Sentiment Analysis of Open Source Software Community Mailing List: A Preliminary Analysis

Jumoke Abass Alesinloye
National University of Ireland Galway
Galway, Ireland
j.abassalesinloye1@nuigalway.ie

Eoin Groarke
National University of Ireland Galway
Galway, Ireland
e.groarke1@nuigalway.ie

Jaganath Babu
National University of Ireland Galway
Galway, Ireland
j.babu1@nuigalway.ie

Subathra Srinivasan
National University of Ireland Galway
Galway, Ireland
s.srinivasan1@nuigalway.ie

Greg Curran
Intel (Shannon), Ireland
greg.curran@intel.com

Denis Dennehy
National University of Ireland Galway
Galway, Ireland
denis.dennehy@nuigalway.ie

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/10.1145/3306446.3340824

Abstract

Open source software has become increasingly popular with companies looking to create business value through collaboration with distributed communities of organizations and software developers who rely on mailing lists to review code and share their feedback. This preliminary study reports on the sentiment analysis of the Data Plane Development Kit (DPDK.org) mailing list to identify and interpret patterns of sentiment during a release-cycle in 2018.

Author Keywords

Open Source; Open Source Community; Sentiment Analysis; Mailing list; DPDK.

ACM Classification Keywords

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

Introduction

Open source software (OSS) has become a feasible alternative to proprietary software and has an integral part in the business models of companies ranging from the technology to telecommunications sectors [1].

DPDK Release Process

A typical release should be complete within three months.

The first version of a new feature should be submitted before the proposal deadline, which is 'V1' in figure 1.

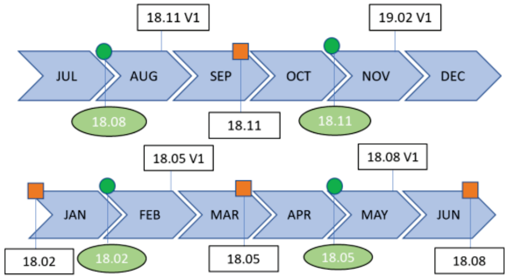
Updated versions (V2, V3 etc.) will be submitted to address any comments that have been received. Any new features must be reviewed, tested and accepted by the community before the integration deadline which is 'Merge' in figure 1.

The final month-long period prior to a release is dedicated to fixing any bugs [5].

OSS has undergone a transformation distancing itself from its free software antecedent, leveraging open source communities to increase development productivity and increased functionality [2] [3]. Furthermore, strategic utilisation of open source can create and capture business value from both the use and engagement with open source communities [4]. Key to the success of OSS is the developer community surrounding it and the activities undertaken by them, including peer reviews which substantially improves the quality of a patch (a software change) and the time it takes for a patch to be approved [1] [3]. By nature, developers of OSS are geographically dispersed, with the primary channel of communication being mailing lists.

Background

This paper focuses on the mailing lists of the Development Plane Data Kit (DPDK), which consists of libraries to accelerate packet processing workloads running on a variety of CPU architectures. This allows for companies specialising in telecommunications to move performance-sensitive applications to virtualised environments running on standard hardware, removing the need for expensive, dedicated single-purpose hardware appliances. In 2013 the open source community was established which is responsible for the full lifecycle of each quarterly DPDK release, including planning, development, test and release, along with the implementation of associated quality and governance practices. Code reviews through the mailing list are the primary quality check that is carried out on each patch set [5].



	V1	Final versions of patches for community review
🟧	Merge	Patch reviews based on community feedback, patches ACK'd
●	Release	Validated patches, fixes docs

Figure 1 DPDK 2018 Release Cycle

Through this form of communication, developers can express their emotions relating to the patch they are reviewing, which can include their satisfaction with the project or difficulties that have arisen and are hindering progress. Analysis of these emotions is important as they highly impact productivity, group morale, the quality of work, potential for further engagement and activity with the community and job satisfaction [6].

Sentiment alludes to lexical opinions, attitudes or emotions expressed through text. When people convey a message or their understanding, sentiment and emotion occurs. This can be identified in open source projects via mailing list analysis and can indicate positive or negative opinions within the context of a message that is conveyed [7]. Analysis of the contributions to the GENTOO open community found that when strong positive or negative emotions are expressed or when a community member deviates from the expected value of emotions, they are more likely to become inactive [8].

DPDK 2018 Dataset

The 2018 dataset contains 34,986 unstructured mails, which required extensive cleaning. The data was converted to CSV files, which were imported into python to be processed. Python scripts were written to segregate headers for the data based on email headers. The headers are as follows: Sender, Date, Subject, Receiver, MessageID, Body of mail, Acknowledgers name, Acknowledgers email. Using custom built regular expressions several elements were removed from the body of mails such as code, file paths and special characters to leave the natural conversation between community members.

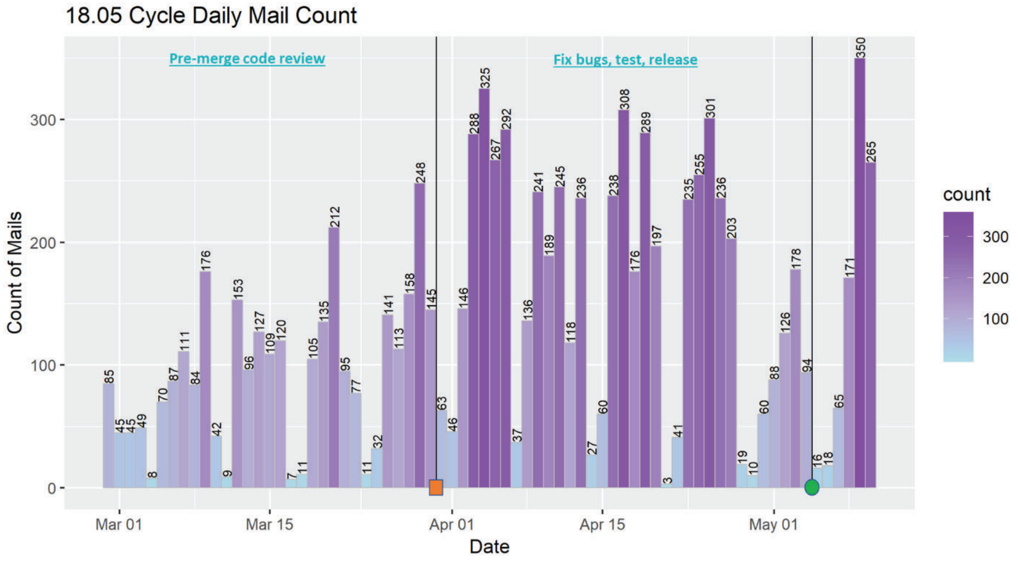


Figure 2 The 18.05 Release Cycle

Our research has been guided by the following questions:

1. What is the busiest period (18.02, 18.05, 18.08, 18.11) during the 2018 DPDK release cycle?

2. What type of sentiment is evident during the busiest cycle and does it fluctuate?

Results and Discussion

What is the busiest period (18.02, 18.05, 18.08, 18.11) during the 2018 DPDK release cycle?

By selecting the data for two months prior to each release (during the merge and bug fixing periods), we plotted the frequency of mails exchanged each day for each of the four cycles. This identified the 18.05 cycle as being the busiest, with 8679 mails being exchanged between community members (see figure 2). A noticeable trend across all four cycles is that the level communication on the mailing list substantially increased during the final month of each cycle as community members increase efforts to ensure a stable on time release.

What type of sentiment is evident during the busiest cycle and does it fluctuate?

The sentiment model used works by breaking an email body into lists of sentences, and further breaking these sentences into lists of words. Using the Natural Language Toolkit (NLTK) library, the words are tagged

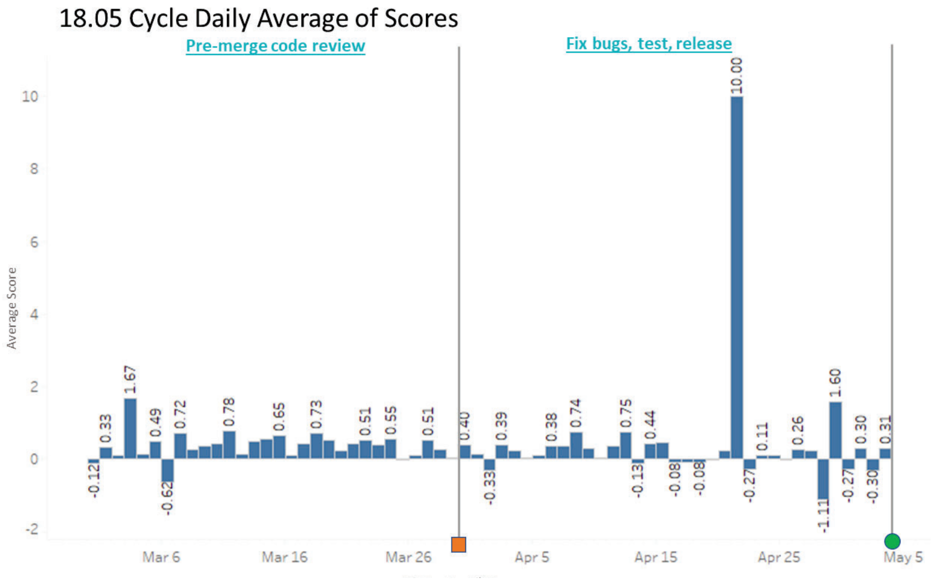


Figure 3 Average Sentiment Scores for 18.05 Cycle

with parts of speech tags (subject, verb, noun etc.). The model tags each word into positive or negative words using an updated custom library with DPDK software terminology being added to the respective dictionaries. Finally, a score is generated based on the count of positive words and the count of negative words.

The results of the sentiment analysis revealed that the sentiment of the DPDK mailing list community was slightly positive, with an average of +0.21 throughout the 18.05 cycle, with a maximum score of +33 and a minimum of -35. Figure 3 shows the daily averages throughout the cycle. This illustrates that in the month prior to the merge there are only two negative days in comparison to the month prior to release, where there are 10 negative days. This indicates that the emotions of some community members changed over the course of the cycle, changing from positive to slightly negative as the cycle ended.

The code used for this analysis could be modified and applied to previous years of DPDK release cycles to identify if the sentiment score change is a common trend across all cycles. Additionally, analysis could identify if sentiment changes occur across the entire community, or just some organisations involved in the project.

References

[1] L. Morgan and P. Finnegan, "Benefits and Drawbacks of Open Source Software: An Exploratory Study of Secondary Software Firms," in *IFIP International Conference on Open Source Systems*, Springer, Boston, MA., 2007.

[2] P. J. Ågerfalk, A. Deverell, B. Fitzgerald and L. Morgan, "Assessing the Role of Open Source Software in the European Secondary Software Sector: A Voice from Industry," in *1st International Conference on Open Source Software,*, Genoa, Italy, 2005.

[3] B. Fitzgerald, "The Transformation of Open Source Software," *MIS Quarterly ,* vol. 30, no. 3, pp. 587-598, 2006.

[4] L. Morgan and P. Finnegan, "Beyond free software: An exploration of the business value of strategic open source," *Journal of Strategic Information Systems ,* vol. 23, no. 3, pp. 226-238, 2014.

[5] DPDK, [Online]. Available: https://www.dpdk.org/about/. [Accessed 8 Feb 2019].

[6] M. De Choudhury and C. Scott, "Understanding Affect in the Workplace via Social Media.," in *2013 conference on Computer supported cooperative work - CSCW*, San Antonio, Texas, USA, 2013.

[7] P. Tourani, Y. Jiang and B. Adams, "Monitoring sentiment in Open Source Mailing Lists - Exploratory stude on the Apache Ecosystem," in *24th annual international conference on computer science and software engineering* , Ontario, Canada, 2014.

[8] D. Garcia, M. Serrano Zanetti and F. Schweitzer, "The Role of Emotions in Contributors Activity: A Case Study on the GENTOO Community," in *3rd International Conference on Cloud and Green Computing*, Karlsruhe, Germany, 2013.