Datasheet For Dataset (Template) (v. 1.1; revised 2 Feb. 2022)

Download a copy of this template to use for the Solo Assignment 3 due in Class 11 of English 146DS at UC Santa Barbara (instructor: Prof. Alan Liu, with co-instructor Rebecca Baker)

Each member of a team *individually* creates a "datasheet" for their team's chosen dataset. Use this Template. as a model. (Choose only one dataset to report on if the team is working with a combination of more than one.) For the rationale and examples of datasheets, see Timnit Gebru et al., "Datasheets for Datasets" (2019). *[The structure and questions included in this template are selectively quoted or borrowed with adaptations from this article.]*

This is a solo writing assignment. Teams will have already discussed their dataset together. But each team member must write a datasheet individually without borrowing directly from anyone else's writing. It is fine, however, to draw on collective team discussion that has already occurred so long as there is a clear footnote or endnote crediting the team (e.g., "This idea comes from our team discussion," or, "I borrow with variation an idea that came up in our team discussion"). Also, if you consult external sources about your dataset (e.g., if you search out what other scholars or users say about it online), please add citations or other forms of attribution in notes.

When you are done, export your datasheet as a PDF. (In Google Docs: File → Download → PDF Document.) Then submit the assignment on the course GauchoSpace site: here.

Grading Rubric for this Assignment

Datasheet for Your Dataset

Please fill out this datasheet as best as you can based on inspection of your dataset, its documentation, and any other information available to you. (You may also want to search online for mentions or reviews of the dataset in case there is any relevant discussion of it on the web or in social media. If you use external sources, add citations in notes and/or links.)

Many answers can be brief and to the point. But provide longer, more detailed answers (e.g., a paragraph) when you come to anything that seems to deserve it (is complex, challenging, hard to figure out, possibly controversial).

Important:

- *By default*, we will assume that if you provide an answer or description about some aspect of the dataset then you have verified it either
    - directly through observation of what is in the dataset

  - through documentation you have seen in the dataset (i.e., explanations or metadata included in the dataset's accompanying material about itself).

- *In other cases* where your answer or description rests on other bases (e.g., your best guess through deduction, or reading about the dataset in other materials online or on social media), please identify the basis of your answer (e.g., give a citation or link to where you read about the dataset; or say something like, "there is no direct indication in the dataset or its explanatory material, but I think it's likely that [something is the case] because [your reason]."

- *Missing information:* Many datasets will not provide information about some (or even most) of the questions here, in which case you need only answer "unknown" or "no available information."

Grading Rubric for this Assignment

The sections and questions below are quoted or adapted in revised language (and sometimes revised examples) from Timnit Gebru et al., "Datasheets for Datasets" (2019). Some sections and questions for which the article gives examples are omitted or simplified to create a more compact datasheet assignment for students.

Your name:

Citation of dataset being reported on in this datasheet:

(*For citation format for datasets or statistics, use either Chicago Style or APA Style. Since citation conventions for data and digital materials are rapidly evolving and may not fit every instance, it is only important that you get the citation for your dataset approximately right with as complete information as possible.)*

1. Motivation

a. 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. (If you cannot learn the purpose from explicit information related to the dataset, note that. If you nevertheless have a good hunch about the purpose, state your inference and what it is based on.)

b. Who created this dataset? (e.g., what persons, team, research group) and on behalf of which entity (e.g., company, institution, organization?)

c. Who (if anyone) funded the creation of the dataset?

d. Other comments on the motivation or background of the dataset?

2. Composition

a. What does the data in the dataset represent (e.g., data about documents, photos, social media posts, people, states, countries)? Are there multiple types of data in the dataset (e.g., about not just movies but viewer and ratings)? **Please provide a description of the data**.

b. How many data instances (individual items) are there in total (and of each type, if there are multiple types of data)? (In a spreadsheet where each row is the record of an item, for example, the number of data instances would be the number of rows. If necessary because of problems in the dataset or its format, provide an approximate answer.)

c. Does the dataset contain all possible data items 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 (e.g., representative in geographic coverage)? If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable, etc.).

d. What is the state or nature of the data in the dataset? E.g., “raw” data (e.g., unprocessed text or images), data after it has been processed (e.g., text that has been cleaned or had a stopwords list applied), quantitative data, qualitative data, unstructured data, structured data, or data format (e.g., text, image files)? **Please provide a description**.

e. Is there a label associated with each data item (e.g., labels for photos or people)? If so, please provide a description of the nature of labels.

f. Is any information missing from individual data items? If so, please provide a description of missing information, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information (e.g., removal of non-relevant information for the intended purpose of the dataset), but might include, e.g., redacted text.

g. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

h. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Please provide descriptions of the most important external resources and any restrictions associated with them.

i. Does the dataset contain data that might be considered confidential (e.g., 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.

j. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

k. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

l. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions or proportions within the dataset.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

m. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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.

n. Any other comments?

3. Collection Process

a. How was the data acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

b. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

c. If the dataset is a sample from a larger set, what was the sampling strategy?

d. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated?

e. Over what timeframe was the data created? If the timeframe in which data collection occurred is different (as would be the case in a recent web collection crawl of old news articles), please also describe the collection timeframe.

f. Were any ethical review processes conducted (e.g., by a university's "institutional review board" or IRB overseeing "human subjects" research)? (For information about what an IRB is, see UCSB's "For Researchers in Human Subjects" page.)

g. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

h. Was the data collected from the individuals in question directly, or obtained via third parties or other sources (e.g., websites)?

i. Were the individuals in question notified about the data collection?

j. Did the individuals in question consent to the collection and use of their data?

k. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

l. Has an analysis been conducted of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)?

4. Distribution

a. How is the dataset distributed (e.g., compressed files on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

b. Is the dataset distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe the main conditions of this license and/or ToU, including any fees associated with these conditions.

5. Maintenance

a. Who hosts/maintains/supports the dataset?

b. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

c. Is the dataset being updated (e.g., to correct labeling errors, add new instances, delete instances)?

6. Summary Evaluation

a. After filling out as much of the information as you can find answers for in the above sections of this datasheet, please write here a brief (one paragraph) summary evaluation of the strengths and weaknesses of your dataset.