| Name of Dataset Dataset: [Link to dataset]  Data Card Authors: [name, name] | Write a summary describing your dataset (limit 200 words). Include information about the content and topic of the data, its sources, the motivation behind the dataset, and the problems or use cases it is suitable for. What is the benefit of using this dataset? |
| --- | --- |

| Authorship | | |
| --- | --- | --- |
| Publishers | | |
| PUBLISHING ORGANIZATION | INDUSTRY SECTOR | PUBLISHER CONTACT |
| Write the names of the institution or organization responsible for publishing the dataset. | Bold to select all applicable.  🖐️ Do not delete any unselected choices. | Provide publisher contact details.For dataset owners, see next row. |
| Organization Name | Corporate  Academic  Not-for-profit  Individual  Others (please Specify) | * POC Name, Affiliation, Contact * Group Email: Mailing-list@website.com * Website: www.website.com |
| Dataset Owners | | |
| DATASET TEAM(S) | DATASET CONTACT | DATASET AUTHORS |
| Write the names of the groups or team(s) that own the dataset. | How can dataset owners be contacted for questions about the model? See previous row for publishing institution. | Write the names of all authors associated with the dataset. Provide the affiliation and year if different from publishing institutions or multiple affiliations: |
| Name of Group or Team(s)  E.g. Delightful Wanderers | * Name, Affiliation, Contact * Group Email: Mailing-list@website.com * Website: www.website.com | * Name, Title, Affiliation, YYYY * Name, Title, Affiliation, YYYY * Name, Title, Affiliation, YYYY |
| Funding Sources | | |
| FUNDING INSTITUTION(S) | FUNDING DETAILS | |
| Write the names of the funding institutions. | Provide a short summary of funding sources and other support, including details such as programs or projects that may have funded the creation, collection, or curation of the dataset. Include links to relevant documents where applicable. | |
| Institution 1  Institution 2  Institution 3 | E.g. Institution 1 and institution 2 jointly funded this dataset as a part of the XYZ data program, funded by XYZ grant awarded by institution 3 for the years YYYY-YYYY. | |

###### 

| Dataset Overview | | |
| --- | --- | --- |
| DATASET SUBJECT | DATASET SNAPSHOT | DESCRIPTION OF CONTENT |
| Bold to select all applicable.🖐️ Do not delete any unselected choices. | Fill out details as indicated, adding rows as needed. Include links to additional table(s) with more detailed breakdowns in the caption. | Provide a short summary of the dataset content. Include links where applicable. |
| Sensitive Data about peopleNon-Sensitive Data about peopleData about natural phenomenaData about places and objectsSynthetically generated dataData about systems or products and their behaviorsUnknownOthers\* (\*please specify) | | Size of dataset | 123456 MB | | --- | --- | | Number of Instances | 123456 | | Number of Fields | 123456 | | Labeled Classes | 123456 | | Number of Labels | 123456789 | | Average labels per instance | 123456 | | Algorithmic Labels | 123456789 | | Human Labels | 123456789 | | Other | 123456 |   <write here> | E.g. bounding-box annotations and labels in images of coarse-and fine-grained objects. |
| DESCRIPTIVE STATISTICS | | |
| Add basic statistics for each field here, as relevant. If there is insufficient space, focus on the most important or critical fields for this dataset. E.g., some statistics will be relevant for numeric data, but not for strings. | | |
| | Statistic | Field\_name | Field\_name | Field\_name | Field\_name | Field\_name | Field\_name | | --- | --- | --- | --- | --- | --- | --- | | count |  |  |  |  |  |  | | mean |  |  |  |  |  |  | | std |  |  |  |  |  |  | | min |  |  |  |  |  |  | | 25% |  |  |  |  |  |  | | 50% |  |  |  |  |  |  | | 75% |  |  |  |  |  |  | | max |  |  |  |  |  |  | | mode |  |  |  |  |  |  |   Caption for table: Provide links to extended tables where relevant. | | |
| Sensitivity of Data | | |
| SENSITIVE DATA | FIELDS WITH SENSITIVE DATA | SECURITY AND PRIVACY HANDLING |
| Bold to select all applicable.🖐️ Do not delete any unselected choices. | Please indicate which features or fields might contain sensitive or personally identifiable information, and if or not collection was intentional using the format below: | Provide a short summary of measures or steps to handle sensitive data in this dataset. Include links and metrics where applicable. |
| User ContentUser MetadataUser Activity DataIdentifiable DataSensitive DataBusiness DataEmployee DataPseudonymous DataAnonymous DataHealth DataChildren’s DataNoneOthers\* (\*please specify) | Intentionally Collected Sensitive Data <Field>: [type of S/PII]  <Field>: [type of S/PII]  <Field>: [type of S/PII] Unintentionally Collected Sensitive Data <Field>: [type of S/PII]  <Field>: [type of S/PII]  <Field>: [type of S/PII]  ... | <write here>   * <method>: [description] * <method>: [description] * <method>: [description] |
|  | RELEVANT LINKS | RISKS AND MITIGATIONS |
|  | Provide link(s) to documents that describe any S/PII where available: | Provide a short summary of how risks from PII or sensitive information have been mitigated in the dataset. Include links and metrics where applicable. |
|  | * <Link type>: [Link] * <Link type>: [Link] * <Link type>: [Link] * ... | <write here>   * <Risk> : [description + mitigation] * <Risk> : [description + mitigation] * <Risk> : [description + mitigation] |
| Dataset Version and Maintenance | | |
| VERSION STATUS | DATASET VERSION | MAINTENANCE PLAN |
| Bold to select ONE.🖐️ Do not delete any unselected choices. | Provide details about this version of the dataset. | Provide a short summary of how the dataset is maintained, including information about refreshes, versioning criteria, errors, feedback and/or recourse. Include links and metrics where applicable. |
| Regularly Updated New versions of the dataset have been or will continue to be made available. Actively Maintained No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data. Limited Maintenance The data will not be updated, but any technical issues will be addressed. Deprecated This dataset is obsolete or is no longer being maintained. | Current Version 1.0  Last Updated MM/YYYY  Release Date MM/YYYY | <write here>   * Versioning: [Description and criteria] * Update: [Description and criteria] * Errors: [Description and criteria] * Feedback: [Description and criteria] |
|  | NEXT PLANNED UPDATE | EXPECTED UPDATES OR CRITERIA |
| ⚠️ Fill this if this dataset isRegularly updatedActively maintained and another version is planned | Provide details about the next planned update. | Provide a short summary for readers to understand updates to the dataset and/or data. Include links, charts, and visualizations as appropriate. |
|  | Version affected 1.0  Next data update MM/YYYY  Next Version 1.1  Next Version update MM/YYYY | <write here>   * Updates to data: <write here> * Updates to dataset: <write here> |

| Example of Data Points | | |
| --- | --- | --- |
| PRIMARY DATA MODALITY | SAMPLING OF DATA POINTS | DATA FIELDS |
| Bold to select ONE (primary modality).🖐️ Do not delete any unselected choices. | Link to multiple data points or exploratory demos. If access is restricted, consider adding a fake example that provides a realistic description of data points in the dataset. | Provide a list of fields in data points, including a description and notes on how to interpret fields in an example of data in this dataset. |
| Image DataText DataTabular DataAudio DataVideo DataTime SeriesGraph DataGeospatial DataMultimodal (Please specify)Others (please specify)Unknown | * Link to demo * Link to a typical or within normal distribution example * Link to an outlier or out-of-distribution example * Link to other example * Link to other example | * Field: description and interpretation * Field: description and interpretation * Field: description and interpretation * ... |
|  | EXAMPLE: TYPICAL DATA POINT | EXAMPLE: OUTLIER DATA POINT |
|  | Copy-and-paste a typical data point. Include a description of what makes it typical. If restricted, consider adding a fake example that provides a realistic description of data points in the dataset. | Copy-and-paste an atypical or outlier data point. Include a description of what makes it atypical. If restricted, consider adding a fake example that provides a realistic description of data points in the dataset. |
|  | <write description and what makes this datapoint typical here>  E.g. of Data Point:   | en,es,1,1-1,"Kaisa-Leena Mäkäräinen (born 11 January 1983) is a Finnish former world-champion and 3-time world-cup-winning biathlete, who currently competes for Kontiolahden Urheilijat.","Kaisa-Leena Mäkäräinen (nacida el 11 de enero de 1983) es excampeona mundial finlandesa, tres veces ganadora de la copa mundial de biatlón y actualmente compite para el Kontiolahden Urkeilijat.",Female,Kaisa Mäkäräinen,https://en.wikipedia.org/wiki/Kaisa\_M%C3%A4k%C3%A4r%C3%A4inen | | --- | | <write description and what makes this datapoint atypical here>  E.g. of Data Point:   | en,es,1,1-3,"Her team coach is Jonne Kähkönen, while Jarmo Punkkinen is her ski coach.","El entrenador de su equipo es Jonne Kähkönen, mientras que Jarmo Punkkinen es su entrenador de esquí.",Female,Kaisa Mäkäräinen,https://en.wikipedia.org/wiki/Kaisa\_M%C3%A4k%C3%A4r%C3%A4inen | | --- | |

###### 

| Motivations & Use | | |
| --- | --- | --- |
| Motivations | | |
| DATASET PURPOSE(S) | KEY DOMAINS AND APPLICATION(S) | PRIMARY MOTIVATION(S) |
| Bold to select ONE.🖐️ Do not delete any unselected choices. | Use comma-separated tags to indicate the key domains for this dataset. | List the primary motivations for creating or curating this dataset: |
| MonitoringResearchProductionOthers (please specify) | Domains Machine Learning, Object Recognition, Computer Vision Problem Space Gender accuracy in text translations that describe occupations. | E.g.   * Bring demographic diversity to imagery training data * Encourage academics to take on second-order challenges |
| Intended Use | | |
| DATASET USAGE | INTENDED AND/OR SUITABLE USE CASE(S) | UNSUITABLE USE CASE(S) |
| Bold to select ONE.🖐️ Do not delete any unselected choices. | Summarize the intended and known use cases of this dataset: | Summarize any known problematic use cases of this dataset: |
| Safe for production useSafe for research useConditional use- some unsafe applicationsOnly approved useOthers (please specify) | * Use Case 1 * Use Case 2 * Use Case 3 ... | * Problematic Use Case 1 * Problematic Use Case 2 * Problematic Use Case 3 ... |
|  | PROBLEM SPACE AND RESEARCH QUESTIONS(S) | PUBLICATION GUIDELINES |
|  | Describe the specific problem space that this dataset intends to address. Include any specific research questions. | Include any guidelines and steps for citing this dataset in research and/or production work. |
|  | E.g. Depth estimation in static images. | E.g. To reference this dataset in your paper, please follow the following guidelines:   * Reference any prior publications that have referenced the dataset, * Please ask dataset authors to review the use of the dataset in your research |

| Access, Retention, & Wipeout | | |
| --- | --- | --- |
| Access | | |
| ACCESS TYPE | DOCUMENTATION LINKS | ACCESS PREREQUISITES |
| Bold to select ONE.🖐️ Do not delete any unselected choices. | Provide links that describe documentation to access this dataset: | Please describe any required training or prerequisites to access: this dataset. |
| UnrestrictedConditionalOpen AccessOthers (please specify) | * Website: [link] * Github Readme: [link] * Colab Example: [link] | For e.g.  This dataset requires membership in [specific] database groups:   * Complete the [mandatory training] * Read [Data Usage Policy] * Initiate a Data Requesting by filing [a bug] |
|  | DIRECT LINKS TO DATASET | ACCESS POLICY |
|  | Provide links to access this dataset: | Summarize the access policy associated with this dataset. Use this space to include any other information or links that might be relevant to accessing the dataset. |
|  | * Direct download link: [link] * Other repository: [link]   Code to download data   | # … | | --- | | * <list here>   <write additional notes here> |
| Retention | | |
|  | RETENTION DURATION | RETENTION POLICY |
|  | Specify the duration for which this dataset can be retained: | Summarize the retention policy for this dataset. Use this space to include any other information or links that might be relevant. |
|  | <Specify time> | <write here> |
|  | RETENTION STEPS | EXCEPTIONS AND EXEMPTIONS |
|  | Summarize any additional requirements and related steps to retain the dataset. | Summarize any additional exceptions and related steps to retain the dataset: |
|  | <write here> | <write here> |
| Wipeout and Deletion | | |
|  | WIPEOUT DURATION | DELETION EVENT |
|  | Specify the duration after which this dataset should be deleted or wiped out: | Summarize the sequence of events and allowable processing for data deletion: |
|  | <Specify time> | <Write here> |
|  | ACCEPTABLE MEANS OF DELETION | POST-DELETION OBLIGATIONS |
|  | List the acceptable means of deletion: | Summarize the sequence of obligations after a deletion event: |
|  | <List here> | <Write here> |
|  | OPERATIONAL REQUIREMENTS | EXCEPTIONS AND EXEMPTIONS |
|  | List any wipeout integration operational requirements: | Summarize any additional exceptions and related steps to a deletion event: |
|  | <List here> | <Write here> |

| Dataset Provenance | | |
| --- | --- | --- |
| Data Collection & Sources | | |
| DATA COLLECTION METHODS | DATA SOURCES | DESCRIPTION OF DATA SOURCE(S) |
| Bold to select all applicable.🖐️ Do not delete any unselected choices. | Describe the source for each collection method. Add rows as meaningful. Refer to guidance onDuplicate for each collection method as necessary. | Provide a brief description of each Data Source by type. Include appropriate breakdowns if data sources contain data from other sources. Include links to more information, metrics, visualizations, etc. |
| APIArtificially GeneratedCrowdsourced - PaidCrowdsourced - VolunteerVendor Collection EffortsScraped or CrawledSurvey, forms or pollsTaken from other existing datasetsUnknownTo be determinedOthers (please specify) | [Collection method used] [Source & Link]: [description]  Date of Collection: [MMM YYYY - MMM YYYY] E.g. API [Unsplash Image API:](https://unsplash.com/developers) A JSON API to access an open collection of images contributed by photographers and associated search keywords.  Date of Collection: [MM YYYY - MM YYYY]  Sensitive or High Risk [Yes / No]  Access Policy [link]  Wipeout Policy [link]  Retention Period [NN days]  Primary Data Modality [Image Data / Text Data / Tabular Data / Audio Data / Video Data / Time Series / Graph Data / Geospatial Data / Multimodal (Please specify) / Others (Please specify) / Unknown]  Update Frequency [Yearly / Quarterly / Monthly/ Biweekly / Weekly / Daily / Hourly / Static / Others (Please specify) | [Data Source]  <write description here>  [Data Source]  <write description here> |
| DATASET TYPE | COLLECTED DATA | DATA PROCESSING |
| Bold to select all applicable.🖐️ Do not delete any unselected choices. | List or describe any fields or data that were collected for this dataset, and indicate if they were included in the dataset or excluded from the dataset. Include links, descriptive statistics, and visualizations where relevant.Duplicate for each collection method as necessary. | If multiple methods were used to collect data, how was the data aggregated, processed, or connected? Include relevant descriptions, statistics, metrics or visualizations, links and libraries in your response.Break down by source type. |
| Static Data was collected once from single or multiple sources. Streamed Data is continuously acquired from single or multiple sources. Dynamic Data is updated regularly from single or multiple sources. Others\* (\*please specify) | [Collection method used] Collected and included   * List here * List here   Collected and excluded   * List here * List here  E.g. API Collected and included   * Photo\_url: permalink URL to image * Photo\_location\_name: where the photo was taken * Photo\_latitude: latitude of photo * Photo\_longitude: longitude of photo * Photo\_description: description of photo written by the photographer * Keyword: keyword or search term   Collected and excluded   * Photo\_id: ID of unsplash image, replaced in this dataset * EXIF data: removed from all images | <write here, 100 words max> [Data Source] <write descriptions, statistics, metrics or visualizations, links and libraries here> [Data Source] <write descriptions, statistics, metrics or visualizations, links and libraries here> |
| Criteria | | |
| SELECTION CRITERIA | INCLUSION CRITERIA | EXCLUSION CRITERIA |
| Please describe the data selection criteria. Break down by method as applicable. Include links, descriptive statistics, and visualizations where relevant. | Please describe the data inclusion criteria. Break down by method as applicable. Include links, descriptive statistics, and visualizations where relevant.. | Please describe the data exclusion criteria. Break down by method as applicable. Include links, descriptive statistics, and visualizations where relevant. |
| [Collection method used] <write here> E.g. API Only images containing visual information describing known categories were used. | [Collection method used] <write here>  E.g. API  100k randomly sampled images containing at least one identifiable object box associated with the search term. | [Collection method used] <write here>  E.g. API   * Quality: Illustrated images, overly stylized, hybrid or distorted images * Content: Images with partial objects, and ambiguous images, images with S/PII |
| Relationship to Source | | |
| USE | BENEFITS AND VALUE | LIMITATIONS AND TRADE-OFFS |
| If at all, how is the resulting dataset aligned with the purposes, motivations, or intended use of the upstream source(s)? Break down by source type. | What are the benefits of the resulting dataset to its consumers, compared to the upstream source(s)? Break down by source type. | What are the limitations of the resulting dataset to its consumers, compared to the upstream source(s)? Break down by source type. |
| <write here> [Data Source] <write here> [Data Source] <write here> | <write here> [Data Source] <write here> [Data Source] <write here> | <write here> [Data Source] <write here> [Data Source] <write here> |
| Updates to Dataset | | |
| ⚠️ Fill this next row if: this is not the first version of the dataset, and there is no data card available for the first version. | | |
|  | FIRST VERSION | NOTES ON FIRST VERSION |
|  | Provide a basic description of the first version of this dataset. | Optional.Provide a short summary describing caveats or nuances of the first version of this dataset. Include links, charts, and visualizations as appropriate. |
|  | Release date MM/YYYY  Link to dataset [link]  Status [Actively Maintained/Limited Maintenance/ Deprecated]  Size of Dataset 123 MB  Number of Instances 123456 | E.g. ~2B images from the first edition of this dataset have been removed due to license expiry. |
| DATASET UPDATE FREQUENCY | DATASET UPDATE SCHEDULE | CHANGES ON UPDATE |
| Bold to select ONE🖐️ Do not delete any unselected choices. | Please describe the update schedule | What happens when the dataset is refreshed? Break down by sources as necessary. Include any applicable policies and changes to the dataset that occur during a refresh. |
| YearlyQuarterlyMonthlyBiweeklyWeeklyDailyHourlyStaticOthers\* (\*Please specify) | [Source Type ]  Date of last update  DD/MM/YYYY  Frequency of Updates [Yearly / Quarterly / Monthly/ Biweekly / Weekly / Daily / Hourly / Static / Others (please specify)  Data points affected 123456  Data points updated 123456  Data points added 123456  Data points removed 123456  Date of next update DD/MM/YYYY | <write here>  [Data Source]  <write here>  [Data Source]  <write here> |

| Human and Other Sensitive Attributes | | |
| --- | --- | --- |
| SENSITIVE HUMAN ATTRIBUTES | INTENTIONALITY OF COLLECTIONS | RATIONALE FOR COLLECTING HUMAN ATTRIBUTES |
| Bold to select ALL ATTRIBUTES that are present in the dataset.🖐️ Do not delete any unselected choices. | For each human attribute indicated, specify if this information was collected intentionally or unintentionally: | Briefly describe the motivation, rationale, considerations or approaches that caused this dataset to include the indicated human attributes. Summarize why or how this might affect the use of the dataset. |
| RaceGenderEthnicitySocio-economic statusGeographyLanguageSexual OrientationReligionAgeCultureDisabilityExperience or SeniorityOthers (please specify) | Intentionally Collected Attributes  (human attributes that were labeled or collected as a part of the dataset creation process)   * Attribute 1: Fields * Attribute 2: Fields * …   Unintentionally Collected Attributes  (human attributes that were not explicitly collected as a part of the dataset creation process but can be inferred using additional methods)   * Attribute 1: Fields * Attribute 2: Fields * ... | <write here> |
|  | SOURCE(S) OF HUMAN ATTRIBUTES | COLLECTION METHODS |
|  | Indicate the source of the sensitive attributes using the format provided. | Describe the methods used to collect human attributes in the dataset. Break down by human attribute as necessary. Include information related to the tasks, platforms, visualizations, links to additional documentation as applicable. |
|  | [Human Attribute]: Source | <list here> |
| DISTRIBUTION OF HUMAN ATTRIBUTES | | |
| Duplicate and populate the following row for each human attribute previously selected. Include the key takeaways in the caption. | | |
| | [Human attribute] | Label or Class | Label or Class | Label or Class | Label or Class | | --- | --- | --- | --- | --- | | Count | 123456 | 123456 | 123456 | 123456 | | [Statistic] | 123456 | 123456 | 123456 | 123456 | | [Statistic] | 123456 | 123456 | 123456 | 123456 | | [Statistic] | 123456 | 123456 | 123456 | 123456 |   Caption for table above | | |
|  | KNOWN CORRELATIONS | RISK, TRADE-OFFS AND CAVEATS |
|  | List or describe any known correlations with the indicated sensitive attributes in this dataset. Summarize why or how this might affect the use of the dataset. Include visualizations, metrics, or links where necessary. | Provide a statement, list or summarize any expectations, systemic or residual risks, trade-offs and caveats due to human attributes in this dataset. Break down by human attribute if necessary. |
|  | <write here> | <write here> |

| Extended Use | | |
| --- | --- | --- |
| Use with Other Data | | |
| SAFETY OF USE WITH OTHER DATA | KNOWN SAFE DATASETS OR DATA TYPES | BEST PRACTICES FOR JOINING OR AGGREGATING WITH DATASET |
| Bold to select ONE.🖐️ Do not delete any unselected choices. | Which known datasets or data can this dataset be safely joined or aggregated with? Describe any relevant transformation types. | Summarize best practices for using this dataset in conjunction with other datasets or data type. Links to demonstrative examples where available. |
| Safe to use with other dataConditionally safe to use with other dataShould not be used with other dataUnknownOthers\* (Please specify) | Dataset or Data  Summary of safe / healthy transformations  Dataset or Data  Summary of safe / healthy transformations  Dataset or Data  Summary of safe / healthy transformations |  |
|  | KNOWN UNSAFE DATASETS OR DATA TYPES | KNOWN LIMITATIONS AND RECOMMENDATIONS |
| ⚠️ Fill out this row if you selected “Conditionally safe to use with other datasets” or “Should not be used with other datasets”: | Which known datasets or data should this dataset not be joined or aggregated with? List and describe any relevant transformation types. | Describe limitations of the dataset that might introduce foreseeable risks to intended use when the dataset is conjoined with other datasets. Include any suggested recommendations. |
|  | Dataset or Data  Summary of unsafe / risky transformations  Dataset or Data  Summary of unsafe / risky transformations  Dataset or Data  Summary of unsafe / risky transformations | * <write limitation> : <describe risk + mitigations> * <write limitation> : <describe risk + mitigations> * <write limitation> : <describe risk + mitigations> |
| Forking & Sampling | | |
| SAFETY OF FORKING / SAMPLING | ACCEPTABLE SAMPLING METHODS | BEST PRACTICES FOR FORKING AND SAMPLING |
| Bold to select ONE.🖐️ Do not delete any unselected choices. | Bold to select all applicable.🖐️ Do not delete any unselected choices | Summarize best practices for forking or sampling this dataset. Links to demonstrative examples where available. |
| Safe to fork and/or sampleConditionally safe to fork and/or sampleShould not be forked and/or sampledUnknownOthers\* (\*Please specify) | Cluster Sampling  Haphazard Sampling  Multi-stage Sampling  Random Sampling  Retrospective Sampling  Stratified Sampling  Systematic Sampling  Weighted Sampling  Unknown  Unsampled  Others\*  (\*Please Specify) | <write here> |
|  | KNOWN RISKS TO SAMPLING | KNOWN LIMITATIONS AND RECOMMENDATIONS |
| ⚠️ Fill out this row if you selected “Conditionally safe to fork and/or sample” or “Should not be forked and/or sampled”. | What known or residual risks are associated with forking and sampling methods when applied to the dataset? List and describe. | Describe limitations of the dataset that might introduce foreseeable risks to intended use when the dataset is forked or sampled. Include any suggested recommendations. |
|  | Sampling method   * Risk 1 * Risk 2 | * <write limitation> : <describe risk + mitigations> * <write limitation> : <describe risk + mitigations> * <write limitation> : <describe risk + mitigations> |
| Use in Machine Learning or AI Systems | | |
| DATASET USE(S) | DATASET SPLITS | USAGE GUIDELINES OR POLICIES |
| Bold to select all applicable.🖐️ Do not delete any unselected choices. | Describe and name the splits in the dataset (if more than one), and include any criteria for splitting the data. | Describe any usage guidelines or policies that users of the dataset should be aware of. Summarize documents and link to them as relevant. |
| TrainingTestingValidationDevOthers\* (\* Please Specify) | Train 123456  Test 123456  Validation 123456  Dev 123456 | For e.g. Potential users must satisfy the following conditions to access and use this dataset:   * Condition 1 * Condition 2 * …   Detailed guidelines are available [here]. |
|  | FEATURE DISTRIBUTIONS | KNOWN CORRELATIONS |
|  | Describe any notable feature distributions in the dataset. Include links to servers where readers can explore the data on their own. | List or describe any known correlations with the indicated features in this dataset. Summarize why or how this might affect the use of the dataset. Include links where necessary. |
|  | Feature 123456  Feature 123456  Feature 123456  Feature 123456 | Correlated feature & feature: Summary  Correlated feature & feature: Summary  Correlated feature & feature: Summary |
|  | SPLIT STATISTICS | |
|  | Provide the sizes of each split. As appropriate, provide any descriptive statistics for features. | |
|  | | Statistic | Train | Test | Valid | Dev | | --- | --- | --- | --- | --- | | Count | 123456 | 123456 | 123456 | 123456 | | Descriptive Stat | 123456 | 123456 | 123456 | 123456 | | Descriptive Stat | 123456 | 123456 | 123456 | 123456 | | Descriptive Stat | 123456 | 123456 | 123456 | 123456 |   Caption for table above | |

| Dataset Transformations | | |
| --- | --- | --- |
| ⚠️ Fill this section if any transformations were applied in the creation of your dataset. | | |
| TRANSFORMATIONS APPLIED | FIELDS TRANSFORMED | LIBRARIES AND METHODS USED |
| Bold to select all applicable🖐️ Do not delete any unselected choices. | What were the data types that fields were transformed to? Break down by transformations applied | List any relevant libraries used to process the data, as applicable. |
| Anomaly DetectionCleaning Mismatched ValuesCleaning Missing ValuesConverting Data TypesData AggregationDimensionality ReductionJoining Input SourcesRedaction or AnonymizationOthers\* (\*Please specify) | [Transformation Applied]  * Source Field: Target Field * Source Field: Target Field * Source Field: Target Field * ...  [Transformation Applied]  * Source Field: Target Field * Source Field: Target Field * Source Field: Target Field * …  [...] | * Transformation type: Libraries used, corresponding method(s) * Transformation type: Libraries used, corresponding method(s) * Transformation type: Libraries used, corresponding method(s) * Transformation type: Libraries used, corresponding method(s) * ... |
| Breakdown of Transformations | | |
| Fill out relevant rows. | | |
| CLEANING MISSING VALUES | METHODS USED | COMPARATIVE SUMMARY |
| Which fields in the data were missing values? How many? | How were missing values cleaned? What other choices were considered? | Why were missing values cleaned using this method (over others)? Provide comparative charts showing before and after missing values were cleaned. |
| <write here> | <write here> | <write here> |
| CLEANING MISMATCHED VALUES | METHODS USED | COMPARATIVE SUMMARY |
| Which fields in the data were corrected for mismatched values? | How were incorrect or mismatched values cleaned? What other choices were considered? | Why were incorrect or mismatched values cleaned using this method (over others)? Provide a comparative analysis demonstrating before and after values were cleaned. |
| <write here> | <write here> | <write here> |
| ANOMALY DETECTION | METHODS USED | OUTLIERS HANDLING |
| How many anomalies or outliers were detected? | What methods were used to detect anomalies or outliers? | If at all, how were anomalies or outliers handled? Why or why not? |
| <write here> | <write here> | <write here> |
| DATA AGGREGATION | METHODS USED | COMPARATIVE SUMMARY |
| Which fields in the dataset were aggregated? | What methods were used to aggregate the data? Include the aggregating operator. What other choices were considered? | Why was the data aggregated using this method (over others)? Provide comparative charts that demonstrate the choices of aggregators. |
| <write here> | <write here> | <write here> |
| DIMENSIONALITY REDUCTION | METHODS USED | COMPARATIVE SUMMARY |
| How many original features were collected and how many dimensions were reduced? | What methods were used to reduce the dimensionality of the data? What other choices were considered? | Why were features reduced using this method (over others)? Provide comparative charts showing before and after dimensionality reduction processes. |
| <write here> | <write here> | <write here> |
| JOINING INPUT SOURCES | METHODS USED | RESIDUAL RISKS AND APPROVALS |
| What were the distinct input sources that were joined? | What are the shared columns of fields used to join these sources? | What are the differential privacy or other residual risks from this join? Include links to relevant approvals and documentation. |
| <write here> | <write here> | <write here> |
| REDACTION OR ANONYMIZATION | METHODS USED | RESIDUAL RISKS AND APPROVALS |
| Which features were redacted or anonymized? | What methods were used to redact or anonymize data? | What are the differential privacy or reidentification risks to redacted data or anonymization? Include links to relevant approvals and documentation. |
| <write here> | <write here> | <write here> |
| OTHERS (PLEASE SPECIFY) | METHODS USED | RESIDUAL RISKS & COMPARATIVE SUMMARY |
| What was done? Which features or fields were affected? | What methods were used? | What are the residual risks associated with this transformation? Include links to relevant approvals and documentation. Why were features reduced using this method (over others)? Provide comparative charts showing before and after this transformation. |
| <write here> | <write here> | <write here> |

| Annotations | | |
| --- | --- | --- |
| ⚠️ Fill this section if any human or algorithmic annotation tasks were performed in the creation of your dataset. | | |
| ANNOTATION WORKFORCE TYPE | ANNOTATION CHARACTERISTICS | ANNOTATION DESCRIPTION |
| Bold to select ALL APPLICABLE🖐️ Do not delete any unselected choices. | Describe relevant characteristics as indicated. For quality metrics, consider including accuracy, consensus accuracy, IRR, XRR at the appropriate granularity (e.g. across dataset, by annotator, by annotation, etc.). Duplicate for each annotation type if multiple methods were used. | Briefly describe the annotations applied to the dataset, including but not limited to: Creation of data, authoring of data, labeling, annotation, rating, etc. Include links, and indicate platforms, tools or libraries used wherever possible. Break down by annotation type as applicable. |
| Annotation Target in DataMachine-generated AnnotationsHuman Annotations - ExpertHuman Annotations - Non-expertHuman Annotations - EmployeesHuman Annotations - ContractorsHuman Annotations - CrowdsourcingHuman Annotations - Outsourced / Managed TeamsUnlabeledOthers\* (\*Please specify) | [Annotation Type] Number of unique annotations 123456789  Total number of annotations 123456789  Avg. Annotations per example 123456789  Number of annotators per example 123456789  [Quality metric per granularity] 123456789  [Quality metric per granularity] 123456789  [Quality metric per granularity] 123456789 | <write summary here>[Annotation type] <description of annotations produced, platforms, tools, libraries, etc.> [Annotation type] <description of annotations produced, platforms, tools, libraries, etc.> [Annotation type] <description of annotations produced, platforms, tools, libraries, etc.> |
|  | ANNOTATION DISTRIBUTION(S) | ANNOTATION TASK AND INSTRUCTIONS |
|  | Provide a distribution of annotations for each annotation or class of annotations using the format below. Duplicate for each annotation type if multiple methods were used. | Briefly summarize the annotation task and instructions provided to annotators or methods employed for machine annotations. Include the inter-annotation adjudication policy, and any golden questions if applicable. Add links wherever possible. Break down by annotation type as applicable. |
|  | [Annotation Type] Annotations (or Class) 12345 (20%)  Annotations (or Class) 12345 (20%)  Annotations (or Class) 12345 (20%)  Annotations (or Class) 12345 (20%)  Annotations (or Class) 12345 (20%) | <write summary here> [Annotation type] <description of tasks, instructions, and policies> [Annotation type] <description of tasks, instructions, and policies> [Annotation type] <description of tasks, instructions, and policies> |
| Description of Human Annotators | | |
| ⚠️ Fill this section if human annotators were used. | | |
|  | ANNOTATOR BREAKDOWN | ANNOTATOR DESCRIPTION |
|  | Provide a description of the annotators. Add more rows as meaningful. For inapplicable rows, refer to guidance on slide 38 of go/recommended-by. Duplicate for each annotation type if multiple methods were used. | Provide a brief description of the annotator pool(s). Elaborate on the annotator type, training provided, selection criteria, and anything else that might affect the quality of annotations. Break down by annotation type. |
|  | [Annotation type] Annotator type [Paid - Expert / Paid - Non-expert / Volunteer - Expert / Volunteer - Non-expert / Vendor Annotation Efforts / Crowdsourced / Others (Please Specify)]  Total unique annotators 12345  Average cost/task/ annotator $$$  Expertise of annotators [ Professional / Trained for task / No training provided] | <write summary here> [Annotation type] <annotator description> [Annotation type] <annotator description> [Annotation type] <annotator description> |
| LANGUAGE(S) OF ANNOTATORS | LOCATION(S) OF ANNOTATORS | GENDER(S) OF ANNOTATORS |
| Provide distributions as available. Duplicate for each annotation type if multiple methods were used. | Provide distributions as available. Duplicate for each annotation type if multiple methods were used. | Provide distributions as available. Duplicate for each annotation type if multiple methods were used. |
| [Annotation type] <language> [percentage %]  <language> [percentage %]  <language> [percentage %] | [Annotation type] <location> [percentage %]  <location> [percentage %]  <location> [percentage %] | [Annotation type] <gender> [percentage %]  <gender> [percentage %]  <gender> [percentage %] |

| Validation Methods | | |
| --- | --- | --- |
| ⚠️ Fill this section if the data in dataset was validated during or after the creation of your dataset. | | |
| VALIDATION METHOD(S) | VALIDATION BREAKDOWN | DESCRIPTION OF VALIDATION |
| Bold to select ALL APPLICABLE🖐️ Do not delete any unselected choices. | Describe the fields and data points that were validated. Duplicate for each validation type if multiple methods were used. | Briefly describe the methods used to validate the dataset. Include tools, frameworks, libraries, platforms used. Indicate results, outcomes, actions and visualizations. Include links wherever possible. Break down by validation type if multiple methods were used. |
| Data Type ValidationRange and Constraint ValidationCode/cross-reference ValidationStructured ValidationConsistency ValidationNot ValidatedOthers\* (\*Please specify) | [Validation Type] # of data points validated 12345  Fields Validated:  Field 123456  Field 123456  Field 123456 | <write summary here> [Validation type] Method: summarize here  Tools: summarize here  Results: write here  Outcomes: write here |
| Description of Human Validators | | |
| ⚠️ Fill this section if the dataset was validated using human validators | | |
|  | VALIDATORS CHARACTERISTIC(S) | VALIDATORS DESCRIPTION(S) |
|  | Describe the following about the validators. Add more rows as meaningful. Duplicate for each validation type if multiple methods as necessary. | Provide a brief description of each validator pool. Elaborate on the annotator type, training provided, selection criteria, and anything else that might affect the quality of annotations. Break down by validation type as necessary. |
|  | [Validation Type] Unique validators 12345  # of examples per validator 123456  Average cost/task/ validator $$$  Training provided Y/N  Expertise required Y/N | <Write summary here> [Validation type] <validator description> |
| LANGUAGE(S) OF VALIDATORS | LOCATION(S) OF VALIDATORS | GENDER(S) OF VALIDATORS |
| Provide the following distribution as available. Duplicate for each validation type as necessary. | Provide the following distribution as available. Duplicate for each validation type as necessary. | Provide the following distribution as available. Duplicate for each validation type as necessary. |
| [Validation Type] <language> [percentage %]  <language> [percentage %]  <language> [percentage %] | [Validation Type] <location> [percentage %]  <location> [percentage %]  <location> [percentage %] | [Validation Type] <gender> [percentage %]  <gender> [percentage %]  <gender> [percentage %] |

###### 

###### 

| Sampling Methods | | |
| --- | --- | --- |
| ⚠️ Fill out the following block if your dataset employed any sampling methods. | | |
| SAMPLING METHOD(S) | SAMPLING CHARACTERISTIC(S) | SAMPLING CRITERIA |
| Bold to select ALL APPLICABLE🖐️ Do not delete any unselected choices. | Provide the following for each sampling method used. Add additional sampling statistics as relevant. Duplicate for each sampling type, if multiple methods were used. | Describe any criteria used to sample the data. Break down by sampling methods as relevant. Include links and metrics where necessary. |
| Cluster SamplingHaphazard SamplingMulti-stage SamplingRandom SamplingRetrospective SamplingStratified SamplingSystematic SamplingWeighted SamplingUnknownUnsampledOthers\* (\*Please specify) | [Sampling Type] Upstream Source [write here]  Total data sampled 123m  Sample size 123  Threshold applied 123k units at property  Sampling Rate 123  Sample Mean 123  Sample Std. Dev. 123  Sampling Distribution 123  Sampling Variation 123  Sample Statistic 123 | <write summary here>   * Sampling method: Criteria * Sampling method: Criteria * Sampling method: Criteria |

###### 

###### 

| Known Applications & Benchmarks | | |
| --- | --- | --- |
| ⚠️ Fill out the following section if your dataset was primarily created for use in AI or ML system(s) | | |
| ML APPLICATION(S) | EVALUATION - RESULTS | EVALUATION - PROCESS |
| ✏️ Write tags separated by commas. Focus on key tasks performed by the model | Enumerate the models on which this dataset was used and corresponding performance metrics. Link to model cards or model documentation. Duplicate for each model. | Describe any notable factors in your process for evaluating your model’s overall performance or assessing how the dataset contributes to the model’s performance.Break down for each model. Include links, metrics, charts, and visualizations. |
| <Write here>  E.g.  Classification, Regression, Object Detection | [Method or model (link)] Accuracy 123  Precision 123  Recall 123  Performance metric 123 | <write summary here> [Method or model (link)]  * Process: <write here> * Factors: <write here> * Considerations: <write here> * Results: <write here> |
|  | MODEL DESCRIPTION(S) AND STATISTICS | EXPECTED PERFORMANCE AND  KNOWN CAVEATS |
| Bold to select ONE🖐️ Do not delete any unselected choices. | Briefly describe the model(s) and tasks that this dataset was used in. Include links where necessary. Duplicate for each model. | Expected performance: Briefly summarize the application and expected performance when using this dataset.Known Caveats: Describe the known caveats, trade-offs and consequences when using this dataset.Duplicate for each model. Include links wherever possible. |
| Duplicate this row as necessary for each model type | [Method or model] <write description here>  E.g. Used to train a lightweight palm detection to detect human palms in an image on a smartphone. Palm detector returns bounding boxes around the palm in an image.  Model Card [Link]  Model Size 123  Model Weights 123  Model Layers 123  Latency 123 | [Method or model] Expected performance: <write summary>   * Finding 1 * Finding 2 * …   Known Caveats: <write summary>   * Finding 1 * Finding 2 * … |

###### 

###### 

| Terms of ArtConcepts and Definitions referenced in this Data Card | | |
| --- | --- | --- |
| ✏️ Use this space to include the expansions and definitions of any acronyms, concepts, or terms of art used across the Data Card. Use standard definitions where possible. Include the source of the definition where indicated. If you are using an interpretation, adaptation, or modification of the standard definition for the purposes of your data card or dataset, include your interpretation as well. | | |
| [TERM OF ART] | [TERM OF ART] | [TERM OF ART] |
| Definition: <write here>  Source: <write and link>  Interpretation: <write here> | Definition: <write here>  Source: <write and link>  Interpretation: <write here> | Definition: <write here>  Source: <write and link>  Interpretation: <write here> |
| [TERM OF ART] | [TERM OF ART] | [TERM OF ART] |
| Definition: <write here>  Source: <write and link>  Interpretation: <write here> | Definition: <write here>  Source: <write and link>  Interpretation: <write here> | Definition: <write here>  Source: <write and link>  Interpretation: <write here> |
| [TERM OF ART] | [TERM OF ART] | [TERM OF ART] |
| Definition: <write here>  Source: <write and link>  Interpretation: <write here> | Definition: <write here>  Source: <write and link>  Interpretation: <write here> | Definition: <write here>  Source: <write and link>  Interpretation: <write here> |

###### 

| Reflections on Data | | |
| --- | --- | --- |
| ✏️ Use this space to include any additional information about the dataset that has not been captured by the Data Card. For example, Does the dataset contain data that might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please contact the appropriate parties to mitigate any risks. | | |
| [Title] | Write notes here | |
| [Title] | Write notes here | |
| [Title] | Write notes here | |