Towards Understanding and Exploiting Developers’ Emotional Variations in Software Engineering

Md Rakibul Islam
University of New Orleans, USA
Email: mislam3@uno.edu

Minhaz F. Zibran
University of New Orleans, USA
Email: zibran@cs.uno.edu

Abstract—Software development is highly dependent on human efforts and collaborations, which are immensely affected by emotions. This paper presents a quantitative empirical study of the emotional variations in different types of development activities (e.g., bug-fixing tasks) and development periods (i.e., days and times), in addition to in-depth investigation of emotions’ impacts on software artifacts (i.e., commit messages) and exploration of scopes for exploiting emotional variations in software engineering activities. We study emotions in more than 490 thousand commit comments across 50 open-source projects. The findings add to our understanding of the role of emotions in software development, and expose scopes for exploitation of emotional awareness in improved task assignments and collaborations.

I. INTRODUCTION

Emotions are inseparable part of human nature, which influence people’s activities and interactions, and thus emotions affect task quality, productivity, creativity, group rapport and job satisfaction [2], [8], [21]. Software development, being highly dependent on human efforts and interactions, is more susceptible to emotions of the practitioners. Hence, a good understanding of the developers’ emotions and their influencing factors can be exploited for effective collaborations, task assignments [6], and in devising measures to boost up job satisfaction, which, in turn, can result in increased productivity and projects’ success [4].

Several studies have been performed in the past for understanding the role of human aspects on software development and engineering. Some of those earlier studies address when and why employees get affected by emotions [2], [12], [13], [23], [27], whereas some other work address how [10], [14], [15], [20], [28] the emotions impact the employees’ performance at work. Despite those earlier attempts, software engineering practices still lack theories and methodologies for addressing human factors such as, emotions, moods and feelings [11], [13]. Hence, the community calls for research on the role of emotions in software engineering [14], [21], [25].

Some software companies try to capture the developers’ emotional attachments to their jobs by means of traditional approaches such as interviews and surveys [28]. Capturing emotions with the traditional approaches is more challenging for projects relying on geographically distributed team settings and voluntary contributions (e.g., open-source projects) [5], [12]. Thus, to supplement or complement those traditional sources, software artifacts such as the developers’ commit comments/messages have been identified for the extraction of important information including developers’ emotional states [12], [13], [23].

In this work, we study the polarity (i.e., positivity, negativity, and neutrality) of emotions expressed in commit messages as posted by developers contributing to open-source projects. In particular, we address the following four research questions.

RQ1: Do developers express different levels (e.g., high, low) and polarity (i.e., positivity, negativity, and neutrality) of emotions when they commit different types (e.g., bug-fixing, new feature implementation, refactoring, and dealing with energy related concerns) of development tasks?
— If we can distinguish development tasks at which the developers express high negative emotions, low positive emotions, or an overall low emotional involvements, stipulating measures can be introduced to emotionally influence the emotions of the developers working on those particular types of development tasks resulting in higher success rate.

RQ2: Can we distinguish a group of developers who express more emotions (positive or negative) in committing a particular type (e.g., bug-fixing) of tasks?
— Programmers who develop in them positive emotions while carrying out a given development task can be more efficient and quicker in completing the task [20] resulting in reduced software cost. Thus, distinguishing a group of practitioners having positive emotional attachment to a particular task can be useful in effective task assignments.

RQ3: Do the developers’ polarity (i.e., positivity, negativity, and neutrality) of emotions vary in different days of a week and in different times of a day?
— If we can identify any particular days and times when developers express significant negative emotions, then managers can take motivating steps to boost up the developers positive feelings on those days and times. Guzman et al. [12] reported that commit comments posted on Mondays tend to have more negative emotions. We also want to verify their claim using a substantially larger data-set.

RQ4: Do the developers’ emotions have any impact on the lengths of commit comments they write?
— Commit messages are pragmatic means of communication among the developers contributing to the same project. Ideally, commit comments contain important information about the underlying development tasks, and the length of developers’
work description is an indication of the description quality [16]. If any relationship can be found between the developers’ emotional state and the lengths of commit comments, then project managers can take steps to stimulate the developers emotional states to get high quality commit comments containing enough contextual information.

II. METHODOLOGY

To address the aforementioned research questions, we extract emotions from the developers’ commit messages using SentiStrength [26], which is a state-of-the-art sentiment analysis tool. SentiStrength was previously used for similar purposes [9], [13], [27] and was reported to be good candidate for analyzing emotions in commit comments [12].

In the following subsections, we first briefly introduce sentiment analysis with SentiStrength (Section II-A) and then, we describe the metrics (Section II-B), tuning of SentiStrength (Section II-C) for software engineering context, and data collection approaches (Section II-D) used in our study. The procedural steps of our empirical study are summarized in Figure 1.

A. Sentiment Analysis

Sentiment analysis using SentiStrength on a given piece of text (e.g., a commit message) $c$ computes a pair $(\rho_c, \eta_c)$ of integers, where $+1 \leq \rho_c \leq +5$ and $-5 \leq \eta_c \leq -1$. Here, $\rho_c$ and $\eta_c$ 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 $(1, -1)$. Further details about the sentiment analysis algorithm of SentiStrength and the interpretation of its outputs can be found elsewhere [26].

B. Metrics

To carry out our analyses for deriving the answers to the research questions, we formulate the following metrics. Given a set $C$ of commit messages, we can obtain two subsets $C^+$ and $C^-$ defined as follows:

$C^+ = \{ c | c \in C, \rho_c > +1 \}$ and $C^- = \{ c | c \in C, \eta_c < -1 \}$.

Mean Positive Emotional Score for a set $C$ of commit messages, denoted as $P(C)$, is defined as:

$$P(C) = \frac{\sum_{c \in C^+} \rho_c}{|C^+|}$$  \hspace{1cm} (1)

Mean Negative Emotional Score for a set $C$ of commit comments, denoted as $N(C)$, is defined as follows:

$$N(C) = \frac{\sum_{c \in C^-} |\eta_c|}{|C^-|}$$  \hspace{1cm} (2)

Cumulative Emotional Score for a particular commit message $c$, denoted as $T(c)$, is defined as follows,

$$T(c) = \rho'_c + \eta'_c$$  \hspace{1cm} (3)

where,

$$\rho'_c = \begin{cases} \rho_c, & \text{if } \rho_c > +1. \\ 0, & \text{otherwise}. \end{cases}$$

$$\eta'_c = \begin{cases} |\eta_c|, & \text{if } \eta_c < -1. \\ 0, & \text{otherwise}. \end{cases}$$

C. Tuning of SentiStrength

The sentiment analysis tool SentiStrength was reported to have 60.7% precision for positive texts and 64.3% for negative texts [26]. To the best of our knowledge, all such sentiment analysis tools including SentiStrength are highly dependent on the polarities of individual words in a given text in computation of its emotional scores. SentiStrength was originally trained on documents on the social web. In a technical field such as software engineering, commit messages include many keywords which have polarities in terms of dictionary meanings, but do not really express any emotions in their technical context. For example, ‘Super’, ‘Support’, ‘Value’ and ‘Resolve’ are English words with known positive emotions, while ‘Dead’, ‘Block’, ‘Default’, and ‘Garbage’ are known to have negative emotions, but neither of these words really bear any emotions in software development artifacts. Those are simply some domain specific technical terms with especial contextual meanings.

SentiStrength provides the flexibility to modify its existing lexicons’ emotional interpretation to customize it for a target context (i.e., software engineering, in this work). For our purpose, we neutralize SentiStrength’s interpretation of the aforementioned technical jargons, as such was also suggested in earlier studies in the area [23], [27].

Having SentiStrength tuned according to the procedure described above, we manually verify the impact of the tuning using a random sample of 200 commit messages extracted from Boa [7], and we found a 26% increase of precision (checked by comparing SentiStrength’s computation of emotional polarities with subjective human interpretation over each of the 200 commit messages). Thus, for our work, we use this improved instance of SentiStrength tuned for use in software engineering context.

D. Data Collection

We study commit messages for open-source projects obtained through Boa [7]. Boa is a recently introduced infrastructure with a domain specific language and public APIs to facilitate mining software repositories. We use the largest (as of February 2016) data-set from Boa, which is categorized as “full (100%)” and consists of more than 7.8 million projects collected from GitHub before September 2015.

From this large data-set, we select the top 50 projects having the highest number of commits. We study all the commit messages in these projects, which constitute 490,659 commit comments. Associated information such as, committers, commit timestamps, types of underlying work, revisions and project IDs are kept in a local relational database for convenient access and query. For each of the commit messages, we compute the emotional scores using the tuned SentiStrength tool. Table I shows some examples of emotional and neutral commit comments in our dataset and computation of their emotional scores.

III. ANALYSIS AND FINDINGS

The research questions RQ1, RQ2, RQ3 and RQ4 are respectively addressed in Section III-A, Section III-B, Section III-C, and in Section III-D.

A. Emotional Variations in Different Task Types

We investigate whether developers’ emotions vary based on their involvements in four different types of software development tasks: (a) bug-fixing tasks, (b) new feature implementation, (c) refactoring, and (d) energy-aware development. We consider that the first three types of tasks mentioned above are self-explanatory. The fourth one (i.e., energy-aware development) deals with software issues with consumption of energy, measured in terms of usage of resources such as processing power and memory. Energy-aware development is a recent important topic in the area of green computing research. Categorization of development tasks in this manner are also found in earlier studies [1], [3], [22] in software engineering research.

Task-based Characterization of Commits: To distinguish commits dealing with bug-fixing tasks, we rely on Boa’s public APIs, which readily indicate whether a commit message is associated with bug-fixing task, or not.

To identify energy-aware commit messages, we select a list of keywords and search those keywords in commit messages. A commit message will be considered as energy-aware commit, if the commit message contains any of the selected keywords. The identified keywords are: *energy consum*, *energy efficien*, *energy sav*, *save energy*, *power consum*, *power ecien*, *save power*, *energy drain*, *energy leak*, *tail energy*, *power efficien*, *high CPU*, *power aware*, *drain*, *no sleep*, *battery life* and *battery consum*. The character ‘*’ in each keyword works as a wildcard, i.e., a query will select those commits messages, which contain at least one of these keywords, regardless of the beginning or the end of the commit message. Note that, these keywords were also used in earlier studies [3], [17], [19], [22] for similar purposes.

To recognize commit messages dealing with new feature implementation and refactoring tasks, we select those keywords, which were used by Ayalew and Mguniin [1] in their work. Keywords *add* and *new feature* are used to categorize commit messages, which are related to new feature development. And *refactor* and *code clean* keywords are used to distinguish those commit messages, which are posted by developers during code refactoring tasks. Note that, a developer may perform refactoring while fixing a bug. Thus, a commit message can be characterized relevant to more than one categories of tasks.

![Distribution of mean positive, negative, and cumulative emotional scores in commit messages dealing with different types of tasks](image)

**Fig. 2.** Distribution of mean positive, negative, and cumulative emotional scores in commits messages dealing with different types of tasks

Investigation: The numbers of commit messages found relevant to each of the four categories of development tasks are presented in the second column from left in Table II. The boxplot in Figure 2 presents the distribution of mean positive, negative, and cumulative emotional scores in each type of task for each of the 50 projects. An ‘x’ mark in a boxplot indicates the mean emotional scores over all the projects.

As observed in Figure 2, emotional scores (positive, negative and cumulative) for energy-aware commit messages are much higher than those in commit messages for three other tasks, and there is not much variations in the emotional scores.
among these three tasks. To verify the statistical significance of these observations, we conduct Mann-Whitney-Wilcoxon (MWW) tests [24] (with $\alpha = 0.05$) between the distributions of mean cumulative emotional scores in commit messages for each possible pair of development tasks. The results of the MWW tests are presented in Table III. The $P$-values reported by the tests, as compared with $\alpha$, suggest statistical significance of our observations.

Again, looking at Figure 2, we see that the commit messages, which are posted during the new features implementation tasks, show more negative emotions than positive ones. Opposite observations are evident for commit messages for three other types of tasks. To verify the statistical significance of our observations in the variations of polarity (positivity and negativity) of emotions, for each of the four types of development tasks, we separately conduct MWW tests between the mean positive and negative emotional scores of commit messages. The results of the MWW tests are presented in the right-most two columns in Table II. The $P$-values of tests, as compared with $\alpha$, suggest statistical significance of our observations for bug-fixing, new feature implementation and refactoring tasks, but not for the energy-aware development tasks.

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

**Ans. to RQ1:** Developers express significantly high positive and negative emotions almost equally in committing energy-aware tasks. For bug-fixing and refactoring tasks, positive emotions are significantly higher than negative emotions. And surprisingly, for new feature implementation tasks, negative emotions are significantly higher than positive polarity.

### B. Emotional Observations in Bug-Fixing Tasks

It is natural that different developers have different expertise, comfort-zones, and interests with respect to types of tasks. The research question RQ2 addresses the possibility of distinguishing a set of developers who particularly express positive emotions at the particular type of task at hand. In addressing the research question RQ2, we choose the bug-fixing tasks as a representative to any particular type of tasks and continue as such.

Across all the projects, we distinguish 20 developers, who are the authors of the bug-fixing commit messages having the highest positive mean emotional scores. Let $D_p$ denote the set of these 20 developers. Similarly, we form another set $D_n$ consisting of 20 developers, who are the authors of bug-fixing commit comments having the highest negative mean emotional scores. By the union of these two sets, we obtain a set $D$ of 30 developers who are authors of bug-fixing commits with the highest mean positive or negative emotional scores. Mathematically, $D = D_p \cup D_n$.

These 30 developers are the authors of 112,462 commits messages among which 32,088 are bug-fixing commits. For each of these 30 developers, we compute a ratio $R(d)$ as follows:

$$R(d) = \frac{\mathcal{P}(\mathcal{C}_d)}{\mathcal{N}(\mathcal{C}_d)} , \text{ where } d \in D$$

Here, $\mathcal{C}_d$ denotes the set of bug-fixing commit comments posted by developer $d$. Notice that, for a particular developer $d$, the ratio $R(d)$ close to 1.0 indicates that the positive and negative emotions are almost equal for the developer $d$. If $R(d)$ is much higher than 1.0, the developer $d$ shows more positive emotions at bug-fixing tasks compared to negative emotions. The opposite holds when $R(d)$ is much lower than 1.0. However, a threshold scheme seems necessary to determine when the value of $R(d)$ can be considered significantly close to or distant from 1.0.

**Clustering Analysis:** Instead of setting an arbitrary threshold, we apply unsupervised Hierarchical Agglomerative Clustering for partitioning the values of $R(d)$. The dendrogram produced from this clustering is presented in Figure 3. In the dendrogram, we identify three major clusters/groups, two
marked (by us) with dotted rectangles and the third left unmarked in the middle. This middle cluster, denoted as $G_n$, represents the set of those developers, who equally express positive and negative emotions during bug-fixing. We have, $0.992 \leq R(d) \leq 1.0, \forall d \in G_b$.

The set of developers who are included in the right-most cluster exhibit more positive emotions compared to negative emotions during bug-fixing. Let $G_p$ denote the cluster of these developers. Here, $1.005 \leq R(d) \leq 1.178, \forall d \in G_p$. The set of developers who render more negative emotions towards bug-fixing belong to left-most cluster, denoted as $G_n$. We have, $0.919 \leq R(d) \leq 0.982, \forall d \in G_n$.

### Table IV

<table>
<thead>
<tr>
<th>Cluster</th>
<th>$G_p$</th>
<th>$G_n$</th>
<th>$G_b$</th>
<th>$P$-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant?</td>
<td>Yes ($P &lt; \alpha$)</td>
<td>Yes ($P &lt; \alpha$)</td>
<td>No ($P \geq \alpha$)</td>
<td>0.00798</td>
</tr>
</tbody>
</table>

### Statistical Significance

For each of the three clusters, we separately conduct $MWW$ tests between the mean positive and negative emotional scores of the commit messages to verify the statistical significances of their differences. The results of the separate $MWW$ tests (with $\alpha = 0.05$) over each of the clusters are presented in Table IV. The $P$-values in Table IV indicate statistical significance in the differences in positive and negative emotions for clusters $G_p$ and $G_n$. As expected, no such significant difference found for the cluster $G_b$ as in this cluster, positive and negative emotions are expressed equally. Thus, our clustering of the developers appears to be accurate with statistical significance. Hence, we answer the research question $RQ2$ as follows:

**Ans. to $RQ2$:** We have been able to distinguish sets of developers who show either high positive or high negative emotions in bug-fixing commit messages while some other developers are found to express both positive and negative emotions almost equally. The same approach can be applied to distinguish such groups of developers for other types of development tasks.

### C. Emotional Variations in Days and Times

For each of the projects, we group all the commit messages into seven disjoint sets in accordance with the days of the week those are committed.

Figure 4 plots the average (over each project) positive, negative, and cumulative emotional scores in commit messages posted in different days of a week. Among all the seven days of a week, negative emotions appear to be slightly higher in commit messages posted during the weekends (i.e., Saturday and Sunday) than those posted in weekdays (i.e., Monday through Friday). Not much differences are visible in the emotional scores for commit messages posted in the five weekdays. $MWW$ tests (with $\alpha = 0.05$) between the distributions of emotional scores in each possible pair of the days of a week suggest no statistical significance in the differences of emotions. $P$-values of the $MWW$ tests are presented in Table V. As can be seen in Table V, for all values of $P$’s, $\alpha < 0.11 \leq P$.

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 again organize the commit messages into three disjoint sets based on their timestamps of posting.

Figure 5 presents the mean (over each project) positive and negative emotional scores (computed using Equation 1 and Equation 2) in commit messages posted in these three periods. Again, in Figure 5, we do not see much variations in the emotional scores of commit messages posted at different periods. $MWW$ tests (with $\alpha = 0.05$) between the distributions of mean positive and negative emotional scores in each possible pair of the periods indicate no statistical significance in their differences. $P$-values of the $MWW$ tests are presented in Table VI. Hence, we derive the answer to the research question $RQ3$ as follows:

**Ans. to $RQ3$:** There is no significant variations in the developers’ emotions in different times and days of a week.
verify the significance of the ship between emotional scores and commit lengths. Next, we also strongly indicates statistical significance of the relation-
scores. A the four groups and having different cumulative emotional
comments in groups of different lengths

Fig. 6. Distribution of mean cumulative emotional scores of commit
messages in groups of different lengths

<table>
<thead>
<tr>
<th>Groups</th>
<th>01-10</th>
<th>11-20</th>
<th>21-40</th>
<th>41-50</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Commits Comments with ( n ) (c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>46,486</td>
<td>32,144</td>
<td>27,352</td>
<td>4,255</td>
</tr>
<tr>
<td>03</td>
<td>2,734</td>
<td>5,967</td>
<td>5,008</td>
<td>409</td>
</tr>
<tr>
<td>04</td>
<td>2,558</td>
<td>4,627</td>
<td>5,008</td>
<td>1,275</td>
</tr>
<tr>
<td>05</td>
<td>245</td>
<td>876</td>
<td>1,555</td>
<td>399</td>
</tr>
<tr>
<td>06</td>
<td>28</td>
<td>145</td>
<td>203</td>
<td>84</td>
</tr>
<tr>
<td>07</td>
<td>12</td>
<td>16</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>08</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

TABLE VII

<table>
<thead>
<tr>
<th>Commit</th>
<th># of Commits Comments with ( n ) (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>02</td>
</tr>
<tr>
<td>01-10</td>
<td>46,486</td>
</tr>
<tr>
<td>11-20</td>
<td>32,144</td>
</tr>
<tr>
<td>21-40</td>
<td>27,352</td>
</tr>
<tr>
<td>41-50</td>
<td>4,255</td>
</tr>
</tbody>
</table>

D. Emotional Impacts on Commit Lengths

To investigate the existence of any relationship between emotions and lengths of commit messages, across all the 50
projects, we distinguish 141,033 commit comments, which are one to 50 words in length having cumulative emotional
scores (computed using Equation 3) higher than one. For each project, we organize these emotional commit messages
into four disjoint groups based on their lengths as shown in
Figure 6, which plots the mean (over each project) cumulative emotional scores of commit messages in the four groups. As
seen in the figure, the emotional scores are strictly higher for
the groups with lengthier commit messages.

Fig. 6. Distribution of mean cumulative emotional scores of commit
comments in groups of different lengths

Table VII presents the frequencies of commit messages in
the four groups and having different cumulative emotional
scores. A Chi-squared [24] test \( (P = 2.2 \times 10^{-16}, \alpha = 0.05) \)
also strongly indicates statistical significance of the relation-
ship between emotional scores and commit lengths. Next, we
verify the significance of the direction of relationship (i.e., if
one increases or decreases with the increase of another).

Fitting of a Generalized Linear Model [24] on the emotional
score and length of every emotional commit message confirms
(with \( \beta = +0.01134, P = 2 \times 10^{-16}, \alpha = 0.001 \)) the positive
correlation between emotional scores and commit lengths.
Based on the analyses, we now derive the answer to the
research question RQ4 as follows:

Ans. to RQ4: Developers’ emotions have statistically sig-
nificant impacts on the lengths of commit messages they
write. Developers post longer commit comments when they
are emotionally active.

IV. Threats to Validity

In this section, we discuss the limitations of our work, the
threats to the validity of our findings, and our attempts to
minimize those threats.

Internal Validity: The internal validity of our work depends
on the accuracy of the tool’s computation of emotional scores. SentiStrength was reported to be effective in sentiment
analysis [26] and suitable for extraction of emotions from commit comments [12]. SentiStrength has relatively
high accuracy compared to other tools of its kind and thus
SentiStrength was used for sentiment analysis in earlier
work in software engineering research [3], [9], [12], [13],
[27]. Moreover, for use in our work, we increased it accuracy
in emotion extraction by 26% through tuning the tool for
application in software engineering context (Section II-C).

Nevertheless, the tuned tool is not 100% accurate in deter-
mining emotional polarities of commit messages, and it was
not possible to perform manual sanity check by going through
each of the 490,659 commit messages included in our work.
We are aware of this threat, although we minimized it by
contextual customization of SentiStrength.

Construct Validity: The choice of the 30 developers in
examining the relationships between emotions and bug-fixing
tasks (Section III-B) can be questioned. Note that, these
30 developers are the authors of more than 112 thousand
commit messages (22.85%), which is a large sample of data
for dependable analysis. The objective was to check if it
was possible to distinguish a group of developers who are
emotionally more active towards a particular type of task. If
we chose a fair number of developers other than our choice
of 30, we would still be able to distinguish a set of target
developers. In that case, the size of the set of developers might
be different from what we found using the 30 developers, but
this does not invalidate the findings of the work.

One may also question the validity of our categorization
of the developers’ commits in different days and periods
pants might be uncomfortable in disclosing their negative emo-
Based approaches suffer from the possibility that the partici-
ments and design performance. Similarly, self-
developers’ positive emotions. The participants self-assessed
watching those movie clips would induce different levels of
did not employ any measurement to extract and quantify the
project before and after regular working hours. However, they
have more negative emotions compared to Sunday,
museum and thus are unlikely to reflect his or her emotions.

For the statistical tests of significance in the variations of
different distributions, we used the Mann-Whitney-Wilcoxon
MWW test [24]. The MWW test is a non-parametric test,
this particular test suits well for our purpose. Moreover,
which for this parameter that enables 95% confidence in the
Since the data in our work do not conform to normal distribu-
This large data-set yields high confidence on the generalizability of the results.

External Validity: The findings of this work are based on
on our study on more than 490 thousand commit messages
who identified developers’ emotions from 60,658 commits and 54,892
We also report that commit comments written on Mondays
But none of these work tuned the tool before application in

Guzman and Bruegge [13] identified emotions in collabora-
to relate them with different development topics. In a separate study, using SentiStrength, Guzman
expressed in 60,425 commit messages and reported that commit comments written on Mondays
tend to have more negative emotions compared to Sunday, Tuesday, and Wednesday. However, from the investigation
of the same phenomenon using a substantially larger dataset of 490,659 commit messages, our study does not identify
any statistically significant variations of emotions in commit comments posted in different days of a week.

Using a Natural Language Toolkit (NLTK), Pletea et al. [23]
minted developers’ emotions from 60,658 commits and 54,892
pull requests for GitHub projects. They analyzed emotional
mined developers’ emotions from 60,658 commits and 54,892
pull requests for GitHub projects. They analyzed emotional
variations in discussions on different topics and reported
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 and larger analysis based on a larger
number of commit messages and diverse aspects of emotional implications.

Using SentiStrength, Tourani et al. [27] extracted emotions from emails of both developers and system users.
They observed the differences of emotional expressions be-
tween developers and users of a system. Using the same
tool, Garcia et al. [9] extracted developers’ emotions from
their email contents to analyze any relationships between
developers’ emotions and their activities in an open source software projects. Although the studies of Tourani et al. [27] and Garcia et al. [9] also used the same sentiment analysis tool we used, the source of their emotional content are different and the objectives of those work are also orthogonal to ours.

VI. CONCLUSION

In this paper, we have presented a quantitative empirical study on the characteristics and impacts of emotions extracted from developers’ commit messages. We have studied more than 490 thousand commit comments over 50 open-source projects. Although the majority (65%) of the commit messages are found to be neutral in emotion, surprisingly, positive emotions are found in relatively much smaller portion (13%) of the commit comments than the commits (22%) containing negative emotions.

In our study, we found that the polarities of the developers’ emotions significantly vary depending on the type of tasks they are engaged in. The developers express equally high positive and negative emotions in committing in energy-aware tasks compared to other tasks. With respect to the polarities of commit messages, positive emotions are found to be significantly higher than negative emotions in commits for bug-fixing and refactoring tasks. Surprisingly, the opposite scenario is found for new feature implementation tasks.

We also found significant positive correlation between the lengths of commit messages and the emotions expressed in them. When the developers remain emotionally active, they tend to write longer commit comments. However, we did not find any significant variations in the developers’ emotions in commit messages posted in different times and days of a week.

Based on emotional contents in commit messages, we have also been able to distinguish a group of developers who express more positive emotions at bug-fixing commit messages, another group with the opposite trait, and a third group of developers who equally render both positive and negative emotions at bug-fixing activities. Same approach can be applied for other types of tasks to distinguish potential developers for improved tasks assignment.

The findings from this work are validated in the light of statistical significance. Although more experiments can be conducted to verify or confirm the findings, the results from this study significantly advance our understanding of the impacts of emotions in software development activities and artifacts, and we exemplify how emotional awareness can be exploited in improving software engineering activities.

For automatic computation of emotional polarities in commit messages, we have used a state-of-the-art tool, SentiStrength, while alternatives exist. Moreover, before applying the tool, we tuned it for our work in the context of software engineering. In future, we plan to replicate this study using other tools and subjects to further validate the findings of this study. We also have plan to conduct more studies on the impacts of emotions extracted from diverse artifacts including program comments, development forums and email groups.

Acknowledgement: Thanks to Rahul Chatterjee for helping in the statistical analyses.

REFERENCES