Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

DATASHEETS FOR DATASETS

This template contains a set of questions covering the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions.

The questions are grouped into seven sections that roughly match the key stages of the dataset creation, maintenance, and distribution process. By grouping the questions in this way, we encourage dataset creators to reflect on the process of creating, distributing, and maintaining datasets, and even to modify this process in response to that reflection. We recommend that dataset creators read through the questions in all sections prior to any data collection so as to flag potential issues early on, and then provide answers to the questions in each section during the relevant stage of the process.

We emphasize that the questions are intended to be used as a starting point for dataset creators to customize. Not all questions will be applicable to all datasets, and dataset creators will likely need to add, revise, or remove questions to better fit their specific circumstances and needs.

To prompt dataset creators to provide sufficient information, all questions are worded so as to discourage yes/no answers. The questions are not intended to serve as a checklist, and dataset creators must be as transparent and forthcoming as possible for datasheets to be useful to dataset consumers.

Answer each question to the best of your ability. Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question.

1. MOTIVATION

1.1 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.

Answer:
Comments:
Do you currently document this information? If so, where?

1.2 Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Answer:

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

Comments:
Do you currently document this information? If so, where?

1.3 Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Answer:
Comments:
Do you currently document this information? If so, where?

1.4 Any other comments?

Answer:
Comments:
Do you currently document this information? If so, where?

2. COMPOSITION

2.1 What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Answer:
Comments:
Do you currently document this information? If so, where?

2.2 How many instances are there in total (of each type, if appropriate)?

Answer:

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

Comments:
Do you currently document this information? If so, where?

2.3 Does the dataset contain all possible instances or is it a subset (not necessarily random) of instances from a larger set? If the dataset is a subset, then what is the larger set? Is the subset representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Answer:
Comments:
Do you currently document this information? If so, where?

2.4 What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Answer:
Comments:
Do you currently document this information? If so, where?

2.5 Is there a label or target associated with each instance? If so, please provide a description.

Answer:
Comments:
Do you currently document this information? If so, where?

2.6 Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Answer:
Comments:

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

Do you currently document this information? If so, where?

2.7 Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Answer:
Comments:
Do you currently document this information? If so, where?

2.8 Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Answer:
Comments:
Do you currently document this information? If so, where?

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

Answer:
Comments:
Do you currently document this information? If so, where?

2.10 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Answer:
Comments:

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

Do you currently document this information? If so, where?

2.11 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.

Answer:
Comments:
Do you currently document this information? If so, where?

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

Answer:
Comments:
Do you currently document this information? If so, where?

2.13 Does the dataset include data about people and/or generated by people (e.g., photos, videos, audio recordings, resumes, news articles, medical records, survey responses, search logs, clicks, emails, social media content, biometric data)? If not, you may skip the remaining questions in this section

Answer:
Comments:
Do you currently document this information? If so, where?

2.14 Does the dataset identify any demographic groups, groups that are protected by law, or groups that may be considered sensitive (e.g., age, gender, disability status)? If so, please describe how these groups are identified and provide a description of their respective distributions within the dataset.

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

2.15 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.

Answer:
Comments:
Do you currently document this information? If so, where?

2.16 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.

Answer:
Comments:
Do you currently document this information? If so, where?

2.17 Any other comments?

Answer:
Comments:
Do you currently document this information? If so, where?

3. COLLECTION PROCESS

3.1 How was the data associated with each instance 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.

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

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

Answer:
Comments:
Do you currently document this information? If so, where?

3.3 If the dataset is a subset from a larger set of possible instances, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Answer:
Comments:
Do you currently document this information? If so, where?

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

Answer:
Comments:
Do you currently document this information? If so, where?

3.5 Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

3.6 Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Answer:
Comments:
Do you currently document this information? If so, where?

If this dataset does not include data about and/or generated by people (see question 2.13), please skip the remaining questions in this section.

3.7 Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Answer:
Comments:
Do you currently document this information? If so, where?

3.8 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.

Answer:
Comments:
Do you currently document this information? If so, where?

3.9 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.

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

3.10 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).

Answer:
Comments:
Do you currently document this information? If so, where?

3.11 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., 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.

Answer:
Comments:
Do you currently document this information? If so, where?

3.12 Any other comments?

Answer:
Comments:
Do you currently document this information? If so, where?

4. PREPROCESSING/CLEANING/LABELING

4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

4.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Answer:
Comments:
Do you currently document this information? If so, where?

4.3 Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Answer:
Comments:
Do you currently document this information? If so, where?

4.4 Any other comments?

Answer:
Comments:
Do you currently document this information? If so, where?

5. USES

5.1 Has the dataset been used for any tasks already? If so, please provide a description.

Answer:
Comments:
Do you currently document this information? If so, where?

5.2 Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Answer:

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

Comments:
Do you currently document this information? If so, where?

5.3 What (other) tasks could the dataset be used for?

Answer:
Comments:
Do you currently document this information? If so, where?

5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Answer:
Comments:
Do you currently document this information? If so, where?

5.5 Are there tasks for which the dataset should not be used? If so, please provide a description.

Answer:
Comments:
Do you currently document this information? If so, where?

5.6 Any other comments?

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

6. DISTRIBUTION

6.1 Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Answer:
Comments:
Do you currently document this information? If so, where?

6.2 How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

Answer:
Comments:
Do you currently document this information? If so, where?

6.3 When will the dataset be distributed?

Answer:
Comments:
Do you currently document this information? If so, where?

6.4 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

6.5 Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Answer:
Comments:
Do you currently document this information? If so, where?

6.6 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Answer:
Comments:
Do you currently document this information? If so, where?

6.7 Any other comments?

Answer:
Comments:
Do you currently document this information? If so, where?

7. MAINTENANCE

7.1 Who will be supporting/hosting/maintaining the dataset?

Answer:
Comments:
Do you currently document this information? If so, where?

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

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

Answer:
Comments:
Do you currently document this information? If so, where?

7.3 Is there an erratum? If so, please provide a link or other access point.

Answer:
Comments:

7.4 Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Answer:
Comments:
Do you currently document this information? If so, where?

7.5 If the dataset includes data about and/or generated by people (see Composition section, question 13), are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Answer:
Comments:
Do you currently document this information? If so, where?

7.6 Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Answer:
Comments:
Do you currently document this information? If so, where?

Answer each question to the best of your ability. If you currently document the information in an answer elsewhere (e.g., compliance procedures, other documentation), please note it in the third box below each question

Use the Comments box to document your thought process and provide feedback on each question's clarity, usefulness, effort required to answer, etc.

7.7 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Answer:
Comments:
Do you currently document this information? If so, where?

7.8 Any other comments?

Answer:
Comments:
Do you currently document this information? If so, where?