# ComSum - 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 is described in "ComSum: Commit Messages Summarization and Meaning Preservation" by Leshem Choshen and Idan Amit.

The data set is mainly aimed for text summarization. The data set fill few gaps

1) It is significantly larger than other text summarization data sets, enabling investigating summarization in scale.
2) It is the first data set that uses commit messages and address software engineering.
3) The data set leverages the commit taxonomy and enables training and evaluation with meaning preserving and not lexical similarity.

*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?*

The data set can be used for summarization. It can also support auxiliary tasks such as active learning, topic modeling, and commit type classification.

*D. Who funded the creation dataset?*

The Hebrew University.

*E. Any other comment?*

NA.

## 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)*

All instances are GitHub commit messages, text that developers write in order to describe modifications they did to the source code. The target of the summarization is the commit "subject". The source is the rest of the commit message.

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

The data set has 7,540,026 instances.

*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?*

The instances are text messages of commits. Each commit is identified by its hash and the repository (project), it belongs too. For summarization purposes, the commit has a subject and message. The message without the subject (an auxiliary column we created) should be used for summarization.

The commits are written by open source developers. We do not include the developer profile. It is common to use nicknames and not real names, so authors that wish their identity hidden are able to do so. There is no available data about a developer's age and gender.

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

This data set is not a supervised classification learning data set but a text summarization one. There are no labels, but there are target texts. One can see the "subject" summary as the label for a given example.

*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 commits are provided with the project they belong to. We do not provide author identifiers. The paper contains evaluation on the usefulness of related commits as predictors. The performance is much lower than that of summarization models.

*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).*

The data set contains all large enough commits of 19,720 projects, until 2021. No sampling was used. Naturally, most of these projects kept evolving and have more commits now. Also, more projects exist but not suitable ones (see paper for project selection).

*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 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 commit messages were written by 317,423 developers. While GitHub product design, community pressure and developer professionalism advocate writing proper commit messages, the developer is free to write any text. We manually labeled thousands of commit messages and this threat is uncommon.

More threats are listed in the paper 'Limitations and Threats to Validity' section. Regarding the data, we also list the existence of 'squash' commits, many commits melt into one. They tend to be rather long and less coherent.

A commit might appear in many related projects. 16% of the commits appear more than once and we removed them. A message or a subject might appear more than once. For example, the most common subject is "Updating submodules", appearing 6612 times. However, 96% of the messages are unique. We left the multiple appearing message since this represents software development. We provide appearance distribution and common messages. This way the data set fits researchers wanting to represent software development. Researchers looking for uniqueness can remove multiple appearing texts.

In 0.8% of the cases the subject appeared in the rest of the message. We extract the common ones and they seem to be due to a generic subject. The leading ones are 'WebCore:' (1554 times), 'Updated Spanish translation.' (809 times), 'Updated Norwegian bokmål translation.' (347 times). Again, in order to enable maximal future flexibility we left them. An interested researcher can filter these few commits out.

*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. This is the only manual part. Google created a BigQuery schema for some OSI compliant projects. We extracted the data from there and applied massive filtering to choose proper projects (described in paper).

*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.

*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 created the commits as part of their contribution to open source projects development. 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.*

The supplementary materials have the number of commits per year. Their vast majority were from 2020 to 2008, the year of GitHub establishment. Some commits go to the nineties, due the projects started elsewhere and ported to GitHub.

## 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.*

The test is provided as is, without processing.

*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.*

Skipped.

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

Skipped.

*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?*

Skipped.

*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 `https://github.com/evidencebp/comsum` (DOI 10.5281/zenodo.5090706)

Data is available at `ttps://figshare.com/articles/dataset/CumSum_data_set/14711370` (DOI 10.6084/M9.FIGSHARE.14711370).

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

The data set will be released once accepted. 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 licences.

*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 there are no plans to update the data set. In the future, it might be extended with new commits.

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

No updates are planned.

*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?*

Given the paper, Google scholar can be used to find papers referencing it.

*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. 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 data set 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 doctorpatient 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. Commit messages are 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*

Discussed in the paper: "Note that 7K commits were identified[1] to contain swearing and 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.*

Discussed in the paper: "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.

[1]Using classifiers from https://github.com/evidencebp/commit-classification

*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.*

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

*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. The data was collected to enable open source development.

*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.*

Discussed in the paper: "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?*

No.