Sentiment Analysis of Open Source Communities: An Exploratory Study

Jennifer Ferreira[†]
Michael Glynn[‡]
David Hunt[‡]
Jaganath Babu[†]
Denis Dennehy[†]
Kieran Conboy[†]

†Lero | The Irish Software
Research Centre
NUI, Galway
Galway, Ireland
jennifer.ferreira@nuigalway.ie
j.babu1@nuigalway.ie
denis.dennehy@nuigalway.ie
kieran.conboy@nuigalway.ie

‡Intel Shannon
Shannon, Ireland
michael.j.glynn@intel.com
david.hunt@intel.com

OpenSym '19, August 20–22, 2019, Skövde, Sweden
© 2019 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-6319-8/19/08.
https://doi.org/10.1145/3306446.3340816

Abstract

Open Source Software (OSS) mailing lists have become popular targets for mining sentiment and emotions, as they provide a centralized communication hub between the distributed OSS community. Sentiment and emotions within communities can provide insights into how a community responds to certain events, who are the key members and how their behaviours impact the rest of the community. Such insights can inform initiatives aimed at fostering positive interactions between OSS community members, strengthening social ties, and helping the community accomplish its tasks. This poster presents our initial results from sentiment analysis of an OSS mailing list, and answers two key questions: (1) Given that the mailing list is used for peer-review of code, is the community sentiment negative overall? (2) Is community sentiment related to the month of the release cycle?

Author Keywords

Open Source; Sentiment Analysis; Mailing List; DPDK.

ACM Classification Keywords

H.1.m Information Systems: Miscellaneous; H.3.m Information Storage and Retrieval: Miscellaneous.

Data Plane Development Kit (DPDK)

DPDK consists of data plane libraries and network interface controller drivers to accelerate the performance of telecommunications and data networks. DPDK is currently managed as an open-source project under the Linux Foundation and licensed under the Open Source BSD License. The project was started by Intel as commercial software in 2010 and became open source in 2013. https://www.dpdk.org/

DPDK Mailing List Statistics for 18.05 Release

- The total number of messages exchanged was 8,585.
- Mailing list activity averaged 130 messages per day.
- The longest thread appeared in the month of April and contained 31 messages.

Introduction

Open Source Software (OSS) communities are a network of people and organisations associated with an OSS project who in various ways contribute to its growth and maintenance. The well-being of these communties is increasingly being recognised as critical to the success of OSS projects, due to their highly collaborative, communication-intensive nature [4]. In line with emerging research into emotion awareness in software development [5], the aim of our study is to explore how community sentiment impacts the software development process. Sentiment analysis is a collection of techniques for studying affective states inherent in human communication [3] and mailing lists are a popular tool for discussions within the community, thus offering a rich source of data regarding community practices and social structure [1]. Our results will contribute to best practices in the management of open source code reviews and to management strategies for collaborative and distributed software development environments such as outsourced and globally distributed teams. This poster presents our initial results, which answer two key questions: (1) Given that the DPDK mailing list is used for peer-review of code, is the community sentiment negative overall? (2) Is community sentiment related to the month of the release cycle? Answers to these questions can improve software development planning and scheduling of work.

Data Plane Development Kit (DPDK) Community

As new networking hardware is developed for improved communications and connectivity, new software features are needed to enable those new hardware capabilities and support on-going improvements in performance. New features are continually being added (committed) to the DPDK codebase by way of the community contributing, reviewing, and approving the software code for these features.

The DPDK community has been steadily growing since its establishment in 2013. Figure 1 shows how the growth from 200 contributors with 20 commits in 2013 to 1,600 with 160 commits in version 18.05 in 2018. The growing community has implications for planning processes at Intel since it takes more time to gain community acceptance of software features.

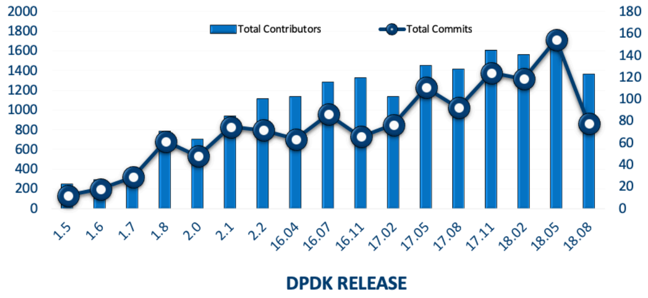

Figure 1: Growth of the DPDK community relative to the number of commits.

DPDK Mailing List

All code contributions to the DPDK project are submitted to the dpdk-dev mailing list for community review. If a patch is approved by the community, it is considered ready to be merged with the main codebase. If a patch requires improvements to the code, the developer will implement the suggestions and resubmit the patch to the mailing list for another round of reviews. If a patch is rejected, then the patch is not accepted by the community and will not be merged with the main codebase.

Community-specific language

During the analysis it emerged that certain phrases, “LGTM” (Looks Good To Me), “nacked” (not accepted) and “nits” (small problems or annoyances) for example, repeatedly appeared in the messages. To determine whether to assign positive, neutral or negative sentiment to such phrases DPDK community members were consulted to clarify the meanings.

Research Method References

[R1] Athanasios-Ilias Rousinopoulos, Gregorio Robles, and Jesus Gonzalez-Barahona. 2014. Sentiment Analysis of Free/Open Source Developers: Preliminary Findings from a Case Study. *Revista Eletrônica de Sistemas de Informação* 13(2), 6.1-6.21.

Research Method

The research method breaks down into 3 phases: Extraction, Pre-processing, and Sentiment Analysis. The method emerged from the close collaboration between the researchers from Lero and Intel. Our data covers the full 66 days of the DPDK 18.05 release cycle.

Phase 1 Extraction: Comprised extracting messages from the dpdk-dev mailing list archived at http://mails.dpdk.org/archives/dev/. A total of 13,461 messages were extracted in RAR file format.

Phase 2 Pre-Processing: Executed using Python scripts, the messages were converted from RAR file format into CSV file format and messages dated outside the release cycle removed. This resulted in 8,585 messages included in this study. The message content was cleaned for analysis using regular expressions to ensure that only the message body and natural language remained; all message headers, code, file paths, and non-alphanumeric symbols/characters were removed. This is an important step to reduce misclassifications.

Phase 3 Sentiment Analysis: The rule-based algorithm for annotating message content with positive, neutral, or negative scores proceeded according to the description given in [R1]. The overall sentiment of a message was computed as the sum of all scores assigned to that message. We used two popular sentiment analysis dictionaries – Opinion Lexicon and Comparative Words (Available at https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon). The algorithm was applied iteratively, refining the language dictionaries and improving the cleaning of the data with each iteration in collaboration with representatives from the DPDK community. Domain-specific terms can pose a significant challenge to sentiment analysis. Our close collaboration with the DPDK community helped us to identify and clarify terms that appear in the mailing list and augment the language dictionaries with community-specific language (see sidebar).

Preliminary Results

Is the DPDK community sentiment negative?

Figure 2 shows that the majority of messages scored 0 (n=3,775, median=0) confirming previous findings that technical communication tends to be neutral [2]. There were 2,942 positively scored messages, 1,868 negatively scored messages, and an overall mean score of 0.21 for the release cycle. Therefore, our analysis suggests that the overall sentiment on the mailing list was not strongly negative for the 18.05 release cycle, despite the mailing list being used for expressing critique and identifying code defects. Future work will extend this analysis to data on earlier and later release cycles to determine whether this finding holds for the other release cycles.

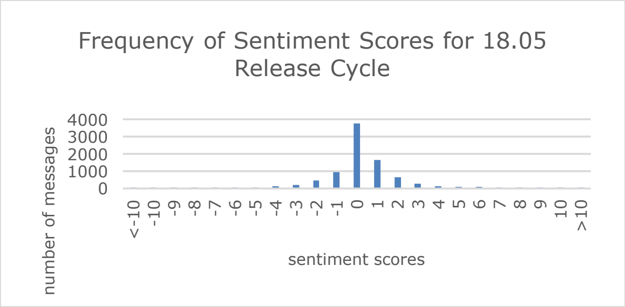

Figure 2: Frequency of sentiment scores for 18.05 DPDK release cycle.

18.05 Cycle Statistics

Mean	0.21
Median	0.00
Variance	6.52
Std.Deviation	2.55

February

Mean	-0.12
Median	0.00
Variance	8.89
Std.Deviation	2.98

March

Mean	0.34
Median	0.00
Variance	6.53
Std.Deviation	2.56

April

Mean	0.15
Median	0.00
Variance	6.53
Std.Deviation	2.56

May

Mean	0.21
Median	0.00
Variance	6.52
Std.Deviation	2.55

Is community sentiment related to the month of the release cycle?

Figure 3 below shows that mailing list sentiment changes over time with variations in the strength of positive and negative sentiment assigned to messages. A large proportion (86%) of the sentiment scores are within one standard deviation (std. dev=2.55), indicating a relatively stable sentiment on the mailing list for the duration of the release cycle. However, there are variations in monthly means (as can be seen in the side bar). The variation in mean sentiment score could potentially be explained by the "phases" of the development work within the release cycle. Initially the cycle begins with scoping activities and concludes with bug-fixing. Future work will compare the sentiment scores with phases of the development work within the cycle to determine whether a relationship exists.

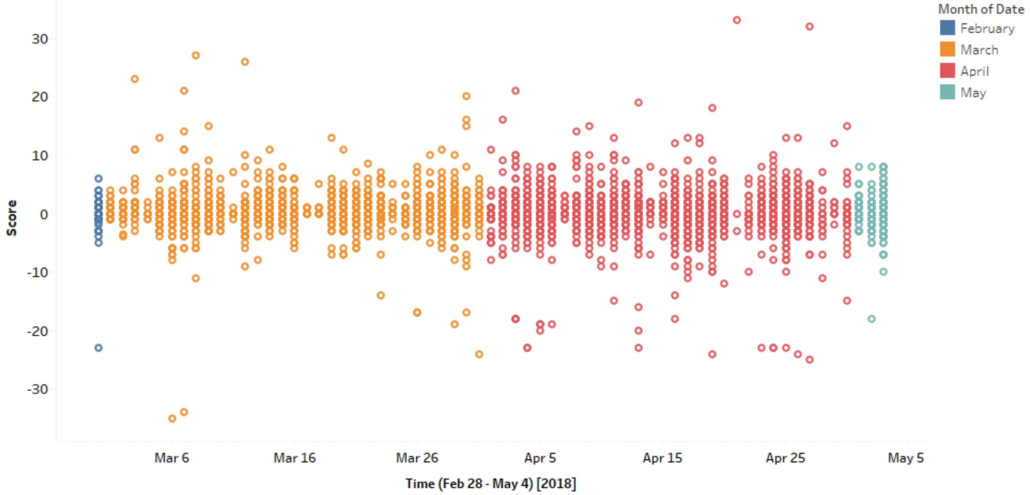

Figure 3: Distribution of sentiment scores for the DPDK release cycle 28 February 2018 to 4 May 2018.

Acknowledgements

This work was financially supported with the Science Foundation Ireland grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero | The Irish Software Research Centre (www.lero.ie).

References

1. Christian Bird, David Pattison, Raissa D'Souza, Vladimir Filkov, and Premkumar Devanbu. 2008. Latent social structure in open source projects. Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering - SIGSOFT '08/FSE-16, ACM Press, 24.

2. Emitza Guzman, David Azócar, and Yang Li. 2014. Sentiment analysis of commit comments in GitHub: an empirical study. Proceedings of the 11th Working Conference on Mining Software Repositories - MSR 2014, ACM Press, 352–355.

3. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1–2: 1–135.

4. Parastou Tourani, Yujuan Jiang, and Bram Adams. 2014. Monitoring sentiment in open source mailing lists: exploratory study on the apache ecosystem. In Proceedings of 24th Annual International Conference on Computer Science and Software Engineering (CASCON '14). IBM Corp., Riverton, NJ, USA, 34-44.

5. Nicole Novielli, Daniela Girardi, Filippo Lanubile (2018) A Benchmark Study on Sentiment Analysis for Software Engineering Research. 2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR), Gothenburg, 2018, pp. 364-375.