

Datasheet Usage and Quality Analysis*

Eshta Bhardwaj

April 21 2023

There has been a prominent uptake of datasheets for datasets in machine learning research to increase transparency in the dataset development process. Using the datasheets published in the NeurIPS 2021 datasets and benchmarks track, I analyze whether the use of datasheets can aid in mitigating risks inherent within large datasets that contribute to biased machine learning models. Although, datasheets are now being used when producing new datasets, there is lack of reflective practice demonstrated while completing these datasheets. In the paper, I discuss how the use of datasheets can be better applied and how datasheets can be improved for better adoption by practitioners.

Table of contents

1	Introduction	3
2	Data Collection	4
2.1	Missing Data	5
2.2	Data Preparation and Cleaning	5
3	Methodology	6
3.1	Overview	6
3.2	Data Processing	7
4	Results	8
5	Discussion	8
5.1	Findings {sec-findings}	8
5.2	Insights	8
5.3	Limitations	8

*Code and data available at: https://github.com/eshtab/datasheet_quality_analysis

5.4	Challenges	9
6	Conclusion	9
6.1	Future Work	9
	References	10

1 Introduction

Predictive algorithms can operate as black boxes and cause widespread harm to the subjects of the model (O’Neil 2017). Data scientists, data engineers, annotators, and other practitioners have ethical responsibilities to help mitigate bias and unfairness. Machine learning research (MLR) has pinpointed the data underlying predictive models to be the largest contributor in introducing bias (Paullada et al. 2021; Sambasivan et al. 2021; Scheuerman, Hanna, and Denton 2021).

Recent publications have identified the importance of prioritizing “data work” as a key issue in machine learning research (Sambasivan et al. 2021). Data work in this context refers to performing data tasks, investigating data quality, applying frameworks around data practices and the preparation of data prior to its use within a model. Sambasivan et al. argue that ignoring data work leads to data cascades which are “compounding events causing negative, downstream effects from data issues, resulting in technical debt over time” (Sambasivan et al. 2021). Data work therefore enables increased focus towards stewardship, quality, accountability, and transparency. Bender et al. emphasize that documentation of data collection practices can aid in mitigating risks inherent within large, biased ML models (Bender et al. 2021). Additionally, the documentation should include the researcher’s positionality and motivation in developing the model and potential risks to the users and stakeholders (Bender et al. 2021).

Emerging research has started to address this lack of data work with the introduction of context documents. Context documents provide documentation for datasets or machine learning (ML) models by detailing aspects of provenance and data collection and are particularly geared to answering ethical questions about the data (Boyd 2021). One of the most popularly used context documents is datasheets. Datasheets provide documentation for machine learning datasets by addressing the needs of two primary user groups: dataset creators and dataset consumers (Gebru et al. 2021). For the creators, it facilitates reflection on the processes of data creation, distribution, and maintenance and allows them to highlight assumptions and potential risks. While consumers benefit from this documentation because it provides the transparency required to make key decisions.

Since the original publication of Datasheets for Datasets in 2018, there has been large uptake of its usage by researchers and practitioners (Gebru et al. 2018). In fact, various forms of context documentation have since emerged. For example, data statements for natural language processing (NLP) datasets contain specifications on demographic information about the dataset annotator, quality of the dataset, provenance, etc. (Bender and Friedman 2018). Similarly, an AI fairness checklist was developed to aid practitioners by providing a structured framework for identifying and addressing issues within their projects (Madaio et al. 2020). Another context document proposed in recent years is model cards which aim to “standardize ethical practice and reporting” within ML models. Model cards include details about the models, their intended use, impacts of the model on the real-world, evaluation data, details on the training data, and ethical considerations (Mitchell et al. 2019). On the other hand, explainability fact sheets are used for similar documentation but are specifically geared towards

the method applied in a predictive model. Therefore, the fact sheet contains an evaluation of the method’s functional and operational requirements, the criteria used for the evaluation, any security, privacy or other vulnerabilities that may be introduced by the method, and the results of this evaluation (Sokol and Flach 2020). However, a review investigating the quality of such context documents remains to be performed.

In this paper, I review 21 datasheets published as part of the papers in the 2021 NeurIPS datasets and benchmarks track. An analysis of these datasheets is performed to analyze how practitioners and researchers fill the datasheets, what areas they choose to focus on, how they answer the questions to determine their level of reflection while completing these datasheets. This review will therefore contribute in 2 ways: 1) it will provide a novel dataset on datasheet quality and 2) it will provide a summary on how practitioners approach the completion of a datasheet.

The remainder of the paper is structured as follows. In Section 2, I discuss details about the data source, provenance, variables, and important ethical implications and biases present within the data collection process. In Section 3, I discuss the methods used to analyze the datasheets, specifically looking at what and how the text analysis is performed. In Section 4, the results of the analysis are presented and subsequently discussed in Section 5 along with a review of limitations of the analysis performed. Section 6 summarizes the paper with a look at potential future work.

2 Data Collection

The source of the datasheets analyzed in this paper is the 2021 NeurIPS datasets and benchmarks track. 2021 was the first year of this track which was introduced to provide datasets to help the NeurIPS community evaluate their algorithms (Vanschoren and Yeung 2021). Given that the NeurIPS community focuses on algorithmic design, the lack of documentations for datasets used for evaluation prevented reuse or rather created situations of inappropriate reuse, where datasets were not fit for the tasks they were applied to. The lack of evaluation of representativeness further added to the misuse of existing datasets and thereby contributed to biased models and biased results (Vanschoren and Yeung 2021).

The papers, although published in NeurIPS2021, did not always have the datasheets available on the publication website. As such, the datasheets were sourced in 1 of 3 ways. Firstly, the supplemental section of each paper was downloaded to check whether the author uploaded an appendix or datasheet file ([link](#)). If there was no file found, then the arxiv version of the publication was sourced which often included the datasheet within an appendix or linked to on another website (like github). In some cases, there was no link or attached datasheet within the publication but it was mentioned in the paper as being completed. In this case, a general search was performed to retrieve any github, personal website, or other source where the datasheet was published.

2.1 Missing Data

There were a total of 174 papers published in the NeurIPS 2021 benchmarks and datasets track. Of the 174 papers, there were 3 main types of contributions: creation of a benchmark model, creation of a dataset, and discussion on the importance of data curation in machine learning dataset development. There were 93 papers that created a new dataset, however only 21 of those papers had a datasheet. It is potentially possible to fill out a datasheet based solely on the information available within the research paper, however that was not done for this analysis.

The publications considered in this analysis are as follows: (will complete this later-todo)

Datasheet Number	Publication Title	Reference
1	A sandbox for prediction and integration of DNA, RNA, and proteins in single cells	(Luecken et al. 2021)
2	Generating Datasets of 3D Garments with Sewing Patterns	(Korosteleva and Lee 2021)
3	PASS An ImageNet replacement for self-supervised pretraining without humans	(Asano et al. 2021)

2.2 Data Preparation and Cleaning

Each of the 21 datasheets was formatted in a unique manner, based on the authors' preferences. This included the use of single or double columns, placing section headers within a box, numbering the questions numerically or alphabetically, and many more. The datasheets were also based on either the 2018 or 2021 version of the publication. In the 2018 version, there was a section on 'Legal and ethical considerations' which was later removed in the 2021 version and a new section called 'Uses' was added. All these differences meant that a certain level of standardization had to be introduced to each datasheet pdf file in order to enable its analysis within R.

To create this standardization, I started by downloading blank datasheet templates for both the 2018 and 2021 versions. To the blank templates, I added 2 types of tags. The first type of tag was [question#]. This tag was placed at the start of each question and the # was replaced for each question (for a total of 50 questions in the 2018 version and 57 questions in the 2021 version). I repeated the same with the answer tag [answer#].

Then, I downloaded all 21 datasheets in a pdf format. For each datasheet, I copied the responses and placed them in the appropriate [answer#] tag. I also removed images, tables, and any other type of content other than text. I changed all section headers to be capitalized (this would also aid in analysis later, as discussed in Section 3.1). Lastly, I added 'No' to

any questions in the datasheet that asked ‘Any other comments’. This was added so that completion of each question could be calculated correctly.

A sample of the original datasheet (direct downloaded from source) can be found here: [link](#)

The standard formatted datasheets can be found here: (yes only 7 now, will have all eventually) [link](#)

3 Methodology

- this work required creating multiple datasets from the sourced datasheet pdfs
- metrics of interest (high level)
 - question completed?
 - overall length of datasheet
 - length of answer for each question
 - score (based on first 3 metrics)
 - * will need to code overall length and length of answer into numeric “grade” between 1-5
 - list of top 10 most common words per question (excluding stopwords)
- todo: add references for R and the packages used

3.1 Overview

- go over each variable for both individual datasheet dataset and summary dataset
 - values, range
- each dataset for datasheet
 - `datasheet_version`
 - `qnum` (ID num between 1 to either 50 or 57 based on datasheet version)
 - `completion` (yes or no for whether the question was completed)
 - `length_words` (response length in words)
 - `top_5_words` (most frequently occurring words in the response (list of 5 words excluding stopwords) but eventually only consider responses that had more than 10 words)
- summary dataset
 - `datasheet_ID` (ID num between 1 to 21)
 - `title_long`
 - `title_short`

- datasheet_version (2018 or 2021)
- total_length_wrds (total response length of all questions)
- question_completion_pct (number of questions with response of any length divided by 50 or 57 questions – unless skipa and skipb value is not related to humans then the questions are counted as complete regardless)
- avg_response_length (avg response length of all questions)
- max_response_length (max response length of all questions)
- max_response_qnum (corresponding question number that had max length)
- min_response_length (min response length (can't be 0) of all questions)
- min_response_qnum (corresponding question number that had min length)
- overall_top_5_words (most common top 5 words in overall datasheet excluding stopwords)
- score

3.2 Data Processing

This section discusses the processing involved in generating each datapoint in the 22 datasets (21 for each datasheet, 1 for summary). This section also specifically references packages used.

- automated cleaning/processing
 - removed rows with blank strings
 - created a column identifying header, question, answer based on the tags (created a loop to go over entire text to do this)
 - total number of words in datasheet (created loop for counting)
- generating datasheet dataset
 - each dataset had 50 or 57 rows based on 2018 or 2021 template
 - completion check (check for length of string after prefix)
 - count # of words for the answer (created loop for counting)
 - list for top 10 most occurring words excluding stopwords (loop plus referenced stopwords package in R)
- summary dataset
 - created dataset by calling various pieces of info from each datasheet dataset
 - added info on ID, title, shortcut title, datasheet version, total words
 - created a question completion % based on # of questions completed out of 50 or 57 based on 2018 vs 2021
 - avg, min and max length of answers for each datasheet
 - if it makes (check for most common words from the top 10 list for each question - if its too generic, omit) this column in dataset)

- overall score (average calculated based on % completed and avg, min, max length of responses)

4 Results

The 22 datasets can be found here (yes, currently only 7 datasheets but will eventually be 21) [link](#).

5 Discussion

5.1 Findings {sec-findings}

- discuss findings from results
- note: is it possible to tell that datasheet1 didn't have a good maintenance section based on data alone?
- note: datasheet 2,5 say it has no repo which inherently means that not enough consideration was given to documentation and reusability- is this reflected in their overall automated score?

5.2 Insights

- inherent missing things within datasheets such as consideration of limited consideration of FAIR principles, ethical considerations, threats to validity, environmental and financial footprint, domain knowledge, context awareness,
- github is not a good way to preserve/maintain data. eg. the actual 2021 datasheet paper that links to a github repo for a sample datasheet doesn't work anymore

5.3 Limitations

- risky to calculate "score" to investigate the 2 research questions proposed
- original 2018 paper had ethical and legal section and no uses section which is in 2021 paper which makes comparison problematic
- automated completion considers questions like any other comments which are optional
- authors often link to other sections of their appendix or attached supplementary material which isn't considered in this analysis
- go over potential issues with the calculation of each variable
- length of response doesn't indicate quality documentation
- some datasheets added their own questions (in some cases that was the purpose of their paper and in other cases, it was unknown - ie datasheet3)

5.4 Challenges

- everyone co-opts the datasheet in their own format (to circumvent this i did manual recoding of the datasheets but that can introduce error)
- some datasheets paraphrased the questions or instead of leaving questions blank, completely removed them altogether (ie datasheet14) - another reason why it was important to manually standardize the texts into one format

6 Conclusion

6.1 Future Work

- analyze 2022 neurips papers
- write datasheets for the generated papers and analyze them
- for the 21 sheets analyzed here, do a manual coding and analysis and compare with automated
- usage of datasheets has to be standardized, and guidelines have to enforce standard formatting to ensure that future analyses can be performed + also retire the 2018 version

References

- Asano, Yuki M, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. 2021. “PASS: An ImageNet Replacement for Self-Supervised Pretraining Without Humans.” In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.
- Bender, Emily M., and Batya Friedman. 2018. “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science” 6: 587–604. https://doi.org/10.1162/tacl_a_00041.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. ACM. <https://doi.org/10.1145/3442188.3445922>.
- Boyd, Karen L. 2021. “Datasheets for Datasets Help ML Engineers Notice and Understand Ethical Issues in Training Data.” *Proceedings of the ACM on Human-Computer Interaction* 5: 1–27. <https://doi.org/10.1145/3479582>.
- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. “Datasheets for Datasets,” no. arXiv:1803.09010. <http://arxiv.org/abs/1803.09010>.
- 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. <https://doi.org/10.1145/3458723>.
- Korosteleva, Maria, and Sung-Hee Lee. 2021. “Generating Datasets of 3D Garments with Sewing Patterns.” In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.
- Luecken, Malte D, Daniel B Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, et al. 2021. “A Sandbox for Prediction and Integration of DNA, RNA, and Protein Data in Single Cells.” In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.
- Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. “Co-Designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. ACM. <https://doi.org/10.1145/3313831.3376445>.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. “Model Cards for Model Reporting.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–29. ACM. <https://doi.org/10.1145/3287560.3287596>.
- O’Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.
- Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. “Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research.” *Patterns* 2 (11): 100336. <https://doi.org/10.1016/j.patter.2021.100336>.

- Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “‘Everyone Wants to Do the Model Work, Not the Data Work’: Data Cascades in High-Stakes AI.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. ACM. <https://doi.org/10.1145/3411764.3445518>.
- Scheuerman, Morgan Klaus, Alex Hanna, and Emily Denton. 2021. “Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development.” *Proceedings of the ACM on Human-Computer Interaction* 5: 1–37. <https://doi.org/10.1145/3476058>.
- Sokol, Kacper, and Peter Flach. 2020. “Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches.” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. ACM. <https://doi.org/10.1145/3351095.3372870>.
- Vanschoren, Joaquin, and Serena Yeung. 2021. “Announcing the NeurIPS 2021 Datasets and Benchmarks Track.” *Medium*. <https://neuripsconf.medium.com/announcing-the-neurips-2021-datasets-and-benchmarks-track-644e27c1e66c>.