Examples of applying our DSL over ML datasets: Technical report

Abstract. This report is an appendix of the paper "A domain-specific language for describing machine learning datasets" and extends the preliminary evaluation presented in section 5. In this report, we describe three benchmark datasets used as an example by the recent works of the Machine Learning community concerning the dataset's definition.

1 Introduction

This report shows a set of examples using the DSL proposed in the paper *A domain-specific language for describing machine learning datasets*. We have selected to describe three benchmark datasets with different compositions and provenance that have been used by other relevant proposals in the machine learning field concernings the definition of the datasets [1, 3–6].

The described datasets are the *Gender Inclusive Coreference* [2], the *Movie Reviews Polarity* [7], and the *SIIM-ISIC Melanoma Classification* [8]. We describe each one in every section of these report, and in every section we describe the *Metadata* part, the *Composition* part, and the *Provenance and Social Concerns* part.

The examples shown in this work are built with the Visual Studio Code plugin presented together with the main paper. We present excerpts of the datasets' definitions, but the full report can be found at the public repository [1] inside the /examples/evaluation folder. On the other side, and to facilitate the repeatability of the experiment, the data of the Melanoma dataset (3) can be found in the folder **/examples/data**. Besides, the data of the other described dataset are publicly available by its original authors.

2 SIIM-ISIC Melanoma Classification Challenge

The SIIM-ISIC Melanoma classification dataset was created to enable a machine learning competition to identify melanoma in lesion images. Has been created between 1998 and 2019 by several health institutions worldwide. In addition, have been used by the project *Dataset Nutrition Labels* [5] as benchmark to evaluate their dataset description guidelines.

To develop our description, we have used the documentation of the *Dataset Nutrition Labels*, the data itself, and the documentation provided by the International Skin Imaging Collaboration (ISIC) [2]. In summary, we found the

[1] https://github.com/ReviewInstrumental/DSL-dataset-description
[2] https://challenge2020.isic-archive.com/

information to build a complete description document, but to extract the information, we have had to extract insights from the textual definitions. Being these processes time-consuming. With our DSL, this information is well-structured and in a format that is easy to be computed.

2.1 Metadata

The first part of the description is the metadata part. In Listing 1, we have an excerpt of the Melanoma metadata part. The interesting point of this part is *Description* in line 6. The information for the three attributes of the description (purpose, tasks, and gaps) was there. Still, we had to read and extract insights from different parts of the documentation to organize the description in its subconcepts.

In terms of *Applications*, in Line 27, we could extract the past uses, in this case, the Kaggle competition, the recommended uses of the dataset, and the non-recommended uses. We have extracted the information about these parts through a qualitative analysis of different parts of the documentation. In terms of *Distribution*, the information is stored as a complement of the *License*. If we observe these in more datasets, we can potentially join the concepts of *Distribution policies* and *Licenses*.

Listing 1. Metadata melanoma excerpt

```
Metadata:
  Title: "2020 SIIM-ISIC Melanoma Classification Challenge Dataset"
  Unique-identifier: SIC_Melanoma_Classification_Challenge_Dataset
  Version: v0001
  Release Date: 08-10-20
  Description:
    Purposes:
      "Advance medical image innovation. The 2020 SIIM-ISIC Melanoma
      Classification challenge dataset was created for the purpose
      of conducting a machine learning competition to identify
      melanoma in lesion images. [...]"
    Tasks:
      "Classification"
    Gaps:
      "Improve melanoma detection of ML models. As the leading
      healthcare organization for informatics in medical imaging,
      the Society for Imaging Informatics in Medicine (SIIM)'s
      mission is to advance medicine imaging informatics through
      education, research, and innovation in a multi-disciplinary
      community. SIIM is joined by the International Skin Imaging
      Collaboration (ISIC), an international effort to improve
      melanoma diagnosis. "
  Licenses: CC BY-NC 4.0 (Attribution-NonCommercial 4.0 International)
  Area: HealthCare
  Tags: Images, Melanoma, diagnosis, Skin Image
  Applications:
    Past Uses:
        "Yes. The 2020 SIIM-ISIC Melanoma Classification challenge
        dataset was created for the purpose of conducting a machine
        learning competition to identify melanoma in lesion[...]"
    Recommended:
        "Melanoma skin detection"
        "Predict incidence of melanoma in a population"
    Non-recommended:
        "Due to low population prevalence and challenges with access to
        care in different populations, the images gathered for large
        datasets such as this for AI classification [...]"
  Distribution:
    Distribution Licenses: " In addition to the CC-BY-NC license, the
    dataset is governed by the ISIC Terms of Use. Learn more about the
    terms of use here:https://datanutrition.org/labels/isic-2020/ "

```

This particular dataset is co-authored by multiple authors and multiple founded. The information about who authored and funded the dataset is publicly available, and we evaluate if our DSL can support complex authoring scenarios such as the present one. Despite this, there are some relevant information of the

authors we couldn't express easily. For example, the contribution of every author was different. Every hospital has donated a different set of images. It was not clear what each author's contribution was in the dataset documentation, but allowing the DSL to support a specific contribution for every author may be a potential improvement for the DSL in the future.

In listing 2, we have an example of the Authoring part of the Melanoma dataset description. In line 10, in *Funders* part, *funders*, such as HPS, are a coalition of many companies and research centers. We miss a way to express these complex scenarios using our DSL. This is also a way to improve the DSL in these scenarios. Finally, these datasets do not provide *Erratums* or *Contribution Guides*, and there is no known update policy.

Listing 2. Metadata melanoma excerpt III

```
1  Authoring:
2    Authors:
3      Name "Skin Imaging Collaboration ISIC"
4          Email "email@email.com"
5      Name "Hospital Clinic de Barcelona"
6          Email "email@email.com"
7      Name "Medical University of Vienna"
8          Email "email@email.com"
9      [...]
10   Funders:
11     Name "The University of Queensland"  type mixed
12         Grantor "National Health and Medical Research Council (NHMRC)
13            Centre of Research Excellence Scheme" GrantId: APP1099021
14     Name "HPS" type private
15         Grantor "NHMRC MRFF Next Generation Clinical Researchers
16         Program Practitioner Fellowship" GrantId: APP1137127
17         [...]
18   Maintainers: [...]
19   Contribution guidelines: "To contribute to this dataset, no
       contribution guidelines are provided."
20   Erratum: "There is no erratum known."
```

2.2 Composition

Inside the composition part, we provide a qualitative description of the data: the type, attributes, number of records, and some relevant statistical values. This information was not present in the different sources of documentation available, so we have to extract it directly from the data. We perform an exploratory data analysis, matching the results from insights extracted from the Dataset Nutrition Label documentation. In Listing 3, we have an example of Melanoma's dataset description composition.

As we can see in lines 14 and 15, the dataset comprises seven attributes, and, 33126 instances. We extract the rationale description from the gathering process information. These represent a misconception in structure and one of the pain

points of adopting our DSL. The composition of the dataset is the result of the gathering process, but also of the labeling process and data preprocessing. So, we understand there is a need for a separate rationale inside the composition part.

In terms of statistical information, at the attribute level, we were able to express relevant information of the dataset, such as, in line 25, the gender distribution. Although, we were not able to express the problems with skin color issues as the data itself does not contain information about that. We believe that beyond expressing a potential racial bias in the data, dataset creators may express these classes' unbalance of potential bias in terms of data statistics. In line 38, we can point the importance of age and gender relationship, and

Otherwise, no consistency rule is provided, as these may not be relevant in that case. As such, quality metrics are not provided, and these are inferred in the natural text explanations of the documentation.

Listing 3. Composition Melanoma Excerpt

```
1 Composition:
2     Rationale:
3         "There is one instance that represents images from the skin of
4         patients in coalition with some label identifying the
5         patient. Metadata for each image included approximate patient
6         age at the time of image capture, biological sex, general
7         the anatomic site of the lesion anonymized patient
      identification
8         number..."
9         [...]
10    Data Instances:
11      Instance:  skinImages
12        Description: "Skin images of the patients"
13        Type: Record-Data
14        Size: 33126
15        Attribute number: 7
16        Attributes:
17          Attribute: ImageId [...]
18          Attribute: patientID [...]
19          Attribute: benignant_malignant
20            Description: "Medical diagnosis of the patient"
21            Labelling process: DiagnosisLabel
22            OfType: Categorical
23            Statistics:
24              Categorical Distribution:
25                "benignant": 88% - "malignant": 12%
26          Attribute: sex
27            Description: "Sex of the patient"
28            Count: 33126
29            OfType: Categorical
30            Statistics:
31                Mode: "Famela"
32                Categorical Distribution:
```

```
                    "Male: 40%
                     Famela: 55%
                     Unknown: 5%"
          Attribute: ageGroup
              Description: "The age group of patients"
              OfType: Categorical
              Statistics:
                Mode: "40-50"
                Categorical Distribution:
                  "0-10": 15% - "10-20": 12% [...]
          Attribute: anatom_site_general_challenge [...]
          Attribute: diagnosis [...]
    Statistics:
        Pair Correlation:
            Between age and benignant_malignant
            Between age and sex
            Between age and external source
                From: "Official population indicator of an
                       hypothetical public institution"
                Rationale: "Similar age distribution"
        Quality Metrics:
            Completeness: 100%
    Consistency rules:
        Inv: skinImages: (ageGroup>=0)
    Is sample:
        "Yes, the sample dataset used accurately reflects the intent or
        desired outcome and has been evaluated to ensure that
        it is fit for purpose. [...]"

    Data Splits:
        "Yes. To test algorithm generalizability, a subset of
        images from six sites (five geographic locations) were
        allocated for the training dataset of the 2020 ISIC Grand
        Challenge. There is a separate test dataset available without
        target information."
```

2.3 Provenance and Social Concerns Part

The provenance and social concert parts were easy to fulfill as many of the documentation proposals of ML field, pay a big attention to it. As the main issue, the curation rationale was typically used to describe the final data composition. We were able to extract the main insights from the documentation of the dataset regarding the gathering, labeling, and preprocessing of the data. It could be noted that the preprocess of the data was omitted from the excerpts of the main paper. It was omitted for the sake of explanation and space reasons.

Data preprocessing has a similar structure to labeling and gathering, but we observe that it contains a set of semantics that should be considered in the future. For example, the type of quality assurance tool or the specific preprocesses applied (such as sample techniques). Regarding the social part, we were able to annotate the issues present in the different documentation sources and relate them with the gathering and labeling process, respectively.

Listing 4. Provenance and Social Concerns melanoma excerpt

```
Data Provenance:
        Curation Rationale: "Collaboration among hospitals of several
                             countries. The curation process has been
                             conducted by several health
                             institutions..."
        Gathering Processes:
            Process: Melanoma_Institute_Australia
                Description:
                    "Practitioners taking pictures from patient
                    melanoma's skin."
                Type: Manual Human Curators
                Source: imagePictures
                    Description: "Practitioners taking pictures in
                                  hospital environments"
                    Noise:
                        "Pictures were taken using cameras. Inconsistent
                        lighting in images may alter skin type."
                        "Duplicates: Due to a clerical error during the
                        data ingestion process to the ISIC Archive, 425
                        pixel-wise identical duplicate images were
                        ingested and included in the dataset. [...]"
                [...]
                Related Instances: skinImages
                Social Issues patientsPrivacy, skinColorRepresentative
                When data was collected:
                    "Images were originally collected by imaging
                    centers during 1998 - 2019; this dataset was
                    curated from those image databases in 2019 - 2020."
                Process Demographics:
                    Countries: 'Australia' [...]
                Gather Requirements
                  Requirement: "1) We queried clinical imaging
```

```
                databases across the six centers to generate
                a multicenter imaging dataset."
                [...]

      LabelingProcesses:
        Process: DiagnosisLabel
          Description: "Medical staff visualizing
          images and annotiating the diagnosis"
          Type: Image & video annotations
          Labels: skinImages.benignant_malignant
          Labeling Team:
              Description: "Internal Medical staff"
              Type: Internal
          Labeling Requirements
              Requirement: "1) Images containing any potentially
                            identifying features, such as
                            jewelry or tattoos, or from patients
                            without at least three qualifying
                            images were excluded during quality
                            assurance review."
              Requirement: "2) When multiple..." [...]
          [...]
      Preprocesses:
           Preprocess: QualityAssesment
              Description:
              "A software annotation tool, called 'Tagger,' was
              developed internally to review diagnostic labeling
              of grouped images 13. [...]"

  Social Concerns:
    Rationale: "There are several social concerns attached..."
    Social Issue: patientsPrivacy
      IssueType: Privacy
      Related Attributes: ageGroup, sex
      Description: "There are privacy patients concerns regarding
                    the age and sex of the patients."
    Social Issue: skinColorRepresentative
      IssueType: Bias
      Related Attributes: ImageId
      Description: "Dataset is not representative with respect to
                   darker skin types."
    [...]
```

3 The *Movie Review Polarity*

The Movie Review Polarity dataset has been used by the Datasheets for datasets [3] proposal as a benchmark and is widely known inside the machine learning community. We described this dataset using the information delivered by the original authors [7] and the documentation generated in the Datasheets for datasets [3] paper as an example. First released in 2004, this dataset is also a benchmark for natural language sentimental classification tasks. In comparison with Melanoma's dataset, is composed mainly of natural language text instead of images.

3.1 Metadata

In Listing 5, we have an excerpt of the Metadata Polarity dataset description. As the proposal of Gebru et al. has used these datasets, the main part of the metadata was easy to translate into our DSL, as it has a similar structure. In that case, in Line 6, the three attributes of the description were easy to complete following the documentation generated by Datasheet for datasets proposal example, as well as the *Applications* and *Distribution* part.

In the Authoring part, in line 42, in contrast with Melanoma dataset, the syntax of the DSL has allowed expressing all the relevant information. At the *Founders* at line 46, we see that the grantors and the specific grantID are missing in comparison with the Melanoma's dataset.

Finally, from line 52, the dataset contains an *Erratum*, a policy to *Version support* and a *Contribution guide*. These concepts are mostly qualitative, and it is not clear if they share a specific semantic between the different cases. Therefore, we express it as a textual description to fit all the possible causes.

Listing 5. Metadata polarity excerpt

```
1
2 Metadata:
3     Title: "Movie Review Polarity"
4     Unique-identifier: Movie_Review_Polarity
5     Version: v0001
6     Description:
7         Purposes:
8             "The dataset was created to enable research on predicting
9             sentiment polarityi.e., given a piece of English text,
10            predict whether it has a positive or negative effector
11            stancetoward its topic[...]."
12        Tasks: "Classification"
13        Gaps:
14            "To fill the Gender Reference gap inside research
15            conference"
16    Licenses: CC BY 4.0 (Attribution 4.0 International)
17    Applications:
18        Past Uses:
```

```
          "At the time of publication, only the original paper
          (http://xxx.lanl. gov/pdf/cs/0409058v1) [...]"
      Recommended:
          "The dataset could be used for anything related to modeling
          or understanding movie reviews. [...]"
      Non-recommended:
          "This data is collected solely in the movie review domain,
          so systems trained on it may or may not generalize to
          other sentiment prediction tasks.[...]"
  Distribution:
      Is public?: yes
      How is distributed:
          "The dataset is distributed on Bo Pang's webpage at Cornell:
          http://www.cs.cornell.edu/people/pabo/movie-review-data.
          The dataset does not have a DOI, and there is no redundant
          archive. The dataset was first released in 2002."
      Distribution Licenses:
          "The crawled data copyright belongs to the authors of the
          reviews unless otherwise stated. There is no license,
          but there is a request to cite the corresponding paper if
          the dataset is used: [...]"
  Area: Sentiment
  Tags: Movie Review Sentiment Classification
  Authoring:
      Authors:
          Name "Bo Pong" Email "XXXX@email.com"
          Name "Lillian Lee"  Email "XXXX@email.com"
      Funders:
          Name "National Science Fundations" type mixed
          Name "Deparment of the Interior" type public
          [...]
          Maintainer:
              Name "Bo Pong" Email "XXXX@email.com"
          Erratum:
              "Since its initial release (v0.9), there have been three
              later releases (v1.0, v1.1, and v2.0). There is not an
              explicit erratum, but updates and known errors[...]"
          Version lifecycle
              "The dataset has already been updated; older versions
              are kept around for consistency."
          Contribution guidelines:
              "The curators of the dataset, Bo Pang and Lillian lee,
              can be contacted at [...]"
```

3.2 Composition

The composition part is less relevant than the other examples because it comprises raw text with some annotations. Anyway, it allows for communication of the dataset's structure, which is not trivial. For instance, following Listing 6, in line 9, the description explains the folder structure and logic of the dataset.

On the other hand, and as a *Statistical* example, in line 22, we can see the categorical distribution of the labeled attribute "tag." This means that probably the dataset has been sampled to balance the different tags. The sampling process can be seen in Line 28. Finally, there is an explanation regarding the sampling process used and some recommendations regarding the data split.

Listing 6. Composition polarity excerpt

```
1 Composition:
2   Rationale:
3     "The instances are movie reviews extracted from newsgroup postings,
4     together with a sentiment polarity rating for whether the text
5     corresponds [...]"
6   Total size: 64702
7   Data Instances:
8     Instance: MovieReviews
9       Description:
10        "Each instance consists of the text associated with the review,
11        with obvious rating information removed from that text (some
12        errors were found and later fixed). [...]"
13      Type: Record-Data
14      Attribute number: 6
15      Attributes:
16        Attribute: fold_id [...]
17        Attribute: text [...]
18        Attribute: tag
19          Description: "The label annotated by the reviewers."
20          OfType: Categorical
21        [...]
22    Statistics:
23      Quality Metrics:
24        Completeness: 100
25        Class Balance "attribute 'tag': 50% positive, 50% negative"
26    Dependencies:
27      Description: "The dataset is entirely self-contained."
28    Is sample:
29      "The sample is from instances collected in English movie reviews
30      from the rec.arts.movies.reviews newsgroup, from which a number
31      of stars rating could be extracted. The sample is [...]"
32    Data Splits:
33      "The instances come with a cross-validation tag to enable
34      replication of cross-validation experiments; results are measured
35      in classification accuracy
```

3.3 Social Concerns

The data of the movie review dataset was directly gathered by the authors from social media. Therefore, the gathering process is simple but presents some privacy issues we will describe in the last part of the description document. On the other hand, the labeling process followed by the taggers was also well documented, and the requirements of the process were shared publicly. This dataset represents a good benchmark in terms of provenance information, and the DSL has been able to express all the relevant concepts. In Listing 7, we have an excerpt of the Provenance and Social Concerns description of the Movie Review Dataset.

Following Listing 7, in line 8, we see the *Gathering Process*. The relevant points here are the Social Issues related to these processes. In these cases, the *privacyAware* issue raises an alert regarding the lack of authorization by the original users to use the data. Humans Curators gathered the data, and the Team demographics are not important because they are the same authors, and the process demographics are not defined.

In line 29, we see the *Labeling Process*. This process has a set of requirements shared between annotators that are defined in the *Requirement*, in line 39. With the provenance information and with these requirements, it's easy to collect new data and label it to help with, for example, experiments' replicability.

At last, we see that beyond the privacy issue, the dataset raises other issues, such as the possibility of having offensive content in the text, as there has not been filtered, or the possibility of having information that can reveal identities, such as emails, in the extracted text.

Listing 7. Provenance and Social Concerns polarity excerpt

```
1  Data Provenance:
2         Curation Rationale:
3             "The data was mostly observable as raw text, except that the
4             labels were extracted by the process described below. The
5             data was collected by downloading reviews from the IMDb
6             archive of the rec.arts.movies.reviews newsgroup, at
7             http://reviews.imdb.com/Reviews."
8         Gathering Processes:
9           Process: IMdbGather
10            Description:
11            "The data was collected by downloading reviews from the
12            IMDb archive of the rec.arts.movies.reviews newsgroup,
13            at http://reviews.imdb.com/Reviews."
14            Type: API
15            Source: IMDb
16                Description:
17                    "The sample of instances collected in English
18                    movie reviews from the rec.arts.movies.reviews
19                    newsgroup, from which [...]"
20                Noise: "unknown"
21            Related Instances: MovieReviews
```

```
        Social Issues privacyAware
        How data is collected: Manual Human Curator
        When data was collected:
        "There are 1,400 instances in total in the
        original (v1.x versions) and 2,000 instances
        in total in v2.0 (from 2014)."

Labeling Processes:
  Process: labelprocess1
    Description:
      "A Rating between 0 a 5 fives stars following the
       Requirements are listed below to determine whether a review
       was positive or negative.
     The original HTML  [...]"
    Type: Entity annotation
    Labels: moreReviews.tag
    Label Requirement
      Requirement:
        "- In order to obtain more accurate rating decisions, the
        maximum rating must be specified explicitly, both for the
        numerical ratings and star ratings.  ('8/10', 'four out of
        five', and 'OUT OF ****: *** are examples of rating
        indications we recognize.)."
        [...]

Preprocesses:
    Preprocess: Cleaning
      Description:
        "Instances for which an explicit rating could not be found
        were discarded. Also only instances
        [...]"

 Social Concerns:
    Social Issue: privacyAware
      IssueType: Privacy
      Description:
        "Individuals were not aware of data collection. The data
        was crawled from public web-sources, [...]"

    Social Issue: inappropiateContent
      IssueType: Social Impact
      Related Attributes: text
      Description:
        "Some movie reviews might contain moderately inappropriate
        or offensive language, but we do not expect this to be the
        norm"
      Instance belong to peopl:
        Are there proctected groups? "No"

    Social Issue: personalInformation
```

```
72          IssueType: Privacy
73          Description:
74            "Some personal information is retained from the newsgroup
75            posting in the raw form of [...]"
```

4 Gender Inclusive Coreference Dataset

The Gender Inclusive Coreference Dataset is a dataset that has used the proposal of *Datasheet for Datasets* [3] to document itself. It is a dataset intended to study the coreference resolution in English for gender-variant people. We described this dataset using the documentation generated by the authors. An interesting point of this dataset is a dataset with a deep commitment to the data provenance and the part of the social concern. Instead, it provides only little information in the composition part.

4.1 Metadata

In Listing 8, we have an excerpt of the Metadata part of the gender dataset. The authors clearly describe the datasets' purposes, tasks, and gaps and the recommended and non-recommended applications. It is distributed under and free documentation license and provides specific information about the authors, funders, and maintainers.

Listing 8. Metadata gender excerpt

```
Metadata:
 Title: "Gender Inclusive Coreference Dataset"
 Unique-identifier: Gender_Inclusive_Dataset
 Version: v0001
 Description:
   Purposes:
       "This dataset was created to study coreference resolution in
       English on documents discussing people who are, in some ways,
       gender-variant (and generally selected so that this variance
       shows up linguistically). Previously coreference resolution
       datasets contain nearly no such examples, largely due to
       the ways in  which those datasets were collected (see paper
       for details)."
   Tasks: "Classification"
   Gaps:
       "As part of a study making coreference systems more gender
       inclusive, we collected and annotated a dataset of documents
       by and about non-binary and binary trans people."
 Licenses:  FDL 1.3 (GNU Free Documentation License 1.3)
 Applications:
   Past Uses: "The dataset has been used to understand the human
               annotation biases and to test existing coreference
               systems. See the the paper linked at the top for
               more details."
   Recommended:
       "The dataset could possibly be used for developing or testing
       systems for referring expression generation."
   Non-recommended:
       "This dataset should not be used for any sort of 'gender
```

```
        prediction.'' First, anyone using this dataset  (or any
        related dataset, for that matter), should recognize that
        'gender' doesn't mean any single thing, and furthermore that
        pronoun != gender. Furthermore, because of the fluid and
        the temporal notion of gender- -and of gendered referring
        expressions like pronouns and terms of address--just
        because a person is described in this dataset in one
        particular way does not mean that this will always
        be the appropriate way to refer to this person."
Distribution:
  Is public?: yes
  How is distributed:
      "The dataset is free for download at
      github.com/hal3/gicoref-dataset. The dataset is distributed as
      of June 2020 in its first version."
  Distribution Licenses:
      "The dataset is licensed under a BSD license."
Area: Gender
Tags: Conference, Trans, NonBinary, Inclusive
Authoring:
  Authors:
      Name "Yang Trista Cao" Email "XXXX@email.com"
      Name "Hal Daume III"   Email "XXXX@email.com"
  Funders:
      Name "UMD CS department" type mixed
      Name "MSR" type private
  Maintainer:
      Name "Yang Trista Cao" Email "XXXX@email.com"
      Name "Hal Daume III"   Email "XXXX@gmail.com"
  Erratum:
      "Currently, no. As errors are encountered, future versions of
      the dataset may be released (but will be versioned). They will
      all be provided in the same GitHub location."
  Version lifecycle
      "All data will be versioned."
  Contribution guidelines:
      "Errors may be submitted via the bug tracker on GitHub. More
      extensive augmentations may be accepted at the authors'
      discretion."
```

4.2 Composition

In Listing 9, we have an excerpt of the Composition part of the gender dataset. This part is the part with less information provided by the authors, and some of the information about its composition was, instead, described during the provenance rationale description. The interesting point is the *Sparsity* of the dataset, which is the number of 0 along the data. A higher number means that there are numerous values equal to 0. Moreover, the other interesting point is the recommendations about sampling and data splits that the authors provide.

Listing 9. Composition gender excerpt

```
Composition:
  Rationale:
    "Each instance is an English document, annotated with coreference
    links only for person entities. In particular, each entity is a
    set of mentions, which are annotated as contiguous text spans.
    Each mention is assigned a numeric identifier to group them into
    mentions that all refer to the same entity."
   Total size: 95
   Data Instances:
     Instance: AnnotatedDocuments
        Description:
            "Each instance consists of text that has been sentence
            separated, tokenized, and annotated with mentions and entity
            identifiers. It is in a CoNLL-style format with bracketing
            for entities"
        Type: Record-Data
        Attribute number: 2
        Attributes:
            Attribute: document
                Description: "The analyzed text."
                Count: 65440
                OfType: Categorical
            Attribute: person_annotation
                Description: "Word tagged as a people entity (note
                that a word could refer to more than one person) "
                Count: 65440
                OfType: Categorical

        Statistics:
            Quality Metrics:
                Sparsity: 8.9
                Noisy labels "Please consider some humans errors in
                the annotation"
        Is sample:
            "It is a sample of all possible documents. It is not
            intended to be representative  (in fact, it is known
            to be quite non-representative): it was specifically
            designed  to focus on documents that contain mentions
```

```
            of people whose gender in some way does not fall
            within the gender binary."
        Data Splits:
            "We expect this data to be used solely for testing purposes.
            We do not explicitly provide a training/validation/testing
            split; however, we recognize that people may wish to do
            this or to do some form of cross-validation. We would
            suggest cross-validation, given that some phenomena only
            occur in a few documents and are likely to be lost in
            any random split.

            Warning: If you do use any of the data for training/testing,
            either in a cross-validation setup or otherwise, you may
            wish to be careful to ensure that documents on very similar
            topics are always in the same split. For instance, there
            are two documents about Leslie Feinberg in the dataset;
            you should ensure that these are in the same split or
            evaluation scores are likely to be inflated."
```

4.3 Provenance and Social Concerns

In Listing 10, we have an excerpt of the Provenance and Social Concerns part of the gender dataset. This dataset has a special commitment to the provenance and the social concerns. The interesting point in these parts is the multiple sources and the description of each source. These represent a complex scenario where our DSL has been able to express this complexity without problems.

On the other hand, the Labeling process was done with a professional software called TagEditor. The authors refer to the software repository (which is open-source) and its documentation.

Lastly, the authors raise a set of social concerns relevant to the dataset, providing a deep social and cultural point of view to the documentation. We observed that the structure of our DSL helped organize the ideas of the original authors.

Listing 10. Provenance gender excerpt

```
Data Provenance:
        Curation Rationale:
            "The data was all downloaded directly from associated
            webpages (Wikipedia, periodicals, or AO3). Tokenization
            and sentence segmentation was initially done automatically
            but corrected by hand.

            Finally, to make annotation more feasible, we truncated
            every document at 1000 space-separated  'words' prior to
            tokenization, so after tokenization, the documents may be
            somewhat longer than 1000 tokens. Finally, there were three
            documents which we forgot to truncate, and so made their
```

```
            way into the dataset at full length (these are all from
            AO3)."
        Gathering Processes:
          Process: Wikipedia
            Description:
                "Articles in English Wikipedia, in
                particular drawn from Wikipedia's List of People
                with Non-binary Gender Identities. These documents
                comprise about 2/3 of the dataset."
            Source: Wikipedia
              Description: "a description"
              Noise: "Already defined"
            Related Instances: AnnotatedDocuments
            How data is collected: Manual Human Curator


          Process: PeriodicalsProcess
            Description: ""
            Source: Periodicals
            Description:
                "Articles from periodicals which provide
                coverage of LGBTQ+ issues or are aimed at the LGBTQ+
                community, mostly drawn from Wikipedia's List of LGBT
                Periodicals. These were collected by issuing a number
                of search queries to Google of the form site:sss
                ppp, where sss is the webpage for one of the
                periodicals and ppp is a non-binary pronoun from the
                set {ze+hir, ze+zir, xey+xem, ey+em, zey+zem, xe+xir}.
                The top te results from each site was read by one of the
                authors to ensure that the use of ppp was actually a use
                of that pronoun, and included if so. Note that because
                these articles were specifically returned as a result of
                queries for that specific set of six neo-pronoun pairs,
                this part of the data is unlikely to include many
                examples of other pronouns, and few examples of singular
                they.

                The complete set of periodicals searched were:
                thetransgentimes.com digitaltransgenderarchive.net
                ifge.org lotl.com qnews.com.au starobserver.com.au
                dailyxtra.com fugues.com wayves.ca thebuzzmag.ca
                pinkplaymags.com pink-pages.co.in attitude.co.uk
                mag.bent.com divamag.co.uk fyne.co.uk gaytimes.co.uk
                pridelife.com boyz.co.uk scotsgay.co.uk
                connextionsmagazine.com curvemag.com hellomrmag.com
                instinctmagazine.com lavendermagazine.com
                metrosource.com omgmag.com out.com
                therainbowtimesmass.com rfdmag.org pride.com
                travelsofadam.com plentitudemagazine.ca, though
                in the end we only found/included documents from
```

```
                The Advocate, The DailyXtra, DigitalTrans, GLReview,
                LambdaLiterary and Lavender. These documents comprise
                about 1/6 of the dataset."
                Noise: "See the potential impact in the social concerns
                        section"
            Related Instances: AnnotatedDocuments
            How data is collected: Manual Human Curator

        Process: AO3Process
            Description: ""
            Source: AO3
                Description:
                "Fan-fiction stories drawn from Archive
                Of Our Own (AO3), a fan-fiction site created and run
                by fans, which includes a large number of stories
                centering around binary and non-binary trans
                characters. We selected stories by first considering
                the Gender-Neutral Pronouns tag (around 3700 stories),
                and then filtering the results to include only General
                Audience stories (vs, eg, 'Mature'), to include
                only No Archive Warnings Apply (vs, eg, 'Graphic
                Descriptions of Violence' or non-consentual sex), and
                limited the results to 'Complete Works Only.'' After
                that, we sorted the remaining approximately 900
                stories by word count, and took the shortest ones that
                were actually stories (some were descriptions
                of audiobooks). These documents comprise about 1/6 of
                the dataset."
                Noise: "See the potential impact in the social concerns
                        section"
            Related Instances: AnnotatedDocuments
            How data is collected: Manual Human Curator
    LabelingProcesses:
      Process: Labeling1
        Description:
            "Our annotation process was for the authors to first
            independently annotate the same five
            documents (all Wikipedia), then compare and discuss
            differences in the annotation in the adjudication process.
            We then independently annotated the rest of the documents,
            followed by a final adjudication step.
            The pre-adjudication documents are also included."
        Type: Entity annotation
        Labels: AnnotatedDocuments.person_annotation
            Label:
            Description:
                "The labels are generated with the TagEditor software,
                and tag the words referencing entity persons.
                A numerical order has been used to identify other
                reference to the same entity. In case of more than one
```

```
            entity person code, an notation as: 1(1), 1(0), 1(2)
            are provided "
    Labeling Team:
        Description: "The two authors"
    Label Requirement:
        Requirement: "The annotation was performed by TagEditor
        software: https://github.com/d5555/TagEditor"

Social Concerns:
    Social Issue: privacyIndividual
        IssueType: Privacy
        Related Attributes: document
        Description: "Data could identify individuals. All raw data
          in the dataset is from public sources (Wikipedia, online
          periodicals, or online open fan-fiction). This is not
          explicitly identified, though many of the articles
          explicitly mention the gender of the people
          described/discussed. Their names could be given in running
          text."

    Social Issue: senstiveInfo
        IssueType: Sensitive_Data
        Related Attributes: document
        Description:
            "All documents deal with people, some of whom have had
            traumatic events in their
            lives that are discussed, particularly around their
            gender identities. All should be read with
            that in mind. Some of the articles, especially from
            fan-fiction, have explicit and intentional
            misgendering of characters, and a few describe (or hint
            out) intimate relationships (though we filtered for
            only stories that did not have 'adult themes')."
        Instance belong to peopl:
            Are there proctected groups? "Yes, most of the articles
            relate to real people (except the fan-fiction
            articles)."
            Might be offensive "The gender information given may be
            sensitive in certain situations, at least for those
            articles dealing with real people. However, all these
            articles are from either Wikipedia or periodicals,
            and is therefore already public information."

    Social Issue: underrepresented
        IssueType: Bias
        Description:
            "While the dataset was specifically constructed to be
            gender-inclusive, it undoubtedly fails in some ways
            to fully achieve that goal. Part of this is due to the
            nature of the underlying texts (e.g., Wikipedia's
```

```
        frequent use of deadnames) and part of it is due
        to plain difficulties in the collection (automatically
        distinguishing specific singular uses of 'they'
        from other uses is currently not possible, and so
        despite 'they' being currently likely, the most
        commonly used non-binary pronoun, it is perhaps
        underrepresented in this dataset). Preprocessing
        hopefully did not introduce errors (in fact, we
        corrected for many tokenization errors, for instance,
        that the default tokenizer did not know to split
        'xe'll' into two tokens as it would 'she'll').

        All of these sources have their own biases. Wikipedia
        texts tend to overuse people's deadnames,
        and tend to treat a person's use of pronouns as a
        'surprise' at the end of the document.
        As mentioned above, the periodical documents are
        specifically constrained to contain certain pronouns.
        And the AO3 stories are also specifically selected to
        have non-binary pronouns."
```

5 Conclusions

During the description of the mentioned dataset, we can state that all elements of the datasets were properly modeled with our DSL. Although, we observe that some datasets have relevant missing information. For instance, in the Gender and Polarity datasets, relevant statistical information and quality metrics in the data composition were missing. We had to do a manual exploratory data analysis to populate this part. Moreover, the Polarity dataset has incomplete information regarding the gathering process, which we see as highly important given the topic of the dataset.

Sometimes the information was there, but *hidden* inside descriptions focused on other aspects of the dataset. For instance, the Polarity and Melanoma datasets use the gathering rationale to express essential details about the data composition. The melanoma dataset also missed detailed information regarding social concerns, not enough to make it operational as part of the dataset description.

In conclusion, our DSL can express all the relevant concepts present in the documentation and could be a tool to prompt authors to complete the relevant missing parts in the dataset documentation.

References

1. Bender, E.M., Friedman, B.: Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics **6**, 587–604 (2018)
2. Cao, Y.T., Daumé III, H.: Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. Computational Linguistics **47**(3), 615–661 (2021)
3. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM **64**(12), 86–92 (2021)
4. Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P.S., Anuoluwapo, A., Bosselut, A., Chandu, K.R., Clinciu, M., Das, D., Dhole, K.D., et al.: The gem benchmark: Natural language generation, its evaluation and metrics. arXiv preprint arXiv:2102.01672 (2021)
5. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The dataset nutrition label. Data Protection and Privacy, Volume 12: Data Protection and Democracy **12**, 1 (2020)
6. McMillan-Major, A., Osei, S., Rodriguez, J.D., Ammanamanchi, P.S., Gehrmann, S., Jernite, Y.: Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. In: Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics. pp. 121–135. ACM, Online (2021)
7. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. pp. 271–es (2004)
8. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Lioprys, K., Malvehy, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., Soyer, H.P.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data **8**(1), 34 (Jan 2021)