SUBMITTED TO EMPIRICAL SOFTWARE ENGINEERING
(— version January 10, 2018)

An Exploratory Qualitative and Quantitative Analysis of Emotions in Issue Report Comments of Open Source Systems

Alessandro Murgia · Marco Ortu · Parastou Tourani · Bram Adams · Serge Demeyer

Last modified on January 10, 2018

Abstract Software development —just like any other human collaboration— inevitably evokes emotions like joy or sadness, which are known to affect the group dynamics within a team. Today, little is known about those individual emotions and whether they can be discerned at all in the development artifacts produced during a project. This paper analyzes (a) whether issue reports —a common development artifact, rich in content— convey emotional information and (b) whether humans agree on the presence of these emotions. From the analysis of the issue comments of 117 projects of the Apache Software Foundation, we find that developers express emotions (in particular gratitude, joy and sadness). However, the more context is provided about an issue report, the more human raters start to doubt and nuance their interpretation. Based on these results, we demonstrate the feasibility of a machine learning classifier for identifying issue comments containing gratitude, joy and sadness. Such a classifier, using emotion-driving words and technical terms, obtains a good precision and recall for identifying the emotion love, while for joy and sadness a lower recall is obtained.

This work was sponsored by (1) the Institute for the Promotion of Innovation through Science and Technology in Flanders by means of a project entitled Change-centric Quality Assurance (CHAQ) with number 120028, and (2) the Regione Autonoma della Sardegna (RAS), Regional Law No. 7-2007, project CRP-17938, "LEAN 2.0"

Alessandro Murgia
University of Antwerp, Belgium — E-mail: alessandro.murgia@uantwerpen.be
Present address: Middelheimlaan 1; 2020 Antwerpen; Belgium

Marco Ortu
University of Cagliari, Italy — E-mail: marco.ortu@diee.unica.it

Parastou Tourani · Bram Adams
MCIS, Polytechnique Montréal, Canada — E-mail: {parastou.tourani,bram.adams}@polymtl.ca

Serge Demeyer
University of Antwerp, Belgium — E-mail: serge.demeyer@uantwerpen.be

Keywords Emotion Mining · Issue Report · Text Analysis · Parrott's Framework

1 Introduction

> *The major problems of our work are not so much technological as sociological in nature.* Tom De Marco (DeMarco and Lister 1999)

In July 2013, the Linux kernel mailing list was shaken up by an agitated discussion between Linus Torvalds and a senior developer (Brodkin 2013): "I am serious about this. Linus, you're one of the worst offenders when it comes to verbally abusing people and publicly tearing their emotions apart." Other people joined her, noting "scolding people [...] is not likely to encourage people to want to become senior developers" and "Thanks for standing up for politeness/respect. If it works, I'll start doing Linux kernel dev. It's been too scary for years." On the other hand, Linus Torvalds defended himself, claiming "not telling people clearly enough that I don't like their approach, they go on to re-architect something, and get really upset when I am then not willing to take their work." The senior developer recently left the kernel project Gold (2015) because "she could no longer work within a developer culture that required overworked maintainers to be rude and brusque in order to get the job done."

This anecdote illustrates a case where developers are underperforming or participating less actively than they usually would, because they do not feel happy. In other words, a purely rational view of software development and its stakeholders that does not consider people's emotions and feelings, only provides a partial explanation of software development productivity. For example, positive emotions like happiness help people to be more creative (Fredrickson 2001), which is essential for successful software design (Brooks 1987). If not, fear, or absence of courage, could refrain developers from changing or refactoring their code (Ambler 2002). These effects of emotions are similar to other domains, where people have found that feelings and emotions dictate to a large extent our actions and decisions (Plutchik and Van Praag 1989). For example, consumer opinions on retailer sites are highly influential for buyer decisions (Piller 1999). The mood of people, evaluated through tweets, correlates with changes in stock market activity (Bollen et al 2011).

Since even the most talented developer could underperform and eventually leave the project just because she is unhappy with her environment or colleagues, it is important to support managers and project leads in detecting emotions (especially negative ones) in their team, after which traditional project management activities could be used to defuse the situation. Especially in today's globally distributed development (both open and closed source) or with the increasing percentage of employees working from home, projects have reduced the personal interaction except for limited conference calls. In such

environments, gauging emotions across geographical locations is hard. If managers would be aware of problems, they could swiftly take proactive actions to limit their impact. Leveraging emotion awareness within the software team could enhance software development and quality, mood regulation within a project team, and improve interactions with all stakeholders (semotion 2016).

When face-to-face meetings are not feasible or efficient to quickly grasp emerging emotions in scattered teams, mining emotions from discussion boards like issue reports, mailing list communication and fora are the only viable alternative. Such boards are used on a daily basis by all development stakeholders to coordinate and manage development activities, and feature extensive logs of developer communication. Such textual logs could be exploited to automatically identify occurrences of emotions, using off-the-shelf or custom analysis. Even though such analyses would not be perfect (as it involves natural language processing), they at least could provide an indication that something is wrong.

As a first step towards mining developer emotions from project discussion boards, we performed a pilot study (involving four of the authors) and a full user study (with sixteen participants involving master students, PhD students and post-docs) to determine whether emotions can actually be detected from issue comments —comments attached to issue reports, and, if so, whether humans can agree on the emotions identified. For this initial exploration, we restricted ourselves to issue reports, since, in our previous work, we found that issue report comments are a rich source of information carrying the full range of emotions (Murgia et al 2014). We also investigated how much information about earlier comments of an issue (context) humans need for identifying emotions, analyzing a significant sample of 792 developer comments (400 in the pilot study, 392 in the full study) of the Apache projects using Parrott's emotional framework (Parrott 2001). Based on this knowledge, we built proof-of-concept machine-learning classifiers to classify issue comments according to the three emotions where human raters agree the most: gratitude ("love"), joy and sadness. To evaluate their performance, we manually analyzed an additional 1,600 developer comments labeled by the tool as containing gratitude, joy, sadness or being emotion-free. On top of that, we investigate which keywords are the best indicators of emotions in issue comments. As such, we address the following research questions:

RQ1 — Rater's Agreement. *Can human raters agree on the presence or absence of emotions in issue comments?*

> We found that raters agree the most on the absence of an emotion, followed by the presence of `love` (i.e., gratitude) and (less strongly) `sadness` and `joy`. Involving additional raters and using majority voting does not significantly improve agreement.

RQ2 — Context Influence. *Does context improve the agreement of human raters on the presence of emotions in issue comments?*

We found that context does not play a significant role in the rating of emotions in issue comments, but when it does, it seems to cast more doubt than confidence (i.e., nuances) in the identified emotions.

RQ3 — Classifier Feasibility. *To what extent can a machine learning classifier identify emotions in issue comments?*

We found that a machine learning classifier can achieve a good precision when identifying comments with emotions love, joy and sadness. However, only for the love emotion the classifier exhibits a high resilience to false negatives, i.e., a high precision.

RQ4 — Important Keywords. *What keywords does a machine learning classifier rely upon to identify emotions?*

We found that the emotions love, joy and sadness are conveyed through specific keywords like "thanks" and "sorry".

This paper extends our earlier work (Murgia et al 2014), which was the first feasibility study of emotion mining in development artifacts like issue reports, in the following ways:

- We extended the number of issue comments in the dataset. The authors manually analyzed 1,600 additional comments.
- We detail the design of a model, based on machine learning classifiers, to identify issue comments exhibiting the emotions love, joy or sadness.
- We build and empirically evaluate this model on the 1,600 issue comments.
- We analyze which keywords are relevant for the identification of emotions in issue comments.
- We perform a more detailed discussion of related work.

Based on our findings, issue comments have potential as data source and it is possible to build a reasonable emotion classifier. However more work is needed to fully understand the role of context on the identification of emotions, and to improve classifier performance.

In the remainder of this paper, we first describe the background notions for emotion mining (Section 2). Next, we show up front some of the emotions that we identified in issue comments (Section 3). The bulk of the paper starts by presenting the case study design (Section 4), followed by qualitative and quantitative answers to the research questions (Section 5) and a discussion of our findings (Section 6). After a discussion of the threats to validity (Section 7) and the related work (Section 8), we finish with conclusions (Section 9).

2 Background

This section provides background about emotion mining, the Parrott emotional framework, and the development artifacts (i.e., issue reports) studied in this paper.

Table 1 Parrott's emotion framework.

Primary emotions	Secondary emotions	Tertiary emotions
Love	Affection	Compassion, Sentimentality, Liking, Caring, ...
	Lust/Sexual desire	Desire, Passion, Infatuation
	Longing	
Joy	Cheerfulness	Amusement, Enjoyment, Happiness, Satisfaction, ...
	Zest	Enthusiasm, Zeal, Excitement, Thrill,Exhilaration
	Contentment	Pleasure
	Optimism	Eagerness, Hope
	Pride	Triumph
	Enthrallment	Enthrallment, Rapture
Surprise	Surprise	Amazement, Astonishment
Anger	Irritability	Aggravation, Agitation, Annoyance, Grumpy, ...
	Exasperation	Frustration
	Rage	Outrage, Fury, Hostility, Bitter, Hatred, Dislike, ...
	Disgust	Revulsion, Contempt, Loathing
	Envy	Jealousy
	Torment	Torment
Sadness	Suffering	Agony, Anguish, Hurt
	Sadness	Depression, Despair, Unhappy, Grief, Melancholy, ...
	Disappointment	Dismay, Displeasure
	Shame	Guilt, Regret, Remorse
	Neglect	Embarrassment, Humiliation, Insecurity, Insult, ...
	Sympathy	Pity, Sympathy
Fear	Horror	Alarm, Shock, Fright, Horror, Panic, Hysteria, ...
	Nervousness	Suspense, Uneasiness, Worry, Distress, Dread, ...

2.1 Parrott's Framework

Emotion is a "psychological state that arises spontaneously rather than through conscious effort and is sometimes accompanied by physiological changes" (Heritage Dictionary 2005). General types of emotions are `joy`, `sadness`, `anger`, `surprise`, `hate` and `fear`. However, many other categories and sub-categories can be identified. Since there is not one standard emotion word hierarchy, many studies in the cognitive psychology domain have focused on emotions, resulting in various proposals for categorizing emotions (Shivhare and Khethawat 2012; Robinson 2004; Plutchik 2001; Parrott 2001).

One of the most recent classifications of emotions is Parrott's framework (Parrott 2001), which classifies human emotions into a tree structure with 3 levels, as is shown in Table 1. Each level refines the granularity of the previous level, making abstract emotions more concrete. For example, level-1 of this classification consists of six primary-emotions, i.e., `love`, `sadness`, `anger`, `joy`, `surprise` and `fear`. By selectively including or excluding the second and third level for certain emotions, the tree structure allows to zoom in or out of emotions to a desired level of detail. A sentence like "sorry for the delay" can then be classified as `guilt` (level 3) or `shame` (level 2) or ultimately `sadness` (level 1).

Section 3 provides detailed illustrations of each of the primary emotions in terms of emotions expressed by developers during software development. The

concise and intuitive nature of the primary emotions makes Parrott's classification easy to understand by different stakeholders. Indeed, the classification is not just aimed at the people rating a particular artifact as describing a particular emotion, but also appeals to people like team leads trying to benefit from the emotional classification to understand the emotions of their team members. Although this paper only considers the six primary emotions, future work can extend our results to the secondary and tertiary emotions of the most prevalent primary emotions.

2.2 Emotion Mining

Emotion mining tries to identify the presence of human emotions from textual, voice and video artifacts produced by humans. As such, it is different from sentiment analysis, which instead evaluates a given emotion as being positive or negative (Pang and Lee 2008). Ideally, sentiment and emotion analysis should be combined, since this provides more detailed insight into the behavior of people. Since emotion and sentiment analysis affect the decision-making process of companies (Pang and Lee 2008), a diverse range of actors, from marketing departments and investors to politicians are in need of techniques to mine and analyze emotions and sentiments.

In software engineering, emotion mining applied to textual development artifacts could be used to provide hints on factors responsible for `joy` and satisfaction amongst developers (e.g., new release), or `fear` and `anger` (e.g., deadline or a recurring bug). Development artifacts like mailing lists or the discussion board of an issue tracking system could be a promising source for mining developer emotions during software evolution, especially since several studies show that it is possible to "contract" emotions from others through computer-mediated communication systems (Guillory et al 2011; Hancock et al 2008).

Finally, emotion mining can also give a different perspective on increasing productivity and job satisfaction. Emotion mining, by offering new ways to measure emotional states of developers, can be exploited to better understand developers' activities and interactions, and ultimately identify obstacles that hinder their productivity. The latter is a fundamental problem, given the shortage/lack of developers to fulfill market demands (Fritz and Müller 2016).

2.3 Issue Tracking System

An issue tracking system is a repository used by software organizations to coordinate software maintenance and evolution. Such repositories —Jira [`https://www.atlassian.com/software/jira`] being a prime example— provide a shared environment where team members can submit and discuss issues (e.g., bugs and feature requests), ask for advice and share opinions useful for maintenance activities or design decisions. These discussions reveal a team member's

view on a bug, feature, project or even other members of the community. As such, they are a rich source of information to study group dynamics, like implementation and technical topics, project status or even social interactions during software development (Guzzi et al 2013). Issue tracking systems are a rich and diverse source for mining emotions, as was confirmed by our earlier work (Murgia et al 2014).

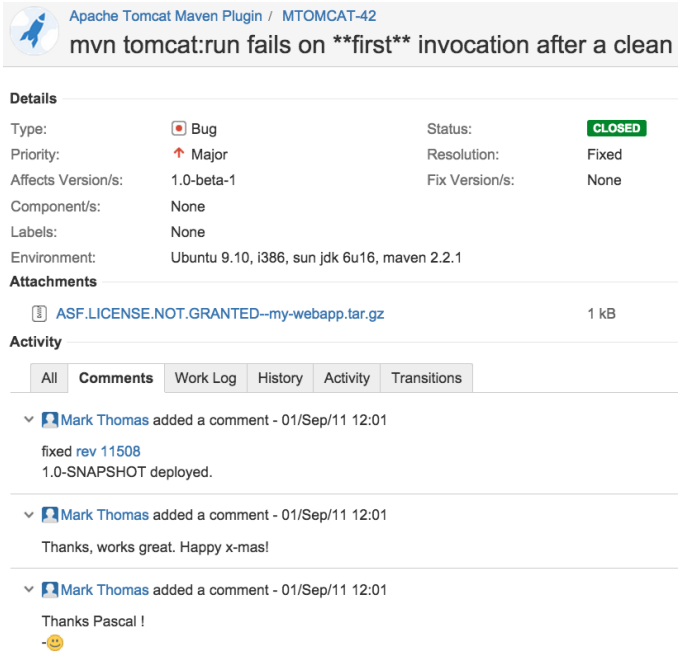

Fig. 1 Example of issue report in Tomcat

An issue report (see for example Figure 1) is characterized by standard fields useful for resolving the issue, such as its priority, status and a list of comments used by developers to discuss and share ideas about the issue resolution. These issue comments are the ones analyzed in our study.

3 Developer Emotions Identified from Issue Comments

This section presents the emotions that we identified in issue comments during the pilot user study. Here, for the sake of clarity, we report only short representative sentences. However, the analysis that we perform targets issue comments that may have multiple sentences. We opted to discuss these emotions up front in order to provide a better understanding of the emotional content of issue comments. For each of the six primary Parrott emotions, we

report the most representative issue discussion snippets as well as an explanation of why the snippet contains that emotion. Whenever appropriate, we also report the secondary (e.g., `shame`) or tertiary (e.g., `guilt`) emotion.

Love

1. Thanks for your input! You're, like, awesome
2. Thanks very much! I appreciate your efforts
3. I would love to try out a patch for [...]

`Love` is generally presented in sentences that express gratitude (i.e., exhibiting a `liking`), toward a person (example 1) or person's activity (example 2), which is a tertiary emotion associated to `Love`. `Love` can also be expressed via its associated secondary emotion `desire` (example 3). In issue comments, `love` primarily is oriented towards co-workers.

Joy

1. I'm happy with the approach and the code looks good
2. great work you guys!
3. Hope this will help in identifying more usecases

`Joy` is normally associated to positive achievements, in the form of `satisfaction` (example 1) or `enthusiasm` (example 2). In the first case, the text reports keywords like "good" or "great". In the second case, the phrase ends with a "!". A less common way to manifest `joy` is via a positive outlook for a successful achievement, namely the developer expresses `optimism`, a secondary emotion associated to `joy` (example 3). `Joy` is expressed either towards software artifacts or co-workers.

Surprise

1. I still question the default, which can lead to surprisingly huge memory use
2. I also documented an unexpected feature with the SlingServletResolver
3. Oops. It needs to be added to Makefile

`Surprise` is expressed for unexpected, generally negative, behavior of a software system (examples 1 and 2). A second case is represented by mistakes introduced accidentally by a developer and discovered later on (example 3). We did not discover any case where `surprise` referred to co-workers.

Anger

1. I will come over to your work and slap you
2. WTF, a package refactoring and class renaming in a patch?
3. This is an - ugly - workaround

`Anger` generally goes along with menaces (e.g., "slap" or "kill"), negative adjectives (e.g., "ugly") or profanity (e.g., "WTF"). These emotions reveal `hostility` and bullying towards co-workers (example 1) or `dislike` towards software artifacts (examples 2 and 3).

Sadness

1. Sorry for the delay Stephen.
2. Sorry of course printStackTrace() wont work
3. wish i had pay more attention in my english class .... now its pay back time .... :-(
4. Apache Harmony is no longer releasing. No need to fix this, as sad as it is.

`Sadness` is generally expressed by developers that feel `guilty`, i.e., they apologize for a delay (example 1) or for the unsatisfactory code produced (example 2). `Sadness` can be expressed also for reasons not dependent on the issue handled (example 3), or outside a developer's influence (example 4).

Fear

1. I'm worried about some subtle differences between char and Character
2. I'm most concerned with some of the timeouts
3. I suspect that remove won't work either in this case.

`Fear` is expressed by a developer in a state of `worry` or `anxiety`. This emotion is expressed explicitly using the keyword "worry" or its synonyms like "concern" (examples 1 and 2). Another common case is to express a negative outlook with respect to a particular development choice (example 3). Like `surprise`, we did not discover any case where `fear` referred to co-workers.

Issue comments are sometimes rich and diverse in emotional content. When they do, these emotions pertain to software artifacts and co-workers (e.g., `joy`, `anger` and `sadness`), while others target only software artifacts (e.g., `surprise` and `fear`) or co-workers (e.g., `love`).

4 Case Study Design

The exploratory case study described in this paper investigates whether issue comments convey emotional information and whether humans agree on the presence of these emotions. Assuming that this is indeed the case, we investigate to what extent a machine learning based classifier can automatically identify issue comments containing gratitude, joy and sadness.

For these analyses, we followed the case study design depicted in Figure 2. We first conducted a pilot study with a sample of issue comments (named

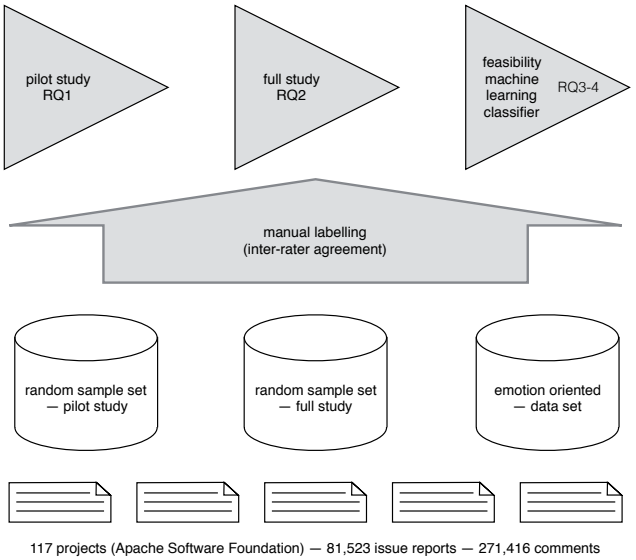

Fig. 2 Overview of the case study design.

the "random sample set — pilot study") collected from the Apache Software foundation, and used this to address RQ1 — Rater's Agreement. Then, we conducted a full study with an extended set of issue reports (named the "random sample set — full study") where we included more context in the issue comments to address RQ2 — Context Influence. This was necessary to verify how raters label the same comment with and without context. The first four authors are involved in both the pilot and full study, for this reason the comments used in both studies do not overlap.

We explain the creation of these data sets in Section 4.2 and proceed with an explanation of the particular set-up used to pair raters involved in the pilot study (Section 4.2.1) and the full study (Section 4.2.2). To address RQ3 — Classifier Feasibility and RQ4 — Important Keywords, we extended the set of issue comments even further (resulting in the "emotion oriented — data set") as explained in Section 4.3. For replication purposes, we also discuss in detail the construction of the machine learning classifiers in Section 4.4.

4.1 Subject Systems

Issue repositories are known to contain a significant amount of information related to a project's development process (Guzzi et al 2013). Therefore, this paper chooses the issue repository as the primary subject of analysis. More concretely, we mined the issue repository of 117 open source projects of the Apache foundation software[`https://issues.apache.org/jira/secure/Dashboard.jspa`]. These systems, randomly selected among the 525 Apache's projects, range from large, long-lived projects such as Tomcat [`http://tomcat.`

apache.org], to smaller projects like RAT [http://creadur.apache.org/rat]. Given this variety of projects, it is reasonable to think that our results would be at least relevant also for issue reports in other open source contexts, other than the projects that we analyzed.

We parsed Apache's Jira-based repository in July 2013, fetching all the issue reports since the 19th of October 2000. For each issue report, we extracted the developers' comments, as well as the standard issue report fields mentioned in Section 2.3. Table 2 reports the statistics of our data set. Since an issue report can have multiple issue comments, and an issue comment multiple sentences (median value of 2.18), we decided to perform our analyses at the level of issue comments. This avoids a too coarse granularity, and yields more than 270,000 comments that can be analyzed. Note that all these comments come from publicly available communication data, hence can be studied without privacy concerns.

[...] I'm not so convinced that moving all the static methods out is useful	(Fear)
How is a bunch of static methods on a utility class easier than a bunch of static methods within the HtmlCalendarRenderer better?	(Anger)
[...] the risk of introducing new bugs for no great benefit	(Fear)
Specific feedback regarding this specific patch: (1) There is significant binary incompatibility	(Neutral)
[...] Previously almost all these helper methods were private; this patch makes them all public [...]	(Neutral)

Fig. 3 Example of issue comments with identified emotions for each sentence

Table 2 Subject System Statistics

Apache Software foundation	
Projects:	117
Issues:	81,523
Comments:	271,416
Users:	20,537
Start Date:	10/2000
End Date:	07/2013

Figure 3 shows an example comment belonging to issue #1235 of the Tomahawk project, where a developer reveals his opinions about the risk of moving towards static methods (which he believes would be useless). To show his dislike, he uses wordings associated with anger and fear, interspersed with neutral phrases where he expresses an more impartial evaluation of the patch. Since we have no ground truth (i.e., knowledge about the actual emotions a reporter had while writing a given comment), we used external raters to assess the presence or absence of a given emotion, as explained below.

4.2 Random Sample Set — Pilot Study and Full Study

As a first step towards verifying whether humans agree on the presence or absence of emotions in issue comments (i.e., RQ1 — Rater's Agreement), we set out for a pilot study involving the first four authors of this paper as raters. For this pilot study, we created an initial random sample of 400 comments

(out of 271,416 comments) of the Apache issue reports, which we refer to as **random sample data set – pilot study** in the remainder of the paper. Given the large number of issue comments, we sampled enough issue comments to obtain at least a confidence level of 95% and confidence interval (error) of 5%. This means that a proportion of X% in our sample of issue comments manually rated as exhibiting a certain emotion, actually corresponds to $X \pm 5$% in the whole population of issue comments exhibiting that emotion. To achieve a confidence interval of 5% for a confidence level of 95%, the size of the random sample should contain a minimum of 384 issue comments. However, to make assignment of comments to raters more straightforward, ensuring that each comment is considered by two or more raters (depending on the research questions), we used as sample sizes 392 (full study) and 400 (pilot study) in order to reduce the rater bias[1].

Since the comments are randomly sampled across all projects' comments, the more issue comments a project has, the higher the probability that we analyzed some of its comments in the sample, i.e., large projects are represented more in the data set. Conversely, the smaller the project, the lower the chance that the analyzed sample has comments of such project, if any. This decision was taken since we valued sample representativeness more than diversity (Nagappan et al 2013).

Once the pilot study confirmed that humans agree on the presence or absence of emotions in issue comments, we conducted a more extensive study involving extra raters (four authors of this paper plus master students and PhD students working in the respective labs) and comments providing extra context (i.e., earlier comments of a given issue) that might influence the impression of emotions. Similar to the pilot study, we created a random sample data set containing 392 comments without any overlap with the random sample data set – pilot study. Just as with the pilot study, this sample provides a confidence level of 95% and confidence interval of 5%. Contrary to the pilot study, we only analyzed the closing comments of issue reports, since those have a higher chance of having context.

To select these extra 392 comments, we restricted ourselves to issue reports having more than one comment and then selected the closing comments, since those have a higher chance of providing sufficient context. In the remainder of this paper, we refer to the resulting data set as **random sample data set – full study**. All data[2] has been made public by the authors (Ortu et al 2016b).

4.2.1 Pairing Raters — Pilot Study

[1] In the pilot study, 4 raters were permuted to label 400 comments —200 comments per rater (cf. section 4.2.1), while in the full study, 16 raters were permuted to label 392 comments —98 comment per rater (cf. section 4.2.2). In all studies, each rater paired up with each other rater the same number of times.

[2] Data set can be downloaded for replication purposes at a web-site hosted by the University of Antwerp: `http://ansymore.uantwerpen.be/system/files/uploads/artefacts/alessandro/MSR16/archive3.zip`.

During the pilot study, we arbitrarily assigned each of the 400 comments in the "random sample data set – pilot study" to the raters, randomly making sure that raters are permuted during the labeling. Eventually, each of the four raters received a file containing 200 issue comments, then went through his or her list of comments to mark all Parrot emotions that he or she was able to identify.

Table 3 Interpretation of Cohen and Fleiss κ values.

κ value	interpretation
<0	poor
0–0.20	slight
0.21–0.40	fair
0.41–0.60	moderate
0.61–0.80	substantial
0.81–1.0	almost perfect

To measure the degree of inter-rater agreement on identified emotions, we calculate either Cohen's κ value (Cohen 1960) (two raters) or Fleiss' κ value (Fleiss 1971) (more than two raters). Both values can be interpreted according to Table 3. In order to determine whether inter-rater agreement values differ statistically significantly, we also provide the values' corresponding confidence interval (with α value of 0.05). If this interval does not overlap with another value's interval, we can reject the null hypothesis and conclude that the two agreement values are significantly different. In addition to these statistical agreement values, we also provide the more basic percentage of cases for which raters agree on a particular emotion or set of emotions.

4.2.2 Pairing Raters — Full Study

Table 4 Comments assigned to person 1 of Group A (P1A) and person 1 of Group B (P1B). Their assignments for round 2 switch presence/absence of context.

ID	Group A		Group B		Round 1 (A)	Round 1 (B)
1	P1A	P2A	P1B	P2B	context	no context
2	P1A	P2A	P1B	P2B	no context	context
...	...	...	...	...	...	...
14	P1A	P2A	P1B	P2B	no context	context
15	P1A	P3A	P1B	P3B	no context	context
16	P1A	P3A	P1B	P3B	context	no context
...	...	...	...	...	...	...
28	P1A	P3A	P1B	P3B	context	no context
29	P1A	P4A	P1B	P4B	no context	context
...	...	...	...	...	...	...
98	P1A	P7A	P1B	P7B	no context	context

During the full study, we had 16 raters evaluating the 392 comments in the **random sample data set – full study**. Those raters consisted of 4 Master's students, 10 PhD students and 2 research associates from Polytechnique Montréal and University of Antwerp. Master and PhD participants were selected in both universities according to their availability during the time of the

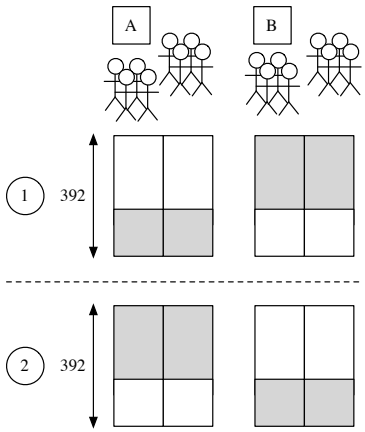

Fig. 4 Design of the full study, consisting of two rounds and two groups (A and B). In each round, each group rates 392 comments twice, where the 392 comments consist of a mixture of comments with context (grey areas) and comments without context (white areas). Groups A and B have eight raters. Each group is divided into two subgroups of four because each question is evaluated by two humans simultaneously. Each rater evaluates 98 comments per round. Note that group A in round 1 and group B in round 2 have the same set of comments, with a one-to-one mapping between the assignments of corresponding raters in the groups. The same holds for group A in round 2 and group B in round 1

experiment (December 2014). All participants were contacted via email (See Appendix 9) and performed the labeling activity independently. The authors did not communicate the reason of the experiment and the participants were free to label any comment without any knowledge of what the other raters had done. As such, no moderation activity was required.

Training was kept as minimal as possible. We were only interested in analyzing how human beings familiar with software development activities perceive the emotions arising in a potential developer from reading her issue comments. Since every human being has been recognizing emotions from birth, we did not perform any training on how to recognize emotions. Hence, participants received an explanation about the Parrot framework as well as examples of each emotion based on the results of the pilot study (Section 3). Beyond the comments, no other information were provided (e.g., parts of the bug report like the issue title or priority). All participants of the experiment have sufficient programming skills and were familiar with software development due to their studies and specializations.

As shown in Figure 4, we organized the raters into two groups A and B, both having the same number of Master's and PhD students. For the sake of clarity, Table 4 describes also how comments assigned to one rater of group A are paired with all other comments assigned to other raters during the case study. The organization of the case study is based on five criteria.

First, in order to compare both groups' ratings, we mapped each member in group A to a member in group B. Since in both groups we tried to have the same number of master students, PhD students and research associates (also

taking into account their maturity), person 1 of group A (p1A) and person 1 of group B (p1B) could for example both be Master's students. Then, in order to compare the ratings between two groups, each couple (p1A, p1B) received the same assignment (modulo random reordering). Second, since we want to verify the influence of context on emotion rating, we divided the case study in two rounds where the raters assess each of the comments assigned to them twice: once with context and once without. So, given a particular couple's assignment, we randomly added context for some of the comments in one round, while we added context for the other comments in the second round, as shown in Table 4.

Third, while each group should rate each comment twice, we also wanted to limit the bias caused by the wide variety in experience, nationalities and culture of raters. For this reason, each group member rated 14 comments in common with each other group member. This is the reason why the full study considers 392 comments instead of 400. Both sample sizes are still large enough to obtain a confidence level of 95% and confidence interval of 5%.

Fourth, to reduce the impact of first seeing a comment with or without context, we designed our experiment such that the assignment of p1A (after randomly adding context) for round 1 corresponds to the assignment of p1B in round 2, while the assignment of p1A in round 2 corresponds to the assignment of p1B in round 1. In each round, all raters assess 98 comments (and each group 392 comments).

Finally, to counter the learning effect and at the same time obfuscate the goal of the study, the two rounds were separated by a time gap of at least 6 days in between submitting the results of the first round and starting the second round.

Similar to the pilot study, in each round each rater received a file with issue comments. For the first round, the raters had two weeks of time to complete the task, for the second round just one week. The raters analyzed the list of comments in order to mark only the primary emotions of Parrot's framework that they were able to identify. For comments with context, the raters were asked to assess *only* the emotions in the comment under analysis, and not the other comments of the context.

4.3 Emotion Oriented Data Set

Once both the pilot study and full study confirmed that issue comments convey human emotions, we conducted a feasibility study to investigate to what extent a machine learning classifier can automatically recognize these emotions. Such a machine learning classifier needs a training set, yet the random sample data sets from Section 4.2 proved inadequate for this purpose. Indeed, the comments in the random sample data sets are biased towards neutral comments and, as such, a classifier would be unable to accurately pick up the words representing particular emotions (i.e., "sorry" for sadness). For this reason, we extended the random sample data sets with 3,413 additional comments.

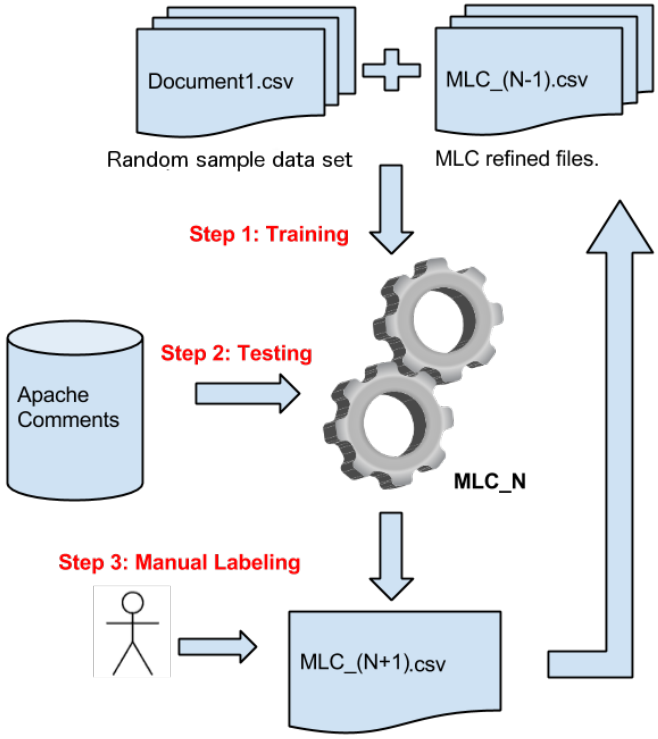

Fig. 5 Iterative process for creating the emotion oriented data set.

This data set was constructed iteratively, as depicted in Figure 5. The iteration started with the random sample data sets mentioned in Section 4.2, which were used to train three classifiers: one for "love", "joy" and "sadness". We did not attempt to train a classifier on the emotions "surprise", "fear" and "anger", since in the pilot and full studies we encountered several cases where these emotions were identified only 0, 1 or 2 times out of hundreds of analyzed comments. Although theoretically a model could be built, the unbalanced nature of these emotions made us decide not to build models for them in our study.

Once a classifier for "love", "joy" and "sadness" was built on a particular version of the training set (details on building a classifier are provided in the next section), we then applied it to the 271,416 comments in the original Apache data set, excluding the training set comments of any iteration. The second author then manually validated the top 400 comments that the classifier reported as having the highest likelihood of exhibiting the emotion under analysis. The goal of this validation was to obtain the true positives amongst these comments, which could then be added to the training set of that emotion. Hence, each iteration in this process added new comments to the training sets of that emotion. At the same time, comments without any emotion were added to the list of "emotion-free" comments, namely comments

without the previous three emotions. These emotion-free comments were used as true negatives for the three classifiers.

The manual verification was performed based on the second author's experience with the pilot study, complemented by the WordNet-affect corpus, which is a crowd-sourced database tagging English words with applicable emotions (Strapparava et al 2004). By repeating this process 10 times for the three emotions and filtering out duplicates amongst the top 400 comments, we ended up with 3,413 additional, manually labeled comments for our training sets.

At the end of the final iteration, the obtained classifiers were ready to be used for creating the **"emotion oriented dataset"**. We used the 3 classifiers for finding the top 400 comments per emotion ("love", "joy" and "sadness"). Then, using the comments of the emotion-free training set, we trained an emotion-free classifier, i.e., a classifier aimed to identify neutral comments. This classifier was finally used for finding 400 emotion-free comments from the original Apache data set (excluding training set data). In the rest of the paper, we refer to these 1,600 comments (400 per emotion) as the emotion oriented data set. In this dataset, all comments are distinct and none of them was previously used for training a classifier. Later on in the paper, we validate the correctness of the labeling with a manual validation (cf. RQ3).

As is common when working with machine learning classifiers, the emotion oriented data set was pre-processed for convenient manipulation for text mining using the python library CLIPS [`http://www.clips.ua.ac.be/pages/pattern`]. Stop-words were removed and words were reduced to their base form by means of the stemming library. Finally, for each comment, we extracted unigrams (individual tokens) and bigrams (pairs of successive tokens) as input for the classification. The choice of using bigrams in particular was driven by the positive results achieved in another study (Ortu et al 2015a), where the authors show that bigrams were able to identify negations like "don't like", which otherwise would not be considered with unigrams. At the end of all these preprocessing steps, each issue comment is transformed in a bag of words. The resulting comments have on average 8.3 tokens, while the number of features (i.e., unique unigrams/bigrams) for comments of type love, joy, sadness and emotion-free is 2,690, 3,907, 3,889 and 5,952, respectively.

4.4 Constructing the Machine Learning Classifier

As listed in the related work, several researchers apply existing sentiment analysis tools such as SentiStrength and NLTK to study sentiments in software artifacts. Yet, these tools have been trained on texts unrelated to software (product reviews, movie rating, ...) and Jongeling et. al. demonstrated that such tools are unreliable for technical prose within software artifacts (Jongeling et al 2015). Consequently, we set out to construct a machine learning classifier specifically trained to identify emotions in issue comments. For replication purposes, we provide a short description on how we constructed the machine learning classifier (classifier hereafter).

For each emotion (`love`, `joy`, `sadness` and emotion-free/neutral), we built a separate classifier. Given an issue comment, such a classifier provides the *likelihood* that the comment contains a particular emotion, where the likelihood is a number between 0 (the comment does not contain the emotion) and 1 (the comment contains the emotion). Moreover, for each classifier we constructed five variants based on different, popular classification algorithms: Support Vector Machine (SVM), Naive Bayes (NB), Single Layer Perceptron (SLP), K-Nearest Neighbor (KNN) and Random Forest (RF).

To evaluate the performance of each classifier, we use the bootstrap validation approach, which provides more stable measures for accuracy, precision and recall, compared to other validation techniques like cross-validation or leave-one-out validation Witten et al (2011). The bootstrap validation divides a data set in training and test set according to a defined ratio —in our set-up 90% training - 10% testing— and generates N sets (1000 in our set-up) uniformly sampled with replacement from the initial dataset. For each generated data set, a classifier is trained and evaluated.

Tables 5, 6 and 7 report for each of the 10 iterations discussed in section 4.3 (in which a classifier is built, then applied to find new occurrences of a given emotion) the best accuracy, precision, recall, F1 and AUC obtained by one of the five classification algorithms, for `love`, `joy`, and `sadness` respectively. We omit the data for the classifier detecting emotion-free comments. Based on these results, we chose the SVM-based classifier for the remainder of our study, since it is the best classifier starting from the eight (`love`), tenth (`joy`) and fourth (`sadness`) iteration, obtaining a precision up to 0.84 (`sadness`) and a recall up to 0.83 (`love`). KNN, RF, NB and SLP (in that order) had lower performance.

Note that the performance reported in Tables 5, 6 and 7 are only provided in order to show how the authors selected the optimal algorithm for emotion classifiers. To fully evaluate the performance of the resulting SVM classifiers, RQ3 compares it to that of three human raters based on the emotion-oriented dataset, which is a *test set* containing only issue comments that were not used to train the classifier (nor to compute the performance reported in Tables 5, 6 and 7). Finally, note that the first 10 rows in Table 5 (and similarly for Tables 6 and 7) are the 10 iterations described in Section 4.3, while the last row is the one from which the final 400 comments were taken to construct the emotional oriented dataset.

5 Case Study Results

For each research question, we first discuss its motivation, followed by the specific approach we used and our results.

Table 5 Performance of the best `love` classifier in each iteration.

Iteration	Best Classifier	Percentage of love comments	accuracy	precision	recall	F1	AUC
0	KNN	6.5%	0.91	0.67	0.68	0.67	0.91
1	KNN	9.03%	0.90	0.72	0.73	0.72	0.94
2	KNN	12.18%	0.89	0.74	0.77	0.75	0.90
3	KNN	14.43%	0.87	0.74	0.76	0.74	0.93
4	KNN	17.12%	0.86	0.75	0.78	0.76	0.94
5	KNN	19.66%	0.84	0.75	0.79	0.77	0.94
6	KNN	23.14%	0.83	0.76	0.80	0.78	0.93
7	KNN	23.14%	0.77	0.69	0.80	0.79	0.93
8	SVM	32.73%	0.81	0.71	0.84	0.82	0.93
9	SVM	35.46%	0.81	0.72	0.83	0.82	0.94
10	SVM	38.2%	0.81	0.70	0.83	0.82	0.94

Table 6 Performance of the best `joy` classifier in each iteration.

Iteration	Best Classifier	Percentage of joy comments	accuracy	precision	recall	F1	AUC
0	SLP	11.48%	0.52	0.52	0.52	0.52	0.82
1	SLP	11.72%	0.57	0.57	0.57	0.57	0.82
2	SLP	13.37%	0.57	0.57	0.57	0.57	0.77
3	KNN	13.47%	0.60	0.58	0.60	0.58	0.80
4	KNN	15.22%	0.63	0.57	0.63	0.62	0.80
5	KNN	17.94%	0.66	0.61	0.66	0.66	0.84
6	KNN	19.20%	0.67	0.67	0.67	0.68	0.84
7	KNN	21.49%	0.69	0.69	0.69	0.70	0.85
8	KNN	24.19%	0.68	0.71	0.68	0.70	0.85
9	KNN	26.97%	0.67	0.72	0.67	0.68	0.88
10	SVM	27.35%	0.68	0.70	0.68	0.70	0.91

Table 7 Performance of the `sadness` classifier in each iteration.

Iteration	Best Classifier	Percentage of sadness comments	accuracy	precision	recall	F1	AUC
0	NB	7.65%	0.84	0.62	0.63	0.62	0.92
1	NB	12.61%	0.82	0.66	0.72	0.69	0.95
2	NB	14.98%	0.79	0.65	0.71	0.68	0.95
3	NB	16.73%	0.77	0.65	0.70	0.67	0.93
4	SVM	18.36%	0.65	0.70	0.65	0.67	0.92
5	SVM	36.22%	0.88	0.89	0.86	0.87	0.92
6	SVM	34.52%	0.87	0.88	0.84	0.86	0.93
7	SVM	35.95%	0.85	0.86	0.82	0.84	0.95
8	SVM	37.01%	0.84	0.84	0.80	0.82	0.94
9	SVM	36.20%	0.83	0.84	0.80	0.82	0.92
10	SVM	35.51%	0.83	0.84	0.80	0.82	0.92

RQ1 — Rater's Agreement. Can human raters agree on the presence or absence of emotions in issue comments?

Motivation. Emotion mining from software development artifacts like issue reports, emails or change logs is not trivial, since such artifacts consist of unstructured, natural language text (Bacchelli et al 2010, 2012). I.e., they are relatively short, written in an informal way (e.g., containing emoticons) and, contrary to regular text parsed in sentiment analysis, they typically contain technical content like stack traces or code snippets interleaved with regular text. Therefore, identifying emotional content in technical prose within software development artifacts is a challenge, let alone agreeing on this content between different human raters.

Approach. We use the random data sets of the pilot and full study to address this research question. As a first step, we measured the percentage of agreement on the presence and absence of emotions. As a second step, we used Cohen's

Table 8 Percentage of agreement (absolute number in parentheses) and Cohen κ values (with confidence intervals) for each emotion in RQ1 — Rater's Agreement(pilot study).

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
`love`	90.75±5 (363)	5.75±5 (23)	85.00±5 (340)	0.38	0.51	0.64
`joy`	69.75±5 (279)	6.50±5 (26)	63.25±5 (253)	0.11	0.19	0.27
`surprise`	96.75±5 (387)	0.00±5 (0)	96.75±5 (387)	-0.02	-0.01	0.00
`anger`	90.75±5 (363)	0.50±5 (2)	90.25±5 (361)	-0.07	0.06	0.19
`sadness`	80.75±5 (323)	3.50±5 (14)	77.25±5 (309)	0.06	0.18	0.29
`fear`	93.25±5 (373)	0.50±5 (2)	92.75±5 (371)	-0.07	0.10	0.26

κ to calculate agreement across all raters of a study for each comment. For the full study, we measured the agreement (a) for each combination of (round, group) separately (i.e., two raters per comment), and (b) across both groups (i.e., more than two raters per comment). In the latter case, we calculate agreement across the union of ratings of *(round 1, group A)* and *(round 2, group B)*, and the union of *(round 1, group B)* and *(round 2, group A)*, as these pairs of ratings consider the same comments. Comparing cases (a) and (b) allows to evaluate whether agreement changes when requiring two, three or four agreeing raters for a comment.

Findings.
[I] Only for `love`, the raters achieved moderate agreement, while `joy` and `sadness` obtained fair agreement. Tables 8 and 9 show the percentage and Cohen κ values of agreement (with confidence interval) for each emotion individually, for the pilot and full study respectively. `Love` clearly obtains the highest κ agreement, corresponding to a moderate value. Except for the pilot study, `joy` and `sadness` have a strong, fair agreement. `Fear`, `anger` and (especially) `surprise` only obtained poor/slight agreement. These numbers are more or less stable across the five cases, with some fluctuations. For example, group A did not have any agreement at all for `surprise` in round 2, contrary to most of the other cases.

The percentage of agreement for `joy` in the pilot study was the lowest, with 30.25±5% of the comments containing disagreement. However, all other emotions and cases had less disagreement than 19.25±5% (`sadness` in pilot study).

[II] At most 7.91±5% (`love`) of the comments agreed on the presence of a particular emotion, whereas up to 96.75±5% (`surprise`) agreed on the absence of a particular emotion. Tables 8 and 9 indeed show that most of the comments were rated as not having a particular emotion (an agreed presence of 0% means that there were no comments where an emotion was present). This is the reason why, despite the high percentage of general agreement, the corresponding κ values are low. The emotions with the lowest κ values (`fear`, `anger` and `surprise`) sometimes have only 0, 1 or 2

Table 9 Percentage of agreement (absolute number in parentheses) and Cohen κ values (with confidence intervals) for each emotion in RQ1 — Rater's Agreement(full study).

(a) (round one, group A)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
`love`	89.03±5 (349)	7.91±5 (31)	81.12±5 (318)	0.40	0.53	0.65
`joy`	86.48±5 (339)	3.06±5 (12)	83.42±5 (327)	0.10	0.24	0.38
`surprise`	89.80±5 (352)	1.28±5 (5)	88.52±5 (347)	-0.00	0.15	0.30
`anger`	90.82±5 (356)	1.28±5 (5)	89.54±5 (351)	0.00	0.17	0.33
`sadness`	93.62±5 (367)	2.04±5 (8)	91.58±5 (359)	0.16	0.36	0.55
`fear`	93.11±5 (365)	1.28±5 (5)	91.84±5 (360)	0.05	0.24	0.43

(b) (round one, group B)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
`love`	92.35±5 (362)	5.61±5 (22)	86.73±5 (340)	0.41	0.55	0.69
`joy`	82.40±5 (323)	3.83±5 (15)	78.57±5 (308)	0.07	0.20	0.33
`surprise`	89.54±5 (351)	0.77±5 (3)	88.78±5 (348)	-0.06	0.07	0.21
`anger`	90.56±5 (355)	0.26±5 (1)	90.31±5 (354)	-0.10	0.00	0.10
`sadness`	91.58±5 (359)	3.57±5 (14)	88.01±5 (345)	0.25	0.41	0.58
`fear`	87.76±5 (344)	1.02±5 (4)	86.73±5 (340)	-0.05	0.08	0.21

(c) (round two, group A)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
`love`	85.71±5 (336)	6.38±5 (25)	79.34±5 (311)	0.27	0.40	0.52
`joy`	82.91±5 (325)	4.08±5 (16)	78.83±5 (309)	0.11	0.23	0.36
`surprise`	90.82±5 (356)	0.00±5 (0)	90.82±5 (356)	-0.05	-0.03	-0.01
`anger`	92.86±5 (364)	0.77±5 (3)	92.09±5 (361)	-0.04	0.14	0.32
`sadness`	93.88±5 (368)	1.79±5 (7)	92.09±5 (361)	0.14	0.34	0.54
`fear`	93.62±5 (367)	0.51±5 (2)	93.11±5 (365)	-0.06	0.11	0.28

(d) (round two, group B)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
`love`	92.35±5 (362)	3.57±5 (14)	88.78±5 (348)	0.27	0.44	0.61
`joy`	82.40±5 (323)	5.61±5 (22)	76.79±5 (301)	0.16	0.29	0.41
`surprise`	89.29±5 (350)	0.77±5 (3)	88.52±5 (347)	-0.06	0.07	0.21
`anger`	92.86±5 (364)	0.51±5 (2)	92.35±5 (362)	-0.07	0.09	0.25
`sadness`	88.78±5 (348)	3.32±5 (13)	85.46±5 (335)	0.17	0.32	0.47
`fear`	91.58±5 (359)	1.02±5 (4)	90.56±5 (355)	-0.01	0.15	0.32

agreed occurrences, while the most frequently agreed emotion (`love`) had up to 31 occurrences (group A, round 1).

[III] Only for `joy` it makes sense to use three raters instead of two raters. Table 10 breaks down the percentages of agreement between at least three and four raters for each individual emotion, together with the corresponding Fleiss κ values of agreement. Agreement between all four raters is harder to achieve (lower percentage values) than requiring at least three

Table 10 Percentage of agreement for the full study between four (2nd column) and at least three (3rd column) raters, together with the Fleiss κ inter-rater agreement and confidence intervals (4th to 6th column). In parentheses, we added the number of analyzed cases of a particular emotion in which raters agreed on the presence of the emotion.

(a) (round 1, group A) and (round 2, group B)

	%agreement of 4 (#)	%agreement of ≥3 (#)	lower κ	Fleiss κ	upper κ
`love`	82.91±5 (13)	94.39±5 (24)	0.48	0.49	0.50
`joy`	69.90±5 (5)	93.88±5 (19)	0.23	0.25	0.26
`surprise`	78.83±5 (0)	97.19±5 (2)	0.05	0.06	0.07
`anger`	85.20±5 (2)	96.94±5 (2)	0.15	0.16	0.17
`sadness`	84.18±5 (7)	96.17±5 (10)	0.34	0.35	0.36
`fear`	85.97±5 (1)	98.21±5 (7)	0.22	0.23	0.24

(b) (round 1, group B) and (round 2, group A)

	%agreement of 4 (#)	%agreement of ≥3 (#)	lower κ	Fleiss κ	upper κ
`love`	79.59±5 (9)	95.41±5 (32)	0.45	0.46	0.47
`joy`	68.37±5 (6)	92.86±5 (17)	0.22	0.23	0.24
`surprise`	81.63±5 (0)	98.21±5 (2)	0.04	0.05	0.06
`anger`	84.18±5 (0)	97.70±5 (1)	0.06	0.07	0.08
`sadness`	85.71±5 (5)	97.19±5 (11)	0.35	0.36	0.37
`fear`	82.91±5 (1)	97.96±5 (4)	0.12	0.13	0.14

Table 11 Percentage of comments in which raters agreed on presence or absence of all 6 emotions, as well as the number of those comments with at least one emotion present.

	pilot	round 1		round 2	
	study	group A	group B	group A	group B
#common	165	215	193	127	207
%common	41.25±5	54.85±5	49.23±5	32.40±5	52.81±5
#common with ≥ 1 present	20	36	36	16	40
%common with ≥ 1 present	5.00±5	9.18±5	9.18±5	4.00±5	10.00±5

raters to agree. In fact, the Fleiss κ values for four-rater agreement are in the same ballpark as for the case of two raters (Tables 8 and 9). Although the agreement between at least three raters overall is higher than in the case of two (or four) raters, only in the case of `joy` there really is a significant improvement. Hence, trying to enlist more than two raters does not seem to be worthwhile when trying to identify emotions.

[IV] Only in on average 46.11±5% of the comments, both raters had the same rating for all 6 emotions. Table 11 shows for each study the number and percentage of comments for which both raters assigned to the same comment agreed on all 6 emotions. The highest number of such agreement occurred for group A in round 1, while the same group obtained the lowest agreement in round 2, statistically significantly lower than in three of the other cases (except for the pilot study). Since (round 1, group A) and

(round 2, group B) both considered the same comments (and context), that configuration of comments and context seems more easy for raters to agree upon.

Furthermore, on average for 7.47±5% of the comments for which raters agreed on all 6 emotions, at least one emotion was present. Raters typically agreed on absence of emotions for dry comments like "committed" and "done".

While some emotions obtain higher agreement than others, only `love`*,* `joy` *and* `sadness` *obtained at least fair agreement. Although comments clearly contain emotions, raters agree the most on the absence of an emotion. Using more than two raters does not significantly change the results in terms of degree of agreement on emotions.*

RQ2 — Context Influence. Does context improve the agreement of human raters on the presence of emotions in issue comments?

Motivation. The interpretation of a phrase usually depends on the previous discussion (i.e., context) of the conversation (Tepperman et al 2006). For example, the sentence "yeah, right" can have a different meaning (both sarcastically and otherwise) when following a sentence like "with java 8 we fix all problems" versus "breaking backward compatibility is risky". For this reason, rating a comment without its context can be compared to eavesdropping on a group conversation and only catching the last phrase of the conversation. However, due to the technical and unstructured nature of software development artifacts, the impact of context might be different in technical prose than in normal language. Here, we want to analyze the impact of context on agreement between raters.

Approach. This research question only considers the full study data set. Since each group considers 392 comments once without and once with context (randomly distributed across two rounds), here we merge the results of both rounds such that, for each group, we can compare the ratings without and with context. For this comparison, we calculate similar agreement percentage and Cohen κ values as for RQ1 — Rater's Agreement. Furthermore, we measure how often raters made a different decision for a particular emotion when seeing context or not, and whether such different decisions led from agreement to disagreement, disagreement to agreement or did not have any net effect.

Findings.

[I] Adding context slightly reduces rater agreement, but not significantly. Table 12 compares, in both groups, the agreement amongst the rating results of the comments without context (odd rows) and those with context (even rows). Except for `surprise` in group A, the κ agreement is not significantly different (the confidence intervals still overlap) with or without

Table 12 Percentage of agreement and Cohen κ values (with confidence intervals) for comments without and with context (RQ2 — Context Influence). The percentages are relative to the 392 comments without and with context, respectively odd and even rows.

(a) Group A

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	88.78±5 (348)	7.91±5 (31)	80.87±5 (317)	0.40	0.52	0.64
	85.97±5 (337)	6.38±5 (25)	79.59±5 (312)	0.27	0.40	0.53
joy	86.22±5 (338)	3.32±5 (13)	82.91±5 (325)	0.11	0.25	0.39
	83.16±5 (326)	3.83±5 (15)	79.34±5 (311)	0.10	0.22	0.35
surprise	91.07±5 (357)	1.28±5 (5)	89.80±5 (352)	0.01	0.18	0.34
	89.54±5 (351)	0.00±5 (0)	89.54±5 (351)	-0.06	-0.04	-0.02
anger	92.86±5 (364)	1.02±5 (4)	91.84±5 (360)	0.00	0.19	0.37
	90.82±5 (356)	1.02±5 (4)	89.80±5 (352)	-0.02	0.13	0.29
sadness	95.41±5 (374)	1.53±5 (6)	93.88±5 (368)	0.15	0.38	0.60
	92.09±5 (361)	2.30±5 (9)	89.80±5 (352)	0.15	0.33	0.50
fear	93.62±5 (367)	1.02±5 (4)	92.60±5 (363)	0.02	0.21	0.41
	93.11±5 (365)	0.77±5 (3)	92.35±5 (362)	-0.02	0.15	0.33

(b) Group B

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	93.11±5 (365)	4.85±5 (19)	88.27±5 (346)	0.40	0.55	0.70
	91.58±5 (359)	4.34±5 (17)	87.24±5 (342)	0.31	0.46	0.62
joy	83.16±5 (326)	4.85±5 (19)	78.32±5 (307)	0.14	0.27	0.40
	81.63±5 (320)	4.59±5 (18)	77.04±5 (302)	0.10	0.23	0.35
surprise	89.80±5 (352)	0.51±5 (2)	89.29±5 (350)	-0.08	0.04	0.16
	89.03±5 (349)	1.02±5 (4)	88.01±5 (345)	-0.04	0.10	0.24
anger	92.60±5 (363)	0.26±5 (1)	92.35±5 (362)	-0.10	0.03	0.15
	90.82±5 (356)	0.51±5 (2)	90.31±5 (354)	-0.08	0.05	0.18
sadness	91.07±5 (357)	3.32±5 (13)	87.76±5 (344)	0.22	0.38	0.54
	89.29±5 (350)	3.57±5 (14)	85.71±5 (336)	0.20	0.35	0.50
fear	90.05±5 (353)	1.02±5 (4)	89.03±5 (349)	-0.03	0.12	0.27
	89.29±5 (350)	1.02±5 (4)	88.27±5 (346)	-0.04	0.10	0.25

context, even though the actual κ values seem lower with context than without. Similarly, the percentages of agreement seem lower with context, but not in a significant way. For some cases, context provides more evidence of the presence of an emotion than without context, while in some cases the inverse situation holds. Similar to RQ1 — Rater's Agreement, both groups have similar results, except for `anger`, for which group B had a much lower agreement (since less occurrences were agreed upon).

[II] Most of the raters pick the same answer with or without context, yet they tend to switch more from absence to presence of an emotion than the other way around. Table 13 shows for both groups how many raters picked a different answer for an emotion in the absence or presence of context. Clearly, in most cases (between 90.3% to 95.7% of the time) raters did not change their rating, which suggests that (1) rat-

Table 13 The number of times raters changed their rating from the rating in a row (comment without context) to the one in a column (comment with context). A 0 in a row or column for a particular emotion means that that emotion previously was not identified without (row) or with (column) context. A 1 in a row or column for a particular emotion means that that emotion previously was identified without (row) or with (column) context.

	context	`love`		`joy`		`surprise`		`anger`		`sadness`		`fear`	
		0	1	0	1	0	1	0	1	0	1	0	1
A	0	650	28	658	46	711	28	719	29	722	32	730	21
	1	29	77	30	50	32	13	21	15	13	17	21	12
B	0	701	18	644	36	712	28	730	23	691	32	712	25
	1	16	49	32	72	21	23	14	17	23	38	22	25

Table 14 How often raters went from disagreement (d) to agreement (a) or vice versa when comparing the set of comments without context (rows) to the set of comments with context (columns), for groups A, B, and when combining both groups (at least three raters agreeing).

		`love`		`joy`		`surprise`		`anger`		`sadness`		`fear`	
		d	a	d	a	d	a	d	a	d	a	d	a
group A	d	26	18	26	28	9	26	13	15	5	13	6	19
	a	29	319	40	298	32	325	23	341	26	348	21	346
group B	d	15	12	42	24	19	21	16	13	17	18	19	20
	a	18	347	30	296	24	328	20	343	25	332	23	330
3-rater	d	9	8	9	8	1	9	3	8	4	8	1	7
	a	14	361	26	349	7	375	7	374	10	370	6	378

ings for a particular comment are fairly stable, and (2) context does not add substantially new information for the interpretation of a particular comment.

At the same time, we can also see that if a rater changes his or her mind, he or she rather tends to mark a previously (i.e., without context) absent emotion as present, than the other way around (except for `love`, `surprise` and `fear` in group A). We noticed that raters especially changed their mind when the context carried information that put the comment in a new perspective. In the case of `sadness`, the relative difference between both cases goes from 50% to more than 100%. This would suggest that although context does not play a major role in agreement, in cases when it does, raters become less sure and tend to mark an additional emotion as being present, i.e., they change their mind.

[III] The change of mind due to context pushes more pairs of raters from agreement to disagreement than the other way. Table 14 shows for each comment and emotion whether the raters' change of mind has an impact on the agreement between the raters. Even though the vast majority of cases did not change agreement/disagreement, the results also show for all emotions and both groups that more raters went from agreement to disagreement when showing context (row "a", column "d"), than the other way around (row "d", column "a"). Again, context does not seem to have a

major impact, but when it does, it causes more uncertainty (disagreement) than agreement.

This observation is less pronounced when using agreement between at least three raters, as shown in the bottom two rows of Table 14. Only for `love` and `joy`, there is still more agreement turning into disagreement than the inverse, but for the other four emotions, the usage of three or more raters makes the results more robust to fluctuations introduced by context. Hence, even though more raters do not significantly improve agreement (see RQ1 — Rater's Agreement), they make ratings more robust in the presence or absence of context.

Context does not play a significant role in the rating of emotions in issue comments, but when it does, it seems to cast more doubt than confidence, unless more raters are used.

RQ3 — Classifier Feasibility. To what extent can a machine learning classifier identify emotions in issue comments?

Motivation. The previous findings show that `love`, `joy` and `sadness` obtained at least fair agreement among human raters. Unfortunately, existing sentiment analysis tools such as SentiStrength and NLTK are unreliable for assessing emotions in technical prose within software artifacts (Jongeling et al 2015). For this reason, we explore the feasibility of an automatically built machine learning classifier for identifying these three emotions.

Approach. First, similar to the pilot and full user studies, we evaluate the degree to which human raters agree on the presence of emotions in the emotion-oriented data set (Section 4.3). For this, the 1,600 comments of the emotion-oriented data set were analyzed by three of the authors for the presence of both `love`, `joy`, `sadness` and `emotion-free` (i.e., none of the three emotions appeared). In other words, although each of these comments originally occurred in the top-400 comments of one particular emotion's classifier (love, joy, sadness or emotion-free), we also analyzed them for the other two emotions as well as for emotion-free. An alternative would have been to use a random data set for evaluation instead of taking the top 1,600 comments. However, given the overall low percentage of comments with a particular emotion, the corresponding percentage in the random data set would be too low to truly evaluate the classifiers.

The same agreement metrics as for RQ1 — Rater's Agreement are calculated. To evaluate the performance of the classifiers, we then used the comments and emotions of the emotion-oriented data set for which the three human raters agreed. For each classifier, we compute the AUC, precision and empirical recall (i.e., the percentage of 1,600 comments with a specific emotion that were identified as such by a classifier), then we compare its performance with a ZeroR classifier. This is a simple baseline classifier that always selects

the majority outcome. It is a popular means to evaluate whether a model really is useful, or whether one could obtain similar results just by guessing. In addition to comparing to a ZeroR baseline model, the AUC values also help us compare to a random model, i.e., the higher a model's AUC value is compared to 0.5, the better its performance compared to a random model.

Findings.

[I] In the emotion-oriented data set created by the machine learning classifier, human raters can achieve a moderate to substantial level of agreement on the presence and absence of emotions. Table 15 reports how humans classified the 1,600 comments of the emotion-oriented data set. As we can see, humans rate most of the comments as neutral (43.4%), whereas among the actual emotions, `love` (37.4%) is the most common. Note that although only 400 out of the 1,600 comments were obtained by the love classifier, human raters could also label comments belonging to the other groups as containing `love`, since the raters did not know which comments were generated by which classifier. Comparing Tables 15 and 10, we can see that humans achieve a higher level of agreement for `love` and `sadness` (and the absence of emotions), whereas for `joy` the level of agreement remains the same.

Table 15 Fleiss κ inter-rater agreement and number of comments containing emotions in the emotion-oriented data set.

Emotion	κ	# comments
Love	0.57	598 (37.4 %)
Joy	0.24	182 (11.4 %)
Sadness	0.69	125 (7.8 %)
Neutral	0.69	695 (43.4 %)

[II] The machine learning classifiers have a good precision for identifying emotions, ranging from 88% for joy to 98% for sadness, while the recall results fluctuate depending on the emotion analyzed. Table 16 reports the confusion matrix for the machine learning classifiers against the ZeroR classifier. Each row reports how the classifier labels the 1,600 comments of the emotion-oriented data set. To simplify the comparison, Table 17 provides the corresponding performance of the classifiers in terms of precision, recall and AUC.

Precision-wise, we observe high values of 78% up to 98%, with the lowest precision of 65% for identifying the absence of `joy`. When comparing these results to a baseline ZeroR classifier that classifies all comments as having a particular emotion, we find that ZeroR achieves a much lower precision of 41% for `love` and `joy`, and 28% for `sadness`. This shows that the classifier is more precise than the baseline classifier for finding comments containing emotions.

For joy and sadness, the recall fluctuates around 25%, whereas for `love` it accounts for 85%. In other words, the `joy` and `sadness` models miss a large number of classifications (high number of false negatives), which might indicate that our models are too specialized to the training data set (or the

Table 16 Confusion Matrix for the machine learning classifier (MLC) and ZeroR classifiers. The rows and columns "Non Love", "Non Joy" and "Non Sadness" represent the absence of Love, Joy and Sadness respectively.

		Human			Human	
		Love	Non Love		Love	Non Love
Love	MLC	556	51	ZeroR	654	946
Non Love		98	895		0	0

		Joy	Non Joy		Joy	Non Joy
Joy	MLC	163	23	ZeroR	663	937
Non Joy		500	914		0	0

		Sadness	Non Sadness		Sadness	Non Sadness
Sadness	MLC	123	3	ZeroR	443	1157
Non Sadness		320	1154		0	0

Table 17 Precision and Recall for the machine learning classifier (MLC) and ZeroR classifiers.

		Precision	Recall	AUC		Precision	Recall
Love	MLC	92%	85%	0.83	ZeroR	41	100
Non Love		90%	95%			0	0
Joy	MLC	88%	25%	0.93	ZeroR	41	100
Non Joy		65%	98%			0	0
Sadness	MLC	98%	28%	0.94	ZeroR	28	100
Non Sadness		78%	100%			0	0

test set is too different from the training set). Only for `love`, the classifier is more robust to false negatives and positives. We did not compare recall to the baseline classifier, since the recall of ZeroR is always 100%, as it labels all comments as containing an emotion.

Comparison of the performance of the classifiers to random models (instead of to ZeroR models) using AUC shows large increases in performance, with AUC values of 0.83, 0.93 and 0.94 respectively. Again, this shows that our models perform substantially better than simpler baseline models.

Finally, we want to discuss the problem of comments misclassified by the classifiers. As an example, we can refer to the sentence "Oh, I didn't consider one flow [...]. I misunderstood the case and sorry for the confusion. [...] Patch looks good to me.", which is classified as Sadness by the classifier, but as Love by the human raters. Hence, the status "misclassified" often depends on a human judgement given by the selected raters that is hard to capture in terms of textual unigrams or bigrams. Just adding new (textual) features would not guarantee to solve this issue. We report this problem in threats to validity along with potential solutions.

The top comments identified by machine learning classifiers enable relatively high agreement on the presence or absence of emotions `love` *and* `sadness`*. The machine learning classifiers also obtain a good precision and recall for identifying the emotion* `love` *in comments, but for* `joy` *and* `sadness`*, the machine learning classifiers obtain a lower recall.*

RQ4 — Important Keywords. What keywords does a machine learning classifier rely upon to identify emotions?

Motivation. Although the machine learning classifiers offer a decent performance, being able to manually find indications of certain emotions is equally important. As the classifiers ultimately look for certain combinations of keywords in issue comments that convey emotions, this research question aims to identify the most important keywords for each emotion. Apart from providing clues for future manual rating of comments, this also allows to validate the consistency of the results obtained in RQ3 — Classifier Feasibility.

Approach. We focus on the classifier models built in RQ3 — Classifier Feasibility. For each keyword used by the classifier, we compute the information gain, namely the expected reduction in entropy when the keyword would be dropped from the model (Mitchell 1997). Sorting the keywords according to the information gain, we identify the most import ones used for the classification. Apart from interpreting the most important keywords, we also checked the keywords' corresponding entry in the WordNet-affect corpus to identify the emotions with which they have been tagged. This allows us to evaluate whether the keywords are indeed relevant.

Findings.

[I] Emotion-driving keywords such as *thanks* or *sorry* are the most important keywords across all classifiers. Table 18 reports the top 20 keywords used by the classifiers to identify emotions in text. Note that the table mentions stemmed words for unigram and bigrams, since comments were preprocessed before using the classifiers (see Section 4.4).

The results show how, for each emotion, there is one keyword with a predictive power far higher than the others. This *emotion-driving* keyword is *thank* for `love` and `joy` and *sorry* (stemmed form: *sorri*) for `sadness`. In all three cases, the weight of the top keyword is at least six times higher than the weight of the second one. During the pilot case, both keywords were already observed and documented with representative examples in Section 3. From Table 18 we also notice that only one bigram (*look good*) is mentioned. In that sense, bigrams have less impact than unigrams for emotion classification in text messages.

[II] The `love` and `joy` classifiers are similar. Although for `love` and `sadness` the top keyword matches with the emotion tagged by Wordnet-affect,

Table 18 Top 20 keywords used by the classifiers models to identify emotions in text. Keywords used across emotions are reported in italics. Keywords in bold occur in the correct classifier according to their WordNet-affect entry.

Love		Joy		Sadness	
keyword	weight	keyword	weight	keyword	weight
thank	0.65	***thank***	0.39	**sorri**	0.61
commit	0.08	*commit*	0.06	**miss**	0.03
appli	0.05	*look* ***good***	0.04	***thank***	0.03
issu	0.03	*fine*	0.03	*close*	0.03
close	0.03	*appli*	0.03	wa	0.02
expect	0.03	integr	0.02	onli	0.02
revis	0.03	*issu*	0.02	nois	0.02
resolv	0.03	*look*	0.02	wrong	0.02
verifi	0.03	*ha*	0.02	file	0.02
look	0.03	review	0.02	*fix*	0.02
hi	0.03	*close*	0.02	guy	0.02
look ***good***	0.02	ok	0.02	*ha*	0.02
move	0.02	*fix*	0.02	forgot	0.02
fine	0.02	improv	0.01	befor	0.02
review	0.02	test	0.01	ill	0.02
file	0.02	*cheer*	0.01	delay	0.02
version	0.02	patch	0.01	*resolv*	0.02
ha	0.02	*resolv*	0.01	*look*	0.02
rev	0.02	comment	0.01	*issu*	0.02
suggest	0.02	ad	0.01	appli	0.02

for `joy` the top keyword *thank* actually is the same as for `love`. In fact, the emotions `love` and `joy` share many keywords, with 5 keywords in the top 10 being the same. This result was expected since in the emotion-oriented data set, human raters labeled 73 sentences as containing both `love` and `joy`. Similarly, the confusion matrices in Table 16 showed roughly the same number of `love` and `joy` comments (654 vs. 663).

On the other hand, the sadness classifier is different, even though some general terms like *thank* or *close*, and technical words like *fix* are shared with the other classifiers. Terms like *miss*, *wrong* and (to some extent) *nois(e)* intuitively make sense as being linked to `sadness`. These (limited) numbers of terms assume a relevant role in the software engineering domain whereas in other domains this may not have the same value. This result is supported by the literature where it has been observed that general purpose tools for sentiment analysis, such as SentiStrength and NLTK, do not agree with the sentiments expressed by developers (Jongeling et al 2015).

In general, only four of the top keywords were correct according to the Wordnet-affect tags (Strapparava et al 2004). The other terms either were incorrect or were not yet tagged in the Wordnet-affect corpus. One reason for this is that the corpus is not yet complete. Wordnet-affect is based on the labeling of general-purpose sentences. Hence, the data set lacks many keywords belonging to specialized domains like software engineering, such as *commit* or *issue*. Still, it is the closest corpus for evaluating emotions in textual data.

The machine learning classifiers focus on emotion-driving top keywords like thank and sorry, various general and technical terms, and a handful of Wordnet-affect keywords. The machine learning classifiers for `love` *and* `joy` *focus on the same keywords, showing overlap between the two emotions.*

6 Discussion

This section discusses our findings in more detail.

6.1 Impact of Context

At first sight, our findings for RQ2 — Context Influence seem counter-intuitive: while one would expect that the addition of context strengthens agreement due to the availability of more information, we seem to observe that either human raters stuck with the same rating or marked an additional emotion as present, reducing the amount of agreement amongst raters. Although more experiments are needed to confirm and understand this phenomenon, we briefly discuss a couple of hypotheses.

The worst case scenario would be that emotion mining is so subjective and nuanced that even for humans it is impossible to correctly determine the presence of a specific emotion in an issue comment. However, we believe that the truth is more subtle. For example, in RQ2 — Context Influence we only rated the last comment of an issue report, and reports with context contain (by definition) the viewpoint of multiple commenters, for which it is not always clear how they relate to the last commenter's viewpoint.

Consider a hypothetical example of the following three comments by three different commenters: "Class FooBar is a total waste of time, just nuke it!", "We do have users relying on its features, I'm afraid we should fix this bug" and "I share your view, working on it". Although the first comment clearly contains `anger` and the second one `sadness`, the third one is quite ambiguous regarding which view is shared. Without context, the comment might be neutral, while with context it might be `neutral`, `anger`, `sadness` or a combination of these emotions. As such, context does not necessarily filter the set of possible emotions. On the contrary, it enriches the nuances on the emotions perceived by a rater and can lead to different interpretations.

Another hypothesis is that using a simple yes/no decision as rating is too large a simplification. Maybe one should provide multiple ratings, which would allow to model uncertainty in a rating.

6.2 Do Emotions Matter for Issue Reports?

The premise of this paper was that, similar to other domains, emotions could have an impact on software development activities like bug fixing or development of new features. In this section, we perform a preliminary analysis with a sample of 207 comments selected out of the three data sets: 73 for `love`, 62 for `joy` and 72 for `sadness`. Note that the same report can feature in multiple emotions. The goal of the analysis is to check whether reports with certain emotions tend to (1) be fixed faster, (2) have more comments or (3) have more people following ("watching") the issue report.

After we extracted the fix time, number of comments and number of watchers of the issue reports of the 207 comments, we need to check the null hypothesis that the reports for the three emotions either have the same average fix time, number of comments or number of watchers. For this reason, we performed (non-parametric) Kruskal-Wallis tests: if the null hypothesis was rejected (α value of 0.05), i.e., at least one emotion has a different average value for one of the three measured attributes, we performed post hoc tests to determine the emotion with significantly different property values.

We found a significant difference for the number of comments, i.e., reports with a comment rated as `love` tend to have a lower number of comments (median value of 5) than `joy` (median value of 7.5) or `sadness` (median value of 12). Similarly, the number of watchers of reports with a comment rated as `love` has a median value of 0 whereas for `sadness` the median value is 1, i.e., less people monitor the former reports. Although not strictly significantly different, the Kruskal-Wallis test for the fixing time of reports obtained a low p-value of 0.057, with reports containing a `love` comment taking a median number of 20 days to be resolved, compared to 53.5 for `joy` and 68.5 for `sadness`.

Of course, more analysis is needed to fully investigate the link between emotions on software development. In a first follow-up study using our machine learning classifiers, we found that emotions such as joy and love are linked with a shorter (i.e., faster) issue resolution time, whereas emotions such as sadness are linked with a longer issue resolution time (Ortu et al 2015a). In contrast to the preliminary analysis of this section, the differences were statistically significant. Note that the follow-up study combined the emotion data with other metrics (e.g., politeness and sentiment) and fully relied on the machine classification of emotions and comments, whereas our preliminary analysis here only used a smaller set of metrics and manually labeled comments. Future work should revisit this study on a larger, manually tagged data set, as well as explore the other two hypotheses on larger data sets.

7 Threats to Validity

Threats to internal validity concern confounding factors that can influence the obtained results. In this context, such a threat arises if the developer

expresses a comment that does not reflect his or her emotions, or conversely does not express any emotion. We consider this threat possible, but only to a limited extent. First of all, empirical evidence (in another domain) shows a causal relationship between a developer's emotions and what he or she writes in issue comments (Pang and Lee 2008). Moreover, since developer communication has as first goal information sharing, removing or disguising emotions *may* make comments less meaningful and cause misunderstanding.

A similar threat would be that developers knowingly express certain emotions in their comments, for example because they are aware that their comments are visible to colleagues and could be analyzed. Since the comments used in this study were collected over an extended period and comprise developers not aware of being monitored, we are confident that the emotions we mined are genuine. Note that this risk is also why we could not involve the authors of the comments in our study.

Threats to construct validity focus on how accurately the observations describe the phenomena of interest. Mining emotions from textual issue comments presents some difficulties due to ambiguity and subjectivity. To reduce these threats, the authors adopted Parrott's framework as a reference for emotions. Finally, to avoid bias due to personal interpretation, each comment in the case study was analyzed by at least two participants.

In RQ3 and RQ4, we assume that certain combinations of keywords in issue comments convey emotions. However, we limited the analysis to unigrams and bigrams. Other, higher-level features (e.g., grammatical structure of the sentences) potentially could improve the performance of the classifiers, but further investigation is needed for this.

Threats to external validity correspond to the generalizability of our experimental results (Campbell and Stanley 1963). In this study, we manually analyze a sample of issue reports belonging to 117 open source projects. We chose the projects as a representative sample of the universe of open source software projects, as Apache projects are popular, large (both in terms of code and team size) and long-lived. We consider the validity of our results limited to the domain of issue reports. Other domains, such as blog posts and Twitter (Aman and Szpakowicz 2007; Balabantaray et al 2012), show levels of agreement among raters different from the ones we reported. Finally, we advocate for the replications of our analysis on other open source systems and on commercial projects in order to confirm our findings.

Threats to reliability validity correspond to the degree to which the same data would lead to the same results when repeated. This research is the first attempt to manually investigate emotions of developers from issue comments, hence no ground truth exists to compare our findings. We defined the ground truth via agreement or disagreement of the raters. On the one hand, other groups of raters might obtain agreement on different emotions and comments, possibly leading to different results. On the other hand, RQ2 — Context Influence showed that both groups of the full study and, to some extent, the pilot study obtained similar levels of agreement. Nevertheless, replications

of this work with different and larger groups of participants are needed to confirm our findings.

This study focused on text written by developers *for* developers. To correctly depict the emotions embedded in such comments, it is necessary to understand the developers' dictionary and slang. "Kill Bill" might refer to a famous movie, an actual murder, or in the context of software it might as well be the innocuous "Bill.kill()" to stop a thread in the java threadpool. This assumption is supported by Elfenbein and Nalini's work that provided evidence that for members of the same cultural and social group it is easier to recognize emotions than for people belonging to different groups (Elfenbein and Ambady 2002). Since all the participants of this study have a background in computer science, we are confident that participants may interpret the issue comments in the same manner as the developers. We did not involve raters with different background (such as linguists or psychologists), because they may make oversights or misinterpret the terms used by developers.

During the pilot study, the only raters involved were the first four authors. These four raters also participated in the full study, which could introduce some form of bias due to a "learning effect" during the pilot study. To understand the magnitude of this bias, we removed all the comments rated by one of the four authors, then repeated the analysis for RQ1 and RQ2. More specifically, using a dataset of 210 comments labeled by the 12 raters not involved in the pilot case (i.e., who were not author), we re-calculated tables 9, 10, 12, 13 and 14 (Appendix B).

With respect to RQ1, Tables 20 and 21 confirm that `love` is still the emotion found in the highest amount of comments (7.62%) and that this emotion achieves the highest level of agreement among raters (moderate). For the other emotions, raters continue to find a limited presence in comments and achieving at best a fair level of agreement (e.g., `sadness`). We only observe a difference in κ agreement for `joy`, which has become lower than in the study with 16 participants, but still higher than for `surprise` and `anger`.

With respect to RQ2, we confirm that the addition of context reduces rater agreement, or at least cannot increase it (Table 22). Raters generally pick the same answer with or without context and the presence of context pushes more pairs of raters from agreement to disagreement (Tables 23 and 24). Hence, the analysis without four authors confirmed the original results, suggesting that the impact of learning effect between the pilot and the full study is minimal.

8 Related Work

Human emotions, i.e., how humans feel and how they perceive their colleagues (Fowler and Christakis 2008), is a concept that recently has started to attract the interest of the (software engineering) research community. As mentioned earlier, sentiment analysis is related to emotions in a sense that it evaluates a given emotion as being positive or negative (see Section 2.2). Sentiment analysis also plays an important role in different domains. Hence, we divide the

related work into two categories: first, we describe why emotions or sentiment analysis grabbed the interest of researchers, then we focus on its role in the software engineering domain.

The Role of Emotions in Marketing and Finance

The domains of marketing and finance have been interested in studying the sentiment of people for quite some time. Many online markets like mobile app stores or Amazon provide facilities for customers to assess their products and give their reviews. For popular products, the number of reviews can run into hundreds or thousands. Companies are interested in applying sentiment analysis on these reviews, since it is useful to organize marketing strategies.

Hu et al. (Hu and Liu 2004) proposed a set of techniques based on data mining and natural language processing methods for mining and summarizing product reviews, and showed that their methods are useful to find the sentiments of customers and their attitudes towards different features of the products. These results are useful since they may help potential customers to make informed decisions and may help manufacturers to keep track and manage customer reviews. Cataldi et al. (Cataldi et al 2013) presented an approach to extract users' opinions from reviews, about specific features of products and services. They model each sentence as a set of terms in a dependency graph connected through syntactic and semantic dependency relations. Compared to an oracle of 39 human subjects for hotel reviews, the approach obtained high precision and recall on the features (precision and recall higher than 0.85 and 0.83 respectively), with the computed polarity degree slightly below the average human performance. The polarity degree they have applied was a five-point scale with two positive cases, two negative cases, and "not available". This scale in addition to the polarity sign of the writer's sentiment towards a feature, depicts the intensity of the sentiment too.

Pak et al. (Pak and Paroubek 2010) focused on using Twitter to extract people's sentiments. They collected a corpus of 300,000 tweets evenly distributed across positive emotions, negative emotions and absence of emotions. By performing statistical linguistics analysis on the corpus, they build a sentiment classifier that uses the corpus as training data. Finally, they conducted an experimental evaluation on a set of microblogging posts and found that their methods are efficient and have better performance in comparison to previously proposed techniques. In our work, instead of sentiment, we build classifiers for identifying emotions like love and joy in comments. Those classifiers also obtained good a precision, more than 80 percent.

Similar to marketing, analysis of financial issues also uses sentiment analysis on news items, articles, blogs and tweets about companies to drive automated trading systems like StockSonar (Feldman 2013). Sehgal et al. (Sehgal and Song 2007) introduce an approach for stock prediction based on sentiments of online messages. This prediction is based on the correlations between stock values and sentiments. Das et al. (Das and Chen 2007) trained an algorithm for small investor sentiment from stock message boards. Their output

can be used to assess the impact of small investor behaviour on stock market activity.

The Role of Emotions in Software Development

Recently, several studies started to investigate the impact of human factors, including emotions and other affective metrics, on the software engineering process. The growing interest in identifying and addressing challenges posed by emotion awareness in software engineering is also attested by the birth of a workshop, i.e., SEmotion (semotion 2016), dedicated to this research domain.

Rigby et al. (Rigby and Hassan 2007) used Linguistic Inquiry and Word Count (LIWC), a pyschometrically-based linguistic analysis tool, to study the Apache httpd developer mailing list. In their study, they tried to investigate the personality of four top developers. They also attempted to examine the general attitude of developers near critical events, like before and after a release or when they join or leave a project. Finally, they gained insight into the discussions happening in the Apache mailing and also into the people that participate in those discussions and proposed directions for future work. Tourani et al. (Tourani et al 2014) studied mailing lists of open source projects belonging to Apache. By running an automatic sentiment analysis tool on these mailing lists, they show that development mailing lists also carry positive and negative sentiments. The study identifies and categorizes these sentiments for user and developer mailing lists.

Bazelli et al. (Bazelli et al 2013) replicated part of the experiment of Rigby (Rigby and Hassan 2007) on user threads hosted in StackOverflow (the question and answer website for programmers). They explored the personality traits of the users by analysing their answers and questions. They applied LIWC to extract and categorize user personalities and found that users with higher reputation are more extravert and show less negative emotions. Tanveer et al. (Ahmed and Srivastava 2017) analyzed the human point of view of technical users participating in StackOverflow posts. They found that there are several bad practices among technical users of StackOverflow, demonstrating that emotion plays a primary role even in answering posts on online developer fora.

Organizational behavior research (e.g., (Amabile et al 2005)) showed the influence of affect on work outcomes such as creativity, productivity, and task quality. Based on this, De Choudhury et al. (De Choudhury and Counts 2013) noted that the "affective climate" of an enterprise, as one valuable resource, can be used to improve organizational processes and outcome. They explored various emotional expressions of employees of 500 large software corporations. They inspected the posts on an internal Twitter-like microblogging tool, called OfficeTalk, to characterize emotional expression of employees at the workplace. Empirical analysis showed that affective expression in the enterprise can be the result of various workplace factors. These factors can either be exogenous or endogenous, and depend on geographical and hierarchical organization. The authors scored the microblog posts of employees over time using LIWC, and

concluded that affective expression in the workplace can provide an efficient tool for assessing performance relevant outcomes.

Guzman et al. proposed an approach for finding emotional awareness in software development teams (Guzman and Bruegge 2013). For this purpose, they used the latent Dirichlet allocation algorithm to identify the topics discussed in the collaboration artifacts (e.g., text from mailing lists). By applying lexical sentimental analysis, they obtained an average emotion score for each of the topics. Guzman et al. used sentiment analysis to extract emotions expressed in commit comments of 29 open source projects on Github (Guzman et al 2014), then analyzed the correlation of the extracted sentiments with different factors such as programming language, team distribution, project approval and time and day of the week in which the comment has been written.

Jongeling et al. investigated whether general-purpose tools for sentiment analysis, such as SentiStrength and NLTK, agree with the sentiments recognized by human raters (Jongeling et al 2015). They found that these tools (i) do not agree with human raters, (2) have limited agreement with each other, and (3) may drive/lead to contradictory conclusions. Finally, they advocated a sentiment analysis tool specific to software engineering artifacts.

Graziotin et al. describe why the software development process, as a primarily intellectual process, is substantially more complex than other industrial processes (Graziotin et al 2014). They introduced psychological measurements for affect, analytical problem solving, and creativity in empirical software engineering, then investigated the correlation among these three factors on a sample of 42 students. The results show no significant difference in the number of generated creative ideas based on the affects. However, the results show also that happier software developers are more productive when dealing with problem solving.

Tourani et al. (Tourani and Adams 2016) performed a large empirical case study on the OpenStack and Eclipse open source projects to investigate the impact of metrics related to human discussions on the quality of the software. As part of their study, they extracted the sentiment of issue comments and review comments to measure their influence on defect-proneness of issues. Their results show that these sentiment related metrics also play role in the quality of the work measured by defect proneness of issues.

Ortu et al. (Ortu et al 2015a) empirically measured whether affectiveness is correlated with developer productivity. In order to measure the developers' affectiveness, they studied emotions, sentiment and politeness of developers in more than 4,000 sentences reported in the issue tracking system of Apache projects, this led to a public dataset available at (Ortu et al 2016c). Then, they built regression models to evaluate the impact of these metrics on the issue fixing time, while controlling for other issue report metrics. They show that affectiveness metrics have an impact on the issue fixing time: joy and love emotions typically have a lower fixing time, whereas sadness has a higher fixing time. They also found that the politeness of the last comment and the average sentiment expressed in the comments, reduce issue fixing time.

To identify emotions, Ortu et al. used an approach based on the machine learning classifiers developed in this paper. Since their sentiment and politeness tools required this, Ortu et al.'s models take individual sentences as input, whereas in our study the machine learning classifiers use entire comments as input. For example, Ortu et al. used four times 1,000 *sentences* as training data, while we used four times 400 *comments*. Whereas models at the sentence level might be more detailed, with different sentences exhibiting different emotions, lifting up the resulting set of emotions to the comment level might not be straightforware, while comment-level models are able to determine the overall emotion, possibly at the costs of subtle nuances within the comment. More work is necessary to determine the optimal granularity of models.

Furthermore, while we focus on a detailed overview of the construction of the machine learning classifiers, and the *evaluation* of their performance in comparison to a human oracle and ZeroR baseline, Ortu et al. mainly *used* the models to build an oracle from data that was not tagged by humans, in order to obtain an idea of the total set of occurrences of each emotion across the full data set.

Destefanis et al. (Destefanis et al 2016) and Ortu et al. (Ortu et al 2015b, 2016a, 2015c) empirically analysed the politeness and sentiment in software artifacts tracked by the Jira issue tracking system. Their results show that the level of politeness in the communication process among developers does have an effect on the time required to fix issues and, in the majority of the analysed projects, it had a positive correlation with attractiveness of the project to both active and potential developers.

Mäntylä et al. (Mäntylä et al 2016) approached affect in software engineering from the emotional dimensions of Valence, Arousal and Dominance (VAD), their results show that issue reports of different type (e.g., Feature Request vs. Bug) have a fair variation of Valence, while increase in issue priority (e.g., from Minor to Critical) typically increases Arousal. As an issue's resolution time increases, so does the arousal of the individual the issue is assigned to. Finally, the resolution of an issue increases valence, especially for the issue reporter and for quickly addressed issues.

Finally, as mentioned in the introduction, this paper is an extension of our previous work, in which we found that human raters can agree to a certain extent on emotions reported in issue comments (Murgia et al 2014).

9 Conclusion

Emotions influence human behavior and interactions. Software development, as a collaborative activity of developers, cannot be considered exempt from such influence. Emotion mining, applied to developer issue comments, can be useful to identify and monitor the spirit and atmosphere within the development team, allowing project leaders to anticipate and resolve potential threats in their team as well as discover and promote factors that bring serenity and productivity in the group.

In this paper, we evaluate the feasibility of a tool for automatic emotion mining. As a first step, we performed an exploratory study of developer emotions in almost 800 issue comments during software maintenance and evolution. Our study confirms that issue comments do express emotions towards design choices, maintenance activity or colleagues. Regarding agreement amongst human raters, we found that some emotions like love, sadness and (to some extent) joy are easier to agree on, but that additional context can cause doubt for raters, unless more raters are used.

As a second step, we created proof of concept machine learning classifiers to identify emotions love, joy and sadness in issue comments. This proof of concept is capable of identifying issue comments where humans can easily agree on the presence or absence of emotions love and sadness. We show that for love and to a certain extent also for joy and sadness it is feasible to automate emotion mining, which is further supported by application of the models in a follow-up study by Ortu et al. (Ortu et al 2015a). Our findings confirm that it is possible to exploit certain emotion-driving keywords like *thanks* or *sorry* for detecting the emotion content of issue comments.

Given the complexity of identifying emotions, more studies with human oracles are required as well as studies of the impact of emotions and other affects on software productivity and quality. Hence, follow-up studies of the four hypotheses considered by this paper as well as other hypotheses are welcome.

Author's Biographies

Alessandro Murgia is a senior data scientist at Sirris, Brussels and former Post Doctoral Researcher at the University of Antwerp (Department of Mathematics and Computer Science), Belgium. He received his PhD in Software Engineering from University of Cagliari, Italy in 2011. His main research interests are text analysis, refactoring and data mining.

Marco Ortu is a Post Doctoral Researcher at the Department if Electronic and Electrical Engineering (University of Cagliari). He received his PhD in Software Engineering from University of Cagliari, Italy in 2015. His main research interests are text analysis, data mining and micro patterns.

Parastou Tourani received the B.Sc. degree in software engineering from the Amirkabir University of Technlogy, Iran, 2003, and the Pd.D. degree in software engineering from Montreal Polytechnique, Canada, in 2016.

Bram Adams is an assistant professor at Polytechnique Montral (Canada). He obtained his PhD at the GH-SEL lab at Ghent University (Belgium), and was an adjunct assistant professor in the Software Analysis and Intelligence Lab at Queen's University (Canada). His research interests include software release engineering in general, as well as software integration and software build systems in particular. His work has been published at premier software engineering venues such as TSE, ICSE, FSE, ASE, EMSE, MSR and ICSME. In addition to co-organizing RELENG 2013 to 2015 (and the 1st IEEE SW Special Issue on Release Engineering), he co-organized the PLATE, ACP4IS, MUD and MISS workshops, and the MSR Vision 2020 Summer School. He is PC co-chair of SCAM 2013, SANER 2015, ICSME 2016 and MSR 2019.

Serge Demeyer is a professor at the University of Antwerp and the spokesperson for the ANSYMO (Antwerp System Modelling) research group. He directs a research lab investigating the theme of Software Reengineering (LORE - Lab On REengineering). Serge Demeyer is a spokesperson for the NEXOR interdisciplinary research consortium and an affiliated member of the Flanders Make Research Centre. In 2007 he received a Best Teachers Award from the Faculty of Sciences at the University of Antwerp and is still very active in all matters related to teaching quality.

His main research interest concerns software evolution, more specifically, how to strike the right balance between reliability (striving for perfection) and agility (optimising for improvements). He is an active member of the corresponding international research communities, serving in various conference organization and program committees. He has written a book entitled "Object-Oriented Reengineering" and edited a book on "Software Evolution". He also authored numerous peer reviewed articles, many of them in top conferences and journals.

Appendices

Appendix A - Email sent to raters

To ensure that participants understand emotions, yet are not biased during the labeling process, we provide a minimal and dry training. Once they accepted to participate in an "ongoing experiment", we sent them an email to clarify the goal of the experiment. The participants were not aware of how many other participants were involved in the experiment, nor about the underlying goals. All the experiments were carried out via Google Spreadsheets. Here follows the email we sent to participants.

Dear XXXXX,

We are performing an experiment on emotions in bug reports, and we would like you to participate in this experiment.

We have created a dataset containing bug report comments by real open source developers. Your task would be to label these comments using a mixture of 6 emotions: Love, Joy, Sadness, Fear, Anger or Surprise. If no emotion can be observed, then the comment automatically is labeled as Neutral.

Attached to this mail, you can find a document that describes the 6 emotions that we use for the experiment. Moreover, it provides some examples of emotion labeling. Please take a look.

Following the link: XXXXX

You get access to a spreadsheet with 2 pages:

> ExampleLabeling: describes an example on how to label text comments. If you think an emotion can be observed in the comment, there will be an x in the corresponding cell. Multiple cells can be selected if multiple emotions are present. Absence of any x means that that comment is Neutral. You have to label only the emotions in the comment reported in the red-highlighted column (Comment N). The other comments (Comment N-1, until Comment 1), if available, are the preceding comments of an issue report, meant to explain the context of Comment N.

> Round1-SpreadsheetX: this document contains the comments that you have to label in Round 1.

The deadline for the results of round 1 are due XXXXX. Thanks again for participating and for returning your results on time!

Appendix B - Analysis of Full Study Excluding the Authors

To assess the impact of the learning effect between the pilot study and the full study in the ratings made by the first four authors, this appendix analyzes the results of the full study by removing the ratings from the first four authors. As specified in Section 7, we re-analyze RQ1 and RQ2 using a dataset of 210 comments labeled by the 12 raters not involved in the pilot case (i.e., excluding the authors). To simplify the comparison, Table 19 maps the tables reported in the original case study results to the new ones.

Note that in Tables 20, 21 and 22 the confidence interval is ± 7% instead of the ± 5% used in Tables 9, 10 and 12 . This is due to the fact that the sample used for the tables in this appendix only consists of 210 commits.

Table 19 Mapping between appendix and case study tables.

Appendix Table	Original Table
20	9
21	10
22	12
23	13
24	14

Table 20 Percentage of agreement out of 210 comments (absolute number in parentheses) and Cohen κ values (with confidence intervals) for each emotion in RQ1 — Rater's Agreement(full study).

(a) (round one, group A)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	89.05±7 (187)	7.62±7 (16)	81.43±7 (171)	0.35	0.52	0.69
joy	85.71±7 (180)	1.90±7 (4)	83.81±7 (176)	-0.04	0.14	0.32
surprise	87.62±7 (184)	2.38±7 (5)	85.24±7 (179)	0.01	0.21	0.41
anger	90.48±7 (190)	1.43±7 (3)	89.05±7 (187)	-0.04	0.18	0.40
sadness	94.76±7 (199)	1.43±7 (3)	93.33±7 (196)	0.03	0.33	0.62
fear	90.95±7 (191)	1.90±7 (4)	89.05±7 (187)	0.03	0.26	0.48

(b) (round one, group B)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	90.48±7 (190)	5.24±7 (11)	85.24±7 (179)	0.28	0.47	0.67
joy	78.09±7 (164)	2.38±7 (5)	75.71±7 (159)	-0.10	0.05	0.20
surprise	90.00±7 (189)	0.95±7 (2)	89.05±7 (187)	-0.10	0.11	0.31
anger	89.52±7 (188)	0.00±7 (0)	89.52±7 (188)	-0.08	-0.05	-0.03
sadness	91.43±7 (192)	3.33±7 (7)	88.10±7 (185)	0.17	0.39	0.62
fear	86.19±7 (181)	1.90±7 (4)	84.29±7 (177)	-0.05	0.14	0.33

(c) (round two, group A)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	82.86±7 (174)	4.76±7 (10)	78.10±7 (164)	0.11	0.28	0.44
joy	78.09±7 (164)	3.33±7 (7)	74.76±7 (157)	-0.02	0.13	0.27
surprise	88.10±7 (185)	0.00±7 (0)	88.10±7 (185)	-0.07	-0.05	-0.02
anger	91.43±7 (192)	0.95±7 (2)	90.48±7 (190)	-0.09	0.14	0.36
sadness	92.85±7 (195)	1.90±7 (4)	90.95±7 (191)	0.05	0.31	0.57
fear	92.85±7 (195)	0.95±7 (2)	91.90±7 (193)	-0.06	0.18	0.43

(d) (round two, group B)

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	91.91±7 (193)	3.81±7 (8)	88.10±7 (185)	0.23	0.44	0.66
joy	78.57±7 (165)	3.81±7 (8)	74.76±7 (157)	-0.02	0.14	0.30
surprise	90.00±7 (189)	0.48±7 (1)	89.52±7 (188)	-0.12	0.04	0.20
anger	93.33±7 (196)	0.00±7 (0)	93.33±7 (196)	-0.05	-0.03	-0.02
sadness	90.48±7 (190)	2.86±7 (6)	87.62±7 (184)	0.11	0.33	0.55
fear	91.42±7 (192)	1.90±7 (4)	89.52±7 (188)	0.03	0.27	0.50

Table 21 Percentage of agreement out of 210 comments between four (2nd column) and at least three (3rd column) raters, together with the Fleiss κ inter-rater agreement and confidence intervals (4th to 6th column).

(a) (round 1, group A) and (round 2, group B)

	%agreement of 4 (#)	%agreement of $\geq$3 (#)	lower κ	Fleiss κ	upper κ
`love`	83.81±7 (8)	94.29±7 (13)	0.50	0.52	0.53
`joy`	66.67±7 (1)	93.81±7 (7)	0.13	0.15	0.16
`surprise`	76.67±7 (0)	96.19±7 (1)	0.04	0.06	0.07
`anger`	84.76±7 (0)	96.19±7 (0)	0.08	0.09	0.10
`sadness`	86.19±7 (2)	96.19±7 (4)	0.30	0.31	0.33
`fear`	83.81±7 (1)	97.62±7 (6)	0.27	0.28	0.30

(b) (round 1, group B) and (round 2, group A)

	%agreement of 4 (#)	%agreement of $\geq$3 (#)	lower κ	Fleiss κ	upper κ
`love`	76.67±7 (3)	93.81±7 (15)	0.37	0.38	0.40
`joy`	61.90±7 (2)	89.52±7 (5)	0.11	0.13	0.14
`surprise`	79.05±7 (0)	98.10±7 (1)	0.02	0.04	0.05
`anger`	81.43±7 (0)	97.62±7 (0)	0.01	0.02	0.04
`sadness`	84.76±7 (3)	97.62±7 (6)	0.33	0.34	0.36
`fear`	80.48±7 (1)	97.62±7 (4)	0.17	0.18	0.20

Table 22 Percentage of agreement and Cohen κ values (with confidence intervals) for comments without and with context (RQ2 — Context Influence). The percentages are relative to the 210 comments without and with context, respectively odd and even rows.

(a) Group A

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	88.57±7 (186)	7.62±7 (16)	80.95±7 (170)	0.33	0.51	0.68
	83.33±7 (175)	4.76±7 (10)	78.57±7 (165)	0.11	0.28	0.45
joy	84.29±7 (177)	1.90±7 (4)	82.38±7 (173)	-0.05	0.12	0.29
	79.52±7 (167)	3.33±7 (7)	76.19±7 (160)	-0.01	0.15	0.30
surprise	88.57±7 (186)	2.38±7 (5)	86.19±7 (181)	0.02	0.23	0.44
	87.14±7 (183)	0.00±7 (0)	87.14±7 (183)	-0.07	-0.04	-0.01
anger	91.43±7 (192)	1.43±7 (3)	90.00±7 (189)	-0.03	0.21	0.44
	90.48±7 (190)	0.95±7 (2)	89.52±7 (188)	-0.08	0.12	0.33
sadness	95.71±7 (201)	1.43±7 (3)	94.29±7 (198)	0.06	0.38	0.70
	91.90±7 (193)	1.90±7 (4)	90.00±7 (189)	0.03	0.28	0.52
fear	92.86±7 (195)	1.90±7 (4)	90.95±7 (191)	0.06	0.31	0.57
	90.95±7 (191)	0.95±7 (2)	90.00±7 (189)	-0.06	0.14	0.35

(b) Group B

	%agreement (#)	%agreed presence (#)	%agreed absence (#)	lower κ	Cohen κ	upper κ
love	91.90±7 (193)	3.81±7 (8)	88.10±7 (185)	0.23	0.44	0.66
	90.48±7 (190)	5.24±7 (11)	85.24±7 (179)	0.28	0.47	0.67
joy	79.52±7 (167)	4.29±7 (9)	75.24±7 (158)	0.01	0.18	0.34
	77.14±7 (162)	1.90±7 (4)	75.24±7 (158)	-0.13	0.01	0.15
surprise	90.48±7 (190)	0.48±7 (1)	90.00±7 (189)	-0.13	0.04	0.21
	89.52±7 (188)	0.95±7 (2)	88.57±7 (186)	-0.10	0.10	0.29
anger	93.33±7 (196)	0.00±7 (0)	93.33±7 (196)	-0.05	-0.03	-0.02
	89.52±7 (188)	0.00±7 (0)	89.52±7 (188)	-0.08	-0.05	-0.03
sadness	90.48±7 (190)	2.86±7 (6)	87.62±7 (184)	0.11	0.33	0.55
	91.43±7 (192)	3.33±7 (7)	88.10±7 (185)	0.17	0.39	0.62
fear	90.48±7 (190)	1.90±7 (4)	88.57±7 (186)	0.01	0.24	0.46
	87.14±7 (183)	1.90±7 (4)	85.24±7 (179)	-0.04	0.16	0.35

Table 23 The number of times raters changed their rating from the rating in a row (comment without context) to the one in a column (comment with context). A "0" in a row or column for a particular emotion means that that emotion previously was not identified without (row) or with (column) context. A "1" in a row or column for a particular emotion means that that emotion previously was identified without (row) or with (column) context.

context		`love`		`joy`		`surprise`		`anger`		`sadness`		`fear`	
		0	1	0	1	0	1	0	1	0	1	0	1
A	0	344	20	347	32	370	16	381	15	388	17	384	13
	1	21	35	16	25	23	11	15	9	7	8	13	10
B	0	375	12	342	17	382	16	393	13	373	15	372	20
	1	3	30	22	39	12	10	5	9	15	17	13	15

Table 24 How often raters went from disagreement (d) to agreement (a) or vice versa when comparing the set of comments without context (rows) to the set of comments with context (columns), for groups A, B, and when combining both groups (at least three raters agreeing).

		love		joy		surprise		anger		sadness		fear	
		d	a	d	a	d	a	d	a	d	a	d	a
group A	d	14	10	17	16	6	18	8	10	3	6	5	10
	a	21	165	29	151	21	165	12	180	14	187	14	181
group B	d	11	6	28	15	9	11	9	5	8	12	9	11
	a	9	184	20	147	13	177	13	183	10	180	18	172
3-rater	d	6	6	5	6	1	6	2	5	2	4	0	5
	a	7	191	19	180	4	199	4	199	5	199	5	200

Acknowledgements This work was sponsored by (a) the Institute for the Promotion of Innovation through Science and Technology in Flanders by means of a project entitled Change-centric Quality Assurance (CHAQ) with number 120028, as well as (b) the Regione Autonoma della Sardegna (RAS), Regional Law No. 7-2007, project CRP-17938, "LEAN 2.0".

References

Ahmed T, Srivastava A (2017) Understanding and evaluating the behavior of technical users. a study of developer interaction at stackoverflow. Human-centric Computing and Information Sciences 7(1):8

Amabile TM, Barsade SG, Mueller JS, Staw BM (2005) Affect and Creativity at Work. Administrative Science Quarterly 50(3):367–403, DOI 10.2307/30037208

Aman S, Szpakowicz S (2007) Identifying expressions of emotion in text. In: 10th International Conference on Text, Speech and Dialogue (TSD), Springer, pp 196–205

Ambler S (2002) "Agile modeling: effective practices for extreme programming and the unified process". John Wiley & Sons, Inc. New York

Bacchelli A, Lanza M, Robbes R (2010) Linking e-mails and source code artifacts. In: Proceedings of the International Conference on Software Engineering (ICSE), pp 375–384

Bacchelli A, Sasso TD, D'Ambros M, Lanza M (2012) Content classification of development emails. In: Proceedings of the International Conference on Software Engineering (ICSE), pp 375–385

Balabantaray R, Mohammad M, Sharma N (2012) Multi-class twitter emotion classification: A new approach. International Journal of Applied Information Systems 4(1):48–53

Bazelli B, Hindle A, Stroulia E (2013) On the personality traits of stackoverflow users. In: Software Maintenance (ICSM), International Conference on, pp 460–463, DOI 10.1109/ICSM.2013.72

Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. Journal of Computational Science 2(1):1–8

Brodkin J (2013) Linus Torvalds defends his right to shame Linux kernel developers. http://www.webcitation.org/6O2zErgzE

Brooks FP Jr (1987) No Silver Bullet Essence and Accidents of Software Engineering. Computer 20(4):10–19

Campbell DT, Stanley JC (1963) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin

Cataldi M, Ballatore A, Tiddi I, Aufaure MA (2013) Good location, terrible food: detecting feature sentiment in user-generated reviews. Social Netw Analys Mining 3(4):1149–1163

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and psychological measurement 20(1):37–46

Das SR, Chen MY (2007) Yahoo! for amazon: Sentiment extraction from small talk on the web. Management Science 53(9):1375–1388, URL `http://EconPapers.repec.org/RePEc:inm:ormnsc:v:53:y:2007:i:9:p:1375-1388`

De Choudhury M, Counts S (2013) Understanding affect in the workplace via social media. In: Proceedings of the Conference on Computer Supported Cooperative Work, ACM, New York, NY, USA, pp 303–316, DOI 10.1145/2441776.2441812, URL `http://doi.acm.org/10.1145/2441776.2441812`

DeMarco T, Lister T (1999) Peopleware: Productive Projects and Teams. Dorset House Publishing Co. 2nd Edition, Inc., New York, NY, USA

Destefanis G, Marco O, Steve C, Steve S, Michele M, Roberto T (2016) Software development: Do good manners matter? PeerJ CompSci

Heritage Dictionary A (2005) The American Heritage science dictionary. http://dictionary.reference.com/browse/, URL `http://dictionary.reference.com/browse/`

Elfenbein HA, Ambady N (2002) On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological bulletin 128(2):203

Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56(4):82–89, DOI 10.1145/2436256.2436274, URL http://doi.acm.org/10.1145/2436256.2436274

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological bulletin 76(5):378

Fowler JH, Christakis NA (2008) Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. BMJ 337, DOI 10.1136/bmj.a2338

Fredrickson BL (2001) The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. American psychologist 56(3):218

Fritz T, Müller S (2016) Leveraging biometric data to boost software developer productivity. In: International Conference on Software Analysis, Evolution and Reengeneering (Future of Software Engineering Track), s.n.

Gold J (2015) A prominent linux kernel developer is stepping down from her direct work in the kernel community. http://www.networkworld.com/article/2988850/opensource-subnet/linux-kernel-dev-sarah-sharp-quits-citing-brutal-communications-style.html

Graziotin D, Wang X, Abrahamsson P (2014) Happy software developers solve problems better: psychological measurements in empirical software engineering. PeerJ p e289, DOI 10.7717/peerj.289, URL http://dx.doi.org/10.7717/peerj.289

Guillory J, Spiegel J, Drislane M, Weiss B, Donner W, Hancock J (2011) Upset now?: emotion contagion in distributed groups. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI), pp 745–748

Guzman E, Bruegge B (2013) Towards emotional awareness in software development teams. In: Proceedings of the Joint Meeting on Foundations of Software Engineering (ESEC/FSE), pp 671–674

Guzman E, Azócar D, Li Y (2014) Sentiment analysis of commit comments in github: An empirical study. In: Proceedings of the Working Conference on Mining Software Repositories (MSR), ACM, New York, NY, USA, MSR 2014, pp 352–355, DOI 10.1145/2597073.2597118, URL http://doi.acm.org/10.1145/2597073.2597118

Guzzi A, Bacchelli A, Lanza M, Pinzger M, van Deursen A (2013) Communication in open source software development mailing lists. In: Proceedings of the Working Conference on Mining Software Repositories (MSR), pp 277–286

Hancock JT, Gee K, Ciaccio K, Lin JMH (2008) I'm sad you're sad: emotional contagion in CMC. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW), pp 295–298

Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '04, pp 168–177, DOI 10.1145/1014052.1014073, URL http://doi.acm.org/10.1145/1014052.1014073

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

Mäntylä M, Adams B, Destefanis G, Graziotin D, Ortu M (2016) Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity? In: Proceedings of the 13th International Workshop on Mining Software Repositories, ACM, pp 247–258

Mitchell TM (1997) Machine Learning, 1st edn. McGraw-Hill, Inc., New York, NY, USA

Murgia A, Tourani P, Adams B, Ortu M (2014) Do developers feel emotions? an exploratory analysis of emotions in software artifacts. In: Proceedings of the Working Conference on Mining Software Repositories (MSR), ACM, pp 262–271

Nagappan M, Zimmermann T, Bird C (2013) Diversity in software engineering research. In: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, ACM, New York, NY, USA, ESEC/FSE 2013, pp 466–476, DOI 10.1145/2491411.2491415, URL http://doi.acm.org/10.1145/2491411.2491415

Ortu M, Adams B, Destefanis G, Tourani P, Marchesi M, Tonelli R (2015a) Are bullies more productive? empirical study of affectiveness vs. issue fixing time. In: Proceedings of the Working Conference on Mining Software Repositories (MSR), Florence, Italy

Ortu M, Destefanis G, Kassab M, Counsell S, Marchesi M, Tonelli R (2015b) Would you mind fixing this issue? In: International Conference on Agile Software Development,

Springer, pp 129–140
Ortu M, Destefanis G, Kassab M, Marchesi M (2015c) Measuring and understanding the effectiveness of jira developers communities. In: Proceedings of the Sixth International Workshop on Emerging Trends in Software Metrics, IEEE Press, pp 3–10
Ortu M, Destefanis G, Counsell S, Swift S, Tonelli R, Marchesi M (2016a) Arsonists or firefighters? affectiveness in agile software development. In: International Conference on Agile Software Development, Springer, pp 144–155
Ortu M, Murgia A, Destefanis G, Tourani P, Tonelli R, Marchesi M, Adams B (2016b) The emotional side of software developers in jira. In: Proceedings of the 13th International Conference on Mining Software Repositories, ACM, New York, NY, USA, MSR '16, pp 480–483, DOI 10.1145/2901739.2903505, URL http://doi.acm.org/10.1145/2901739.2903505
Ortu M, Murgia A, Destefanis G, Tourani P, Tonelli R, Marchesi M, Adams B (2016c) The emotional side of software developers in jira. In: Proceedings of the 13th International Conference on Mining Software Repositories, ACM, pp 480–483
Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Chair) NCC, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA), Valletta, Malta
Pang B, Lee L (2008) Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135
Parrott W (2001) Emotions in Social Psychology. Psychology Press
Piller C (1999) Everyone is a critic in cyberspace. Los Angeles Times 3(12):A1
Plutchik R (2001) The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist 89(4):344–350
Plutchik R, Van Praag H (1989) The measurement of suicidality, aggressivity and impulsivity. Progress in Neuro-Psychopharmacology and Biological Psychiatry 13:S23–S34
Rigby PC, Hassan AE (2007) What Can OSS Mailing Lists Tell Us? A Preliminary Psychometric Text Analysis of the Apache Developer Mailing List. In: Proceedings of the Working Conference on Mining Software Repositories (MSR), pp 23–
Robinson MD (2004) Personality as Performance Categorization Tendencies and Their Correlates. Current Directions in Psychological Science 13(3):127–129
Sehgal V, Song C (2007) Sops: Stock prediction using web sentiment. In: Proceedings of the International Conference on Data Mining Workshops (ICDMW), IEEE Computer Society, Washington, DC, USA, pp 21–26
semotion (2016) The First International Workshop on Emotion Awareness in Software Engineering, ICSE 2016 Workshop, Austin, Texas (USA)
Shivhare SN, Khethawat S (2012) Emotion detection from text. Computer Science, Engineering and Applications
Strapparava C, Valitutti A, et al (2004) Wordnet affect: an affective extension of wordnet. In: LREC, vol 4, pp 1083–1086
Tepperman J, Traum D, Narayanan SS (2006) "yeah right": Sarcasm recognition for spoken dialogue systems. In: Proceedings of InterSpeech, pp 1838–1841
Tourani P, Adams B (2016) 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, pp 189–200
Tourani P, Jiang Y, Adams B (2014) 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, pp 34–44
Witten IH, Frank E, Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann