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 purpose of the dataset is to control the playing of music. This dataset intended to be use for developing a music control system. The dataset may have been created to address a specific gap in existing music control systems, such as the need for more intuitive gesture recognition. When people are listening to songs and doing other things, they may not be able to touch the keyboard with their hands to control the music playing on the computer. This makes it easier to use gestures to control music to be played. It could also help people with physical disabilities to control the playing of music.

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

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

I downloaded songs from the website (https://tools.liumingye.cn/music/#/) to create the dataset.

- **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.

  N/A

- **Any other comments?**

  N/A

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.

  This sample dataset consists of 14 songs. Randomly select two songs from each group in the larger set

  The larger set divided into 8 groups by different artists (Sandy Lam, Lala Hsu, Adele, Beyoncé, Harry, Justin Bieber, Miley Cyrus and Troye Sivan).

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

  Due to the oversized dataset, I have only provided the sample which includes 14 songs.

  The larger set includes 204 songs of Sandy Lam (24), LaLa Hsu (21), Adele (18), Beyoncé (13), Harry. (20), Justin Bieber (76), Miley Cyrus (14) and Troye Sivan (18).

- **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).

  This dataset is a sample (random) of instances from a larger set. I have a larger set includes 204 songs of Sandy Lam (24), LaLa Hsu (21), Adele (18), Beyoncé (13), Harry. (20), Justin Bieber (76), Miley Cyrus (14) and Troye Sivan (18). The largest set is all songs in the world. It is not representative of the larger set, because some songs didn't get copyrighted. Different countries have different songs and some niche songs do not have resources.

- **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 raw data. Each instance consists of unprocessed songs. Download songs directly from the website.

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

  N/A

- **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.

  Some singers' songs are now not copyrighted and cannot be downloaded.

- **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.

  Songs in the same category in the dataset are sang by one singer. Some songs are also on the same album, which was released at the same time.

- **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 most common data split recommended is the 70/15/15 split, where 70% of the data is used for training, 15% for validation or development, and 15% for testing. The rationale behind this split is as follows:

  Training set: The largest portion of the data is used for training the model. The purpose of the training set is to provide sufficient data to the model to learn the underlying patterns and relationships in the data. This helps to avoid overfitting and ensure that the model generalizes well to new data.

Validation/Development set: A smaller portion of the data is used for validating the model during the training process. The validation set is used to tune the hyperparameters of the model and to monitor its performance during training. By using a separate validation set, the model can be evaluated on data that it has not seen before and identify overfitting.

Testing set: The final portion of the data is used to test the model's performance after the training is complete. The testing set is used to evaluate the model's generalization performance, and it should not be used for any training or hyperparameter tuning. This helps to obtain an unbiased estimate of the model's performance on unseen data.

Overall, these recommended data splits are used to ensure that the model is trained, validated, and tested on separate datasets to ensure the model's performance on unseen data.

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

  No

- **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.

  If a dataset does rely on external resources, there may or may not be a guarantee that these resources will be present and remain unchanged over time. For example, if the external resource is a well-established and widely used data repository or API, it is likely to continue to be available and supported for the foreseeable future. However, if the external resource is more volatile or prone to change, such as social media data or web pages, it is less certain that the data will remain accessible and unchanged.

  In some cases, there may be official archival versions of the complete dataset, including any external resources that existed at the time the dataset was created.
  This will depend on whether the creator of the dataset has taken steps to archive the external resources and make them available to future users.

There may also be restrictions associated with external resources, such as fees or usage restrictions that apply to future users. It is important for the creator of the dataset to provide clear information about any such restrictions and how they affect the use or analysis of the dataset. Links to external resources or other access points should also be provided if they are necessary for the use of the dataset.

Overall, transparency and clarity of any external resources associated with the dataset is important to ensure that the dataset can be used effectively and accurately, both now and in the future.

- **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.

  No

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

  No

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

  No

  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 within the dataset.

  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

- **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.

  No

- **Any other comments?**

  No

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.

  These data were directly observable.

- **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?

  Collect these data by manual human curation. These mechanisms or procedures validated by myself.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

  Stratified Sampling: Stratified sampling is used when there is a class imbalance in the data. In this method, the data is divided into subgroups based on the target variable, and samples are drawn from each subgroup to ensure that the proportion of each class is maintained in the sample.

The larger set divided into 8 groups by different artists (Sandy Lam, Lala Hsu, Adele, Beyoncé, Harry, Justin Bieber, Miley Cyrus and Troye Sivan). This sample dataset consists of 14 songs. Randomly select two songs from each group in the larger set

- **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)?**

  By myself.

- **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.

  10/03/2023-14/03/2023

- **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

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

  No

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

  Websites https://tools.liumingye.cn/music/#/

- **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.

/

- **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.

/

- **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.

/

- **Any other comments?**

  No

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.

  No

- **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.

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

- **Any other comments?**

Uses

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

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

  Some of these songs were used for my mini project.

- **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.

  My mini project used this dataset.

  (**https://github.com/Bryce138675/Mini-project-Gesture-Music-Player**)

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

  Analysing the emotion of music, AI generated music

- **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?

There are several factors that may affect future use and may lead to unfair treatment or other undesirable harm to individuals or groups.

Bias in the dataset. If a dataset is biased, for example, if it under-represents certain groups, this could lead to unfair treatment of individuals or groups. For example, my dataset currently only includes popular music, not classical music, rock music, etc. This may cause confusion to groups who like these music, and I will continue to improve the database music categories in the future..

Inaccurate or incomplete data. If a dataset contains inaccurate or incomplete data, it may lead to incorrect predictions or decisions, resulting in increased risk or other undesirable harm.

To mitigate these undesirable harms, prospective users should carefully evaluate the dataset and consider potential biases and limitations. They should also look at the limitations of the dataset and use it in conjunction with other data sources to minimise bias and improve accuracy. In addition, they should consider using techniques such as data augmentation or synthetic data generation to create more diverse and representative datasets. Finally, they should be aware of the potential impact of their use of the dataset and take steps to mitigate any potential harm.

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

  Production for commercial use

- **Any other comments?**

  No

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.

This dataset will be uploaded online and provided to those who need it to download and use it.

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

  I compressed my dataset and put it on the website (https://artslondon-my.sharepoint.com/personal/m_liang0620221_arts_ac_uk/_layouts/15/onedrive.aspx?view=0).

- **When will the dataset be distributed?**

  16/3/2023

- **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.

- **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.

  A music license is a right held by someone to distribute and/or use a piece of copyrighted music. Often seen in movies, a music license is also used for commercials, television shows, internet videos, and any other visual medium that wants to use a song with permission. A variety of music license types exist, each with their own nuances, sometimes requiring a person to hold more than one for proper usage in a project.

- **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?**

  No

Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**

Minghao Liang

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

m.liang0620221@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)?

  Dataset will be updated (add new instance) once a month by myself. Mailing list will be communicated to users.

- **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.

  No

- **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.

  Dataset continue to be supported. I will update the database once a month.

- **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.

  If other people want to extend/augment/build on/contribute to the dataset, they could send me email. I will verify the reliability and riskiness of the data source and decide whether to extend the database. Once adopted I will reply with an email of thanks and keep updated.

- **Any other comments?**

  No