**Common-Voice Datasets**

* **Link:** <https://commonvoice.mozilla.org/en/datasets>
* **Description:** The dataset includes a unique MP3 file and its corresponding text file and it consist of demographic metadata (age, sex, accent) useful for training speech recognition engines. The dataset currently consists of 20,409 validated hours in 124 languages.
* **Last upgradation Date:** 20/3/2024
* **Size:** 82.4 GB
* **Version:** Common Voice Corpus 17.0
* **Audio Format:** MP3
* **Research Paper:** Common Voice: A Massively-Multilingual Speech Corpus
* **Research Paper Link:** <https://arxiv.org/abs/1912.06670>
* **Research Paper Summary:** The paper discusses a large dataset of voice recordings from people around the world to help improve speech recognition technology. The dataset is unique because it includes many different languages, making it very diverse. The paper also talks about the challenges of ensuring data quality and the efforts to keep adding more languages.

**Unique Metadata Association Model DataSet**

* **Link:**[https://github.com/Amrita-TIFAC-Cyber/Digital Forensics?tab=readme-ov-file](https://github.com/Amrita-TIFAC-Cyber/Digital-Forensics?tab=readme-ov-file)
* **Description:** The dataset is designed for metadata forensics and includes real-world datasets sourced from five types of devices (mobile, laptop, pendrive, desktop, and cloud) and five social media platforms (Telegram, WhatsApp, Instagram, Twitter, and Facebook). The dataset comprises ten experiments with various sources per experiment.
* **Last upgradation Date:** 4/12/2020
* **Size:** 1.6 GB
* **Version:** v1.0
* **File Format:** .rar
* **Research Paper:** [Holistic Analytics of Digital Artifacts: Unique Metadata Association Model](https://www.researchgate.net/publication/354283974_Holistic_Analytics_of_Digital_Artifacts_Unique_Metadata_Association_Model?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoiX2RpcmVjdCJ9fQ)
* **Research Paper Link:** [https://www.igi-global.com/article/holistic-analytics-of-digital-artifacts/2831280](https://arxiv.org/abs/1912.06670)
* **Research Paper Summary:** The paper proposes a new way to connect metadata (information about data) with digital artifacts (digital files or objects). This model looks at all aspects of these digital items to provide a complete and integrated analysis. The goal is to make it easier to track, organize, and analyze digital data.

**arXiv Dataset**

* **Link:** <https://www.kaggle.com/datasets/Cornell-University/arxiv>
* **Description:** The dataset includes information such as the title, authors, submission dates, categories (e.g., subject areas), and other relevant details of the research papers hosted on arXiv. The dataset includes the abstracts of the research papers, which provide a summary of the research and its findings.
* **Last upgradation Date:** 3 days ago (monthly updating)
* **Size:** 4.2 GB
* **Version:** Version 182
* **File Format:** .json
* **Research Paper:** A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers
* **Research Paper Link:** <https://arxiv.org/abs/2105.03011>
* **Research Paper Summary:** The paper introduces QASPER, a dataset comprising 5,049 questions derived from 1,585 Natural Language Processing (NLP) research papers. This dataset is unique because the questions are designed to seek specific information from the full text of research papers, not just the abstract or the title. These questions were created by NLP practitioners who read only the paper's title and abstract and were answered by other practitioners who also provided supporting evidence. The dataset aims to improve the performance of question-answering systems on complex, document-grounded tasks, demonstrating that current models perform significantly worse than humans on these questions, thus encouraging further research in this area

**Analysis of Melanoma Metadata and EffNet Ensemble**

* **Link:** <https://www.kaggle.com/code/datafan07/analysis-of-melanoma-metadata-and-effnet-ensemble/input>
* **Description:** This dataset contains medical images and associated metadata used for the computer-aided diagnosis of skin cancer. The dataset aims to improve early detection and diagnosis of skin cancer, particularly melanoma, which is the most deadly form of the disease. Through the analysis of these images and accompanying clinical data, researchers seek to develop algorithms and models that can assist medical professionals in accurately identifying and diagnosing skin cancer at its early stages.
* **Last upgradation Date:** 25/07/2020
* **Size:** 117.96 GB
* **Version:** version 28
* **File Format:** .csv
* **Research Paper:** Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge
* **Research Paper Link:** <https://www.researchgate.net/publication/344622007_Identifying_Melanoma_Images_using_EfficientNet_Ensemble_Winning_Solution_to_the_SIIM-ISIC_Melanoma_Classification_Challenge>
* **Research Paper Summary:** The paper presents a solution that won the SIIM-ISIC Melanoma Classification Challenge. The authors used an ensemble of convolutional neural network (CNN) models, primarily based on the EfficientNet architecture. They combined models with different backbones and input sizes, some of which utilized image-level and patient-level metadata. Key factors in their success included a stable validation scheme, effective model targeting, a well-tuned training pipeline, and the diversity of the ensemble. Their solution achieved a 0.9600 AUC on cross-validation and 0.9490 AUC on the private leaderboard.

**Dataset Search: metadata for datasets**

* **Link:** <https://www.kaggle.com/datasets/googleai/dataset-search-metadata-for-datasets/data>
* **Description:** This dataset contains metadata from Google's Dataset Search tool, which is designed to help users find datasets stored across the web. The metadata includes t**itle, description, keywords, URL, provider, creator, date published, and license**
* **Last upgradation Date:** 16/10/2020
* **Size:** 5.68 GB
* **Version:** Version 3
* **File Format:** .csv
* **Research Paper:** An Analysis of Online Datasets Using Dataset Search
* **Research Paper Link:** <https://research.google/blog/an-analysis-of-online-datasets-using-dataset-search-published-in-part-as-a-dataset/>
* **Research Paper Summary:** This research paper is probably all about looking at how we can use tools like Google Dataset Search to find and understand datasets that are available online. It likely discusses how these tools help researchers discover different kinds of data on the internet. In this metadata provides essential information about datasets, such as their title, description, creator, publication date, and keywords. This information helps researchers evaluate the relevance and quality of datasets before accessing them.

**Android Application Packets (APK)**

* **Link:** <https://corp.digitalcorpora.org/corpora/mobile/>
* **Description:** This dataset includes directories and files containing data from various Android and iOS devices, including specific versions and models. These data sets can provide information such as the app's version, permissions, package name, and other details used for digital forensics and research purposes.
* **Last upgradation Date:** 22/11/2020
* **Size:** 1.6 GB
* **Version:** -
* **File Format:** .zip

**Email Datasets**

* **Link:** <https://www.cs.cmu.edu/~./enron/>
* **Description:** This dataset contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This datasets contain email timestamps , sender and recipient Information, subject Lines and folder Structure.
* **Last upgradation Date:** 7/5/2015
* **Size:** 423 MB
* **Version:** -
* **File Format:** .tar.gz
* **Research Paper:** Exploration of Communication Networks from the Enron Email Corpus
* **Research Paper Link:** <http://www.casos.cs.cmu.edu/publications/protected/2005-2006/diesner_2005_explorationsenron.pdf>

**IoT-23 Datasets**

* **Link:** <https://www.stratosphereips.org/datasets-iot23>
* **Description:** This dataset provide realistic and diverse IoT network traffic data for analyzing security vulnerabilities, developing intrusion detection systems (IDS), and studying IoT communication patterns. It consists of twenty three captures (called scenarios) of different IoT network traffic. These scenarios are divided into twenty network captures (pcap files) from infected IoT devices (which will have the name of the malware sample executed on each scenario) and three network captures of real IoT devices network traffic (that have the name of the devices where the traffic was captured). On each malicious scenario they executed a specific malware sample in a Raspberry Pi, that used several protocols and performed different actions.
* **Last upgradation Date:** Jan 2020
* **Size:** 21GB
* **Version:** 1.0.0
* **File Format:** .tar.gz
* **Research Paper:** ToN\_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Datasets.
* **Research Paper Link:** <https://ieeexplore.ieee.org/document/9444348>