A Data Set of Extracted Rationale from Linux Kernel Commit Messages

Mouna Dhaouadi
Université de Montréal
Montreal, Canada
mouna.dhaouadi@umontreal.ca

ABSTRACT

Developer's commit messages contain information about decisions taken and their rationale. Extracting this information is challenging since we lack a detailed understanding of how developers express these concepts. Our work-in-progress targets this challenge by producing a labelled data set of commit messages for a Linux Kernel component. We report preliminary analyses which suggest that larger commit messages and more experienced developers commits tend towards having 40% of sentences containing rationale. This may indicate a guideline for developers to target.

CCS CONCEPTS

• **Software and its engineering → Software design engineering**.

KEYWORDS

Software rationale, Linux kernel, Data set, Commit messages

ACM Reference Format:
Mouna Dhaouadi. 2023. A Data Set of Extracted Rationale from Linux Kernel Commit Messages. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '23), December 3–9, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3611643.3617851

1 RESEARCH PROBLEM AND MOTIVATION

Rationale for decisions and changes in software projects is often recorded in the commit messages [1]. Capturing this implicit knowledge is useful to create a record that could be used when revisiting decisions or to produce future recommendations [2].

The Linux kernel is a large-scale open-source project involving many collaborators. Thus, having a shared understanding is necessary to make coherent decisions. Linux kernel commit messages usually contain a description of the rationale behind the introduced changes. However, this rationale information is embedded inside the commits and we lack a systematic process of capturing and organizing decisions and their rationale.

Our prior work has proposed a vision for an end-to-end reconstruction pipeline to explicitly structure this rationale and relate it to the decisions [3]. In this ongoing work, we create a high-quality data set from a subset of Linux kernel commits. This data set will help us develop our pipeline to improve the developer's knowledge of the code base, and better understand how decisions and rationale are recorded.

The two main contributions of this paper are: 1) a labelled data set of 1144 sentences extracted from 160 commits from the Out-Of-Memory Killer (OOM-Killer) kernel component, and 2) initial analyses and insights concerning the abundance (**RQ1**), amount (**RQ2**) and developer experience (**RQ3**) characteristics of rationale.

Table 1: Codebook

Label	Meaning
Decision	An action or a change that has been made, including a description of the patch behaviour
Rationale	Reason for a decision or value judgment
Supporting Facts	A narration of facts used to support a decision
Inapplicable	Pre-processing error or bad sentences (i.e., does not contain English sentences)

2 BACKGROUND AND RELATED WORK

Developers rationale refers to the reasoning behind the decisions that developers make [2]. Prior work has tried to extract rationale from textual artifacts for different projects; e.g., from Python email archives [6], from Apache project commit messages [4] and from Chrome Bug reports [5]. However, to the best of our knowledge, there is no previous work that proposed a data set of extracted rationale from the commit messages for the Linux kernel. In particular, Linux developers are encouraged to describe concisely their rationale/motivation in the commit descriptions[1]. This makes Linux commit messages a comprehensive and very semantically-rich repository of decision/rationale information.

3 APPROACH AND CONTRIBUTIONS

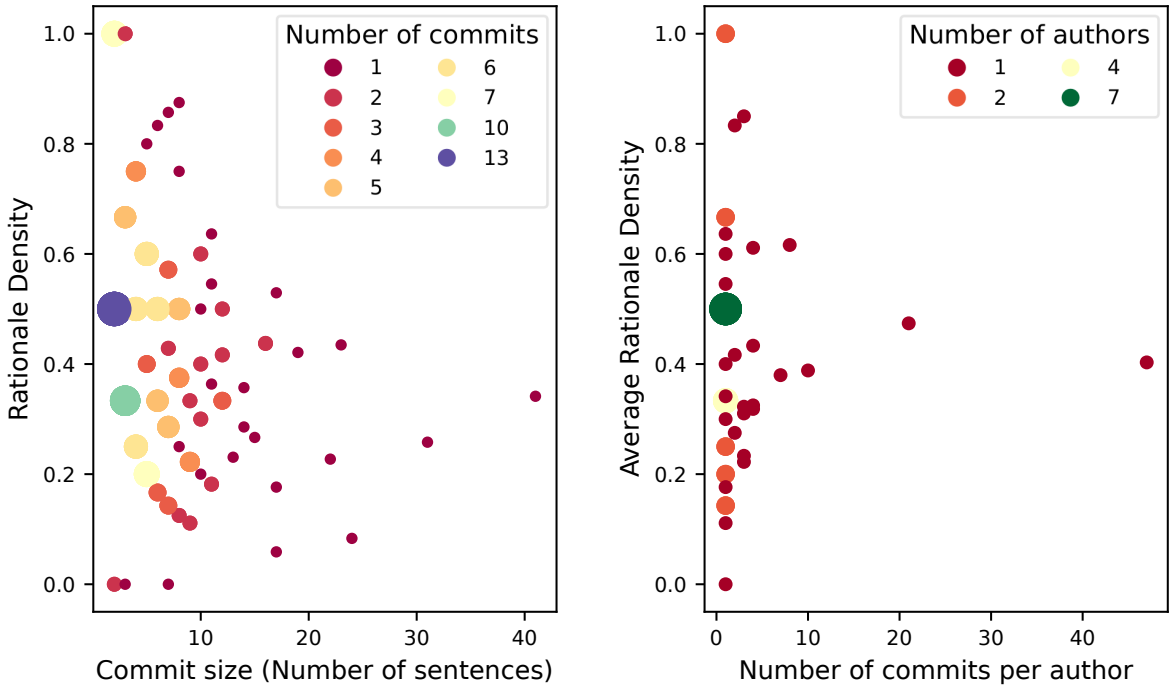
To create our data set, we obtained the commit history (418 commits) of the OOM-killer file[2]. For each commit, we reduced noise (e.g., removed code and meta data), and split it into sentences to analyze.

Three annotators (including myself) iterated over six piloting rounds to reach a consensus regarding the set of labels to use (Table 1). We agreed to consider terse value judgment language (e.g., "fix" or "cleanup") to imply the presence of rationale and decisions. For instance, the sentence "`fix it`" is considered both

ESEC/FSE '23, December 3–9, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0327-0/23/12…$15.00
https://doi.org/10.1145/3611643.3617851

[1] https://www.kernel.org/doc/html/v4.10/process/submitting-patches.html#describe-your-changes
[2] https://github.com/torvalds/linux/commits/master/mm/oom_kill.c accessed on 12/01/2023



(a) (b)

Figure 1: a) Commit size versus rationale density b) Commits per author versus average rationale density

Decision and *Rationale*. We provide an example of labelling a commit message[3] in Table 2.

Table 2: Example of labelling a commit

Sentence	Label
signal: Use SEND_SIG_PRIV not SEND_SIG_FORCED with SIGKILL and SIGSTOP	Decision
Now that siginfo is never allocated for SIGKILL and SIGSTOP there is no difference between SEND_SIG_PRIV and SEND_SIG_FORCED for SIGKILLand SIGSTOP.	Supporting Facts
This makes SEND_SIG_FORCED unnecessary and redundant in the presence of SIGKILL and SIGSTOP.	Rationale
Therefore change users of SEND_SIG_FORCED that are sending SIGKILL or SIGSTOP to use SEND_SIG_PRIV instead.	Decision
This removes the last users of SEND_SIG_FORCED.	Decision

During the labelling process, Fleiss Kappa averaged 0.68 for seven rounds (so far). This indicates strong agreement considering the subjective nature of rationale [2].

RQ1. How many commits contain rationale? 97.5% of the commits contain at least one sentence with rationale information. This suggests that rationale is almost always described.

RQ2. How much of the commit contains rationale? The rationale density is the percentage of commit sentences that contain rationale. Figure 1a shows these values versus the size of commits. The figure shows that as a commit becomes longer, it tends to have 40% of its sentences contain rationale information. Overall, 43.87% of all the commit messages contains rationale information.

RQ3. Does the quantity of rationale reported depend on the developer experience? We visualize the *average rationale density* per author in Figure 1b. Most of developers commits have a density between 30% and 60%, but more experienced developers commits have a density near 40%.

[3] https://api.github.com/repos/torvalds/linux/git/commits/079b22dc9be985c591589fcb94769b8e13518aa0

ACKNOWLEDGMENTS

Partially funded by the Fonds de Recherche du Québec (B2X). I also thank the other project members: Bentley James Oakes and Michalis Famelis.

REFERENCES

[1] Rana Alkadhi, Manuel Nonnenmacher, Emitza Guzman, and Bernd Bruegge. 2018. How do developers discuss rationale?. In *2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. 357–369. https://doi.org/10.1109/SANER.2018.8330223

[2] Janet E Burge, John M Carroll, Raymond McCall, and Ivan Mistrik. 2008. *Rationale-based software engineering*. Springer.

[3] Mouna Dhaouadi, Bentley James Oakes, and Michalis Famelis. 2022. End-to-End Rationale Reconstruction. In *37th IEEE/ACM International Conference on Automated Software Engineering*. 1–5.

[4] Jiawei Li and Iftekhar Ahmed. 2023. Commit Message Matters: Investigating Impact and Evolution of Commit Message Quality. (2023).

[5] Tanmay Mathur. 2015. *Improving Classification Results Using Class Imbalance Solutions & Evaluating the Generalizability of Rationale Extraction Techniques*. Ph. D. Dissertation. Miami University.

[6] Pankajeshwara Nand Sharma, Bastin Tony Roy Savarimuthu, and Nigel Stanger. 2021. Extracting rationale for open source software development decisions—a study of python email archives. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1008–1019.

Received 2023-06-05; accepted 2023-08-11