Science of Computer Programming 231 (2024) 103030

Contents lists available at ScienceDirect

Science of Computer Programming

journal homepage: www.elsevier.com/locate/scico

Original software publication

DescribeML: A dataset description tool for machine learning

Joan Giner-Miguelez [a,*], Abel Gómez [a], Jordi Cabot [b]

[a] Internet Interdisciplinary Institute, Universitat Oberta de Catalunya (UOC), Barcelona, Spain
[b] Luxembourg Institute of Science and Technology (LIST), Esch-Sur-Alzette, Luxembourg

ARTICLE INFO

Article history:
Received 5 December 2022
Received in revised form 4 September 2023
Accepted 6 September 2023
Available online 12 September 2023

Keywords:
Datasets
Machine learning
Model-driven engineering
Fairness
Domain-specific languages

ABSTRACT

Datasets are essential for training and evaluating machine learning models. However, they are also the root cause of many undesirable model behaviors, such as biased predictions. To address this issue, the machine learning community is proposing as a best practice the adoption of common guidelines for describing datasets. However, these guidelines are based on natural language descriptions of the dataset, hampering the automatic computation and analysis of such descriptions. To overcome this situation, we present *DescribeML*, a language engineering tool to precisely describe machine learning datasets in terms of their composition, provenance, and social concerns in a structured format. The tool is implemented as a Visual Studio Code extension.

Code metadata

Code metadata description	
Current code version	v1.2.0
Permanent link to code/repository used for this code version	https://github.com/ScienceofComputerProgramming/SCICO-D-22-00332
Permanent link to Reproducible Capsule	http://hdl.handle.net/20.500.12004/1/A/DML/005
Legal Code License	MIT License
Code versioning system used	git
Software code languages, tools and services used	TypeScript, Langium
Compilation requirements, operating environments and dependencies	Visual Studio Code
If available, link to developer documentation/manual	http://hdl.handle.net/20.500.12004/1/A/DML/002
Support email for questions	jginermi@uoc.edu

1. Motivation and significance

While datasets are becoming more and more important in machine learning, recent research has revealed unintended consequences and negative downstream effects in the entire machine learning (ML) pipeline due to data issues [1]. For

The code (and data) in this article has been certified as Reproducible by Code Ocean: https://codeocean.com/. More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail address: jginermi@uoc.edu (J. Giner-Miguelez).

https://doi.org/10.1016/j.scico.2023.103030

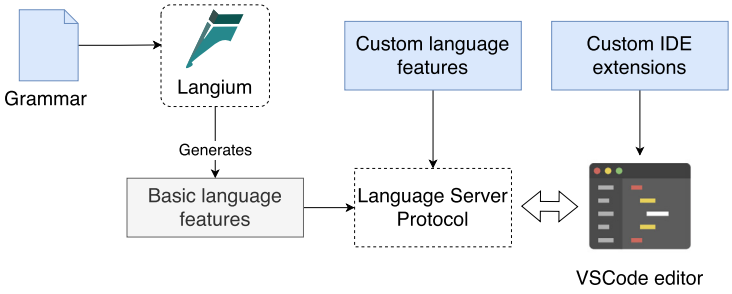

Fig. 1. Architecture overview.

example, face analysis datasets with a low proportion of darker-skinned faces may reduce the accuracy of face analysis models for that group, causing social harm. As another example, because of the differences in language accents and styles, a natural language dataset gathered from Australian speakers may reduce the accuracy of models trained to support users in the United States. In both examples, we see the need to store information about the data provenance or high-level analysis, such as the social impact on specific groups to better understand the quality and social limitations of the trained models.

This situation has triggered the interest of the research community in standardizing data creation processes and implementing best practices around datasets for ML. Recent works such as *Datasheets for datasets*, among others [2–4], have come up with guidelines for the creation of standard dataset documentation. These guidelines help to identify data aspects that may potentially influence how the dataset is used or the quality of the ML models trained with it. Nevertheless, these proposals rely on textual descriptions in natural language, which pose clear challenges when it comes to automatically compute and analyze them, hampering their benefits.

In that sense, current tools for describing structured metadata are focused on initiatives such as the *Data Documentation Initiative*,[1] and the *Data Catalog Vocabulary*,[2] that do not cover the guidelines of the ML community. To overcome these limitations, we introduce *DescribeML*, a tool for precisely describing datasets in accordance with the dimensions specified by the aforementioned proposals. This tool provides a specific notation based on a domain-specific language (DSL) to describe datasets for ML [5], together with common modern language features such as auto-completion, syntactic and semantic highlight, validation, cross-references, etc. Furthermore, the tool eases the dataset documentation process by allowing users to preload the data and generate HTML documentation from a valid dataset description.

We have developed *DescribeML* as an extension for Visual Studio Code (VSCode). The tool has been released in the VSCode Marketplace. It has also been open sourced in a public repository.[3]

2. Software description

This section presents a detailed description of *DescribeML*. We first introduce an overview of the architecture together with the grammar powering the DSL behind the tool, and then, we present its main features.

2.1. Architecture

The main components of *DescribeML* are the DSL grammar, from which we are able to generate a set of core language extensions, and several custom IDE extensions to facilitate the datasets' documentation processes. The tool is implemented on top of Langium,[4] a language engineering toolkit to create textual DSLs, and the VSCode extension API.[5]

In Fig. 1 we can see how these components are linked together. The ***Grammar*** is defined as a concrete textual notation for the mentioned DSL [5] using the Langium Grammar Language.[6] Once defined, Langium takes the grammar and generates the *Basic language features* served by the *Language Server Protocol*[7] to the user. On the other hand, the ***Custom language features*** have been developed using the Visual Studio Code Extension API. As this API acts as a wrapper of the Language Server Protocol, it has allowed us to develop domain-specific language features beyond those provided by Langium out-of-the-box. Finally, the ***Custom IDE extensions*** have also been developed using the same API and provide specific IDE features to facilitate the dataset documentation process, such as preloading the data and generating HTML documentation from a valid description.

[1] https://ddialliance.org/.
[2] https://www.w3.org/TR/vocab-dcat-3/.
[3] http://hdl.handle.net/20.500.12004/1/A/DML/001.
[4] https://langium.org/.
[5] https://code.visualstudio.com/api.
[6] https://langium.org/docs/grammar-language/.
[7] https://microsoft.github.io/language-server-protocol/.

```
1 Metadata:
2      (citation=Citation)?
3      'Description:'
4         (description=STRING |
5         (('Purposes:' descriptionpurpose=STRING)?
6          ('Tasks:' '[' descTasks+=MLTasks (',descTasks+=MLTasks)*']')?)?
7      [...]
8 SocialIssue:
9     ('Related Attributes:'('attribute:'rAtt=[Attribute])*)?
10     [...]
11 MLTasks returns string: 'Text-classification'|'Question-answering'| //...
```

Listing 1: Tool's grammar excerpt.

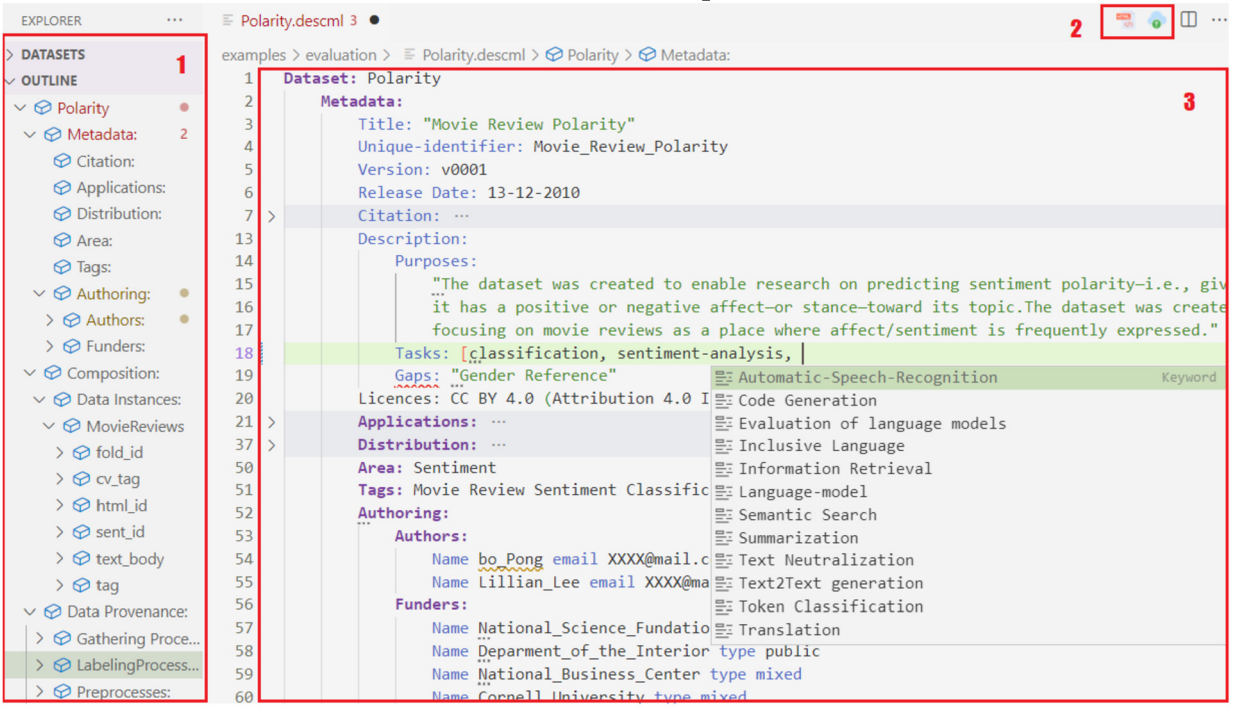

Fig. 2. Tool UI overview where the square marked with "1" is the document outline, the square marked with "2" is the preloader and generation services, and the square marked with "3" is the editor.

In Listing 1, we can see an excerpt of the grammar at the core of the process. We have expressed the optional attributes using the "?" symbols, such as "citation" in line 2, and the OR operator using the "|" symbol, such as in line 3, where *Description* can be a single string or a couple of purposes and tasks. Also, to express zero to many multiplicity relations, we use the "*" symbol. For instance, in line 6, you can provide a set of ML tasks from the *MLTasks* in line 16. Finally, we have expressed the cross-references using brackets, such as in line 14, where we can assign a set of related *Attributes* to a specific *Social Issue*. For example, the attribute "gender" may relate to a gender parity social issue. The complete grammar of the tool can be seen in the tool's open-source repository.[8]

2.2. Main features

DescribeML implements all expected basic language editing features such as syntactic checking and semantic highlighting, autocompletion, and code snippets, similar to other general-purpose language tools. In Fig. 2 we can see an overview of the tool UI with an autocompletion example. Besides, the tool implements a custom validation service for adding custom validations and providing hints to users during the documentation process. The hints are based on the guidelines provided by recent works [2–4] in the ML field and provide context to users, facilitating the consistency and the correct usage of the language terms. Fig. 3 shows an example of a hint.

[8] http://hdl.handle.net/20.500.12004/1/A/DML/004.

Fig. 3. a) Tool's hint example & b) HTML generation implementing Schema.org.

```
1 Composition:
2   Data Instance: MovieReviews
3         Attribute:  textBody [...]
4         Attribute:  tag
5           Description: "The label annotated by the reviewrs"
6           Labelling process: SentimentalLabeling
7           OfType: Categorical
8             Categorical Distribution: ["pos":45%, "neg":55% ]
9 [...]
10 Provenance:
11   Labeling Processes: SentimentalLabeling
12       Description: "A Rating between 0 a 5 fives stars following"
13       Type: Entity annotation  Labels: MovieReviews.tag
```

Listing 2: Example excerpt of the Movie Review Polarity dataset description.

From an end-user perspective, and to avoid users starting from scratch in the data documentation process, the tool offers a *data preloading service*. This service allows users to automatically upload data files and create a first draft of the description document. The service analyzes the provided data by extracting the data structure (the files and attributes) and detecting its type (numerical or categorical). The tool calculates relevant statistics for each attribute depending on the attribute's type. For instance, for categorical attributes, *DescribeML* calculates the distribution, mode, and completeness; for numerical attributes, it calculates the mean and the standard deviation.

Once the data is documented, to facilitate its publication and discoverability across the web, the tool implements a *generation service* that takes a valid document description and generates HTML documentation. This HTML documentation is also populated with the *Schema.org* vocabulary,[9] a vocabulary to facilitate the discoverability of the content by the search engines. In particular, we implement the extension *"@dataset"* of the vocabulary used by the *Google Dataset Search* engine to discover datasets across the web. Fig. 3 shows an example of the generated HTML implementing *Schema.org*.

3. Examples

As a usage example of *DescribeML*, we have released a set of DescribeML-based descriptions of popular datasets in a public repository[10] that can be used as a guide for using the tool. We chose these datasets because they have already been the subject of recent publications in the ML community about dataset documenting practices, and have a diverse provenance and composition. In addition, to flatten the tool's learning curve, we have released a video[11] presenting the tool, and a language reference guide[12] with a set of usage examples.

Listing 2 shows a description excerpt of one of the mentioned datasets; the *Movie Reviews Polarity*[13] dataset. This dataset is a widely used benchmark dataset for sentimental analysis tasks, composed of movie reviews tagged with a sentiment flag (positive or negative) by a group of reviewers. In line 8 of the Listing 2, we can see the *Categorical Distribution* of the attribute *tag* and its link with the *SentimentalLabeling* process. Knowing the distribution of the labels and the type of

[9] Schema.org project homepage: https://schema.org/, visited November 2022.
[10] http://hdl.handle.net/20.500.12004/1/A/DML/003.
[11] http://hdl.handle.net/20.500.12004/1/A/DML/006.
[12] http://hdl.handle.net/20.500.12004/1/A/DML/002.
[13] https://www.cs.cornell.edu/people/pabo/movie-review-data/.

labeling process performed is helpful in evaluating the suitability of the dataset for a particular use case. Moreover, having this information in a structured format enables to build search and comparison applications to facilitate this evaluation.

4. Impact

DescribeML is a tool designed to help standardize dataset description practices. However, standardizing these practices offers a set of challenges [6] that *DescribeML* aims to address. For instance, the need for more interactivity and user assistance during the documentation process is faced by offering a set of language features, such as auto-completion, hints and code snippets, etc. On the other hand, the need to automate several parts of the process is faced with features such as the data preloader and HTML generation services.

Furthermore, we identified several opportunities as a result of the adoption of the tool. When a dataset is described using the tool, it can be subsequently manipulated using a variety of existing engineering tools and methodologies. For instance we could; *(i)* Compare dataset description to highlight how different datasets on the same domain differ so that ML experts can choose the best one for their project. *(ii)* Search for datasets based on (partial) requirements. *(iii)* Generate a test set to ensure that the dataset is still compliant with the documentation (relevant for incremental datasets). Or code (e.g., Python) to facilitate its manipulation by ML libraries.

5. Conclusions

In this work, we have presented *DescribeML*, a tool to describe machine learning datasets. The tool provides a set of language features and enhanced IDE capabilities to facilitate the documentation process. As a usage example, several dataset descriptions are available on the tool's public repository, together with a video introducing the tool and a language reference guide. We believe that *DescribeML* is an important step toward standardizing dataset documentation practices and its future impact in achieving higher-quality machine learning models.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Paullada, I.D. Raji, E.M. Bender, E. Denton, A. Hanna, Data and its (dis) contents: a survey of dataset development and use in machine learning research, Patterns 2 (11) (2021) 100336.
[2] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H.D. Iii, K. Crawford, Datasheets for datasets, Commun. ACM 64 (12) (2021) 86–92.
[3] E.M. Bender, B. Friedman, Data statements for natural language processing: toward mitigating system bias and enabling better science, Trans. Assoc. Comput. Linguist. 6 (2018) 587–604.
[4] S. Holland, A. Hosny, S. Newman, J. Joseph, K. Chmielinski, The dataset nutrition label, in: Data Protection and Democracy, in: Data Protection and Privacy, vol. 12, 2020, pp. 1–25.
[5] J. Giner-Miguelez, A. Gómez, J. Cabot, A domain-specific language for describing machine learning datasets, J. Comput. Lang. 76 (2023) 101209.
[6] A.K. Heger, L.B. Marquis, M. Vorvoreanu, H. Wallach, J. Wortman Vaughan, Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata, Proc. ACM Hum.-Comput. Interact. 6 (CSCW2) (2022) 1–29.