"bedtime" + 0.039*"flow" + 0.022*"listening" + 0.005*"maker" + 0.005*"player" + 0.001*"compilation"')

(3, '0.279*"begin" + 0.146*"gift" + 0.092*"fan" + 0.084*"consider" + 0.083*"enjoyable" + 0.072*"piece" + 0.052*"window" + 0.042*"incredible" + 0.031*"tradition" + 0.020*"outside"')

(4, '0.451*"definitely" + 0.121*"death" + 0.096*"appeal" + 0.084*"though" + 0.070*"nearly" + 0.035*"own" + 0.031*"memorable" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat"')

(5, '0.299*"care" + 0.182*"woman" + 0.151*"grandmother" + 0.114*"various" + 0.050*"charm" + 0.039*"separate" + 0.024*"claim" + 0.004*"ostensibly" + 0.000*"sursa" + 0.000*"timpului"')

(6, '0.093*"book" + 0.055*"read" + 0.034*"story" + 0.030*"love" + 0.029*"child" + 0.018*"get" + 0.017*"make" + 0.017*"time" + 0.017*"kid" + 0.015*"go"')

(7, '0.400*"girl" + 0.237*"home" + 0.165*"sister" + 0.078*"smile" + 0.067*"fantastic" + 0.003*"backyard" + 0.000*"teorii" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa"')

(8, '0.202*"night" + 0.191*"watch" + 0.139*"often" + 0.061*"chance" + 0.045*"poem" + 0.041*"modern" + 0.038*"mark" + 0.032*"famous" + 0.032*"wide" + 0.027*"belong"')

(9, '0.301*"school" + 0.068*"sense" + 0.065*"magic" + 0.061*"present" + 0.058*"relate" + 0.041*"middle" + 0.040*"difference" + 0.039*"notice" + 0.028*"culture" + 0.025*"happiness"')

(10, '0.130*"original" + 0.110*"introduce" + 0.090*"edition" + 0.088*"self" + 0.083*"surprise" + 0.077*"hide" + 0.075*"secret" + 0.047*"steal" + 0.045*"fair" + 0.039*"print"')

(11, '0.210*"absolutely" + 0.199*"away" + 0.116*"mystery" + 0.081*"creature" + 0.076*"town" + 0.045*"solve" + 0.032*"bond" + 0.027*"premise" + 0.024*"gentle" + 0.018*"neighborhood"')

(12, '0.098*"sometimes" + 0.078*"least" + 0.066*"free" + 0.060*"guess" + 0.056*"summer" + 0.048*"situation" + 0.047*"power" + 0.039*"shelf" + 0.038*"entertain" + 0.037*"certain"')

(13, '0.082*"illustration" + 0.075*"thing" + 0.067*"look" + 0.060*"page" + 0.056*"picture" + 0.049*"word" + 0.046*"different" + 0.040*"show" + 0.032*"kind" + 0.032*"share"')

(14, '0.461*"mother" + 0.107*"adore" + 0.096*"niece" + 0.057*"forest" + 0.046*"enchant" + 0.040*"fabulous" + 0.031*"originally" + 0.019*"compassion" + 0.018*"naturally" + 0.011*"hood"')

(15, '0.324*"ever" + 0.216*"classic" + 0.129*"delightful" + 0.064*"encourage" + 0.046*"encounter" + 0.043*"sequel" + 0.037*"comment" + 0.037*"look_forward" + 0.017*"lovable" + 0.003*"somehow_misse"')

(16, '0.163*"cat" + 0.156*"teacher" + 0.141*"aloud" + 0.130*"class" + 0.096*"view" + 0.064*"topic" + 0.033*"poetry" + 0.028*"heavy" + 0.027*"th_grade" + 0.023*"center"')

(17, '0.211*"believe" + 0.175*"appreciate" + 0.154*"voice" + 0.130*"plan" + 0.081*"english" + 0.066*"act" + 0.052*"front" + 0.029*"genius" + 0.009*"essentially" + 0.000*"stiti"')

(18, '0.196*"leave" + 0.150*"let" + 0.140*"lose" + 0.136*"bad" + 0.099*"line" + 0.072*"imagine" + 0.045*"scary" + 0.027*"identify" + 0.023*"somewhat" + 0.016*"medium"')

(19, '0.073*"world" + 0.073*"favorite" + 0.061*"live" + 0.052*"grow" + 0.044*"become" + 0.039*"play" + 0.038*"place" + 0.029*"set" + 0.028*"remind" + 0.026*"bear"')

(20, '0.455*"family" + 0.198*"name" + 0.132*"tale" + 0.109*"father" + 0.021*"dance" + 0.017*"orphan" + 0.011*"adopt" + 0.002*"actress" + 0.000*"sursa" + 0.000*"stiti"')

(21, '0.000*"timpului" + 0.000*"toate" + 0.000*"romanian" + 0.000*"sen" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa" + 0.000*"teorii" + 0.000*"proza"')

(22, '0.116*"big" + 0.108*"fact" + 0.100*"mom" + 0.068*"plot" + 0.053*"soon" + 0.049*"realize" + 0.048*"job" + 0.043*"interested" + 0.025*"car" + 0.024*"awesome"')

(23, '0.121*"true" + 0.088*"imagination" + 0.076*"simply" + 0.062*"main" + 0.056*"grade" + 0.054*"fly" + 0.052*"letter" + 0.052*"top" + 0.052*"actual" + 0.051*"beautifully"')

(24, '0.240*"star" + 0.136*"question" + 0.130*"sweet" + 0.125*"person" + 0.056*"easily" + 0.037*"reviewer" + 0.032*"agree" + 0.029*"study" + 0.027*"college" + 0.024*"humorous"')

(25, '0.292*"wait" + 0.273*"maybe" + 0.131*"hate" + 0.067*"elephant" + 0.061*"egg" + 0.038*"stupid" + 0.016*"crush" + 0.000*"teorii" + 0.000*"toate" + 0.000*"wasteful"')

(26, '0.257*"animal" + 0.105*"detail" + 0.070*"sad" + 0.057*"colorful" + 0.054*"number" + 0.051*"fantasy" + 0.046*"cool" + 0.042*"stick" + 0.035*"count" + 0.030*"sign"')

(27, '0.000*"timpului" + 0.000*"toate" + 0.000*"romanian" + 0.000*"sen" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa" + 0.000*"teorii" + 0.000*"proza"')

(28, '0.701*"daughter" + 0.079*"relationship" + 0.040*"store" + 0.030*"shoe" + 0.023*"grandpa" + 0.017*"kick" + 0.013*"inappropriate" + 0.008*"dollar" + 0.006*"frustrated" + 0.004*"meanwhile"')

(29, '0.150*"chapter" + 0.122*"move" + 0.111*"face" + 0.075*"light" + 0.059*"fill" + 0.055*"sleep" + 0.043*"city" + 0.037*"country" + 0.036*"door" + 0.036*"ready"')

(30, '0.155*"funny" + 0.148*"lesson" + 0.090*"style" + 0.077*"writing" + 0.072*"illustrate" + 0.065*"student" + 0.047*"description" + 0.038*"similar" + 0.036*"witch" + 0.026*"cousin"')

(31, '0.000*"timpului" + 0.000*"toate" + 0.000*"romanian" + 0.000*"sen" + 0.000*"stele" + 0.000*"stiti" + 0.000*"studiat" + 0.000*"sursa" + 0.000*"teorii" + 0.000*"proza"')

(32, '0.073*"part" + 0.069*"back" + 0.045*"quite" + 0.042*"point" + 0.039*"especially" + 0.039*"together" + 0.036*"perfect" + 0.036*"one" + 0.033*"course" + 0.032*"right"')

(33, '0.160*"message" + 0.160*"sit" + 0.149*"special" + 0.103*"month" + 0.099*"glad" + 0.057*"twist" + 0.052*"nephew" + 0.043*"throw" + 0.038*"approach" + 0.021*"plant"')

(34, '0.231*"wonderful" + 0.098*"baby" + 0.098*"receive" + 0.089*"memory" + 0.072*"favourite" + 0.067*"dream" + 0.048*"hour" + 0.045*"touch" + 0.042*"future" + 0.035*"red"')

(35, '0.431*"boy" + 0.124*"cry" + 0.116*"tree" + 0.080*"tear" + 0.074*"perfectly" + 0.046*"single" + 0.017*"boyfriend" + 0.015*"sharing" + 0.013*"sible" + 0.009*"unconditional"')

(36, '0.290*"short" + 0.159*"bed" + 0.144*"base" + 0.081*"fairy_tale" + 0.075*"lie" + 0.045*"mine" + 0.033*"marry" + 0.028*"flip" + 0.013*"ish" + 0.012*"myth"')

(37, '0.146*"talk" + 0.121*"laugh" + 0.088*"draw" + 0.077*"explain" + 0.069*"several" + 0.061*"loud" + 0.060*"opinion" + 0.047*"alone" + 0.034*"space" + 0.033*"afraid"')

(38, '0.203*"add" + 0.152*"language" + 0.136*"list" + 0.107*"struggle" + 0.101*"strong" + 0.094*"event" + 0.073*"charming" + 0.044*"fiction" + 0.009*"non_fiction" + 0.002*"license"')

(39, '0.142*"art" + 0.096*"black" + 0.085*"build" + 0.082*"princess" + 0.080*"white" + 0.078*"hit" + 0.060*"aspect" + 0.052*"fast" + 0.051*"money" + 0.047*"reach"')