How might we compare classification schemes?

Mark H.
Butler
NKOS 2022

Voise

Problem Definition

Compare:

1) Scheme change over time
2) Various schemes covering the same domain.

What are the universal characteristics for comparison of classifications, the way things (classes) are arranged and described (and used)?

My Goal:

Begin asking some questions from my perspective

Are we ready to start giving answers?

What would it mean to compare two schemes?

- We’ve seen examples of comparing same scheme at different times
- Crosswalks alone?
  - Compare vocabulary in two schemes (lexical)
  - Compare triples of term – relation – term (semantic)
  - Compare terms in context of neighbors (semantic)
- Can we use UMLS crosswalks and cross-walking techniques to make comparisons?
- Is mapping comparing?

Why compare schemes?

I'm a **user** of classification schemes for automated classification

My task is to automatically create values for metadata fields for use in downstream processes (finding and organizing)

Some of my questions:

- How do I select appropriate scheme?
- How do I compare candidate schemes?
- How do I know I'm using the scheme (as intended) properly?
- What was the scheme originally designed to do?
- What data is tagged with the candidate scheme?

We need the ability to describe and compare the schemes within a domain

Possible Comparison Criteria

In order to compare classification schemes, we need to have some **set of criteria** that can be the basis for description and comparison.

- Limit to vocabulary and relations?
- What about sub-domain(s)?
- What about intended use(s)?
- What about creator(s)?
- What about user(s)?

How can we identify possible criteria that might matter?

Possible Comparisons

Perhaps look to natural language processing (NLP) and ethical AI.

They have developed, for the entities that matter to them (predictive models and datasets), **a framework for describing the attributes, creation, and uses** of those entities.

These “**model cards**” and “**data cards**” allow them to describe and discuss how the model or dataset was built, what data was used in training, how the model can be evaluated, and how the model's use could affect real people in the real world through bias and exclusion.

Model Cards and Data Cards

What are model cards?

A structured description of the ML model that allows a variety of stakeholders to understand its intended uses as well as potential limitations. They should also describe how the model was created, the data used for training and how it can be evaluated.

What are data cards?

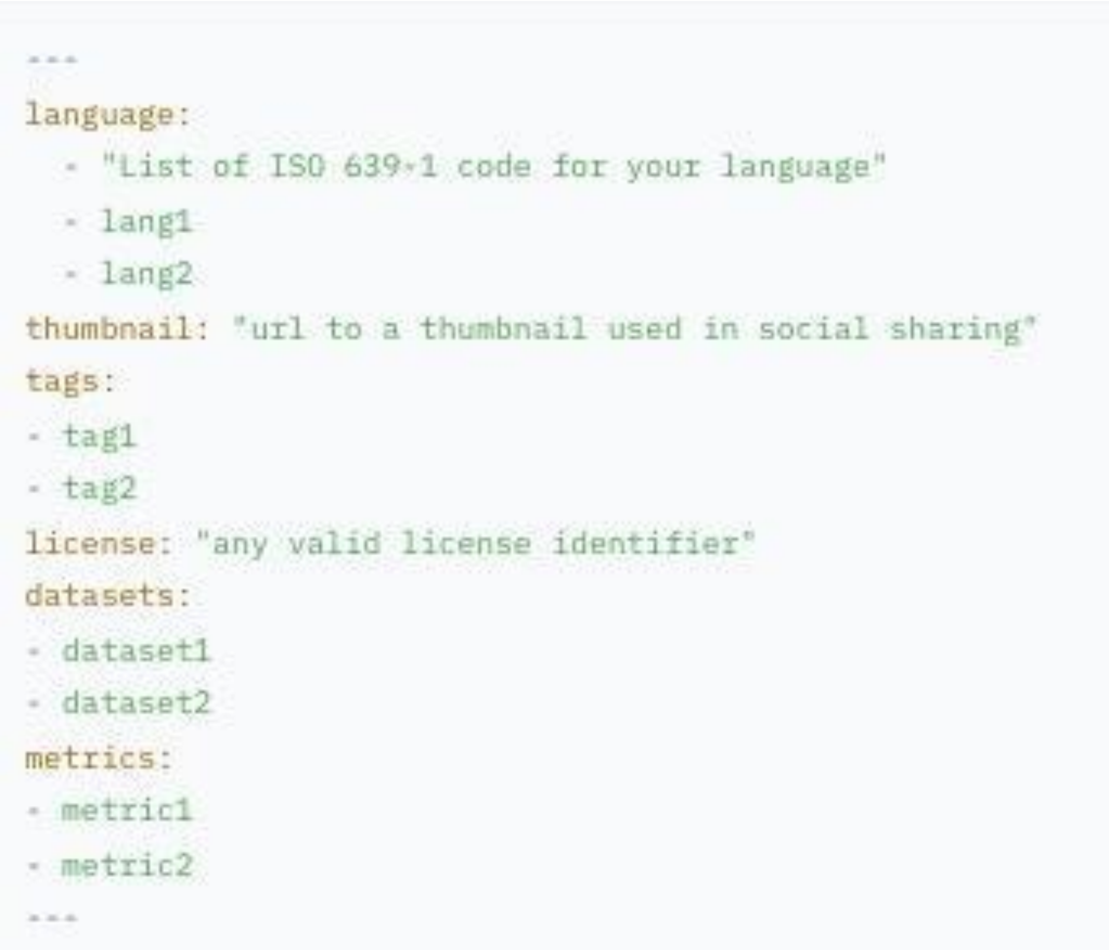
A structured description of the dataset that allows a variety of stakeholders to understand its intended uses as well as potential limitations. They should also describe how the dataset was assembled, known gaps in coverage, and how it can be evaluated.

Model Cards at Hugging Face

The Hugging Face Library provides access to a set of open source pre-trained models.

Model cards allow me to **compare and choose** the best pre-trained model given my goal

Model cards are a structured set of criteria – like a metadata record



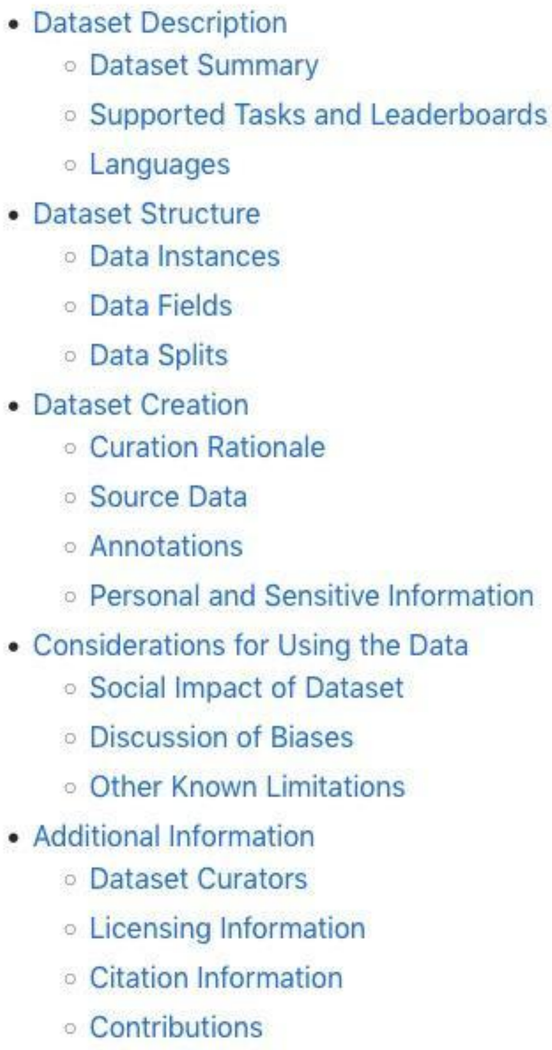
```
---
language:
  - "List of ISO 639-1 code for your language"
  - lang1
  - lang2
thumbnail: "url to a thumbnail used in social sharing"
tags:
- tag1
- tag2
license: "any valid license identifier"
datasets:
- dataset1
- dataset2
metrics:
- metric1
- metric2
---
```

Dataset Cards at Hugging Face

The Hugging Face Library provides access to a set of known datasets for training and testing language and image processing.

Dataset cards allow me to **compare and choose** the best dataset given my goal

Dataset cards are a structured set of criteria – like a metadata record



- Dataset Description
  - Dataset Summary
  - Supported Tasks and Leaderboards
  - Languages
- Dataset Structure
  - Data Instances
  - Data Fields
  - Data Splits
- Dataset Creation
  - Curation Rationale
  - Source Data
  - Annotations
  - Personal and Sensitive Information
- Considerations for Using the Data
  - Social Impact of Dataset
  - Discussion of Biases
  - Other Known Limitations
- Additional Information
  - Dataset Curators
  - Licensing Information
  - Citation Information
  - Contributions

Possible Comparison Criteria

In order to compare classification schemes we need **set of criteria as basis for comparison**.

As noted by model cards, different aspects of development, deployment and use will have different meanings to different stakeholders. It is important to bring those meanings in to the foreground so they can be described, discussed, analyzed, and compared.

1. Could a framework similar to model cards be developed by identifying the criteria that matter, and creating a standardized description (a metadata framework)?

2. Could this set of structured attributes provide the basis for beginning to develop a set of processes for comparing classification schemes?

3. What are some of the criteria we would want to include in our standardized description of classification schemes in the same domain?

Possible Comparison Criteria

One obvious criterion to describe is **structure**: pre-coordinated or post coordinated, faceted, enumerated list, hierarchical, poly hierarchical, graph, etc.

What was the **method of construction**? What were the **sources** used for developing the scheme?

Is the classification being **actively maintained**? If not, how has it been taken up and used. How has that usage changed its ecology?

Possible Comparison Criteria

Another dimension concerns the **stakeholders**: who created it (*creators*), who funded it (*funders*), who used it (*consumers*), what kinds of systems was it used to enable, and who used systems that made use of the classification (*second order consumers*).

What was its intended use or **uses** -- bibliographic control, indexing, controlled description, etc.?

What were its actual uses?

Who were its intended **users**?

What **processes of interpretation** interact with the scheme/uses/users?

How has this changed over time?

Possible Comparison Criteria

Was it built to facilitate **classification or interoperability** with other schemes?

What known **biases** can we identify in its construction?

What known biases can we identify in its application?

What biases have been discovered in its use?

Possible Comparison Criteria

Clearly there are a great many criteria that could matter from a descriptive or analytic or comparative point of view.

The goal here is not to propose the definitive scheme but rather to **begin a conversation** that can lead to identifying a set of criteria for making the comparisons and studying the differences and similarities.

Thank You

Please complete a survey about our session at:

https://forms.gle/Z4gFo8hhMpYD2q556

References

Bender, Emily and Friedman, Batya. (2018). “Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science”. *Transactions of the ACL (TACL)* (2018)

Gebru et. al., (2018) “Datasheets for Datasets”, https://arxiv.org/pdf/1803.09010.pdf

Holland et. al., (2018). “The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards.” CoRR abs/1805.03677 http://arxiv.org/abs/1805.03677

Mai, Jens-Erik, (2010) “Classification in a Social World: Bias and Trust”, *Journal of Documentation* Vol. 66 No. 5 pp. 627-642.

Mitchell et.al., (2019) “Model Cards for Model Reporting”, *FAT ‘19*, https://doi.org/10.1145/3287560.3287596

Tennis, J. T., (2017) “Never Facets Alone: The Evolving thought and Persistent Problem in Ranganathan’s Theories of Classification”, *Facets(Ergon*): 31-38.