DATASHEET TEMPLATE FOR SOUND AND AUDIO RELATED AI MODELS

RAISE (Research on Artificial Intelligence in Sound and Musical Expression)

This datasheet template, designed for sound and audio specific AI models and tools, is based on the paper “Datasheet for Datasets” created by Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford, Published in CACM in December 2021. Source: https://arxiv.org/abs/1803.09010
And on Aubrey Beards Datasheet-for-Datasets template:
https://github.com/AudreyBeard/Datasheets-for-Datasets-Template/blob/master/Datasheet_for_Datasets.pdf

GENERAL DATASHEET
--Model Name --
--Developer Name --
--Institution--

Contents:

1. Short introduction of the Model:
Name of Model, Short description of Function of the model, Developers, Institution, Research Team behind the Model

2. About Datasets:
A. Motivation:
For what purpose was the dataset created?
Was there a gap to be filled?
Who created (which parts of) the dataset?

B. Composition:
What do the instances that comprise the dataset represent (e.g. audio files, vocal samples, percussion samples, file form)?
How big is the Dataset? How many Instances are there in total?

C. Uses:
What (other) tasks can the dataset be used for?
Is there tasks the dataset should not be used for?

D. Distribution:
Will the dataset be distributed to third parties outside of entity?
How will the dataset be distributed (GitHub, API,...)
Will the dataset be distributed under a copyright or other intellectual property license (IP) or other Terms of Use (ToU)
Please describe this license and/or ToU?

E. Maintenance:
Who will be supporting/hosting/maintaining the dataset?
How can the developer of the dataset be contacted?
How will the dataset be updated?
And communicated to users?

3. Datasheets
For datasets created collaboratively by multiple individuals using different workflows, it is essential that each contributor creates a separate Datasheet. This Datasheet should address the following questions to document the creation and composition of the single parts of the dataset.

a. Dataset Name:
What is the name of the dataset?

b. Dataset Creation
Who contributed to the creation of this dataset?

c. Data Representation:
What do the instances in the dataset represent?

d. Source of Pre-Recorded Samples:
If pre-recorded samples are used, where do they originate from, where do they originate from, and under what circumstances were they recorded?

e. Copyright and Licensing of Samples:
If pre-recorded samples are used, are they subject to any copyright or intellectual Property (IP) licenses, or any Terms of Use (ToU)? If so, please provide a detailed description of these licenses or a link to access them.

f. Data Collection Mechanisms:
What mechanisms or procedures were employed to collect the data?
Please specify any software programs, hardware, etc., used.

g. Data Pre-Processing:
Has the data been pre-processed, cleaned, or labeled (e.g., removal of instances) If so, please describe the process.

4. Workflows
For datasets created collaboratively by multiple individuals using different workflows, provide links to descriptions of different workflows e.g., in form of flowcharts.

DATASHEET TEMPLATE FOR SOUND AND AUDIO RELATED AI MODELS

Datasets are crucial to the success of AI, machine learning, and deep learning models. However, detailed and comprehensive information about the datasets used in developing these models is often missing. To address this issue, the concept of 'datasheets' has been proposed.Datasheets aim to remedy the frequently insufficient information about datasets by including details on data collection methods, sources, pre-processing steps, and metadata. They are intended for a wide range of stakeholders, including developers, evaluators, and reviewers. The goal of these datasheets is to improve transparency and accountability in the AI field. Despite challenges in balancing information detail and data privacy, standardized datasheets are proposed to ensure consistency. Their use helps in assessing dataset quality, identifying biases, and addressing ethical concerns. The exemplary template provided in the paper serves as a guideline for creating customized datasheets tailored to specific datasets.