Datasheet for Children's Book Ratings*

Aliyah Maxine Ramos

25 March 2023

This datasheet describes Alex Cookson's data on Children's Book Ratings.

1 Introduction

The dataset of interest is Alex Cookson's data on the ratings of children's books (Cookson 2020a). For the datasheet of this dataset, I follow the recommendations of Gebru (Gebru et al. 2021).

2 Full Datasheet for Children's Book Ratings

2.1 Motivation

- **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.**
  - The dataset was created to address the issue of inconsistency on rating sites. This problem arises when different rating platforms, such as Rotten Tomatoes or Amazon, have identical ratings but have varying implications, causing an uncertainty in which rating to trust. It was also created to estimate more reasonable ratings, specifically for children's books, by using empirical Bayes estimation.
- **Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?**
  - Alex Cookson, while working at the Royal Canadian Mint.

*Code and data are available at: https://github.com/mxnrms/Tutorial_10.git

- **Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.**
  - No direct funding was received for this project.
- **Any other comments?**
  - No.

2.2 Composition

- **What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.**
  - The dataset consists of textual documents. Each row of the overall dataset is a children's book that includes various information about that book.
- **How many instances are there in total (of each type, if appropriate)?**
  - There are 9241 instances.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).**
  - The dataset specifies that it approximately 9000 children's books that have been rated from one to five stars. It is not known whether the dataset contains all the possible instances of children's books that have received a 1-5 star rating.
- **What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.**
  - Each instance consists of a ten-digit International Standard Book Number (ISBN), the book title, the author, the number of individuals that gave the book a rating, and the ratings themselves. The "raw" data consists of the features mentioned previously, as well as unadjusted ratings and some metadata, such as cover type and publisher.
- **Is there a label or target associated with each instance? If so, please provide a description.**

  - No.

- **Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.**
  - No.
- **Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.**
  - Yes, through the unique ten-digit ISBN.
- **Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.**
  - No.
- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.**
  - No.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.**
  - The dataset is linked to a blogpost on Alex Cookson's website, "Rating children's books with empirical Bayes estimation (1 of 2)" (Cookson 2020b). There are no restrictions associated with the external sour.
- **Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.**
  - No, all data were gathered from public sources.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.**
  - No.
- **Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.**
  - No.
- **Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.**
  - It is not possible to identify the individuals who have rated the children's books.
- **Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.**
  - No.
- **Any other comments?**
  - No.

2.3 Collection Process

- **How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.**
  - The dataset is a publicly available dataset.
- **What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?**
  - Web scraping using R. The data was also imported using R.

- **If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**
  - It is not made known whether the dataset is a sample.
- **Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?**
  - Alex Cookson.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.**
  - The timeframe of which the data was collected is not made known.
- **Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.**
  - No.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?**
  - Third parties.
- **Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.**
  - Likely no.
- **Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.**
  - Likely no.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).**

  - Consent was not obtained.

- **Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.**
  - Some forms of statistical bias were noted, such as self-selection bias, when giving a book a rating.
- **Any other comments?**
  - No.

2.4 Preprocessing/Cleaning/Labeling

- **Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.**
  - Yes, the dataset went through cleaning and labelling of relevant variables.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.**
  - The "raw" data was saved in addition to the cleaned data and can be accessed through GitHub.
- **Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.**
  - R was used.
- **Any other comments?**
  - No.

2.5 Uses

- **Has the dataset been used for any tasks already? If so, please provide a description.**
  - The dataset has been used for a blogpost on Alex Cookson's website (Cookson 2020b).

- **Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.**
  - No.
- **What (other) tasks could the dataset be used for?**
  - The dataset could be used for any personal uses.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?**
  - No.
- **Are there tasks for which the dataset should not be used? If so, please provide a description.**
  - No.
- **Any other comments?**
  - No.

2.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.**
  - The dataset is available publicly through GitHub.
- **How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?**
  - The dataset is currently distributed through GitHub.
- **When will the dataset be distributed?**
  - The dataset is available now.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.**

  - No.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.**
  - None that are known.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.**
  - None that are known.
- **Any other comments?**
  - No.

2.7 Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  - Alex Cookson.
- **How can the owner/curator/manager of the dataset be contacted (for example, email address)?**
  - Alex can be contacted through Twitter, where the link is available on his website (Cookson 2020b).
- **Is there an erratum? If so, please provide a link or other access point.**
  - No.
- **Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?**
  - No current plan for updating.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.**

  - No.

- **Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.**
  - No.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.**
  - Pull request on GitHub.
- **Any other comments?**
  - No.

References

Cookson, Alex. 2020a. "Children's Book Ratings." https://github.com/tacookson/data/tree/master/childrens-book-ratings.

———. 2020b. "Rating Children's Books with Empirical Bayes Estimation (1 of 2)." https://www.alexcookson.com/post/rating-childrens-books-with-empirical-bayes-estimation/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.