Metadata, the undiscovered treasure box for book recommender systems for children

What traits can be inferred from book's descriptive and structural metadata which could support the designers of books recommender systems for children?

Mohamad Awab Alkhiami

Supervisor: Dr. Sole Pera

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 29, 2023

Name of the student: Mohamad Awab Alkhiami
Final project course: CSE3000 Research Project
Thesis committee: Dr. Sole Pera

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

This study examines traits that may be derived from book metadata and identifies statistically significant patterns and trends in order to assist recommender system designers for children's books in selecting the most important traits to take into account. This research focuses on descriptive such as publish year and structural metadata such as the chapters' count. We rely on data from two different data sets, Goodreads and Wikisource.

The findings of this study, such as the fact that there is a relationship between the number of pages and the reader's age, may have an impact on the design of recommender systems for children's books.

This is all with the respect to the fact that children are a vulnerable audience, making it challenging to not infringe on their privacy while gathering information that can be used in statistical experiments in many fields such as education and health.

1 Introduction

Kids nowadays are attached to social media. Tweens (ages 8 to 12) spend around 5 hours and 33 minutes on social media daily while teens (ages 13 to 18) spend up to 8 hours and 39 minutes on average [1]. Experts are concerned that when kids spend so much time on screens, they are substituting other useful activities such as reading with social media, and thus can affect their cognitive skills [2]. This should draw our attention to the fact that the cognitive abilities of the future generation might be affected, and society should act rapidly to preserve the next generation.

One measure that might help to do so, is encouraging them to read more, this can be done by offering them books that meet their interests. Recommender systems are an ideal solution when it comes to finding related items based on reactions or reviews of people with the same interests. However, studies show that children's recommender systems may not be optimum and tend to recommend popular books over those that are suited for children's reading skills or interests. [3]. This aligns with the study that mentioned that among many U.S. kids, only 17% of kids said they read for fun [1]. Although many traits are used by recommender systems to rank books. These traits do not meet the desires of underage groups since they are not considered mainstream users. As a result, we must investigate alternative methods of determining features that may be desirable to them.

This process is more complicated than it is expected to be since kids are a vulnerable audience therefore the data collection process has numerous restrictions and would require consent from kids and their guardians as well. Furthermore, the most common method for recommender systems is collaborative filtering, which suffers from a cold start problem when no rating data are given, which makes it difficult for the recommender system to identify books that match the reader's preferences. Those preferences are limited and difficult to obtain for the children's audience, stressing the significance of this research, which aims to supply more characteristics to recommender system designers for them to enhance their algorithms by considering these traits. This research highlights the integral role that metadata of books can play in this process by extending the research done by Milton et al [4] by using different data sets and by considering more traits inferred from the metadata perspective since not all data sources have the information that previous scholars have highlighted as significant traits that are considered worthy. In this study, multiple data sets that are verified to be children's books are used. This variety helps to avoid the passive trends that might occur in one data set, After that, contrasting the statistical results with the data set that contains only adults' books will underline the real significant trends that can be considered by recommender systems designers.

Our approach starts by gathering data and augmenting it to different patches categorized by age groups, inferring traits from books metadata, analyzing them, and identifying the statistically significant results which form a trending curve that is expected to align with the ratings and reviews of that data set once they are available. Knowing that early exposure to books encourages the development of greater language skills in children than waiting till later [5], We believe that recommender systems that use our promoted traits will offer kids a trusted set of recommended books and hence increase the number of readers children to evolve and help the society.

This paper explains the methods that were employed. Additionally, high-level features are examined, data sets are contrasted, and statistical graphs are shown. Additionally, the outcomes are listed and justified. Finally, several restrictions are outlined and suggestions for further research are made.

2 Related work

There have been few studies in this area, but the study by Milton et al. [4] is comparable in that it concentrates on the same high-level topic, but examines different traits such as the emotional analysis of a book's content, the brightness and the colourfulness of a book's cover. The intersection with this study's traits, is the books length trait. They also utilize other data sources and age-groups. Other research looked at other factors, such as the child's gender and the ability of interactions, and found that some traits, such as the preference of folding books, differed across genders [6]. They also found that girls prefer narrative literature while boys favor nonfiction, however this was contradicted by another research [7]. We believe that this inconsistency is driven by the fact that they used different data sets and that the age groups examined differ. They also concluded that the vast majority of first graders favored educational books, particularly animal books.

Looking at all of these researches, we can see that they employed metadata analysis in their research, such as book genre and book interaction type. However, there haven't been much more studies on other metadata traits, thus this study is attempting to fill that need.

Age group	Number of books
0-12	124082
12-18	93398
18+	2360655
Total	2578135

Table 1: Number of books in age-groups

Age groups			
Goodreads	Children		Young-adults
	0-12		12-18
Wikisource	Preschoolers	School-aged	Adolescents
	0-6	6-12	12-18

Table 2: Age-groups per data set

3 Data description and experiments' preparation

The aim of this research project is to find out which high-level type of metadata has an impact in the matter of traits inferring for book recommender systems for children and young adults in age groups ranging from 0 up to 17, Two main data sources were used here namely the Goodreads data set [8; 9; 10] and Wikisource data set [11]. The Goodreads data were gathered in late 2017; by simply scraped users' public shelves. These data were gathered exclusively for scholarly purposes and can not be redistributed for commercial purposes. We are interested in two major subsets (shelves) of this data set: children (ages 0 to 12), and young-adults (ages 12 to 18). where both subcategories have been proven to be books for children. We utilized a third Goodreads category, adult books, to contrast and compare children's traits with those of adults.

The Wikisource data set contains instances of books for children as well, however, based on the distribution of the book's reading ages in this data set, it was beneficial to divide the books into three different groups than the division we used in the Goodreads data set. Preschoolers (ages 0 to 6), school-aged (ages 6 to 12), and adolescents (ages 12 to 18).
Table 1 shows the number of books for different age groups, while table 2 shows the age-groups category of the used data sets.

In the experiments of this study, the used data sets will be mentioned to help the reproducibility aspect for researchers in future studies. We applied several examination methods on these data sets to observe whether readers' preferences are changing between different audiences, and based on the results it was determined whether there was a statistically significant difference in the preferences of the two age groups for that feature. The aforementioned tests can be in one of the following categories: Analysis of variances such as t-tests, correlation analysis, and observations from diagrams. The suitable test was chosen depending on the experiment and the hypothesis that has to be proven or rejected. furthermore, the experiments will be discussed in the next section.

This research is focused on two types of metadata. First, there is structural metadata, which describes how a digital element is organized. Second, there is descriptive metadata, which comprises information about the item, such as the title, creator, and relevant keywords.
We created the following questions based on these two categories:
Structural metadata questions:
Is there a relation between a book's age-group and the structural metadata of that book? including the number of pages (RQ1), and the number of chapters (RQ2).
Descriptive metadata questions:
Is there a relation between a book's age group and the descriptive metadata of that book? including the fact that a book is part of a series (RQ3), the publication year (RQ4), and the cover type of that book (RQ5).

The traits of the experiments which report P-values less than 0.05 can be used in improving the design of recommender systems in multiple aspects as the following [12; 13]:

- A/B testing: In a recommendation algorithm, hypothesis testing can be performed to identify the performance difference between control and treatment groups.
- Testing can be done to evaluate the performance of a recommendation algorithm to that of a baseline method.
- Identifying key aspects: Hypothesis testing may be used to establish which properties are critical for the recommendation algorithm.
- Customization Evaluation: Testing may be performed to assess the performance of personalization algorithms.
- Model comparison: Testing may be used to evaluate the performance of many models in order to choose the best one for the job.

4 Results and discussion

Experiments were conducted to investigate the trends and traits of books that emerged from our examination of the aforementioned resources, and we discussed the significance of our findings and highlighted them so that the recommended children's books by the designers who used our reported patterns could satisfy the tastes and needs of children.

This section examines two high-level metadata categories: descriptive and structural metadata. In addition, many traits from these two groups are presented in this section.

4.1 Number of pages (RQ1)

The projected book length is one of several factors that authors must think about while producing a book. Based on the story line and the developmental stage of the intended readership, authors should determine the length of their books. Numerous research have been done to determine the ideal book length for children of various ages. These studies have provided an approximation of the words per book for different age groups [14; 15; 16; 17].

- **Picture books**, often for ages 2 to 8, are illustrated children's stories with 50-1,000 words and topics pertaining to emotional intelligence and social connections.
- **Chapter books** are lengthier children's stories for ages 7-9 that typically include 4,000–15,000 words and are divided into chapters for easy comprehension.

- Children starting to read independently ages 5-9 should use **Easy Reader books**, also known as Emerging Reader or Beginning Reader books, which have 48–64 pages of brief, straightforward stories with 200–1,500 words.
- Depending on the reader's age, **juvenile novels** can be either fiction or nonfiction and have word counts 2,000-80,000. They are written for children ages 7-18.
- For children aged 8-12, **middle grade novels** often include 20,000–40,000 words with sibling rivalry and story lines connected to fitting in.
- **Young adult novels** are any genre, 40,000–80,000 words long, and intended for readers between the ages 12-18.

We formed the hypothesis that there is no difference in the means of books' lengths over age groups. After evaluating the children and young adults, we reject the null hypothesis with a p-value lower than 0.05, we may draw the conclusion that there is an increase in book length as children become older. However, the analysis of the adult and young adult subsets showed a higher p-value. As a result, we are unable to disprove the null hypothesis using this test.

In Figure 1, we illustrate the number of pages of the studied age-groups. We can see that children's average page count is as low as it can possibly be and is much close to 23 pages. Additionally, young adults prefer to read books considerably longer than the typical adult reader, but as readers become older, the maximum number of pages increases which also aligns with the findings of Milton et. al [4] and indeed there is a relation between a book length and the targeted age group.

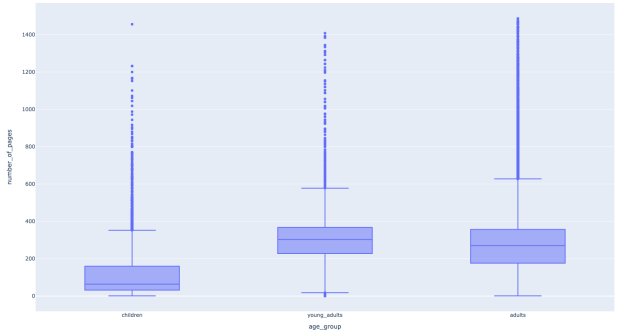

Figure 1: Number of pages of the studied age-groups

4.2 Number of chapters (RQ2)

The relationship between the number of chapters in a children's book and their reading habits is intricate and complicated. While some research implies that children prefer shorter books with fewer chapters because they are more digestible, others say that children prefer longer books with more chapters because of the prolonged storyline and added material[18].

To investigate the association between a book's chapter count and the intended age group, we used a t-test on the means of the number of chapters across the Wikisource data set's age categories, namely preschoolers, school-aged, and adolescents. The stated p-value of 0.02 for the association between the number of chapters in children's books and reading preferences in preschoolers and adolescents is most likely due to the small sample size, rather than a genuine correlation. Furthermore, we cannot emphasize any relationship in this sense as a response to the second research question.

Figure 2 demonstrates how the number of chapters varies by age group. This chart shows that the chapter count range for adolescents is substantially wider than for the other research age groups.

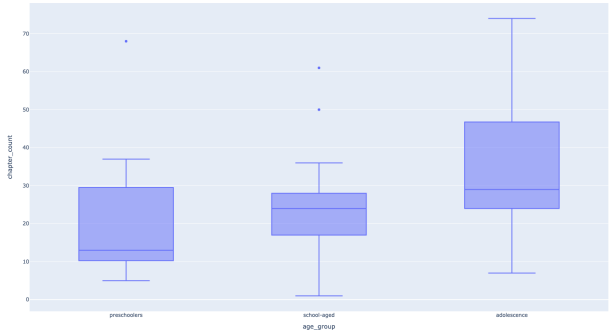

Figure 2: Number of chapters over age groups

4.3 Is part of a series (RQ3)

This subsection shows the statistics of books being part of a series, multiple series or a standalone book that is not a member of any series. Let we first remember that readers who fall in love with characters and places want more of them. Moreover, When children finish one book in a series, they experience a great feeling of accomplishment. And with each completed book, they gain greater confidence in reading as they may simply follow the storyline and characters they have loved over time [19]. Additionally, It is often assumed that children and young people prefer to read publications that are part of a series over solitary books [20]. It is important to stress that this is not a fact, because each child is unique and may have different reading preferences. To compare and determine whether there is a relationship between the two variables, we ran a correlation test and created the correlation heatmap seen below in figure 3. Surprisingly, the data in the table does not reveal any correlation between variables series member and age group. One can assume as a result that readers' choices for series publications do not vary that much as they get older.

4.4 Publication year (RQ4)

In this subsection the null hypothesis states that readers choices is not affected by the publication year. To validate this hypothesis, Multiple T-tests were conducted on the different subsets of the Goodreads data set which results to the following:
When the means of subgroups children and young adults were

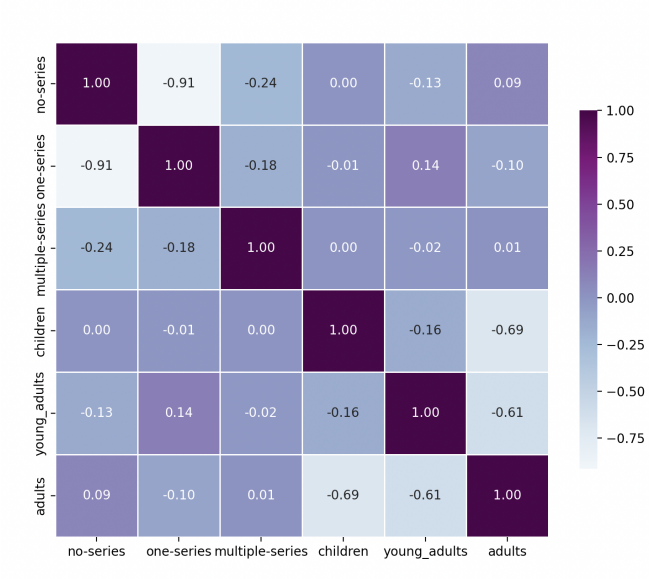

Figure 3: Correlations between series and age-groups of the Goodreads data set

compared, the p-value was 0.04, which is below the level of statistical significance and permits the rejection of the null hypothesis, however, the p-value of the comparison between the young adults and adults subsets was just above the threshold and does not give enough confidence to reject or accept the null hypothesis. Surprisingly, is the p-value below 0.05 when contrasting the children's subset along with the adults' one and therefore a statistically significance can be highlighted. As a conclusion, there is a relationship between reader age and book publication year when children grow to be young adults or adults, while we don't have compelling evidence to state this when young adults grow to be adults. Furthermore, to capitalize on our findings, Figure 4 shows a Box-plot of the publishing year of the researched age groups after selecting the books to be published between 1900 and 2023.

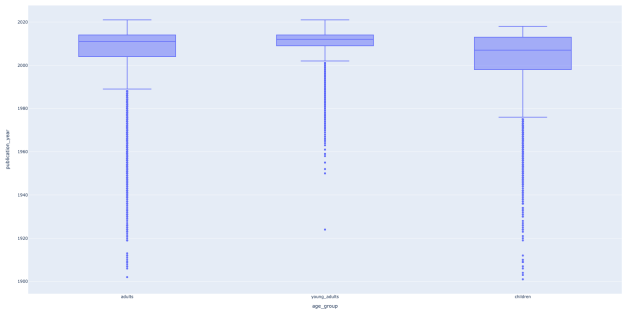

Figure 4: Publication year of the studied age-groups of the Goodreads data set

From this plot one can observe the followings:

- Most books that are written for young adults seemed to be published after 1950.
- Children appear to read more classic books than young adults, despite the common misconception that they will be drawn to newly released books.
- The majority of young adults' books were published between 2009 and 2015.
- Around the time of World War One, there is a noticeable gap in publication dates.

Surprisingly, when comparing the data from the Wikisource data set in figure 5, this contradicts the expected tendencies. We also feel that this is due to the limited sample size of this data set.

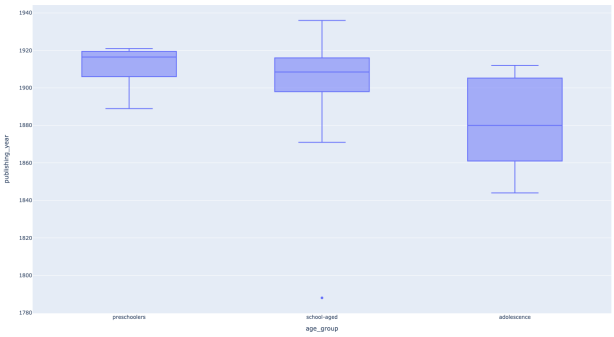

Figure 5: Publication year of the studied age-groups of the Wikisource data set

4.5 Cover type (RQ5)

Publishers must take into consideration the benefits of choosing softcover versus hardcover publications, since softcover books have advantages like being generally less expensive, lighter, and more flexible and thus have higher sales numbers compared to hardcover books [21]; on the other hand, hardcover books can be preferred over softcover books, for being generally more durable and having a more premium feel [22]; however, they are generally more expensive, heavier, and less flexible. From a different angle, it is reasonable to assume that softcover books are more appropriate for adults since they are less expensive and easier to travel than hardcover books, which are robust and able to withstand heavy/rough use [22; 23]. The reads data set contained plenty of book cover types, although the types that were focused on are soft and hard covers books and the other categories are a matter of different traits, therefore multiple cover types were categorized into one of the two categories above. For instance: the type 'hardcover and CD' or 'library-binding book' are hardcover books. In order to determine whether there is a relation between the cover type variable and the age group, a correlation test was performed and a contingency table is plotted in figure 6.

Looking at this table, there is a low correlation between cover types and the three studied age groups and that the variables are not interdependent.

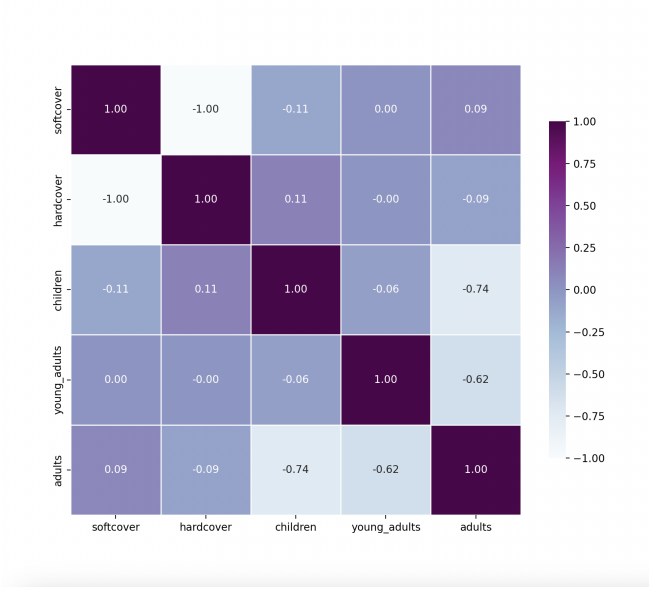

Figure 6: contingency table of cover types and age-groups.

5 Responsible Research

Children's privacy is taken into account in this study since official public data sets from Goodreads are utilized rather of data collected directly from children, moreover, The analyzed data sets were legitimate because they were published for educational purposes.

6 Limitations and Future Work

This section discusses the limitations and difficulties encountered while conducting this study. Such constraints may be associated with the data gathering process, data accuracy, or the way the gathered data can be integrated together. We also considered potential solutions to the constraints we encountered. Finally, A statement on our recommendations for further research is also included after that.

6.1 Data precision and validity limitations

The utilized data sets include certain unexpected values, which may cause the data to be disconnected or to reflect weird patterns when analyzed. Such as the following: Looking at book A in the data set, book B is in the same series as book A, yet book A is not in the same series as book B in book B entity in the data set. We also noticed that many books are released in the far future e.q. after the year 2500.

Beside all that, A book's targeted age group was not clear and representative, resulting in a wide age span for a single book due to the fact that the Goodreads data sets only includes data that corresponds to one of the three stated age groups and do not include targeted age property. For instance the age range of a book that is included in the children data set is 0 to 12 years old. Moreover, The Wikisource data set contains a recommended age field, however it reflects a range beginning with one year and has no upper limit. e.g. A children's book can have this attribute set to 8+, however this is just a starting point and may mislead researchers because an adult reader is older than 8 years old.

As a solution to such limitations, More work has to be done to filter out strange numbers from data sets, validate data, and discover a better technique to tighten the desired age span of a book. Knowing that some validations have already been performed; however, due to time constraints, this step can not indeed be finished.

6.2 Data collection and integration limitations

Finding the right data sources was a challenge knowing the face that reviews of books that have been rated by kids are infrequent due to children's privacy protection procedures. Additionally, The provided ratings in the Goodreads data sets are provided by adults, making them irrelevant for this study. Another issue in this matter is, that due to the possibility of each edition of a book having its unique ISBN, data integration between several data sets from various sources is not always as seamless as anticipated. Moreover, We were unable to combine the Wikisource data collection with the Goodreads data sets since it lacked ISBNs or strong Identifiers.

This highlights the need to support the exported data with unique identifiers such as ISBNs, and to find a better way of collecting consented ratings from children.

6.3 Future work

Aside from the possible solutions we suggested based on the study's limitations, it is vital to offer some future work proposals. These concepts formed from the observations made throughout this research. Researchers may seek for similarities and identify reoccurring patterns in books from the same series. We also strongly advise finding a method for reducing the age range in the analyzed samples and reapplying the statistical tests indicated in this work. Finally, we recommend looking into other factors and qualities to enhance the conclusions of this study with other traits. These characteristics may also be derived from the book's metadata, such as the following: lines per page, publishing country, weight, physical dimensions, edition count, number of available languages, and suggested price. According to the findings and observations of the study, young people own 10% more ebooks than children. Furthermore, the number of pages in a book grows considerably as children grow into young adults, however, there is no conclusive evidence that this pattern continues as adults. Additionally, young adult readers read longer books than typical adult readers. Moreover, the majority of young adult books appear to have been written around 1950, whereas children appear to read books published considerably earlier than young adult literature. Likewise, the majority of the literature on young adults was published between 2009 and 2015. There is also a significant gap in publication dates around the time of World War One. We were unable to find a significant correlation between reader age and the number of chapters, cover type, or the fact that a book is part of a series. We hope that the designers of children's book recommender systems would take them into account, which we expect would enhance their recommenders and hence the quality of the recommended books, which will have a positive influence on the children themselves.

References

[1] Among many U.S. children, reading for fun has become less common, federal data shows. https://www.pewresearch.org/fact-tank/2021/11/12/among-many-u-s-children-reading-for-fun-has-become-less-common-federal-data-shows/, nov 12 2021. [Online; accessed 2023-01-16].

[2] Md. Obaidullah and Molla Azizur Rahman. The impact of internet and social media on the habit of reading books: A case study in the southern region of Bangladesh. *Studies in English Language and Education*, 5(1):25–39, mar 1 2018.

[3] Yashar Deldjoo, Cristina Frà, Massimo Valla, Antonio Paladini, Davide Anghileri, Mustafa Tuncel, Franca Garzotto, and Paolo Cremonesi. Enhancing children's experience with recommendation systems. 08 2017.

[4] Ashlee Milton, Levesson Batista, Garrett Allen, Siqi Gao, Yiu-Kai D Ng, and Maria Soledad Pera. "don't judge a book by its cover": Exploring book traits children favor. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 669–674, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Adam C. Payne, Grover J. Whitehurst, and Andrea L. Angell. The role of home literacy environment in the development of language ability in preschool children from low-income families. *Early Childhood Research Quarterly*, 9(3-4):427–440, 1 1994.

[6] Mingming Zhang, Guanhua Hou, Yeh-Cheng Chen, Tao Zhang, and Jie Yang. A Book Interaction Scheme to Enhance Children's Reading Experiences and Preferences. *Frontiers in Psychology*, 11, oct 14 2020.

[7] Kathleen A. J. Mohr. Children's Choices for Recreational Reading: A Three-Part Investigation of Selection Preferences, Rationales, and Processes. *Journal of Literacy Research*, 38(1):81–104, 3 2006.

[8] Ucsd Book Graph. sites.google.com/eng.ucsd.edu/ucsdbookgraph. [Online; accessed 2023-01-18].

[9] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics, 2019.

[10] Mengting Wan and Julian J. McAuley. Item recommendation on monotonic behavior chains. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM, 2018.

[11] Wikisource:book sources - Wikisource, the free online library. https://en.wikisource.org/wiki/Wikisource:Book_sources. [Online; accessed 2023-01-20].

[12] Daniel J. Benjamin and James O. Berger. Three Recommendations for Improving the Use of p-Values. *The American Statistician*, 73(sup1):186–191, mar 20 2019.

[13] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, mar 10 2012.

[14] Libguides: Children's Literature: Genres in children's literature. https://mie-ie.libguides.com/ChildrensLiterature/genre, nov 29 2022. [Online; accessed 2023-01-08].

[15] Kit Kittelstad M.A. Education. List of Children's Book Types. https://reference.yourdictionary.com/reference/books-literature/list-of-children-s-book-types.html. [Online; accessed 2023-01-08].

[16] How Long Should A Children's Book Be? Word Counts per Age. https://self-publishingschool.com/childrens-books-word-count/, sep 30 2021. [Online; accessed 2023-01-08].

[17] Understanding Children's Book Classifications. https://www.authorlearningcenter.com/writing/children-s-books/w/age-groups/6212/understanding-children-s-book-classifications—article. [Online; accessed 2023-01-08].

[18] Chapter Books and Middle Grade Books: What's the Difference? https://prowritingaid.com/art/1319/differences-between-chapter-books-and-middle-grade-books.aspx. [Online; accessed 2023-01-28].

[19] Five reasons kids love a good series | Better Reading. https://www.betterreading.com.au/kids-ya/five-reasons-kids-love-a-good-series/, may 9 2018. [Online; accessed 2023-01-22].

[20] Reading Rewards and Chris Brogan. The 5 Great Advantages of Book Series for Kids. https://www.reading-rewards.com/blog/the-5-great-advantages-of-book-series-for-kids/, apr 11 2019. [Online; accessed 2023-01-22].

[21] Hardcover vs. Paperback: What Sells More Copies? - Letter Review. https://letterreview.com/hardcover-vs-paperback-what-sells-more-copies/, jul 20 2022. [Online; accessed 2023-01-09].

[22] Should I Get Hardcover or Paperback Children's Books? - Books for Children. https://www.books-for-children.com/articles/should-i-get-hardcover-or-paperback-childrens-book/. [Online; accessed 2023-01-09].

[23] Eevi . Everything You Need To Know About Hardcovers For Your Children's Books - EeviJones. http://www.eevijones.com/hardcovers-for-childrens-books/, aug 23 2020. [Online; accessed 2023-01-17].