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

[*] 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

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 documentation 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 absence 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).

2.1 Overview

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 publicaton was sourced which often included the datasheet within an appendix or linked to another website (such as 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.

2.2 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:

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)
4	Constructing a Visual Dataset to Study the Effects of Spatial Apartheid in South Africa	(Sefala et al. 2021)
5	CSFCube - A Test Collection of Computer Science Research Articles for Faceted Query by Example	(Mysore et al. 2021)
6	Datasets for Online Controlled Experiments	(Liu et al. 2022)
7	PROCAT Product Catalogue Dataset for Implicit Clustering, Permutation Learning and Structure Prediction	(Jurewicz and Derczynski 2021)

Datasheet Number	Publication Title	Reference
8	OmniPrint A Configurable Printed Character Synthesizer	(Sun, Tu, and Guyon 2021)
9	Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research	(Koch et al. 2021)
10	WildfireDB An Open-Source Dataset Connecting Wildfire Occurrence with Relevant Determinants	(Singla et al. 2021)
11	Addressing Documentation Debt in Machine Learning	(Bandy and Vincent 2021)
12	Artsheets for Art Datasets	(Srinivasan et al. 2021)
13	Benchmarking Multimodal AutoML for Tabular Data with Text Fields	(Shi et al. 2021)
14	CREAK A Dataset for Commonsense Reasoning over Entity Knowledge	(Onoe et al. 2021)
15	CSAW-M An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer	(Sorkhei et al. 2021)
16	HumBugDB A Large-scale Acoustic Mosquito Dataset	(Kiskin et al. 2021)
17	Modeling Worlds in Text	(Ammanabrolu and Riedl 2021)
18	Multilingual Spoken Words Corpus	(Mazumder et al. 2021)
19	RedCaps Web-curated image-text data created by the people, for the people	(Desai et al. 2021)
20	The CPD Data Set Personnel, Use of Force, and Complaints in the Chicago Police Department	(Horel et al. 2021)
21	Towards a robust experimental framework and benchmark for lifelong language learning	(Hussain et al. 2021)

2.3 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. In some cases, the datasheet was created as a standalone document and in other cases it was part of the appendix of the original paper which meant it had text preceding and following the datasheet content. 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 statistical programming language (R Core Team 2020).

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 an original datasheet (direct download from source) can be found here: link

All the standard formatted datasheets can be found here: link

3 Methodology

The standardized and consistently formatted datasheets were used to create 2 types of datasets. The first type of dataset contained metrics about each individual datasheet while the second type was a summary of all 21 datasheets. Overall, these datasets helped in examining the following metrics of interest:

- Question completion (any length of response for each question and overall completion)
- Length of datasheet (total length in words, excluding the questions themselves)
- Length of response (length of each response in words)
- Frequently used words (most frequent words in the datasheets excluding stopwords)
- Score (calculated based on total datasheet length, average length of responses, and question completion rate)

In the following sections, the variables and the processing involved to generate them are discussed in detail.

3.1 Overview

Given that 21 datasheets were analyzed in this paper, there were a total of 21 datasets generated (one for each datasheet) as well as an additional summary dataset for a total of 22 datasets. These datasets served as the input for the textual analysis performed in R statistical programming language (R Core Team 2020).

In each of the datasheet datasets, the following variables were created:

- **datasheet_version**: Either 2018 or 2021, based on the template used
- **qnum**: An ID number assigned between 1-50 or 1-57 based on the datasheet version (2018 or 2021, respectively)
- **completion**: Binary 'Yes' or 'No' variable to indicate if a response had been provided to the question
- **length_words**: Total number of words of the response
- **top_5_frequent_words**: The top 5 most frequent words for each question response, left blank if incomplete

In the summary dataset, the following variables were created:

- **datasheet_ID**: An ID number assigned between 1-21
- **datasheet_version**: Either 2018 or 2021, based on the template used
- **title_short**: A short keyword to identify each datasheet
- **total_length_wrds**: Total length of all responses in the datasheet, in words
- **question_completion_pct**: The percentage of questions with responses in the datasheet
- **avg_response_length**: Average response length of all questions, in words
- **max_response_length**: Maximum response length of all questions, in words
- **max_response_qnum**: Corresponding question number with maximum response length
- **min_response_length**: Minimum response length of all questions, in words
- **min_response_qnum**: Corresponding question number with minimum response length
- **overall_top_5_words**: Overall top 5 words from each datasheet
- **score_pct**: Score (in percentage) calculated using **total_length_wrds**, **question_completion_pct**, and **avg_response_length**

3.2 Data Processing

The processing involved for creating each variable is listed below. Packages that were used include: tidyverse (for various data related functions) (Wickham et al. 2019), pdftools (for extracting text from pdf files) (Ooms 2023), vctrs (for vector manipulation) (Wickham, Henry, and Vaughan 2023), reshape2 (for the melt function which was used to reshape dataframes) (Wickham 2007), stopwords (for filtering out stopwords) (Benoit, Muhr, and Watanabe 2021), and randomWords (for generating lists of random strings) (Heckmann 2019).

To start with, each datasheet in pdf format was read in and every new line was separated into a row of a dataframe. Blank rows due to spaces were deleted. The tags that were added in during the manual coding phase were then used to identify each line in the dataframe as a section header, question, or answer.

To create the datasheets' datasets, the following variables were developed:

- **datasheet_version**: Manually coded as 2018 or 2021
- **qnum**: An ID number was generated by looping through rows of each datasheet
- **completion**: The completion was recorded as 'Yes' if the response had a length of words greater than 1. There were also 2 questions in the 2021 version of the datasheet which stated "If the dataset does not relate to people, you may skip the remaining questions in this section.". The responses for these 2 questions were recorded and ensured that if the authors answered the questions could be skipped, then their answer for those questions would still be considered complete.
- **length_words**: The length of words was calculated by looping through the datasheet and adding the length of the responses for each question.
- **top_5_frequent_words**: The top 5 most frequent words for each response was derived by generating the frequencies for all the words in the datasheets and then filtering out question and answer tags, the top 400 stopwords, and keeping only the 5 most frequent words per question.

To create the summary dataset, the following variables were developed:

- **datasheet_ID**: An ID number was assigned between 1-21
- **datasheet_version**: The datasheet version was retrieved from each datasheet's individual dataset
- **title_short**: A short keyword was manually coded for each datasheet
- **total_length_wrds**: Total length of all responses was calculated by summing the lengths of each response as per the datasheet datasets.
- **question_completion_pct**: The completion percentage was calculated by counting the number of questions considered answered (from the datasheet dataset) with the denominator as 50 or 57 questions based on the template used.
- **avg_response_length**: The average length of a response was calculated as the mean of the lengths of each response as per the datasheet datasets.
- **max_response_length**: The maximum length of any given response for each datasheet was retrieved from the datasheet datasets.
- **max_response_qnum**: The corresponding question number with the maximum response length was retrieved by matching the value of the response length with the question number from the datasheet datasets.
- **min_response_length**: The minimum length of any given response for each datasheet was retrieved from the datasheet datasets.
- **min_response_qnum**: The corresponding question number with the minimum response length was retrieved by matching the value of the response length with the question number from the datasheet datasets. However, in many cases, this returned more than 1 question number. These were then formed into a list type.
- **overall_top_5_words**: The overall top 5 words for each datasheet were generated by combining all the top 5 words from each question in a given datasheet into one list (along

with their frequencies). However this list excluded responses that had 10 or less words. Then the 5 words with the highest frequencies were considered as the most overall 5 most used words.

- **score_pct**: The score for each datasheet was calculated by giving a numbered score between 0-5 to **total_length_wrds**, **question_completion_pct**, and **avg_response_length**. The score for total length of a datasheet was 0 for having less than 1000 words, 1 for 1001-1500 words, 2 for 1501-2000 words, 3 for 2001-2500 words, 4 for 2501-3000 words, and 5 for over 3000 words. The score for question completion was 1 if the completion percentage was between 0-50, 2 for 51-75, 3 for 76-85, 4 for 86-95, and 5 for above 95%. The score for average length of the response was 1 for 0-20 words, 2 for 21-30 words, 3 for 31-40 words, 4 for 41-50 words, and 5 for over 50 words. Each of these scores had an equal weight in determining the overall score for each datasheet.

4 Results

The 22 datasets can be found here link. These datasets were used to generate the metrics discussed in Section 3.

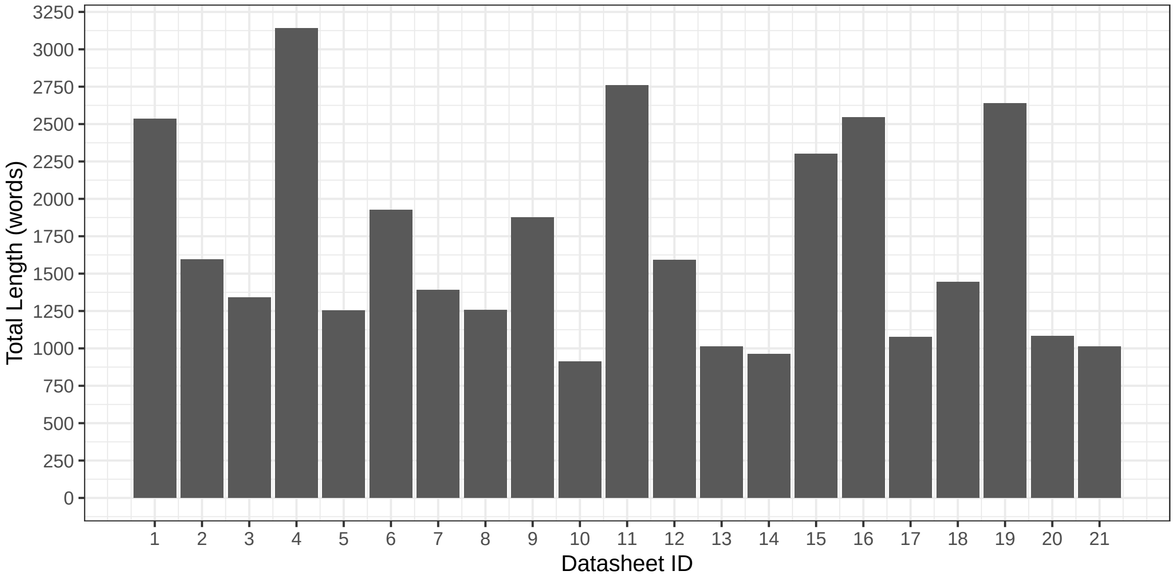

Figure 1: Total Length (words) of 21 Datasheets

Figure 1 shows each datasheet's total length in words (excluding questions and section headers). Out of 21 datasheets, 2 datasheets had less than 1000 words, 13 datasheets had 1000-2000 words, 5 datasheets had between 2000-3000 words and only 1 datasheet had more 3000 words.

The three datasheets with the most words are datasheets 4, 11, and 19 while the three with the least words are 10, 13, and 14.

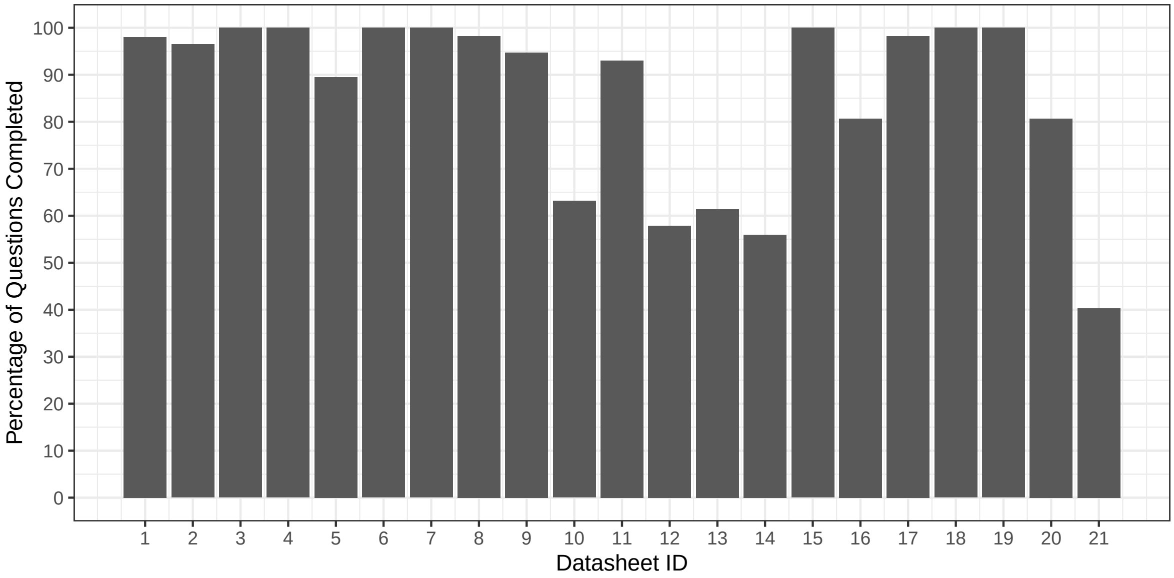

Figure 2: Datasheet Completion Percentage

In Figure 2, the percentage of question completion is demonstrated. Most datasheets had completion of between 80-100%. A few datasheets had completion near 60% and only one datasheet under 50% (datasheet 21). Seven datasheets had full completion - 3, 4, 6, 7, 15, 18, and 19. None of the datasheets with the lowest number of words had a full completion rate indicating that authors that fully filled out the datasheet were likely to do so in greater detail.

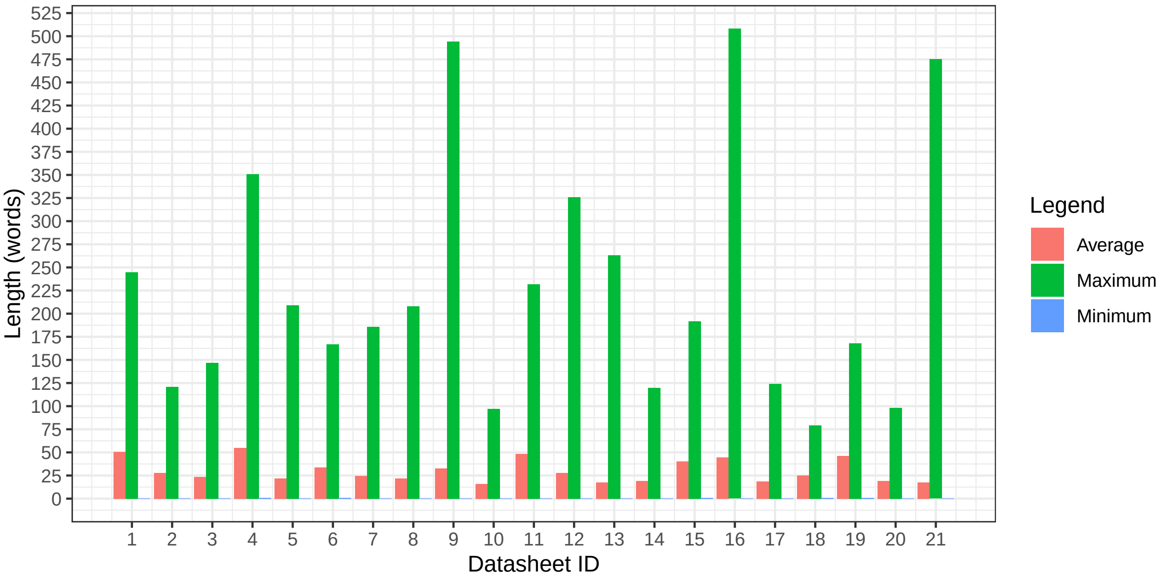

Figure 3: Average, Maximum, and Minimum Length of Responses

In Figure 3, three metrics are shown - average response length in words, maximum response length in words, and minimum length in words. A third of the responses had an average length of 20-30 words, roughly another third of the responses had an average length of 30-40 words, only 2 datasheets had average response length over 50 words and the remaining were under 20 words. Five datasheets (4, 9, 12, 16, 21) had the highest maximum word count ranging from over 325 words to over 500 words. Datasheet 21 with the lowest completion rate also had 3rd highest maximum response length. The minimum response length was either 0 or 1 word. 0 indicated an incomplete response while 1-word responses indicated answers like ‘No’.

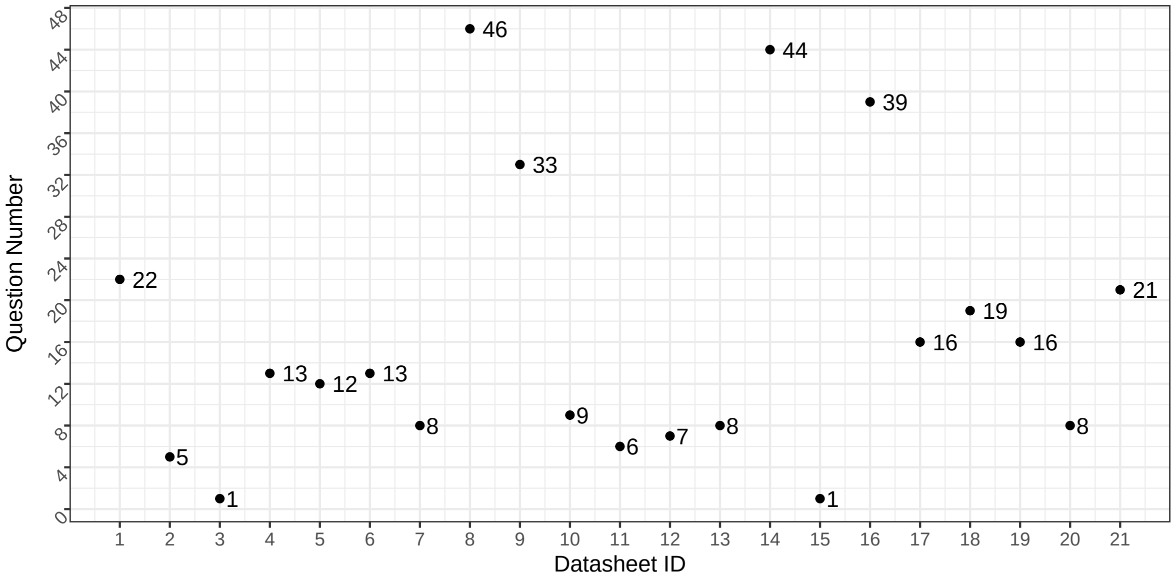

Figure 4: Datasheet Question with Maximum Response Length

Figure 4 shows a scatterplot with the Datasheet ID plotted against the question number to demonstrate the questions with the maximum response length. The labels are the question number. Therefore reading the graph horizontally reveals the question numbers that commonly had the maximum response length. This included questions like 1, 8, 13, and 16.

- **Question 1**: 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.
- **Question 8**: What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.
- **Question 13**: Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
- **Question 16**: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

While the first two questions are standard descriptions included in all documentation of datasets, the second two reflect a deeper reflection on potential negative impacts of datasets.

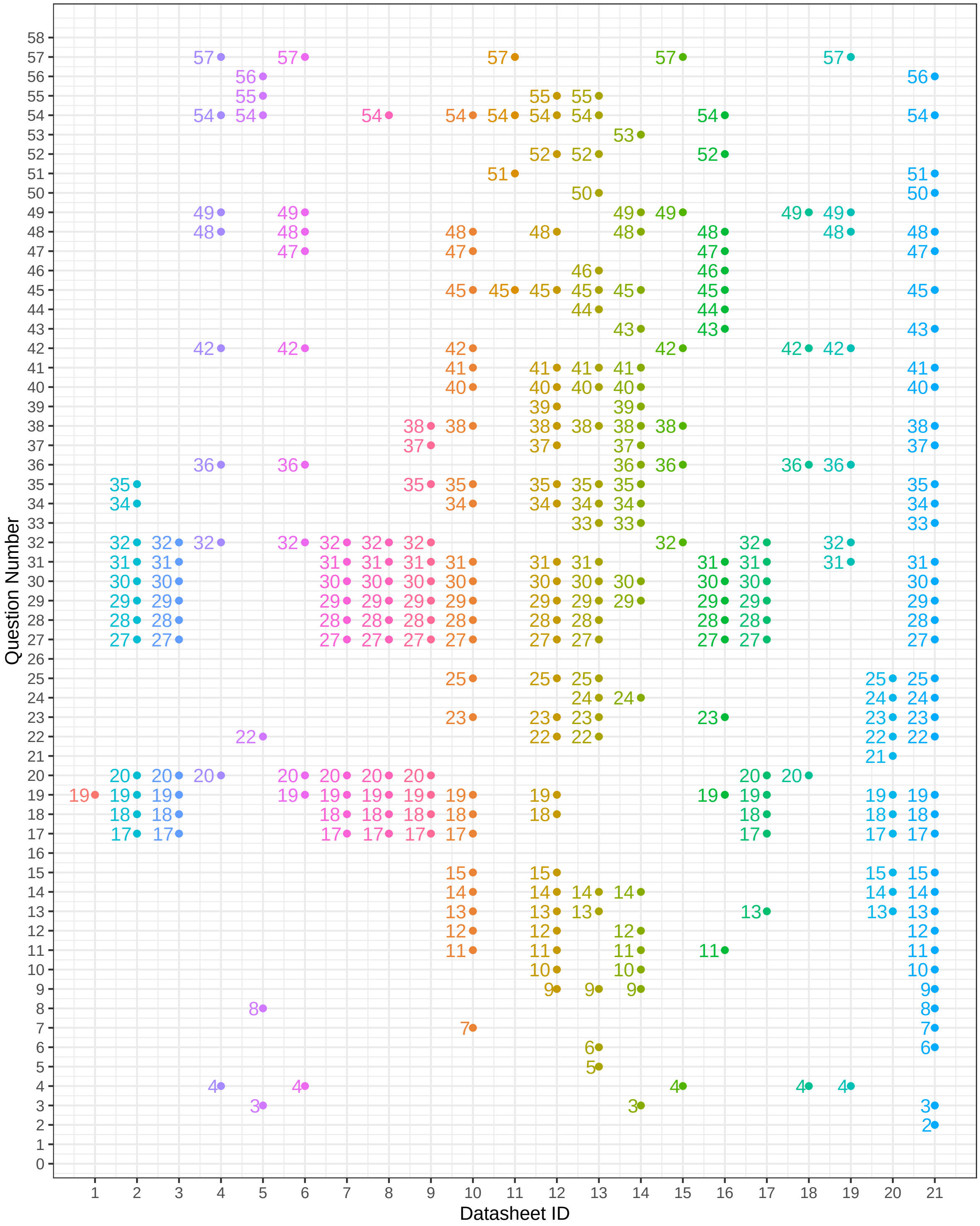

Figure 5: Datasheet Questions with Minimum Response Length

Figure 5 also shows a scatterplot with datasheet ID plotted against question number, however these question numbers represent the responses with the corresponding minimum length. For example, datasheet 1 has a minimum response length of 0 words for question 19. Whereas datasheet 4 has a minimum response length of 1 word for questions 4, 20, 32, 36, 42, 48, 49, 54, and 57. The labels represent the question number (unlike Figure 4 where they represented the datasheet ID). Therefore, reading the graph horizontally shows the question numbers that most commonly had the lowest response lengths across all datasheets, while reading the graph vertically shows the question numbers with the lowest response length for each datasheet.

There were only 4 datasheets that had a minimum response length of 1 word (datasheet 4, 15, 18, 19). The remaining datasheets had a response length of 0, therefore the question completion percentage can also indicate which datasheets had the most incomplete responses.

Reading the graph horizontally shows that the responses that most often had the lowest word count were for questions 18, 19, 27, 28, 29, 30, 31, and 32.

- **Question 18**: 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.
- **Question 19**: 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.
- **Question 27**: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
- **Question 28**: 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.
- **Question 29**: 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.
- **Question 30**: 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).
- **Question 31**: 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.
- **Question 32**: Any other comments?

Questions 18 and 19, however were not blank, but often answered with a 1-word response of

‘No’. Whereas, questions 27-32 often followed the question “If the dataset does not relate to people, you may skip the remaining questions in this section.”. In which case, if the authors skipped the questions, it rendered as a blank answer (although not contributing to their completion percentage).

The responses that had least often had the lowest response length were questions 1, 2, 5, 6, 7, 8, 16, 21, 26, 39, 44, 46, 50, 51, 53, and 56, where questions 1, 16, and 26 never had the lowest response length. For questions 1, 8, and 16, this corresponds to the results seen in Figure 5 because these often had the highest response length. Whereas, for the remaining questions it demonstrates that while these questions did not have the longest responses, they were often answered with an average word count.

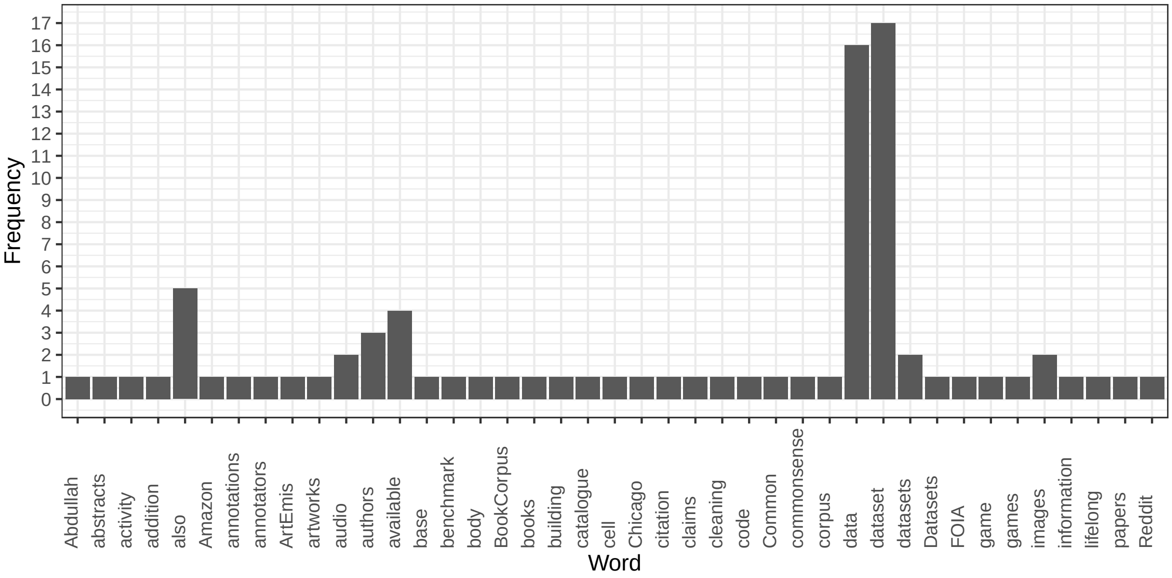

Figure 6: Most frequently used words in the datasheets

The top 5 words across each datasheet contained special characters, numbers, and stopwords that were not filtered out in the earlier preprocessing. In order to further clean and visualize these words, special characters, numbers, and words with 3 or less characters were excluded. Figure 6 shows the remaining top most frequently used words along with their frequencies. Some of the words are still stopwords like ‘also’ while others contain very specific non-corpus related words like ‘Abdullah’ representing an organization name that was often repeated ‘King Abdullah University of Science and Technology’.

The most frequent words used across all datasheets were the words ‘data’ and ‘dataset’. As these are generic words, unspecific to a particular research topic, these were used most often. Other generic words included ‘also’, ‘authors’, ‘available’, and ‘datasets’. Some more specific words used frequently were ‘audio’ and ‘images’.

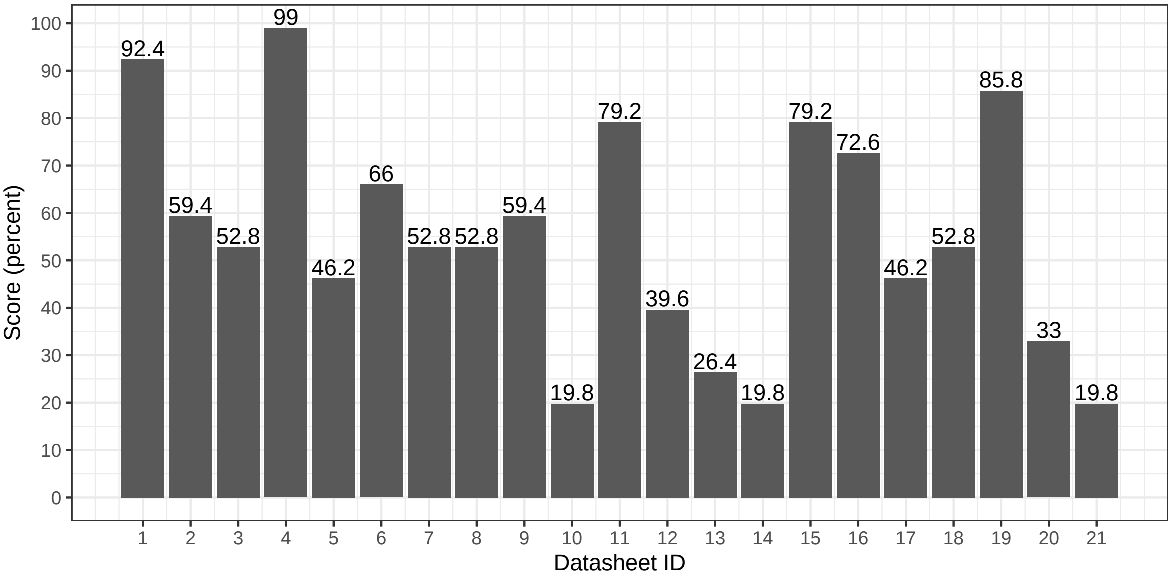

Figure 7: Overall Score for Each Datasheet

Lastly, Figure 7 shows the overall score for each datasheet based on its completion percentage, overall total response length, and average response length. The datasheets with the lowest score was datasheet 10, 14, and 21 with scores less than 20%. While the 2 highest scores of more than 90% were datasheets 1 and 4.

5 Discussion

5.1 Findings

Based on the results from analyzing various metrics regarding the datasheets, it is clear that the current scope of analysis is narrow due to various challenges and limitations. The goal of this analysis was to understand how practitioners and researchers fill the datasheets, what areas they choose to focus on, and how they answer the questions to determine their level of reflection while completing these datasheets. While examining the completion and response length of questions across the datasheets can aid in understanding these aspects, it does not provide a certified, full picture on how practitioners approach the completion of a datasheet. An example of this is that generating a score by summarizing metrics across the entire datasheet can hide information on per question or per section completion. For example, if authors choose to focus on some sections of the datasheet as compared to others, it can heavily influence their score as sections have differing number of questions within them. In the 2021 version, the composition and data collection sections are the largest with 26 questions while the motivation section and processing sections have 8 questions. This raises the doubt whether sections should be examined individually and whether a weighted score should be applied to give some sections more importance than others.

There are also inherent insights, limitations, and challenges that can be gleaned from the methodology of the current analysis, which are discussed further below.

5.2 Insights

Although, datasheets are a strong start for the machine learning community to perform standardized documentation of their dataset creation process, there are missing considerations of important concepts vital in reducing bias and increasing accountability and transparency. This includes limited consideration of the FAIR principles (n.d.). The datasheet contains sections on maintainability and distribution but these can be strengthened by adopting the language and concerns of the FAIR principles. Additionally, in terms of maintenance and distribution, some questions asked whether Github (among other resources) was used for these processes. However, since Github is dependent on its author maintaining the repository, it is not a good method to preserve and store data in the long term. In fact, the 2021 version of the datasheets publication (Gebru et al. 2021) points to a Github repository that no longer exists. Lastly, datasheets need to include more questions concerned with the ethicality of the dataset and the dataset creation process including threats to validity, environmental and financial footprint, whether the creation process required specific domain knowledge and whether its reuse requires similar expertise, and whether the involved individuals documented/demonstrated an awareness of the social, political, historical context of the dataset developed.

5.3 Limitations

There are some limitations to the analysis performed in this paper. To start with, the method of deriving the score that is assigned to each datasheet does not fully capture the researchers' process in completing the datasheet nor can fully be used as an indicator of the datasheet's quality. Primarily, this is because the length of a response does not indicate the quality of the documentation being performed. Furthermore, the score of each datasheet can become problematic if it is used to compare the quality of one datasheet with another or various authors' processes of documentation. For example, comparing 2018 datasheets with 2021 datasheets is an issue because the 2018 version had an 'Ethical and legal' section whereas the 2021 paper did not and had an additional 'Uses' section. Many of the limitations in the analysis were due to the challenges faced in examining varying datasheets, as discussed in the next section.

5.4 Challenges

The primary challenge in evaluating the 21 datasheets was the unique formatting and content used by each set of authors. Each of 21 papers co-opted the datasheet in their own method. Firstly, some of the datasheets had paraphrased questions (i.e., changed the wording of the questions), removed questions altogether, combined multiple questions together, or spliced one question into parts which they answered as individual questions (some examples include: (Onoe et al. 2021; Singla et al. 2021; Bandy and Vincent 2021; Shi et al. 2021)). Some datasheets added their own questions which was reasonable since their research was centered around extending or adopting datasheets for a specific purpose (Srinivasan et al. 2021; Bandy and Vincent 2021), however others had no known purpose for doing so (Asano et al. 2021). Additionally, there were also papers where each question was not answered individually but rather a paragraph of text was provided for each section header from the datasheet and undoubtedly this lead to various questions being unanswered (Singla et al. 2021; Hussain et al. 2021). All these differences lead to performing a manual recoding to introduce some standardization around the challenges and enable a simplistic analysis of each datasheet.

Another common challenge throughout the datasheets was referencing of information outside and within the datasheet. This included referencing supplementary material, Github repositories, appendices, the main research paper content, and other responses within the datasheet, such as: “See question 4”. Material external to the datasheet was outside the scope of analysis while pointing to responses to other responses in the datasheet ultimately reduced points attributed to length.

6 Conclusion

In this paper, 21 datasheets from the 2021 NeurIPS benchmarks and datasets track were analyzed to investigate researchers’ approach in completing datasheets. With the analysis performed and the resulting scores, it is evident that the quality of completion and therefore the level of documentation and reflection is highly dependent on individual researchers as the use of datasheets is not formalized, standardized, and does not have to fulfill any minimum quality metrics. As a resource, datasheets provide a first step to documenting and overcoming bias in datasets and their creation process. However, the ML community must enforce rules and standard practices around its use so that completing datasheets can have beneficial outcomes.

6.1 Future Work

To extend this work, a greater sample size of datasheets should be investigated, such as the datasheets from the 2022 NeurIPS benchmarks and datasets track or by creating datasheets for existing publications retrospectively. Furthermore, the 21 datasheets analyzed using R statistical programming langugage (R Core Team 2020) in this paper can be compared with manual analysis and coding to ascertain differences in results. This can then help improve the methodology of automated analyses of datasheets. Lastly, in order to facilitate the examination of datasheets to yield methods of increasing fairness, accountability, and transparency in the dataset development process, the use and adoption of datasheets must be standardized with guidelines.

References

n.d. *GO FAIR*. https://www.go-fair.org/fair-principles/.

Ammanabrolu, Prithviraj, and Mark O. Riedl. 2021. "Modeling Worlds in Text." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, June.

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*.

Bandy, Jack, and Nicholas Vincent. 2021. "Addressing 'Documentation Debt' in Machine Learning Research: A Retrospective Datasheet for BookCorpus." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, May.

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.

Benoit, Kenneth, David Muhr, and Kohei Watanabe. 2021. *Stopwords: Multilingual Stopword Lists*. https://CRAN.R-project.org/package=stopwords.

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.

Desai, Karan, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. "RedCaps: Web-Curated Image-Text Data Created by the People, for the People." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, November.

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.

Heckmann, Mark. 2019. *randomWords: Generate a Random Word*. https://rdrr.io/github/markheckmann/dissertation/src/R/utils-imports.R.

Horel, Thibaut, Lorenzo Masoero, Raj Agrawal, Daria Roithmayr, and Trevor Campbell. 2021. "The CPD Data Set: Personnel, Use of Force, and Complaints in the Chicago Police Department." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*.

Hussain, Aman, Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2021. "Towards a Robust Experimental Framework and Benchmark for Lifelong Language

Learning." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021.*

Jurewicz, Mateusz, and Leon Derczynski. 2021. "PROCAT: Product Catalogue Dataset for Implicit Clustering," *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021.*

Kiskin, Ivan, Marianne Sinka, Adam D. Cobb, Waqas Rafique, Lawrence Wang, Davide Zilli, Benjamin Gutteridge, et al. 2021. "HumBugDB: A Large-Scale Acoustic Mosquito Dataset." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, October.

Koch, Bernard, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, December.

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.*

Liu, C. H. Bryan, Ângelo Cardoso, Paul Couturier, and Emma J. McCoy. 2022. "Datasets for Online Controlled Experiments." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, January. http://arxiv.org/abs/2111.10198.

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.

Mazumder, Mark, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Ciro, Keith Achorn, Daniel Galvez, et al. 2021. "Multilingual Spoken Words Corpus." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021.*

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.

Mysore, Sheshera, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. "CSFCube – a Test Collection of Computer Science Research Articles for Faceted Query by Example." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, no. arXiv:2103.12906 (November). http://arxiv.org/abs/2103.12906.

O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

Onoe, Yasumasa, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. "CREAK: A Dataset for Commonsense Reasoning over Entity Knowledge." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, September.

Ooms, Jeroen. 2023. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents.* https://CRAN.R-project.org/package=pdftools.

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.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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.

Sefala, R., T. Gebru, Luzango P. Mfupe, N. Moorosi, and R. Klein. 2021. "Constructing a Visual Dataset to Study the Effects of Spatial Apartheid in South Africa." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, December.

Shi, Xingjian, Jonas Mueller, Nick Erickson, Mu Li, and Alexander J. Smola. 2021. "Benchmarking Multimodal AutoML for Tabular Data with Text Fields." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, November.

Singla, Samriddhi, Ayan Mukhopadhyay, Michael Wilbur, Tina Diao, Vinayak Gajjewar, Ahmed Eldawy, Mykel Kochenderfer, Ross Shachter, and Abhishek Dubey. 2021. "WildfireDB: An Open-Source Dataset Connecting Wildfire Spread with Relevant Determinants." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*. https://doi.org/10.5281/ZENODO.5636429.

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.

Sorkhei, Moein, Yue Liu, Hossein Azizpour, Edward Azavedo, Karin Dembrower, Dimitra Ntoula, Athanasios Zouzos, Fredrik Strand, and Kevin Smith. 2021. "CSAW-m: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021*, December.

Srinivasan, Ramya, Emily Denton, Jordan Famularo, Negar Rostamzadeh, Fernando Diaz, and Beth Coleman. 2021. "Artsheets for Art Datasets." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021.*

Sun, Haozhe, Wei-Wei Tu, and Isabelle Guyon. 2021. "OmniPrint: A Configurable Printed Character Synthesizer." *Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS), Virtual, December 2021.* http://arxiv.org/abs/2201.06648.

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.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Lionel Henry, and Davis Vaughan. 2023. *Vctrs: Vector Helpers*. https://CRAN.R-project.org/package=vctrs.