2016 IEEE/ACM 13th Working Conference on Mining Software Repositories

The Emotional Side of Software Developers in JIRA

Marco Ortu[1], Alessandro Murgia[2], Giuseppe Destefanis[3],
Parastou Tourani[4], Roberto Tonelli[1], Michele Marchesi[1] and Bram Adams[4]
[1]DIEE, University of Cagliari
{marco.ortu,michele,roberto.tonelli}@diee.unica.it
[3]Brunel University London, UK
giuseppe.destefanis@brunel.ac.uk
[2]University of Antwerp, Belgium
alessandro.murgia@ua.ac.be
[4]École Polytechnique de Montréal, Canada
{parastou.tourani,bram.adams}@polymtl.ca

ABSTRACT

Issue tracking systems store valuable data for testing hypotheses concerning maintenance, building statistical prediction models and (recently) investigating developer affectiveness. For the latter, issue tracking systems can be mined to explore developers emotions, sentiments and politeness —affects for short. However, research on affect detection in software artefacts is still in its early stage due to the lack of manually validated data and tools.

In this paper, we contribute to the research of affects on software artefacts by providing a labeling of emotions present on issue comments.
We manually labeled 2,000 issue comments and 4,000 sentences written by developers with emotions such as love, joy, surprise, anger, sadness and fear. Labeled comments and sentences are linked to software artefacts reported in our previously published dataset (containing more than 1K projects, more than 700K issue reports and more than 2 million issue comments). The enriched dataset presented in this paper allows the investigation of the role of affects in software development.

Keywords

Mining software repositories; Issue Reports; Affective Analysis

1. INTRODUCTION

The issue tracking system (ITS) is a software repository that hosts all development tasks of a software organization, i.e., new features, bug fixes and other maintenance tasks. For each task, the ITS provides a description, administrative metadata like the state of the issue (e.g., opened or fixed) and the priority, as well as a chronology of comments and attachments by developers to discuss the task at hand. For this reason, ITS is fundamental to explore how developers interact, as well as how they feel about the project and their peers. In issue comments, developers discuss issues by providing technical details and opinions, useful to understand the reason of certain design decisions or about the status of a project. From this textual information it is possible to extract emotions, sentiments and politeness (affects). Murgia et al. [11] showed that developers do express emotions such as love, joy, sadness towards colleagues and software artefacts. Ortu et al. [12] showed that emotions contained in these issue comments have negligible correlation with each other.

Emotions and feelings have a big influence on our actions and decisions [17]. Thus, a purely rational view of software development based on counting the artefacts it produces, provides only a partial explanation of team dynamics and productivity. Recently, the software engineering community started to closely investigate the role of affects in software development [11, 12, 14, 19, 22]. However, being the subject relatively new[1], we are in need of more data and tools for continuing the research. Today, there are no public datasets —manually validated— which link affects to software artefacts, neither standard tools for extracting affects information from software artefacts. Although there exist publicly available tools able to detect sentiment and politeness, they have been created for domains different from software development and software engineering. Therefore, their performances may underachieve the expectations or in the worst case lead to wrong results. Limited to sentiment analysis, Jongeling et al. [6] showed that general purpose tools such as SentiStrength and NLTK were unreliable for assessing sentiments in technical prose within issue comments.

In this paper, we address the lack of data in affects associated to software artefacts, providing a manual labelling of emotions present within issue comments. Using as a baseline our previous published dataset [13], we provide:

- 392 issue comments labeled with emotions love, joy, surprise, anger, fear and sadness [11].
- 1,600 issue comments labeled with emotions love, joy, sadness

[1]The international workshop on emotion awareness in software engineering has been held for the first time in 2016.

MSR'16, May 14-15 2016, Austin, TX, USA

ISBN 978-1-4503-4186-8/16/05…$15.00

http://dx.doi.org/10.1145/2901739.2903505

- 4,000 issue sentences[2] labeled with emotions love, joy, anger and sadness [12].

This data is highly valuable for (i) investigating the impact of affects on software development as well as (ii) training tools for affects detection. Linking affects to software artefacts hosted in ITS, allows easy replication and extension of previous research based on ITS [12, 14, 15]. Beyond the analysis of affects, the dataset allows studies on traditional ITS topics such as bug triage, bug tossing, and bug priority. Finally, by hosting "Agile data" like story points, sprints etc., the dataset can be exploited also for investigations related to Agile practices. Jira is one of the most common ITS technologies adopted by companies. The dataset we enriched hosts more than 1K projects, 700K issue reports and 2 million of comments. This data is collected from the repositories of four open source communities: Apache, Spring, JBoss, and CodeHaus. These ecosystems were selected since they are well known by practitioners.

The rest of the paper is structured as follows. First we describe how the dataset is built and organized (Section 2). Then we report the research opportunities based on its adoption (Section 3) and finally the conclusions (Section 4).

2. DATASET

2.1 The Emotional Annotated Dataset

We extended the dataset by Ortu et al. [13] adding a set of manually annotated comments which have been used in several studies on human aspect in software engineering [11, 12, 14]. Table 1 shows some statistics about the new content. In particular, we provide new information (divided in three groups) for emotion detection with different granularity: comment level and sentence level. During the labeling process, a different number of raters were involved for each group of files. Each file, provided in CSV format, contains a column named *id* which represents the corresponding *comment ID* baseline dataset [13]. These files are available as archive[3].

Being the rating process performed across different experiments and for different purposes, we reported a specific section to describe the methodology adopted to create the three groups of files.

Name	Granularity	Comm. Labeled	Raters	Emotions
Group 1	Comment	392	16	*JOY*, *LOVE*, *SADNESS*, *ANGER*, *FEAR*, *SURPRISE*
Group 2	Comment	1600	3	*JOY*, *LOVE*, *SADNESS*
Group 3	Sentence	4000	3	*JOY*, *LOVE*, *SADNESS*, *ANGER*

Table 1: Emotional Dataset Statistics

Group 1

We used the Parrott's framework as a reference for emotions [16]. According this framework, raters labeled the comments as having (or not) one of the following six emotions: love, joy, surprise, anger, sadness and fear. The labeling depended on (i) rater's personal interpretation of emotions, and (ii) his/her common understanding of Parrott's framework[4]. The raters used for this labeling were four Master students, ten PhD students and two research associates at the Ecole Polytechnique de Montreal and the University of Antwerp. The dataset provided contained 16 files, one for each rater. Raters were organized in two groups in which master and PhD students were evenly distributed. The comments used for the analysis were randomly sampled. Each rater labeled 14 comments in common with each other group member to reduce possible biases due to different nationalities, skills and cultural background of the participants. Moreover, each comment was labeled by four raters. We calculated the degree of agreement for the identified emotions using the Cohen's κ value[5] (two raters) or Fleiss' κ value[6] (more than two raters) [11]. The results showed that raters agreed the most on the absence of an emotion and having more than two raters did not have a big impact on the agreement. Finally, we found that love, joy and sadness obtained at least fair agreement. The interested reader can found more details on the experiment in Murgia et al. [11].

Group 2

After the first analysis, we noticed that only love, joy and sadness obtained fair agreement among raters. For this reason we extended the original dataset only focusing on these emotions. The new dataset accounted for 1,600 comments. These comments were labeled by three of the authors of Murgia et al. [11] for the presence of love, joy, sadness, emotion-free (none of the three emotions present). In these dataset the three raters achieved a level of agreement from moderate to substantial about presence or absence of emotions. Also in this case the majority of the comments were labeled as neutral (43.4%).

Group 3

This group contains the data used by Ortu et al. [12]. The files contain 4K sentences labeled by three raters, at a sentence level, who focused on four emotions: love, joy, sadness and anger. Compared to Group 1 and Group 2, the annotated comments provide a finer-grain labeling. These files contain a different labeling convention, they report disagreement among raters.

2.2 Jira Database Description

The Jira dataset consists of issues extracted from the public Jira repository of four open source (OS) communities: Apache, Spring, JBoss and CodeHaus[7]. These OS communities use Jira for both tracking issues and managing projects. Issues in Jira are divided in categories such as bugs, improvements, feature requests and much more, as described by Ortu et al. [13]. The database contains the following tables:

[2]Each issue comment is a sequence of sentences. An issue's discussion comprises multiple issue comments

[3]http://ansymore.uantwerpen.be/system/files/uploads/artefacts/alessandro/MSR16/archive3.zip

[4]The framework was explained and illustrated to the raters before starting the labeling process

[5]https://en.wikipedia.org/wiki/Cohen%27s_kappa

[6]https://en.wikipedia.org/wiki/Fleiss%27_kappa

[7]https://spring.io,
http://www.apache.org,
http://www.codehaus.org,
http://www.jboss.org

JIRA_ISSUES_REPORT: it stores the information extracted from the issue reports. Issues are associated with comments, attachments and history changes.

JIRA_ISSUES_COMMENTS: it contains all the comments posted by users and developers in a Jira issue report. This table is associated with the *JIRA_ISSUES_REPORT* table. This table is enriched with affective data, such as the emotional content, sentiment [8] and politeness [4]. The following is an example of an extracted comment:

```
Hey <dev_name_a>,
Would you be interested in contributing
a fix and a test case for this as well?
Thanks,
<dev_name_b>
```

JIRA_ISSUE_BOT_COMMENT: it contains automatically generated comments from tools such as Jenkins or Jira itself.

JIRA_ISSUES_FIXED_VERSION: it records the software version of fixed issues.

JIRA_ISSUES_AFFECTED_VERSION: it records which software versions are affected by an issue.

JIRA_ISSUE_ATTACHMENT: it contains all files attached to an issue report.

JIRA_ISSUE_CHANGELOG_ITEM: it contains information about operations performed on a given issue such as editing, updating, status changing, etc.

As an example, the following SQL query describes how to retrieve the number of issue reports from the Apache Software Foundation (ASF) with comments expressing *ANGER*:

```
SELECT count(distinct(i.id))
FROM jira_issue_comment c,
     jira_issue_report i
WHERE i.id = c.issue_report_id
                AND c.anger_count > 0
                AND i.repositoryName = 'ASF'
```

Table 2 shows the statistics of the publicly available dataset.

Description	Values
# Projects	1K
# Issues	700K
# Comments	2M
# Users	100K
# Attachments	60K

Table 2: Dataset statistics

Although the dataset is limited to four open source ecosystems, we are confident that the data extracted is complete and consistent. However, considering that the communication process of software development is held across different media such as mailing list and social media, not all the discussions about an issue are held in the issue tracking system and this represents a major limitation. Future extensions of our dataset will take also these media into account.

[8]Measured using SentiStrength
http://sentistrength.wlv.ac.uk/

3. RESEARCH OPPORTUNITIES

Recently, researchers started to investigate the role of affects in software engineering. Researchers are investigating how the human aspects of a technical discipline such as software engineering can affect the final results [2, 5, 7, 8]. In this context, we believe that our new data can be exploited to:

- consider affects in models for bug fixing time estimation [23]. Ortu et al. [12] showed that bug fixing time correlate with the affects expressed by developers in issue comments. Murgia et al. [10] and Ortu et al. [12] showed that issue fixing time is related respectively to the type of maintenance performed and the affects reported in issue reports.
- study the impact of affects on (i) the learning-curve, (ii) productivity and (iii) project's attractiveness to new developers. For example, Ortu et al. [14] showed that the more polite developers were, the more new developers wanted to be part of a project and the more they were willing to continue working on it over time [14].
- investigate the impact of affects on software quality. For example, Tourani et al. [21] studied the impact of human discussion metrics including affective metrics on software quality. However, this work only touched the tip of the iceberg.
- Bacchelli et al. [1] and McIntosh et al. [9] found that modern reviewing techniques, do not imply high quality reviews. Affects extracted from comments in addition to other important metrics, specifically can be used to investigating code review quality.
- analyze social and technical debt in software development [18, 20] or bug life cycle [3].
- study the impact of affects regarding scheduling of developers. Ortu et al. [12] showed that time spent for issue fixing correlated with affects extracted from discussions. Investigating whether it can be extended to other parts of project scheduling is another research opportunity.

4. CONCLUSION

Data stored in ITS is fundamental for empirical research in software engineering since it can be used for verifying, refuting and challenging previous theory and results in software maintenance and productivity. Recently, analysis on ITS has focused on affects in software development. In this context, the MSR community has started to mine issue comments in order to extract emotions, sentiments and politeness.

This paper provides a manual labeling of emotions reported in 2,000 issue comments and 4,000 sentences written by developers. The data extends our previously dataset extracted from more than 1K projects, 700K issue reports and 2 million comments. We used this dataset for (i) studying emotions expressed by developers towards colleagues and software artefacts and (ii) investigating how affects influence software development. Sharing this repository, we would like to encourage the community to perform replication as well as further studies on affect analysis or classical ITS topics.

5. REFERENCES

[1] A. Bacchelli and C. Bird. Expectations, outcomes, and challenges of modern code review. In *Proc. of the 2013 Intl. Conf. on Software Engineering (ICSE)*, pages 712–721, 2013.

[2] A. P. Brief and H. M. Weiss. Organizational behavior: Affect in the workplace. *Annual Review of Psychology*, 53(1):279–307, 2002. PMID: 11752487.

[3] B. Caglayan, A. T. Misirli, A. Miranskyy, B. Turhan, and A. Bener. Factors characterizing reopened issues: A case study. In *Proceedings of the 8th International Conference on Predictive Models in Software Engineering*, PROMISE '12, pages 1–10, New York, NY, USA, 2012. ACM.

[4] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A computational approach to politeness with application to social factors. In *Proceedings of ACL*, 2013.

[5] A. Erez and A. M. Isen. The influence of positive affect on the components of expectancy motivation. *Journal of Applied Psychology*, 87(6):1055, 2002.

[6] R. Jongeling, S. Datta, and A. Serebrenik. Choosing your weapons: On sentiment analysis tools for software engineering research. In *Software Maintenance and Evolution (ICSME), 2015 IEEE International Conference on*, Sept 2015.

[7] E. Kaluzniacky. *Managing Psychological Factors in Information Systems Work: An Orientation to Emotional Intelligence*. AAA, 2004.

[8] O. Kononenko, O. Baysal, L. Guerrouj, Y. Cao, and M. W. Godfrey. Investigating code review quality: Do people and participation matter? In *Software Maintenance and Evolution (ICSME), 2015 IEEE International Conference on*, pages 111–120. IEEE, 2015.

[9] S. McIntosh, Y. Kamei, B. Adams, and A. E. Hassan. The impact of code review coverage and code review participation on software quality: A case study of the qt, vtk, and itk projects. In *Proc. of the 11th Working Conf. on Mining Software Repositories (MSR)*, pages 192–201, 2014.

[10] A. Murgia, G. Concas, R. Tonelli, M. Ortu, S. Demeyer, and M. Marchesi. On the influence of maintenance activity types on the issue resolution time. In *Proceedings of the 10th International Conference on Predictive Models in Software Engineering*, PROMISE '14, pages 12–21, New York, NY, USA, 2014. ACM.

[11] A. Murgia, P. Tourani, B. Adams, and M. Ortu. Do developers feel emotions? an exploratory analysis of emotions in software artifacts. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, pages 262–271, New York, NY, USA, 2014. ACM.

[12] M. Ortu, B. Adams, G. Destefanis, P. Tourani, M. Marchesi, and R. Tonelli. Are bullies more productive? empirical study of affectiveness vs. issue fixing time. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR 2015, 2015.

[13] M. Ortu, G. Destefanis, B. Adams, A. Murgia, M. Marchesi, and R. Tonelli. The jira repository dataset: Understanding social aspects of software development. In *Proceedings of the 11th International Conference on Predictive Models and Data Analytics in Software Engineering*, PROMISE '15, pages 1:1–1:4, New York, NY, USA, 2015. ACM.

[14] M. Ortu, G. Destefanis, M. Kassab, S. Counsell, M. Marchesi, and R. Tonelli. Would you mind fixing this issue? In *Agile Processes, in Software Engineering, and Extreme Programming*, pages 129–140. Springer, 2015.

[15] M. Ortu, G. Destefanis, M. Kassab, and M. Marchesi. Measuring and understanding the effectiveness of jira developers communities. In *Proceedings of the 6th International Workshop on Emerging Trends in Software Metrics*, WETSoM 2015, 2015.

[16] W. Parrott. *Emotions in Social Psychology*. Psychology Press, 2001.

[17] R. Plutchik and H. Van Praag. The measurement of suicidality, aggressivity and impulsivity. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 13:S23–S34, 1989.

[18] A. Potdar and E. Shihab. An exploratory study on self-admitted technical debt. In *Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on*, pages 91–100. IEEE, 2014.

[19] P. C. Rigby and A. E. Hassan. What can oss mailing lists tell us? a preliminary psychometric text analysis of the apache developer mailing list. In *Proceedings of the Fourth International Workshop on Mining Software Repositories*, page 23. IEEE Computer Society, 2007.

[20] D. Tamburri, P. Kruchten, P. Lago, H. Van Vliet, et al. What is social debt in software engineering? In *Cooperative and Human Aspects of Software Engineering (CHASE), 2013 6th International Workshop on*, pages 93–96. IEEE, 2013.

[21] P. Tourani and B. Adams. The impact of human discussions on just-in-time quality assurance. In *Proceedings of the 23rd IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, Osaka, Japan, March 2016.

[22] P. Tourani, Y. Jiang, and B. Adams. Monitoring sentiment in open source mailing lists – exploratory study on the apache ecosystem. In *Proceedings of the 2014 Conference of the Center for Advanced Studies on Collaborative Research (CASCON)*, Toronto, ON, Canada, November 2014.

[23] H. Zhang, L. Gong, and S. Versteeg. Predicting bug-fixing time: An empirical study of commercial software projects. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, pages 1042–1051, Piscataway, NJ, USA, 2013. IEEE Press.