Reproduction of a GitHub Commit Comments Sentiment Analysis Empirical Study

Eleonora Pura
eleonora.pura@uzh.ch
University of Zurich
Zurich, Switzerland

Christian Skorski
christian.skorski@uzh.ch
University of Zurich
Zurich, Switzerland

Gioele Luca Monopoli
gioeleluca.monopoli@uzh.ch
University of Zurich
Zurich, Switzerland

ABSTRACT

Background. A 2016 paper by Sinha et al.[4], *Analyzing Developer Sentiment in Commit Logs* analyzed the GitHub commit messages of 28'466 projects in order to evaluate their general trend of sentiment, the relationships between sentiment and day of the week as well as between sentiment and number of files changed.

Aim. Reproduce results of the aforementioned paper in sentiment analysis and validate them using more recent data. The goal is to is to reproduce the results of this study to provide new empirical evidence to either confirm or contradict its conclusions.

Method. We want to download a set of commit messages of a yet undefined number of open source projects on GitHub using a data collection script of our own and then classify and analyze them using SentiStrength, an analysis tool for measuring and classifying emotions of comments and messages of online public forums.

Conclusion. With this work, we would like to contribute to the repository mining research field by validating the results found in Sinha, Lazar and Sharif's 2016 paper[4] by using our own data collection method and newer data samples, and using the same analysis tool.

1 INTRODUCTION

Usually when thinking about how to improve software engineering we consider facets such as techniques, processes or means of communication. It is not all that common to think about more humane aspects like developers' sentiments and mood. It's important to note that as software engineering continues evolving into an increasingly social field requiring constant interaction with stakeholders, these behavioral expectations result in an emotional toll that can have concrete effects on productivity [3].

Sentiment analysis - the process of measuring and classifying sentiments of unstructured text written by people online [1] - used in the context of software engineering can therefore be an important source of information about productivity, task quality, popularity and trends of projects and programming languages, relationships between developers and job satisfaction. There are multiple sentiment analysis techniques, but all of them boil down to mining people's messages or comments on public forums, parsing and converting them into structured text in order to finally classify them and evaluate their sentiment [1].

Analyzing Developer Sentiment in Commit Logs[4] analyzed and explored sentiments in the context of GitHub commits. This paper focused on different aspects: from the general sentiment of developers to more specific facets such as sentiment depending on the day of the week and on number of changed files. Specifically, the questions to be answered are the following:

RQ1: What is the general developer sentiment in commit messages for GitHub projects?

RQ2: What is the relationship between developer sentiment in commit messages and the day of the week the commit was made?

RQ3: Is there a correlation between the number of changed files and developer sentiment?

Our goal is to reproduce this existing study in order to provide new empirical evidence to either confirm or contradict the authors' results, as well as investigate differences between the older GitHub repositories used in the study (up to 2016) and more recent ones.

To start with, we will specify a sample (of which size is yet to be determined) of target GitHub projects. We will then develop a data mining script in order to extract the commit messages of the sample's projects. After this we will use SentiStrength, a widely tested and improved upon by previous research [6][5] sentiment analysis tool, to automatically evaluate the messages. SentiStrength is well known and used worldwide because of its easy-to-use implementation and adaptability to numerous kinds of problems. The tool evaluates the sentiment strength in the messages evaluating them in a scale between -5 and +5 [2].

The results found from applying SentiStrength to our data set will be evaluated and displayed into graphs.

REFERENCES

[1] Mitali Desai and Mayuri A Mehta. 2016. Techniques for sentiment analysis of Twitter data: A comprehensive survey. In *2016 International Conference on Computing, Communication and Automation (ICCCA)*. IEEE, 149–154.
[2] Professor Laeeq Khan. [n.d.]. *Sentiment Analysis with SentiStrength.* https://professorkhan.com/2019/03/29/sentiment-analysis-with-sentistrength/
[3] Alexander Serebrenik. 2017. Emotional Labor of Software Engineers.. In *BENEVOL*. 1–6.
[4] Vinayak Sinha, Alina Lazar, and Bonita Sharif. 2016. Analyzing developer sentiment in commit logs. In *Proceedings of the 13th International Conference on Mining Software Repositories*. 520–523.
[5] Mike Thelwall. 2017. The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. In *Cyberemotions*. Springer, 119–134.
[6] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology* 61, 12 (2010), 2544–2558.