# Datasheet Usage and Quality Analysis*

Eshta Bhardwaj

April 20 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



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

# 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

- data sourced by downloading in 1 of 4 ways: supplemental section from neurips site, sourced from arxiv, github, or links provided (either in arxiv paper or supplemental section of paper). the actual neurips paper didn't have it attached/provided
- discuss: an overview of how neurips has exploded in size in the past 10 yrs, its influence on the industry, where neurips fits in this ecosystem and the importance of good datasheets there

## 2.1 Missing Data

- discussion on missing datasheets from research papers that created datasets

## 2.2 Data Preparation and Cleaning

- manual cleaning:
  - created blank datasheet templates for both 2018 and 2021
    - then filled datasheets from their original formatting to standardized formatting

        - this was needed b/c all datasheets used their own formatting like starting the answer right after the question instead of on a new line, 2 column layout, or had other sections/appendices etc before/after it. to be able to automate any type of analysis, the data had to be standardized in one format

    * made it easier to identify diff elements of a datasheet (header, question, answer)
        - numbered the questions
        - added A. prefix for answers
        - made the section headers all caps

# 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)

## 3.1 Overview

- go over each variable for both individual datasheet dataset and summary dataset
    - values, range
- each dataset for datasheet
    - q_number (ID num between 1 to either 50 or 57 based on datasheet version)
    - completed (yes or no for whether the question was completed)
    - length (response length in words)
    - top_10_words (most frequently occuring words in the response (list of 10 words excluding stopwords))
- summary dataset
    - datasheet_ID (ID num between 1 to 21)
    - paper_title_full
    - paper_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 value is not related to humans then the denominator is diff)
  - 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_10_words (most common top 10 words in overall datasheet excluding stopwords)

## 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 (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 occuring 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 (avg of per question score from datasheet dataset)

# 4 Results

The 22 datasets can be found here 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?

## 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

## 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)

# 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

## References


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.

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.