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.

Questions

Motivation

- **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.

  The dataset was prompted by my research and readings about the discrimination in the health industry towards Black and brown people. Studies show that the pain of people of color are less likely to be believed than white people's pain.
  A thing that is described in various papers[1], but also backed up by the lived experience of me and the people I know.
  This all relate to how pain is perceived. What does it mean to be perceived as in pain? How do we communicate about our pain? And which issues is created in the interpolation between these two notions.

  The objective of the project is to map and interrogate the discrepancy between how we actually describe and talk about our illnesses and our pain and the way that we are expected to do so inside institutions.
  In order to do so, I needed to create a dataset of real-life descriptions of pain and illnesses that were created without the pressure of a possible diagnosis or judgement by any system or institution.

- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  The dataset was created as part of the final project for the course, AI for media that is part of the MSc course, Data Science in the Creative Industries at the Creative Computing Institute

[1] https://www.npr.org/templates/story/story.php?storyId=201128359

For more information about these questions and about datasheets for datasets in general, please see T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. *Datasheets for Datasets.* The latest version of this paper can be found online at https://arxiv.org/abs/1803.09010 at University of the Arts London. The ideation process and data gathering were done by and on behalf of a single student.

- **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.

  No funding was provided for the creation of this dataset and no parties involved was paid for the work or time put into it.

- **Any other comments?**

Composition

- **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.

  The dataset consists of texts scraped from the user-centered and anonymous website, Reddit. Each instance is representing a comment made by a user. There is no metadata included, such as in which subreddit the comment were made, the timestamp of the comment or the username of the comment.

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

  The dataset consists of 1511 instances (a single instance representing a single comment or post) from ten different subreddits.

- **Does the dataset contain all possible instances 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., 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).

  The dataset was created by scraping the comments of users on the anonymous platform, Reddit. But the data was collected from 10 different subreddits and not the whole platform, and was only included if it contained at least one key word from a list. Therefore, the dataset is a sample of all the comments from all the subreddits on Reddit containing one of the keywords, and therefore it does not contain all possible instances.

  It is hard to say anything about the representativeness of the data, and not possible to validate due to the anonymous nature of the experiences shared. The data is filtered in the way that it contains all instances that describes a medical condition or something painful.

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

  Each instance consists of preprocessed text.
  The data has been preprocessed as it has been cleaned up and the text were tokenized using the NLTK library. The features are whether each word contains at least one of the words from the 'advice_seeking_words' list.

- **Is there a label or target associated with each instance?** If so, please provide a description.
  The data has been labeled with binary labels, 0 and 1, representing whether the comments contain a question or not.

- **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.
  No information is missing from any of the instances.

- **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.
  Each instance is represented by a comment and its label, and there is no explicit relationship between instances.

- **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.
  The dataset was created to be used for text generation models and therefore is split into a train and a test dataset, each respectively containing 1200 and 311 instances.

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
  All duplicates were removed from the dataset, however there could still be noise if comments contain typos or if a comment for example consisted of a lot of emojis and keywords, it would still be labeled as an instance of data that is giving advice. This is noisy when finetuning and training, but could be useful for other tasks such as qualitative analysis.

- **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.

- **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

  Yes, and No. The dataset contains sensitive material that in certain contexts would be considered confidential. However, the data is also made publicly available by the user itself.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.
  The dataset contains people's personal experiences in the health industry and with their own illnesses and potential conditions. Therefore, it could be triggering to some people if they share common experiences.

- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
  Yes, the data is related directly to people.

- **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

  No.

- **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.
  No, I don't think so. There are no timestamps, no user-names or connection to the subreddit the comment was scraped from. If there was, it could be possible to identify the identity of the person who posted the comments, which would be an ethical issue (see elaboration further down)

- **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.
  Yes, the data could contain all of these things. The possibility that the sensitive information the data contains is related to the individual's medical condition and history is highly probable.

- **Any other comments?**
  However, the data is also highly anonymous and it would be hard to trace the source or identity behind the comments. The data is also collected from a public platform, where the individual chose to publish the given information, knowing that everyone would be able to access the information. The dataset doesn't contain more information than what is made publicly available on the website by the individual themselves.

Collection Process

- **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.

  The dataset was created by using web scraping and no subject has been part of the collection or inferred the data, and the experiences described or information given in each instance of the dataset has not been validated or verified.

- **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?
  The data was collected using Python and PRAW an API created using Reddit.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  The dataset was created based on the presence of certain keywords, which is known as a key-word-based sampling. This is a version of purposive sampling that is part of the qualitative research methods.

- **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)?**
  Only one person was part of the data collection. No parties involved was paid for the work or time put into it.

- **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.
  The timeframe of the data was short and only lasted the amount of time it takes to crawl over 10 subreddits.
  The timeframe more or less matches the creation timeframe, as new comments as well as old were included in the dataset.

- **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.
  No ethical review was conducted by a second party.

- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
  Yes, the dataset relates directly to people, as it consists mostly of people's own descriptions of their personal experiences with being in pain or being ill.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  The data was collected via third parties as it was scraped from the website, Reddit.

- **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.

  No, as the individuals are anonymous there was no way to notify before collecting.

- **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
  No, the people were not given a choice of consent, as the data was scraped from the internet without prior notification.

- **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).

- **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.

  Yes: While the dataset contains sensitive information it is also totally anonymous and I don’t think there could be any directly harmful outcomes or impacts of the data.

- **Any other comments?**
  While it could seem very sensitive that the data contains personal information, it’s use is intended to map the discrepancy between how actual people actually describe and talk about their illnesses and their pain and the way that people are expected to do so inside institutions. It gives an important and heartfelt insight into how people experience their illnesses and how they feel treated in the health industry, which is data that might be hard to get from somewhere else.
  E.g. the first instance in the data bear witness to an experience that is not recorded within the system: “I’m sick of trying to navigate a Healthcare system that doesn’t give a shit about me. ⚛', 'where im at right now 🥲 im just so done”

Preprocessing/cleaning/labeling

- **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.
  The NLTK library was used to tokenize the comments into words in order to check if the keywords were presents in the comment in order to label the data.

- **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.
  Yes, the raw data is saved in the list, ‘pain_all’ that contains all comments that were scraped based on the key-word condition.

- **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.
  Yes, the software is a notebook available on github:
  https://github.com/signeaijing/AIforMedia/blob/main/final_project_web_scraper_FINALVERSION.ipynb

- **Any other comments?**

Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.
  The dataset was created for the purpose of creating a new pain chart for the AI for Media Final Project.
  That is the only task that the dataset has been used for.

- **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.
  Yes, the project for which this data was used can be found in this repository: https://github.com/signeaijing/AIforMedia

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

- **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?

  Not in the way the preprocessing is right now, but always be aware of labels and what we remove and not.

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

- **Any other comments?**

  The punctuation and emojis were not removed on purpose, as emojis is a very real and actual way people communicate emotions and experiences that there might be limited words to describe.

Distribution

- **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.
  No, the dataset will not be distributed anywhere, but it is available in a public repository, that will be made private no more than a month after the project is concluded.

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
  The dataset is located in a repository on GitHub, that will be private no more than a month after the project is concluded.

- **When will the dataset be distributed?**
  Has been available on GitHub since 14.03.2023

- **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.
  No.

- **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.
  No.

- **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.

No.

- **Any other comments?**

Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  The dataset will not be maintained after the date, 16.03.2023

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  Via mail: s.rodkjrhansen0620221@arts.ac.uk

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

  No.

- **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)?

- **If the dataset relates to people, 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.

- **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.

- **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.

- **Any other comments?**