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 information to build a complete description document but to extract the information,

---

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

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 9. 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 sub-concepts.

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 of 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

```
 1 Metadata:
 2   Title: "2020 SIIM-ISIC Melanoma Classification Challenge Dataset"
 3   Unique-identifier: SIC_Melanoma_Classification_Challenge_Dataset
 4   Version: v0001
 5   Release Date: 08-10-20
 6   Description:
 7     Purposes:
 8       "Advance medical image innovation. The 2020 SIIM-ISIC Melanoma
 9       Classification challenge dataset was created for the purpose
10       of conducting a machine learning competition to identify
11       melanoma in lesion images. [...]"
12     Tasks:
13       "Classification"
14     Gaps:
15       "Improve melanoma detection of ML models. As the leading
      healthcare organization for informatics
16       in medical imaging, the Society for Imaging Informatics
17       in Medicine (SIIM)'s mission is to advance medicine
18       imaging informatics through education, research, and
19       innovation in a multi-disciplinary community. SIIM is
20       joined by the International Skin Imaging Collaboration
21       (ISIC), an international effort to improve melanoma
22       diagnosis. "
23   Licenses: CC BY-NC 4.0 (Attribution-NonCommercial 4.0 International)
24   Area: HealthCare
25   Tags: Images, Melanoma, diagnosis, Skin Image
26   Applications:
27     Past Uses:
28         "Yes. The 2020 SIIM-ISIC Melanoma Classification challenge
29         dataset was created for the purpose of conducting a machine
30         learning competition to identify melanoma in lesion[...]"
31     Recommended:
32         "Melanoma skin detection"
33         "Predict incidence of melanoma in a population"
34     Non-recommended:
35         "Due to low population prevalence and challenges with access to
36         care in different populations, the images gathered for large
37         datasets such as this for AI classification [...]"
38   Distribution:
39     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/ "
40
41
```

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 10 and 11, 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 the relevant information of the dataset, for example, in line 20, regarding 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: sex
20             Description: "Sex of the patient"
21             Count: 33126
22             OfType: Categorical
23             Statistics:
24                 Mode: "Famela"
25                 Categorical Distribution:
26                     "Male: 40%
27                      Famela: 55%
28                      Unknown: 5%"
29           Attribute: ageGroup
30               Description: "The age group of patients"
```

```
31              OfType: Categorical
32              Statistics:
33                 Mode: "40-50"
34                 Categorical Distribution:
35                    "0-10": 15% - "10-20": 12% [...]
36          Attribute: anatom_site_general_challenge [...]
37          Attribute: diagnosis [...]
38          Attribute: beningnant_malignant
39            Description: "Medical diagnosis of the patient"
40            Labelling process: DiagnosisLabel
41            OfType: Categorical
42            Statistics:
43               Categorical Distribution:
44                    "beningnant: 88%
45                    malignant: 12%"
46    Statistics:
47        Pair Correlation:
48            Between age and benignant_malignant
49            Between age and sex
50            Between age and external source
51                From: "Official population indicator of an
52                       hypothetical public institution"
53                Rationale: "Similar age distribution"
54        Quality Metrics:
55            Completeness: 100%
56    Consistency rules:
57        Inv: skinImages: (ageGroup>=0)
58    Is sample:
59        "Yes, the sample dataset used accurately reflects the intent or
60        desired outcome and has been evaluated to ensure that
61        it is fit for purpose. [...]"
62
63    Data Splits:
64        "Yes. To test algorithm generalizability, a subset of
65        images from six sites (five geographic locations) were
66        allocated for the training dataset of the 2020 ISIC Grand
67        Challenge. There is a separate test dataset available without
68        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 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

```
1 Data Provenance:
2       Curation Rationale: "Collaboration among hospitals of several
3                            countries. The curation process has been
4                            conducted by several health
5                            institutions..."
6       Gathering Processes:
7          Process: Melanoma_Institute_Australia
8             Description:
9                "Practitioners taking pictures from patient
10               melanoma's skin."
11            Type: Manual Human Curators
12            Source: imagePictures
13               Description: "Practitioners taking..."
14               Noise:
15                  "Pictures were taken using cameras. Inconsistent
16                  lighting in images may alter skin type."
17                  "Duplicates: Due to a clerical error during the
18                  data ingestion process to the ISIC Archive, 425
19                  pixel-wise identical duplicate images were
20                  ingested and included in the dataset. [...]"
21            [...]
22            Related Instances: skinImages
23            Social Issues patientsPrivacy, raceRepresentative
24            When data was collected:
25               "Images were originally collected by imaging
26               centers during 1998 - 2019; this dataset was
27               curated from those image databases in 2019 - 2020."
28            Process Demographics:
29               Countries: 'Australia' [...]
30            Gather Requirements
31              Requirement: "1) We queried clinical imaging
32              databases across the six centers to generate
```

```
33                         a multicenter imaging dataset."
34                     [...]
35
36         LabelingProcesses:
37             Labeling process: DiagnosisLabel
38                 Description: "Medical staff visualizing
39                 images and annotiating the diagnosis"
40                 Type: Image & video annotations
41                 Labels: skinImages.benignant_malignant
42                 Labeling Team:
43                     Who collects the data: "Internal Medical staff"
44                     Type: Internal
45                 Labeling Requirements
46                     Requirement: "1) Images containing any potentially
47                                     identifying features, such as
48                                     jewelry or tattoos, or from patients
49                                     without at least three qualifying
50                                     images were excluded during quality
51                                     assurance review."
52                     Requirement: "2) When multiple..." [...]
53                 [...]
54         Preprocesses:
55             Preprocess: QualityAssesment
56                 Description:
57                 "A software annotation tool, called 'Tagger,' was
58                 developed internally to review diagnostic labeling
59                 of grouped images 13. [...]"
60
61     Social Concerns:
62       Rationale: "There are several social concerns attached..."
63       Social Issue: patientsPrivacy
64         IssueType: Privacy
65         Related Attributes: ageGroup, sex
66         Description: "There are privacy patients concerns regarding
67                       the age and sex of the patients."
68       Social Issue: raceRepresentative
69         IssueType: Bias
70         Related Attributes: attribute: ImageId
71         Description: "Dataset is not representative with respect to
72                       darker skin types."
73       [...]
```

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 7, the three attributes of the description were easy to complete following the documentation generated of Datasheet for datasets, as well as the *Applications* and *Distribution* part.

In the Authoring part, in line 43, in contrast with Melanoma dataset, the syntax of the DSL has allowed expressing all the relevant information. At the *Founders* at line 47, we see that the grantors and the specific grantID are missing in comparison with the Melanoma's dataset.

Finally, from line 53, 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:
```

```
19              "At the time of publication, only the original paper
20              (http://xxx.lanl. gov/pdf/cs/0409058v1) [...]"
21        Recommended:
22              "The dataset could be used for anything related to modeling
23              or understanding movie reviews. [...]"
24        Non-recommended:
25              "This data is collected solely in the movie review domain,
26              so systems trained on it may or may not generalize to
27              other sentiment prediction tasks.[...]"
28    Distribution:
29        Is public?: yes
30        How is distributed:
31              "The dataset is distributed on Bo Pang's webpage at Cornell:
32              http://www.cs.cornell.edu/people/pabo/movie-review-data.
33              The dataset does not have a DOI, and there is no redundant
34              archive. The dataset was first released in 2002."
35        Distribution Licenses:
36              "The crawled data copyright belongs to the authors of the
37              reviews unless otherwise stated. There is no license,
38              but there is a request to cite the corresponding paper if
39              the dataset is used: [...]"
40    Area: Sentiment
41    Tags: Movie Review Sentiment Classification
42    Authoring:
43        Authors:
44              Name "Bo Pong" Email "XXXX@email.com"
45              Name "Lillian Lee"  Email "XXXX@email.com"
46        Funders:
47              Name "National Science Fundations" type mixed
48              Name "Deparment of the Interior" type public
49              [...]
50        Maintainer:
51              Name "Bo Pong" Email "XXXX@email.com"
52        Erratum:
53              "Since its initial release (v0.9), there have been three
54              later releases (v1.0, v1.1, and v2.0). There is not an
55              explicit erratum, but updates and known errors[...]"
56        Version lifecycle
57              "The dataset has already been updated; older versions
58              are kept around for consistency."
59        Contribution guidelines:
60              "The curators of the dataset, Bo Pang and Lillian lee,
61              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 categoric 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 32. 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 unknown.

In line 27, we see the Labeling process. This process have a set of requirements shared between annotators that are defined in the *Requirement*, in line 41. 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
```

```
22              Social Issues privacyAware
23              How data is collected: Manual Human Curator
24              When data was collected:
25              "There are 1,400 instances in total in the
26              original (v1.x versions) and 2,000 instances
27              in total in v2.0 (from 2014)."
28
29      LabelingProcesses:
30          Labeling process: labelprocess1
31          Description:
32            "A Rating between 0 a 5 fives stars following the
33             Requirements are listed below to determine whether a review
34             was positive or negative.
35           The original HTML  [...]"
36          Type: Entity annotation
37          Labels: moreReviews.tag
38          Label Requirement
39            Requirement:
40              "- In order to obtain more accurate rating decisions, the
41              maximum rating must be specified explicitly, both for the
42              numerical ratings and star ratings.  ('8/10', 'four out of
43              five', and 'OUT OF ****: *** are examples of rating
44              indications we recognize.)."
45              [...]
46
47      Preprocesses:
48          Preprocess: Cleaning
49            Description:
50              "Instances for which an explicit rating could not be found
51              were discarded. Also only instances
52              [...]"
53
54       Social Concerns:
55          Social Issue: privacyAware
56            IssueType: Privacy
57            Description:
58              "Individuals were not aware of data collection. The data
59              was crawled from public web-sources, [...]"
60
61          Social Issue: inappropiateContent
62            IssueType: Social Impact
63            Related Attributes:
64              Attribute: text
65            Description:
66              "Some movie reviews might contain moderately inappropriate
67              or offensive language, but we do not expect this to be the
68              norm"
69            Instance belong to peopl:
70              Are there proctected groups? "No"
71
```

```
72          Social Issue: personalInformation
73            IssueType: Privacy
74            Description:
75              "Some personal information is retained from the newsgroup
76              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 a 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

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

```
30        prediction.'' First, anyone using this dataset  (or any
31        related dataset, for that matter), should recognize that
32        'gender' doesn't mean any single thing, and furthermore that
33        pronoun != gender. Furthermore, because of the fluid and
34        the temporal notion of gender- -and of gendered referring
35        expressions like pronouns and terms of address--just
36        because a person is described in this dataset in one
37        particular way does not mean that this will always
38        be the appropriate way to refer to this person."
39  Distribution:
40    Is public?: yes
41    How is distributed:
42        "The dataset is free for download at
43        github.com/hal3/gicoref-dataset. The dataset is distributed as
44        of June 2020 in its first version."
45    Distribution Licenses:
46        "The dataset is licensed under a BSD license."
47  Area: Gender
48  Tags: Conference, Trans, NonBinary, Inclusive
49  Authoring:
50    Authors:
51        Name "Yang Trista Cao" Email "XXXX@email.com"
52        Name "Hal Daume III"   Email "XXXX@email.com"
53    Funders:
54        Name "UMD CS department" type mixed
55        Name "MSR" type private
56    Maintainer:
57        Name "Yang Trista Cao" Email "XXXX@email.com"
58        Name "Hal Daume III"   Email "XXXX@gmail.com"
59    Erratum:
60        "Currently, no. As errors are encountered, future versions of
61        the dataset may be released (but will be versioned). They will
62        all be provided in the same GitHub location."
63    Version lifecycle
64        "All data will be versioned."
65    Contribution guidelines:
66        "Errors may be submitted via the bug tracker on GitHub. More
67        extensive augmentations may be accepted at the authors'
68        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 are the *Sparsity* of the dataset, which is the number of 0 along the data. A higher number means that there are lots of values equal to 0. Also, the other interesting point is the recommendations about sampling and data splits that the authors provide.

Listing 9. Composition gender excerpt

```
1  Composition:
2    Rationale:
3      "Each instance is an English document, annotated with coreference
4      links only for person entities. In particular, each entity is a
5      set of mentions, which are annotated as contiguous text spans.
6      Each mention is assigned a numeric identifier to group them into
7      mentions that all refer to the same entity."
8    Total size: 95
9    Data Instances:
10      Instance: AnnotatedDocuments
11        Description:
12            "Each instance consists of text that has been sentence
13            separated, tokenized, and annotated with mentions and entity
14            identifiers. It is in a CoNLL-style format with bracketing
15            for entities"
16        Type: Record-Data
17        Attribute number: 2
18        Attributes:
19            Attribute: document
20                Description: "The analyzed text."
21                Count: 65440
22                OfType: Categorical
23            Attribute: person_annotation
24                Description: "Word tagged as a people entity (note
25                that a word could refer to more than one person) "
26                Count: 65440
27                OfType: Categorical
28
29        Statistics:
30            Quality Metrics:
31                Sparsity: 8.9
32                Noisy labels "Please consider some humans errors in
33                the annotation"
34        Is sample:
35            "It is a sample of all possible documents. It is not
36            intended to be representative  (in fact, it is known
37            to be quite non-representative): it was specifically
38            designed  to focus on documents that contain mentions
```

```
39              of people whose gender in some way does not fall
40              within the gender binary."
41          Data Splits:
42              "We expect this data to be used solely for testing purposes.
43              We do not explicitly provide a training/validation/testing
44              split; however, we recognize that people may wish to do
45              this or to do some form of cross-validation. We would
46              suggest cross-validation, given that some phenomena only
47              occur in a few documents and are likely to be lost in
48              any random split.
49
50              Warning: If you do use any of the data for training/testing,
51              either in a cross-validation setup or otherwise, you may
52              wish to be careful to ensure that documents on very similar
53              topics are always in the same split. For instance, there
54              are two documents about Leslie Feinberg in the dataset;
55              you should ensure that these are in the same split or
56              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

```
1   Data Provenance:
2       Curation Rationale:
3           "The data was all downloaded directly from associated
4           webpages (Wikipedia, periodicals, or AO3). Tokenization
5           and sentence segmentation was initially done automatically
6           but corrected by hand.
7
8           Finally, to make annotation more feasible, we truncated
9           every document at 1000 space-separated  'words' prior to
10          tokenization, so after tokenization, the documents may be
11          somewhat longer than 1000 tokens. Finally, there were three
12          documents which we forgot to truncate, and so made their
```

```
13            way into the dataset at full length (these are all from
14            AO3)."
15        Gathering Processes:
16          Process: Wikipedia
17            Description:
18                "Articles in English Wikipedia, in
19                particular drawn from Wikipedia's List of People
20                with Non-binary Gender Identities. These documents
21                comprise about 2/3 of the dataset."
22            Source: Wikipedia
23              Description: "a description"
24              Noise: "Already defined"
25            Related Instances: AnnotatedDocuments
26            How data is collected: Manual Human Curator
27
28
29          Process: PeriodicalsProcess
30            Description: ""
31            Source: Periodicals
32            Description:
33                "Articles from periodicals which provide
34                coverage of LGBTQ+ issues or are aimed at the LGBTQ+
35                community, mostly drawn from Wikipedia's List of LGBT
36                Periodicals. These were collected by issuing a number
37                of search queries to Google of the form site:sss
38                ppp, where sss is the webpage for one of the
39                periodicals and ppp is a non-binary pronoun from the
40                set {ze+hir, ze+zir, xey+xem, ey+em, zey+zem, xe+xir}.
41                The top te results from each site was read by one of the
42                authors to ensure that the use of ppp was actually a use
43                of that pronoun, and included if so. Note that because
44                these articles were specifically returned as a result of
45                queries for that specific set of six neo-pronoun pairs,
46                this part of the data is unlikely to include many
47                examples of other pronouns, and few examples of singular
48                they.
49
50                The complete set of periodicals searched were:
51                thetransgentimes.com digitaltransgenderarchive.net
52                ifge.org lotl.com qnews.com.au starobserver.com.au
53                dailyxtra.com fugues.com wayves.ca thebuzzmag.ca
54                pinkplaymags.com pink-pages.co.in attitude.co.uk
55                mag.bent.com divamag.co.uk fyne.co.uk gaytimes.co.uk
56                pridelife.com boyz.co.uk scotsgay.co.uk
57                connextionsmagazine.com curvemag.com hellomrmag.com
58                instinctmagazine.com lavendermagazine.com
59                metrosource.com omgmag.com out.com
60                therainbowtimesmass.com rfdmag.org pride.com
61                travelsofadam.com plentitudemagazine.ca, though
62                in the end we only found/included documents from
```

```
63              The Advocate, The DailyXtra, DigitalTrans, GLReview,
64              LambdaLiterary and Lavender. These documents comprise
65              about 1/6 of the dataset."
66              Noise: "See the potential impact in the social concerns
67                      section"
68          Related Instances: AnnotatedDocuments
69          How data is collected: Manual Human Curator
70
71      Process: AO3Process
72          Description: ""
73          Source: AO3
74              Description:
75              "Fan-fiction stories drawn from Archive
76              Of Our Own (AO3), a fan-fiction site created and run
77              by fans, which includes a large number of stories
78              centering around binary and non-binary trans
79              characters. We selected stories by first considering
80              the Gender-Neutral Pronouns tag (around 3700 stories),
81              and then filtering the results to include only General
82              Audience stories (vs, eg, 'Mature'), to include
83              only No Archive Warnings Apply (vs, eg, 'Graphic
84              Descriptions of Violence' or non-consentual sex), and
85              limited the results to 'Complete Works Only.'' After
86              that, we sorted the remaining approximately 900
87              stories by word count, and took the shortest ones that
88              were actually stories (some were descriptions
89              of audiobooks). These documents comprise about 1/6 of
90              the dataset."
91              Noise: "See the potential impact in the social concerns
92                      section"
93          Related Instances: AnnotatedDocuments
94          How data is collected: Manual Human Curator
95  LabelingProcesses:
96      Labeling process: Labeling1
97      Description:
98          "Our annotation process was for the authors to first
99          independently annotate the same five
100         documents (all Wikipedia), then compare and discuss
101         differences in the annotation in the adjudication process.
102         We then independently annotated the rest of the documents,
103         followed by a final adjudication step.
104         The pre-adjudication documents are also included."
105     Type: Entity annotation
106     Labels: AnnotatedDocuments.person_annotation
107         Label:
108         Description:
109             "The labels are generated with the TagEditor software,
110             and tag the words referencing entity persons.
111             A numerical order has been used to identify other
112             reference to the same entity. In case of more than one
```

```
113              entity person code, an notation as: 1(1), 1(0), 1(2)
114              are provided "
115      Labeling Team:
116          Who collects the data: "The two authors"
117      Label Requirement:
118          Requirement: "The annotation was performed by TagEditor
119          software: https://github.com/d5555/TagEditor"
120
121  Social Concerns:
122      Social Issue: privacyIndividual
123          IssueType: Privacy
124          Related Attributes:
125              Attribute: document
126          Description: "Data could identify individuals. All raw data
127            in the dataset is from public sources (Wikipedia, online
128            periodicals, or online open fan-fiction). This is not
129            explicitly identified, though many of the articles
130            explicitly mention the gender of the people
131            described/discussed. Their names could be given in running
132            text."
133
134      Social Issue: senstiveInfo
135          IssueType: Sensitive_Data
136          Related Attributes:
137              Attribute: document
138          Description:
139              "All documents deal with people, some of whom have had
140              traumatic events in their
141              lives that are discussed, particularly around their
142              gender identities. All should be read with
143              that in mind. Some of the articles, especially from
144              fan-fiction, have explicit and intentional
145              misgendering of characters, and a few describe (or hint
146              out) intimate relationships (though we filtered for
147              only stories that did not have 'adult themes')."
148          Instance belong to peopl:
149              Are there procted groups? "Yes, most of the articles
150              relate to real people (except the fan-fiction
151              articles)."
152              Might be offensive "The gender information given may be
153              sensitive in certain situations, at least for those
154              articles dealing with real people. However, all these
155              articles are from either Wikipedia or periodicals,
156              and is therefore already public information."
157
158      Social Issue: underrepresented
159          IssueType: Bias
160          Description:
161              "While the dataset was specifically constructed to be
162              gender-inclusive, it undoubtedly fails in some ways
```

```
163                         to fully achieve that goal. Part of this is due to the
164                         nature of the underlying texts (e.g., Wikipedia's
165                         frequent use of deadnames) and part of it is due
166                         to plain difficulties in the collection (automatically
167                         distinguishing specific singular uses of 'they'
168                         from other uses is currently not possible, and so
169                         despite 'they' being currently likely, the most
170                         commonly used non-binary pronoun, it is perhaps
171                         underrepresented in this dataset). Preprocessing
172                         hopefully did not introduce errors (in fact, we
173                         corrected for many tokenization errors, for instance,
174                         that the default tokenizer did not know to split
175                         'xe'll' into two tokens as it would 'she'll').
176
177                         All of these sources have their own biases. Wikipedia
178                         texts tend to overuse people's deadnames,
179                         and tend to treat a person's use of pronouns as a
180                         'surprise' at the end of the document.
181                         As mentioned above, the periodical documents are
182                         specifically constrained to contain certain pronouns.
183                         And the AO3 stories are also specifically selected to
184                         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)