End to End Software Engineering Research - Datasheet

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

The data set and its motivation is described in detail in "End to End Software Engineering Research" by Idan Amit.

The dataset is mainly aimed for end to end learning in software engineering. The dataset can serve few goals

1) Apply machine learning to learn common process metrics (detailed later)
2) Build new process metrics by applying NLP to the commit messages and learn them from the code
3) Apply co-change analysis and other techniques based on changes in code and process metrics over time to detect causal relations.
4) Define by examples text similarity and code similarity
5) Define by examples program difficulty
6) Program repair

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

The data set was not used before.

C. What (other) tasks could the dataset be used for?

See "Other Use Cases" in paper.

D. Who funded the creation dataset?

The Hebrew University.

E. Any other comment?

No.

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

There are several instances - files (source code), commits, and enhancement of them (e.g., process metrics computation). A description is given in the paper appendix.

B. How many instances are there in total (of each type, if appropriate)?

The June 2021 extraction of the dataset has more than 5m files (see details in paper "introduction"). We extract data every two months so the data set keeps growing. Note that new extraction also enhances older data (e.g., provides more cases of co-change).

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are sub-populations identified (e.g., by age, gender, etc.) and what is their distribution?

Files, commits and their enhancement. See details in paper appendix.

D. Is there a label or target associated with each instance? If so, please provide a description.

Process metrics (e.g., bug fixes found by commit linguistic analysis, commit duration computed by commits timestamps) are given as labels. The commits are also provided, allowing to compute new labels. See details in the paper appendix.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The relations are explicit and due to the nature of software development. This is a structure that enables improved learning.

We provide the content of the same file on different dates. Files might belong to the same project, written by the same developer or created at the same time. Commits might be shared in a few projects.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

No sampling was used.

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We provided training/validation/test splits in both the commit, file and repository level. The split was done using hash function so it is pseudo-random, reproducible and can be extended to new data (e.g., future commits).

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The projects' selection process is described in [3]. Forks, a great source of project redundancy were removed using the GitHub API. Projects that are not software projects (e.g., board-game whose boards are shared in GitHub) were identified using the lack of bugs, hence having a negative corrective commit probability [3]. Unlike the use of API, the method is not perfectly accurate yet it was validated by the use of pull requests and identification of programming language. Redundant commits were further identified and removed due to more than 50 common commits. Hence, some redundancy might exist but the process cleaned most of it.

Many of the metrics are based on the language model from [2]. Obsessively, none of the models have perfect accuracy. The language models are provided with their protocols (e.g., is a typo considered a bug? and an error in a test file?), labeled samples and performance estimation.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data set is self contained.

Any other comments?

No.

III. Collection Process

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Full process is described in the paper. The developer contributed to open source projects hosted on GitHub. Google created a BigQuery schema for some OSI compliant projects [5]. Using the code that we published in [1], we constructed the dataset and hosted it on GitHub.

B. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The commit, and its message, are stored in GitHub in order to enable software development. The developers create commits in order to modify the code. The files contained in the commit are part of its meta data. The Google BigQuery scheme [5] contains only the HEAD version of the content, its current content. Therefore we extract the HEAD version every two months in order to have versions of the content over time.

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Not sampling was used.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The developers developed open source projects at GitHub. Google created the BigQuery scheme. From this point, other than the researchers, no manual work was needed.

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Data was collected from June 2021, and extracted every two months. The commits, used for process metrics, were created since the project's beginning. Their vast majority has been since 2008, the year of GitHub establishment. Some commits go to the nineties, due the projects started elsewhere and ported to GitHub.

We plan to keep extracting data in the future.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Process metrics were computed, described in the paper appendix. All code is provided [1].

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Not applicable.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

All code is provided [1].

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Not applicable.

E. Any other comments

No.

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

Source code and documentation are available at [1].

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The data set will be published once accepted. It is already publicly available at [1] The license will be 'CC BY 4.0'.

C. Are there any copyrights on the data?

All projects in the data set are open source projects with OSI compliant licenses.

D. Are there any fees or access/export restrictions?

No.

E. Any other comments?

No.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The data set is stand alone and no maintenance is needed.

B. Will the dataset be updated? If so, how often and by whom?

Currently we plan to continue and extract data every two months.

C. How will updates be communicated? (e.g., mailing list, GitHub)

Current data and all new one will be available at GitHub [1]

D. If the dataset becomes obsolete how will this be communicated?

The data set is stand alone and should not become obsolete.

E. Is there a repository to link to any/all papers/systems that use this dataset?

Yes, [1].

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

We provide all the code used for the construction of the data set [1]. Any researcher can use the code and extend the data set. We don't see such extensions as part of the data set. A researcher might reach out to us with an extension and if we find it suitable, we can add a reference to the extension in the data set repository.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Ethical considerations are discussed in the paper. An IRB was not involved as the dataset does not involve humans.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

All the data is already public at GitHub, and therefore on the web. It is known in advance to be public and do not contain confidential information.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

Note that more than 7K commits were identified[1] to contain swearing and more than 325k commits were identified to contain negative sentiment. The true numbers might be higher due to the classifiers' false negatives. As this data is already open we did not filter those, but warn future users of the data to filter profanity if their needs so require.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

While we do not store developers' personal information, each commit is identified by a hash. Given the hash, a look up in the project metadata retrieves the developer's profile. Since it is required from the development process, the developers accept that and we do not ease look up or provide new information about the developer. In any case, the developer controls the data published on them and not us. Moreover, they can remove or alter it in any way that does not violate GitHub's terms. We consider this concern as addressed too.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Neither Github, the project nor us have this information. Many developers use nicknames. Some use real names where gender can be assumed and further data crawling might identify more information about the developer. Note that the developer chose to publish this data and the data is already public on the web.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

See answer to how the dataset relates to people.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The data is from professional activity. Commit messages should describe the source code modification. Sensitive data like these described in the question should not be contained since the do not describe the source code modification.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was collected from the developers by GitHub. Note that this is not a survey and no questions were asked.

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The developers were not notified regarding the or data set. However, the developers agreed to contribute to open source projects. By that the agreed not only to make the commit messages public but also the code itself.

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The agreement was done in the contribution to open source projects. All projects have an OSI complaint license.

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

While we do not store developers' personal information, each commit is identified by a hash. Given the hash, a look up in the project metadata retrieves the developer's profile. Since it is required from the development process, the developers accept that and we do not ease look up or provide new information about the developer. In any case, the developer controls the data published on them and not us. Moreover, they can remove or alter it in any way that does not violate GitHub's terms. We consider this concern as addressed too.

M. Any other comments?

This datasheet is based on the template of [4].

References

[1] I. Amit. End to end software engineering repository (https://github.com/evidencebp/e2ese), Aug 2021.

[2] I. Amit. Natural language processing for software engineering (https://github.com/evidencebp/commit-classification), Aug 2021.

[3] I. Amit and D. G. Feitelson. Corrective commit probability: a measure of the effort invested in bug fixing. *Software Quality Journal*, pages 1–45, Aug 2021.

[4] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford. Datasheets for datasets, 2018.

[5] GitHub. Github activity data, September 2019.

[1] Using classifiers from [2]