Sentiment Analysis of Software Bug Related Commit Messages

Md Rakibul Islam and Minhaz F. Zibran
Department of Computer Science, The University of New Orleans
New Orleans, Louisiana, United States
(mislam3, mfzibran)@uno.edu

Abstract

Software development activities, being highly dependent on human efforts, are affected by individual's emotions. We study the emotional variations in bug-introducing and bug-fixing commit messages. Our quantitative study includes more than 24,000 commit messages over three open-source projects. The results advance our understanding of the extent to which emotions affect tasks resulting in software bugs, and how such bug-introducing commits differ from bug-fixing ones in terms of the emotions expressed in the commit messages.

1 Introduction

Software developers, being humans, are affected by their emotions, which influence their activities and interactions. Thus, emotions affect task quality, productivity, creativity, group rapport and job satisfaction [4, 21]. Several studies have been performed in the past for understanding roles of emotions on software development activities. Some of those earlier studies address *when* and *why* employees get affected by emotions [4, 9, 10, 22, 26].

Some other studies determine correlations between various job performance factors (e.g., productivity, quality and efficiency) and emotions that developers experience during development activities [8, 11, 12, 18, 19, 20, 27]. In addition, a few studies have *successfully* used emotion as a factor in prioritizing applications' features to develop [6] and in predicting qualities of developers' interactions (e.g., asking questions and answering) in technical forums [2, 3, 17] and bug severity [28].

Considering above studies, it deems emotions can be an influential factor to be used in complex machine learning and deep learning systems to predict bugs (i.e., buggy commits) in software. However, before using emotions to predict bugs, we need to empirically evaluate of such possibility in the context. Towards this goal, in this work, we study the polarity (i.e., positivity, negativity, and neutrality) of emotions expressed in two types of commit messages, (i) *bug-introducing* and (ii) *bug-fixing*, which are posted by developers contributing to open-source projects. In particular, we address the following two research questions.

RQ1: *Do developers express different levels (e.g., high, low) and polarity (i.e., positivity, negativity, and neutrality) of emotions in bug-introducing and bug-fixing commits?*
— If we can identify developers express higher level emotions (either positive or negative) during commits, which cause bugs in software compared to other commits (e.g., bug-fixing commits) then expressed emotional levels can be used as a feature to predict bugs in commits. In addition, here we also want to verify the finding of Islam and Zibran [12] where they claim positive emotions are significantly higher in bug-fixing commits compared to negative emotions.

RQ2: *Do the developers' polarity (i.e., positivity, negativity, and neutrality) of emotions vary in different times of a day in bug-introducing and bug-fixing commits*?
— Here we conduct a deeper analysis by grouping bug-introducing and bug-fixing commits according to their commit timestamps. If we can identify any particular times in a day when developers express significant negative emotions, then managers can take required steps to uplift the developers positive feelings at those times.

2 Methodology

The procedural steps of our empirical study are summarized in Figure 1. To address the aforementioned research questions, we extract bug-introducing and bug-fixing commit messages from three selected projects. Then, we compute emotional scores of commit messages using the tool `SentiStrength-SE` [13, 15], which is the first domain specific sentiment analysis tool for software engineering texts. Finally, to answer the research questions we conduct statistical analyses. In the following we briefly describe the procedures of data collection, computation of emotional scores and how we conduct data analysis.

978-1-943436-13-2 / copyright ISCA, SEDE 2018
October 8-10, 2018, New Orleans, Louisiana, USA

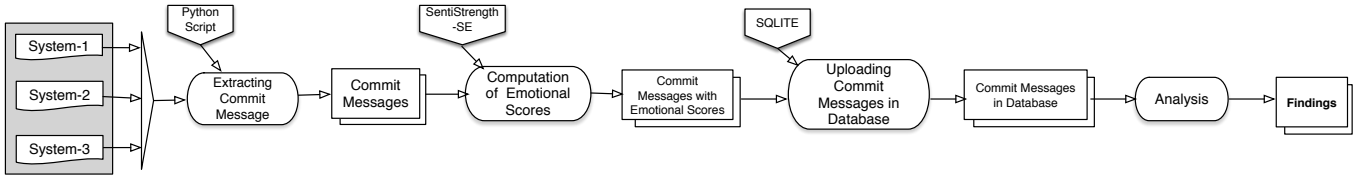

Figure 1: Procedural steps of our empirical study

Table 1: Subject Systems

Systems	Lang.	Domain	BIC*	BFC$^+$
Netty	Java	Network	8,745	8,739
Presto	Java	SQL	2,963	2,963
Facebook-android-SDK	Java	Social Network	740	740

*BIC=Bug-introducing commits, $^+$BFC=Bug-fixing commits

2.1 Data Collection

We collect three open-source projects from GitHub listed in Table 1 along with other information related to those projects. Those three projects are also used in other studies [24, 16]. The bug-introducing and bug-fixing commit messages of those projects are identified in an earlier study [24] that we reuse in this study.

We study all those two types of commit messages in these projects, which constitute 24,890 commit comments. Associated information such as, committers, commit timestamps, revisions and project names are kept in a local relational database for convenient access and query.

2.2 Sentiment Analysis

For each of the commit messages, we compute the emotional scores using the tool `SentiStrength-SE` [13]. Sentiment analysis using `SentiStrength-SE` on a given piece of text (e.g., a commit message) c computes a pair $\langle \rho_c, \eta_c \rangle$ of integers, where $+1 \leq \rho_c \leq +5$ and $-5 \leq \eta_c \leq -1$. Here, ρ_c and η_c respectively represent the positive and negative emotional scores for the given text c.

A given text c is considered to have positive emotions if $\rho_c > +1$. Similarly, a text is held containing negative emotions when $\eta_c < -1$. Note that, a given text can exhibit both positive and negative emotions at the same time, and a text is considered emotionally neutral when the emotional scores for the text appear to be $\langle 1, -1 \rangle$. Further details about the sentiment analysis algorithm of `SentiStrength-SE` and the interpretation of its outputs can be found elsewhere [13].

2.3 Statistical Measurements

To verify the statistical significance of emotional variances in bug-introducing and bug-fixing commit messages, we apply the statistical *Mann-Whitney-Wilcoxon (MWW)* test [1] at the significance level $\alpha = 0.05$. The non-parametric MWW test does not require normal distribution of data, and thus it suits well for our purpose. To measure the effect size, we compute the non-parametric effect size *Cliff's delta* d [1]. We consider *significant difference* exists between distributions of emotional scores in bug-introducing and bug-fixing commits if p-value of a MWW test is found to be less than α and *Cliff's delta* d value is not negligible (i.e., $|d| > 0.15$).

3 Analysis and Findings

The research questions *RQ1* and *RQ2* are respectively addressed in Section 3.1 and Section 3.2.

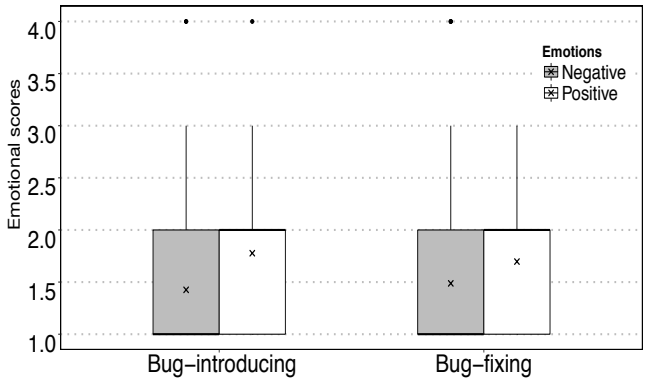

Figure 2: Distributions of positive and negative emotional scores in commits messages dealing with bug-introducing and bug-fixing tasks

3.1 Overall Emotional Variations

Emotional variations in commit types: The boxplot in Figure 2 presents the distributions of positive and negative emotional scores in bug-introducing and bug-fixing commit messages collected from the three projects. An 'x' mark in a box in the box-plot indicates the mean emotional scores over all the commits. As observed in Figure 2, for both bug-introducing and bug-fixing commit messages, emotional scores follow two similar patterns: (i) emotional scores of 75% commit messages remain within one to two in positively and negatively polarized commit messages and (ii) mean

and median emotional scores in positively polarized messages are higher as opposed to mean and median emotional scores in negatively polarized messages.

We conduct a *MWW* test to verify statistical significance of difference between positive and negative emotional scores in bug-introducing commit messages. The test obtains p-value 2.57×10^{-11} where $p < \alpha$ and *Cliff's delta* d value -0.3032 where $|d| > 0.15$. Thus, those values indicate that the difference between positive and negative emotional scores in bug-fixing commit messages is significant.

Similarly, we conduct another *MWW* test between positive and negative emotional scores in bug-fixing commit messages to verify statistical significance of their difference. The test obtains p-value 2.2×10^{-16} where $p < \alpha$ and *Cliff's delta* d value -0.2246. Again, the obtained values indicate that the difference between positive and negative emotional scores in bug-fixing commit messages is significant.

Variation of a particular emotion across commit types: Next, we focus on emotion-wise differences between emotional scores in bug-introducing and bug-fixing commit messages. In other words, we want to verify statistical significances of difference between positive emotional scores found in bug-introducing and bug-fixing commit messages and difference between negative emotional scores found in bug-introducing and bug-fixing commit messages. From Figure 2, we see that the difference in mean positive emotional scores between bug-introducing and bug-fixing commit messages does not differ that much. Similar pattern can also be seen for difference in mean negative emotional scores between bug-introducing and bug-fixing commit messages.

To verify the statistical significance of our observations, we sequentially conduct two *MWW* tests. The first *MWW* test is conducted between positive emotional scores found in bug-introducing and bug-fixing commit messages. The test obtains p-value 0.076 where $p > \alpha$ indicates that the difference between positive emotional scores in bug-introducing and bug-fixing commit messages is not significant.

Similarly, we conduct another *MWW* test between negative emotional scores found in bug-introducing and bug-fixing commit messages, which obtains p-value 0.9561 where $p > \alpha$. Thus, the later also test indicates no significant difference between negative emotional scores in bug-introducing and bug-fixing commit messages.

Based on our observations and statistical tests, we derive the answer to the research question *RQ1* as follows:

Ans. to RQ1: *Both bug-introducing and bug-fixing commit messages have significantly higher positive emotional scores compared to negative emotional scores. However, neither positive nor negative emotional scores differ much between bug-introducing and bug-fixing commit messages.*

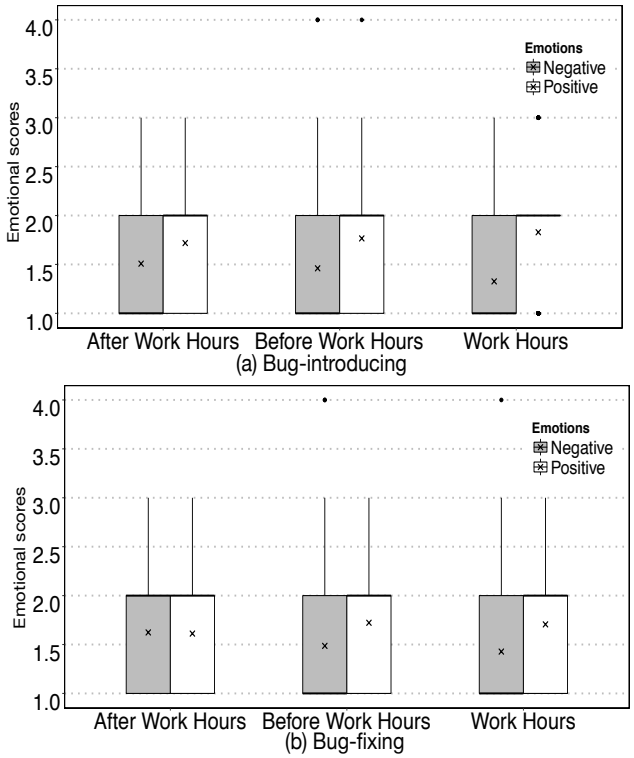

Figure 3: Distributions of positive and negative emotional scores in (a) bug-introducing and (b) bug-fixing commits messages

3.2 Hour-wise Emotional Variations

To study the relationship between developers emotions and times of a day when commit comments are posted, we divide the 24 hours of a day in three periods (a) 00 to 08 hours as *before working hours*, (b) 09 to 17 hours as regular *working hours* and (c) 18 to 23 hours as *after working hours*. Then, for each project, we organize the commit messages into three disjoint sets based on their timestamps of posting.

Variations of emotions in different commit types with respect to work hours: Figure 3(a) and 3(b) present the distributions of positive and negative emotional scores in bug-introducing and bug-fixing commit messages respectively posted in these three periods. We see from these figures that average positive emotional scores are always higher as opposed to average negative emotional scores except in *after work hours* for bug-fixing commit messages. Again, for both sentiments the median values in *after work hours* are equal only for bug-fixing commit messages, whereas

for rest of the cases, the median values are always higher for positive sentiments. Another noticeable pattern can be observed for positive emotional scores in *work hours* for bug-introducing commit messages where the emotional scores are highly centered to value two.

Table 2: Results of *MWW* tests between positive and negative emotional scores of different commit types posted in different times of a day

Commit Types	AWH	BWH	WH
Bug-introducing	**0.0007, -0.2812**	**1.1e-10, 0.6943**	**2.2e-16, -0.1708**
Bug-fixing	0.4430, -0.3882	**6.8e-07, 0.6805**	**3.3e-08, -0.2941**

Here, AWH=After Work Hours, BWH=Before Work Hours, WH=Work Hours

To verify the statistical significance of our observations, we conduct a series of *MWW* tests between positive and negative emotional scores in bug-introducing and bug-fixing commit messages respectively. The obtained p-values and their corresponding d values are presented (as a pair of p-value, d) in Table 2 where significant differences are marked bold. We see the differences are statistically significant in all cases except in *after work hours* for bug-fixing commit messages.

Variation of a particular emotion with respect to work hours: Next, we focus on emotion-wise difference between emotional scores of bug-introducing and bug-fixing commit messages posted in the three working periods. Figure 4(a) and 4(b) present the distributions of positive and negative emotional scores respectively in bug-introducing and bug-fixing commit messages in the three working periods. In those working periods, the averages of positive emotional scores are always higher in bug-fixing commit messages (see Figure 4(a)) whereas, interestingly, the averages of negative emotional scores are always higher in bug-introducing commit messages (see Figure 4(b)).

To verify the statistical significance of our observations, we again conduct a series of *MWW* tests between positive emotional scores in bug-introducing and bug-fixing commit messages and between negative emotional scores in bug-introducing and bug-fixing commit messages. The obtained p-values and their corresponding d values are presented in Table 3 where significant differences are marked bold. We see the only significant difference is between positive emotional scores in bug-introducing and bug-fixing commit messages posted during working hours.

Based on our observations and statistical tests, we now answer the research question *RQ2* as follows:

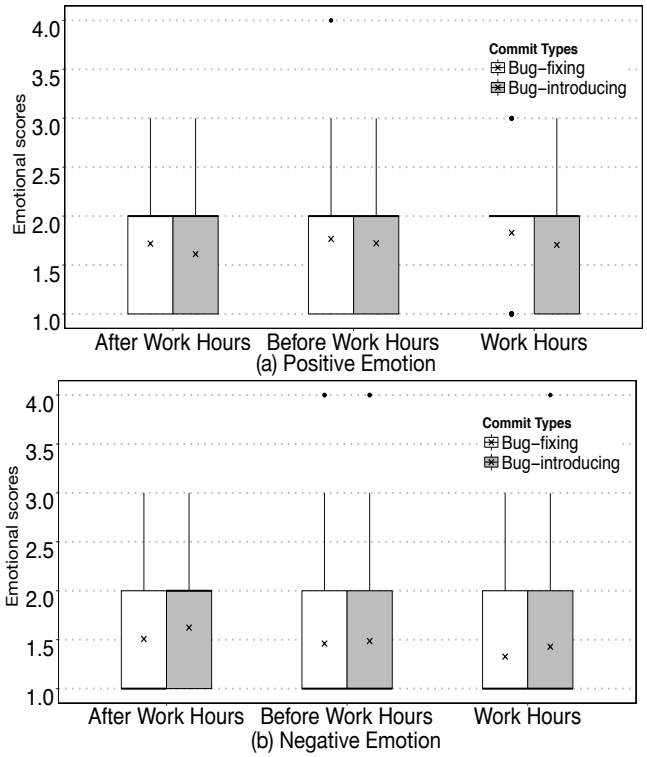

Figure 4: Distributions of (a) positive emotional scores and (b) negative emotional scores, in bug-introducing and bug-fixing commit messages

Table 3: Results of *MWW* tests of emotional scores between bug-introducing and bug-fixing commit messages posted in different times of a day

Emotion Types	AWH	BWH	WH
Positive	0.0866, 0.6640	0.2433, -0.2377	**0.0089, -0.1708**
Negative	0.9084, -0.4921	0.6677, 0.4075	0.9616, -0.6733

Ans. to RQ2: *Both bug-introducing and bug-fixing commit messages have significantly higher positive emotional scores compared to negative emotional scores in commits made in the three working hours, except for the after work hours of bug-fixing commit messages. During working hours, positive emotional scores in bug-introducing commits are significantly higher compared to the positive emotional scores in bug-fixing commits.*

4 Threats to Validity

Construct Validity: One may question the validity of our categorization of the developers' commits in different periods (Section 3.2), considering the possibility that the projects and developers may be physically

located at different geographic locations and time-zones. However, we used the time zone information associated with the commit messages to convert all the timestamps to corresponding local times.

For the statistical tests of significance, we used the *Mann-Whitney-Wilcoxon (MWW)* test [23] and to compute effect size we used *Cliff's delta* [1]. The non-parametric *MWW* test along with non-parametric effect size *Cliff's delta* do not require the data to have normal distribution. Since the data in our work do not conform to normal distribution, this particular test suits well for our purpose. Moreover, the significance level α set to 0.05, which is a widely adopted value for this parameter that enables 95% confidence in the results of the *MWW* tests.

Internal Validity: The *internal validity* of our work depends on the accuracy of the tool's computation of emotional scores. `SentiStrength-SE` was reported to be effective in sentiment analysis [14] and suitable for extraction of emotions from commit comments. Nevertheless, the tool is not 100% accurate in determining emotional polarities of commit messages. We are aware of this threat and manually checked 50 comments and found their emotions were correctly identified by `SentiStrength-SE`.

One may question the accuracy of the approach to determine bug-introducing and bug-fixing commits. However, a manual checking found 96% accuracy of the approach in distinguishing bug-introducing and bug-fixing commits in the projects [24].

External Validity: The findings of this work are based on our study on more than 24,000 commit messages across three open-source projects. Generalizability of the results can be questioned due to small sized dataset.

Reliability: The methodology of this study including the procedure for data collection and analysis is well-documented in this paper. The subject systems being open-source, are freely accessible while the tool `SentiStrength-SE` is also available online. Moreover, all the bug-introducing and bug-fixing commits are also available. Therefore, it should be possible to replicate the study.

5 Related Work

There are several studies exist in literature that compare emotional variations found in various textual artifacts related to software engineering. In such a study, Islam and Zibran [12] presented a quantitative empirical study of the emotional variations in different types of development activities and development periods, in addition to in-depth investigation of emotions' impacts on software artifacts. They found significantly higher positive emotion in bug-fixing commits compared to negative emotion, which is similar to our finding. From commit messages, Guzman et al. [9], Chowdhury and Hindle [5] and Sinha et al. [25] also identified emotions and group those in various dimensions. However, none of the above studies computed emotional scores in bug-introducing commits and compared against emotional scores in bug-fixing commit, that we have performed in our work.

Tourani et al. [26] extracted emotions from emails of both developers and system users. They observed the differences of emotional expressions between developers and users of a system. Garcia et al. [7] extracted developers' emotions from their email contents to analyze any relationships between developers' emotions and their activities in an open source software project. The studies of Tourani et al. [26] and Garcia et al. [7] differ with our work mainly in two ways (i) the source of their emotional content is different than ours and (ii) the objectives of those work are also orthogonal to ours.

Pletea et al. [22] mined developers' emotions commits and pull requests in `GitHub` projects. They analyzed emotional variations in discussions on different topics and reported to have found higher negative emotions in security-related discussions in comparison with other topics. While their objective, approach as well as source of emotional content and method of emotion extraction were different from our work, ours includes a deeper analysis based on emotional variations in bug-introducing and bug-fixing commit messages.

All the above mentioned studies used domain independent tools (e.g., *SentiStrength* and *NLTK*) to compute emotional scores in software engineering textual artifacts, however, for the first time, we have used a domain specific tool *SentiStrength-SE* for this study, which is one of the unique attributes of this work.

6 Conclusion

In this paper, we have presented a quantitative empirical study on the emotional variations between bug-introducing and bug-fixing commit messages. We have studied more than 24,000 commit messages over three open-source projects.

In our study, we find that both bug-introducing and bug-fixing commit messages have overall statistically significantly higher positive emotional scores compared to negative emotional scores. We also observe similar findings while analyzing emotional scores in bug-introducing and bug-fixing commit messages with respect to three working periods. An exception is found in the later case where no significant difference found between positive and negative emotional score in bug-

fixing commit messages posted during *after work hours*.

While comparing with respect to a particular sentiment (i.e., for positive or negative sentiment), we find no significant difference between overall emotional scores in bug-introducing and bug-fixing commit messages. The earlier finding also holds true while analyzing emotional scores in bug-introducing and bug-fixing commit messages with respect to three working periods with an exception where positive emotional scores are significantly higher in bug-fixing commits posted during *working hours*.

The findings from this work are validated in the light of statistical significance. Although more experiments can be conducted in large scale to verify or confirm the findings, the results from this study significantly advance our understanding of the impacts of emotions in software development activities.

Acknowledgement

This work is supported in part by the SCoRe grant at the University of New Orleans.

References

[1] D. Anderson, D. Sweeney, and T. Williams. *Statistics for Business and Economics*. Thomson Higher Education, 10th edition, 2009.

[2] F. Calefato, F. Lanubile, M. Marasciulo, and N. Novielli. Mining successful answers in stack overflow. In *MSR*, pages 430–433, 2015.

[3] F. Calefato, F. Lanubile, and N. Novielli. Moving to stack overflow: Best-answer prediction in legacy developer forums. In *ESEM*, 2016.

[4] M. Choudhury and S. Counts. Understanding affect in the workplace via social media. In *CSCW*, pages 303–316, 2013.

[5] S. Chowdhury and A. Hindle. Characterizing energy-aware software projects: Are they different? In *MSR*, pages 508–511, 2016.

[6] A. Ciurumelea, A. Schaufelbuhl, S. Panichella, and Harald Gall. Analyzing reviews and code of mobile apps for better release planning. In *SANER*, pages 91–102, 2017.

[7] D. Garcia, M. Zanetti, and F. Schweitzer. The role of emotions in contributors activity: A case study on the gentoo community. In *CCGC*, pages 410–417, 2013.

[8] D. Graziotin, X. Wang, and P. Abrahamsson. Are happy developers more productive? The correlation of affective states of software developers and their self-assessed productivity. In *PROFES*, pages 50–64, 2013.

[9] E. Guzman, D. Azócar, and Y. Li. Sentiment analysis of commit comments in github: An empirical study. In *MSR*, pages 352–355, 2014.

[10] E. Guzman and B. Bruegge. Towards emotional awareness in software development teams. In *ESEC/FSE*, pages 671–674, 2013.

[11] M. Islam and M. Zibran. Exploration and exploitation of developers' sentimental variations in software engineering. *International Journal of Software Innovation*, 4(4):35–55, 2016.

[12] M. Islam and M. Zibran. Towards understanding and exploiting developers' emotional variations in software engineering. In *SERA*, pages 185–192, 2016.

[13] M. Islam and M. Zibran. Leveraging automated sentiment analysis in software engineering. In *MSR*, pages 203–214, 2017.

[14] M. Islam and M. Zibran. A comparison of software engineering domain specific sentiment analysis tools. In *SANER*, pages 487–491, 2018.

[15] M. Islam and M. Zibran. SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of System and Software*, 145:125–146, 2018.

[16] M. Islam, M. Zibran, and A. Nagpal. Security vulnerabilities in categories of clones and non-cloned code: An empirical study. In *ESEM*, pages 20 – 29, 2017.

[17] J. Jiarpakdee, A. Ihara, and K. Matsumoto. Understanding question quality through affective aspect in Q&A site. In *SEmotion*, pages 12–17, 2016.

[18] I. Khan, W. Brinkman, and R. Hierons. Do moods affect programmers' debug performance? *Cogn. Technol. Work*, 13(4):245–258, 2010.

[19] T. Lesiuk. The effect of music listening on work performance. *Psychology of Music*, 33(2):173–191, 2005.

[20] A. Murgia, P. Tourani, B. Adams, and M. Ortu. Do developers feel emotions? an exploratory analysis of emotions in software artifacts. In *MSR*, pages 261–271, 2014.

[21] R. Palacios, A. López, A. Crespo, and P. Acosta. A study of emotions in requirements engineering. *Organizational, Business, and Technological Aspects of the Knowledge Society*, 112:1–7, 2010.

[22] D. Pletea, B. Vasilescu, and A. Serebrenik. Security and emotion: Sentiment analysis of security discussions on github. In *MSR*, pages 348–351, 2014.

[23] F. Ramsey and D. Schafer. *The Statistical Sleuth*. Duxbury-Thomson Learning, second edition, 2002.

[24] B. Ray, V. Hellendoorn, S. Godhane, Z. Tu, A. Bacchelli, and P. Devanbu. On the "naturalness" of buggy code. In *ICSE*, pages 428–439, 2016.

[25] V. Sinha, A. Lazar, and B. Sahrif. Analyzing developer sentiment in commit logs. In *MSR*, pages 520–523, 2016.

[26] P. Tourani, Y. Jiang, and B. Adams. Monitoring sentiment in open source mailing lists – exploratory study on the apache ecosystem. In *CASCON*, pages 34–44, 2014.

[27] M. Wrobel. Emotions in the software development process. In *HSI*, pages 518–523, 2013.

[28] G. Yang, S. Baek, J. Lee, and B. Lee. Analyzing emotion words to predict severity of software bugs: A case study of open source projects. In *SAC*, pages 1280–1287, 2017.