Taking the emotional pulse of software engineering — A systematic literature review of empirical studies

Mary Sánchez-Gordón[a*], Ricardo Colomo-Palacios[a]
[a] Østfold University College, Faculty of Computer Sciences, Norway
mary.sanchez-gordon@hiof.no, ricardo.colomo-palacios@hiof.no

Abstract

Context: Over the past 50 years of Software Engineering, numerous studies have acknowledged the importance of human factors. However, software developers' emotions are still an area under investigation and debate that is gaining relevance in the software industry.

Objective: In this study, a systematic literature review (SLR) was carried out to identify, evaluate, and synthesize research published concerning software developers' emotions as well as the measures used to assess its existence.

Method: By searching five major bibliographic databases, authors identified 7172 articles related to emotions in Software Engineering. We selected 66 of these papers as primary studies. Then, they were analyzed in order to find empirical evidence of the intersection of emotions and software engineering.

Results: Studies report a total of 40 discrete emotions but the most frequent were: *anger, fear, disgust, sadness, joy, love, and happiness*. There are also 2 different dimensional approaches and 10 datasets related to this topic which are publicly available on the Web. The findings also showed that self-reported mood instruments (e.g., SAM, PANAS), physiological measures (e.g., heart rate, perspiration) or behavioral measures (e.g., keyboard use) are the least reported tools, although, there is a recognized intrinsic problem with the accuracy of current state of the art sentiment analysis tools. Moreover, most of the studies used software practitioners and/or datasets from industrial context as subjects.

Conclusions: The study of emotions has received a growing attention from the research community in the recent years, but the management of emotions has always been challenging in practice. Although it can be said that this field is not mature enough yet, our results provide a holistic view that will benefit researchers by providing the latest trends in this area and identifying the corresponding research gaps.

Keywords: Systematic literature review, Social aspects of software development, Emotion, Affect, Mood, Behavioral software engineering

1 Introduction

Software Engineering (SE) is concerned with all aspects of software production. Therefore, SE is not just concerned with technical processes of software development, but also includes people and tools [1]. SE is inherently a social activity which involves a large amount of interaction [2–4], as software development team members often need to cooperate with each other [3] in different levels and tasks. Thus, software is a product of human activities that incorporates our social interaction and cognitive aspects [5] which also elicit emotions. Regarding problem solving capabilities and creativity, they are both valuable cognitive processing abilities that software engineers need to possess [6] in order to be competent in their tasks. Consequently, software is intensive in human capital [7,8] and some cognitive processes have been shown to be deeply linked to the affective states of individuals [6]. Even though software is rarely produced by a single person [3], characteristics such as human behavior [9] and affects [10] —emotions and moods— are always present in software work. In fact, recent researches has revealed evidence that software developers experience a wide range of emotions [11–15] throughout the rich ecosystem of communication channels [16]. However, it is worth noting that human factors (called human aspects in SE) such as satisfaction, motivation, affective commitment, and well-being are not affects per se —even happiness is considered a peripheral affect— but affective reactions of the individuals influence all of the human factors [10]. This raises a key question which will facilitate a better understanding of the current status of research and addresses further investigation: "How the scientific literature approach the investigation of the software developers' emotions through software development process?". In this literature study, we attempt to create a comprehensive view of the literature addressing the role of emotions in SE so that we look at emotions —discrete and dimensional approaches—, measures —self-reported moods instrument (e.g., SAM, PANAS), physiological measures (e.g., heart rate, perspiration) or behavioral measures (e.g., keyboard use)—, and tools —machine-learning-based and lexical-based. Thus, we focus only on empirically validated studies and explicitly exclude studies that did not approach the emotional side of software developers in order to focus our research.

To the best of our knowledge, there is not any published secondary study with these research objectives. Previous literature review studies have not studied the entire state of the art in a holistic manner yet. In fact, they have only focused on the definition of behavioral software engineering [17], the definition of happiness [18] or the role of personality in SE [19,20]. Therefore, this paper is aimed to fill this gap by conducting a Systematic Literature Review (SLR) on this topic. Our SLR included 66 primary studies published in 5 major bibliographic databases. Although the time period for the review was not limited, the earliest paper was published in 2005. We believe this study provides a relevant contribution for the field, because affective states seem to be little known or understood from a SE perspective. Although, it is worth noting that they have been a subject of other Computer Science disciplines, such as human–computer interaction (HCI) and computational intelligence, in particular, affective computing. Consequently, this study is valuable for both software practitioners and software researchers alike. For practitioners, summarizing the literature is valuable due to the large amount of academic literature from various

sources. For researchers, our study provides a good starting point for further research since we have identified several areas for promising future research in this area.

This paper is structured as follows. First, authors give background information about emotions and previous related work in Section 2. Next, we describe our methodology including our research goal and questions, search strategy, selection criteria, quality assessment, data extraction strategy and process in Section 3. After that, we introduce the results in Section 4, which we further discuss in Section 5. Finally, we present our conclusions and ideas for future work in Section 6. Authors also present a reproducibility package available as archived open data [21].

2 Background and related work

In the field of psychology, there is no consensus on a unique definition of emotion [22]. There are many and varied definitions in the emotion literature sources. However, in general, they comprise of the basic elements that make up the theoretical conceptualization of the construct. Despite the fact that a deeper debate about it is out of the scope of this paper, we need some definitions in order to guide the review process. In this section, we first provide the definitions and then introduce previous literature studies that are related to the subject.

2.1 Concepts and definitions

There are many and varied definitions of emotions in the literature. According to the study carried out by Kleinginna and Kleinginna [23], there are 92 definitions and nine skeptical statements. Based on that compilation, they suggest a formal definition of emotion as "a complex set of interactions among subjective and objective factors, mediated by neural and hormonal systems, which can (i) give rise to affective experiences such as feelings of arousal, pleasure and displeasure; (ii) generate cognitive processes such as, emotionally relevant perceptual affect, appraisals, and labeling processes; (iii) active widespread physiological adjustments to the arousing conditions; and (iv) lead to behavior that is often, but not always expressive, goal-directed and adaptive". But, in general, the term is taken for granted in itself and is often defined with reference to a list: *anger, disgust, fear, joy, sadness,* and *surprise* [22]. Additionally, emotions as a type of affective state do have valence and intensity [24] on one hand, but on the other hand Scherer [25] stated that the general affective valence or preference should not be treated in the same manner as emotional episodes and it should not be more enduring than attitudes. The controversy continues with Shouse [26], who pointed out that emotions are the expression of affect and/or feelings, whereas Thoits [27] defined emotions as culturally determined types of feelings or affect.

In this context, Graziotin et al. [6] found that many authors have considered mood and emotion to be interchangeable terms, although, it has been also acknowledged that numerous attempts exist to differentiate these terms. Something similar happens with the terms emotion and affect, for instance the Picard's work on affective computing [28] have used them interchangeably. For

example, according to the APA dictionary of clinical psychology cited in [6], feelings have been defined as the conscious subjective experience of emotions, whereas affective states are defined as "any type of emotional state . . . often used in situations where emotions dominate the person's awareness". In turn, Shouse [26] defined affect as a "non-conscious experience of intensity: as a moment of unformed and unstructured potential" that "plays an important role in determining the relationship between our bodies, our environment, and others". Affect was also found as the most general of the terms by Batson et al. [24]. These authors described affect as more phylogenetically and ontologenetically primitive than emotions. But in a broader sense, affect has been thought of as an umbrella term for emotions, feelings, and sentiments [29].

Likewise, the link between emotions and sentiments is not always clear. Sentiments are defined by Gordon [30] as "socially constructed patterns of sensations, expressive gestures, and cultural meanings organized around a relationship to a social object, usually another person (…) or group such as a family". While Murray and Morgan [31] defined sentiment as "a more or less enduring disposition in a personality to respond with a positive or negative affect to a specified entity". A recent study [32] revealed that different terminologies are used according to the scientific communities involved and the targeted objectives. However, the actual studied phenomena overlap among themselves. The Affective Computing community is grounded in the definition of "emotion" provided by Scherer [25]. The natural language processing (NLP) community tends to use more frequently *opinion* and *sentiment*, whereas the embodied conversational agent (ECA) community tends to use *emotions*. According to [33], even with having clear definitions of these terms, there are still some controversial issues at the time of classifying some particular human states as an emotion. For instance, some researchers consider thankfulness or gratitude as an emotion, whereas others consider actions such as greeting, thanking, and congratulating as communicative functions [33]. For the purposes of our SLR, we use the terms *emotion*, *feelings*, *affect* and *affective states* interchangeably, in line with the findings of the primary studies.

On the other hand, there is not a universally accepted model of emotions. However, there are two prevalent approaches in this field: discrete approach and dimensional approach. The first one is based on a set of basic affects, which can be distinguished fundamentally from one another [12]. The second approach, on the contrary, describes each emotion as a point in a continuous multidimensional space where each dimension represents a quality of the emotion. This approach group affects in a smaller set of major dimensions (one or more) where one of them usually relates to intensity of emotions [34]. Such a dimensional approach allows a clear distinction among the dimensions and distinguishes itself from the discrete approach in its fewer elements to evaluate. The dimensions that are used most frequently *are valence, arousal and dominance* (thus the VAD acronym) although some authors refer to these dimensions with different names (e.g. pleasure instead of valence in [35] or activation instead of arousal in [36]). Emotions involve different components, but in the case of basic emotions, it is important to contextualize them into "families". It means that each basic emotion represents a "family" of closely related emotions [37,38]. For instance, the basic-emotion family of sadness would include emotions such as distress and anguish [37]. Moreover, basic emotions can be innate and universally recognized by humans

world-wide [38,39]. In fact, researchers of both sides have proposed lists of emotions that tend to be basic ones.

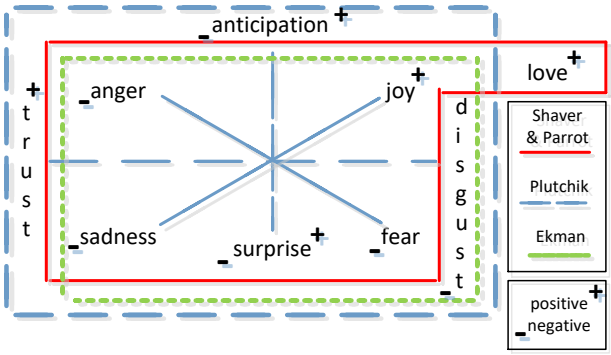

Figure 1. Four well-known emotion models adapted from [33].

Figure 1 shows four well-known emotion models. Ekman, one of the well-known emotion researchers, suggested that those certain emotions that are universally recognized form the set of basic emotions. Ekman et al. [40] in a cross-cultural study found six basic emotions which are *sadness, happiness, anger, fear, disgust, and surprise*. He later expanded his set of emotions by adding 12 new positive and negative emotions [41]. Shaver et al. [42] defined a tree-structured hierarchical classification of emotions where each level refines the granularity of the previous one, thus providing more indication on its nature. The basic level of the emotions hierarchy consists of *love, joy, anger, sadness, fear, and surprise*. Parrott [43] proposed a three layered categorization of emotion. In the first layer, he considered six primary emotions: *love, joy, surprise, anger, sadness, and fear*, followed by 25 secondary emotions in the next layer. In the last layer, more fine grained emotions were categorized. Plutchik [36] proposed an alternative viewpoint called the Wheel of Emotions in which he categorized eight basic emotions as pairs of opposite emotions: *joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation*. These eight basic emotions can vary in intensity and can be combined with one another to form secondary emotions. Apart from them, the "circumplex model", proposed by Posner et al. [44], represents emotions according to a bi-dimensional representation schema capturing the emotion valence (ranging from *pleasant* to *unpleasant*) and arousal (ranging from *calm* to *excited*). A third dimension, called "*dominance*" or "*control*" can be added to this space to signify the subjective feeling of control (*dominant* vs. *submissive*). Dominance can be understood as a social or cognitive interpretation of an affective event.

2.2 Measurement of emotions

Usually, the measurement of emotions has been carried out by the use of surveys. One of the most notable measurement instruments for affective states is the Positive and Negative Affect Schedule (PANAS) [45]. The PANAS is a 20-item survey that represents positive affects (PA) and negative affects (NA). Apart from that, some scales have been proposed to reduce the number of the PANAS scale items and overcome some of its shortcomings. Among these works, we must

mention the Scale of Positive and Negative Experience (SPANE) developed by Diener et al. [46]. The SPANE is a 12-item scale that is divided into two subscales (SPANE-P and SPANE-N) which assesses positive and negative affective states. There are also non-verbal assessment methods. One of the most used is Self-Assessment Manikin (SAM). SAM is based on pictures which measures *valence*, *arousal*, and *dominance* —i.e. *happy vs. unhappy, excited vs. calm, and controlled vs. in-control*— associated with a person's affective reaction to a stimulus [47]. Another one is the affect grid proposed by Russell [48]. It is a scale designed as a quick means of assessing affect along hedonic —*pleasure vs. displeasure*— and arousal —*sleepy vs. activated*— dimensions on a 1-9 scale. The Geneva Emotion Wheel (GEW) [49] is another grid that organizes 20 emotion words in a wheel-like format along valence —*unpleasant vs. pleasant*— and power —*low control vs. high control*— dimensions, with opposite points of the spikes of the wheel representing the intensity of the associated subjective feeling (distance from origin).

The behavioral measures include body expressions and measurement of voice [50]. Emotion influences bodily motions, such as gesture (facial scaling), posture, and keyboard and mouse movements. In order to capture these emotions, Microsoft Face API is a HTTP REST API which provides a number of resources related to both face detection and emotion detection —*anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise*— of the face itself [51]. EMOVoice identifies emotions —*joy, sadness, and anger*— by using acoustic voice analysis based on a number of specific physical measures of voice, such as pitch and intensity [52]. In turn, physiological measures require biometric sensors to measure the changes in the body caused by emotions, e.g., Neurosky MindBand sensor to capture electroencephalography (EEG) data, Empatica E3 wrist band to record skin- and heart-related signals, and Eye Tribe eye tracker to capture eye-related measures, such as pupil size [53].

A different approach is to identify affective states reported in sentences which is called sentiment analysis [34]. In its basic usage scenario, sentiment analysis is used to classify the *subjectivity* —neutral vs. emotionally loaded— and *polarity* —positive vs. negative— of a text. It relies on sentiment lexicons, which means large collections of words, each annotated with its own positive or negative polarity. The overall sentiment of a text is therefore computed upon the prior polarity of the contained words. Although, there have been some sentiment analysis tools such as SentiStrength, Alchemy, Stanford NLP sentiment analyser and NLTK [54], more recently, some customized sentiment analysis tools have appeared specifically within the SE field as well (e.g. SentiStrength-SE and SentiCR). On the automatic classification of texts, machine learning has been shown to be a promising approach to find links between low-level data capture (e.g. collected data from text and biometrics) and high level phenomena of interest (i.e. affects) [55]. Therefore, building classifiers to identify affects from a set of training dataset (i.e. gold standard) is known to be another approach.

2.3 Previous related works

No secondary study has yet been reported in the large scope of affects, moods and emotions in SE. Although previous literature studies did not approach this topic in a holistic manner, still a few

secondary studies in the more specific areas have been reported. We were able to identify two of these studies throughout the selection process of this SLR, in particular, by applying the exclusion criteria about secondary studies. Lenberg et al. [17] published in 2015 a definition of the Behavioral Software Engineering (BSE) research area and performed a SLR based on 55 related concepts. The definition emphasizes that BSE is the study of cognitive, behavioral and social aspects at different levels relating to the work of software engineers. The main result from the SLR indicates that there are some knowledge gaps in the existing BSE research. Moreover, that earlier research has been focused on a few concepts, which have been applied to a limited number of software engineering areas. In particular, personality and stress were two of the concepts included, but other concepts such as emotions, mood or affect are not included. Barros-Justo et al. [18] published in 2018 a relative small literature review that included 8 out of 619 studies. It summarizes the existing definitions of happiness from 4 studies as well as the metrics to assess its level within software engineers from the remaining studies. The authors concluded that further research is needed to consolidate our understanding about the relationships between happiness and software engineering.

Moreover, given that personality refers to individual differences among people's behavior, cognition, and emotion patterns [56], we reviewed two SLRs [19,20] on the influence of personality on individual performance or team work in programming. The review by Cruz et al. [19] included 42 publications between 1970 and 2010. The goal of the second review [20] was to increase the sensitivity of the first one by expanding the search string to include synonyms of the search terms and adding a ''snowball'' search strategy in the second stage of the search process. As a result, 19000 papers were found and 90 relevant papers were included in the review. Although these reviews present important results for the research in personality, they do not reveal any empirical evidence about emotions. Despite that fact, we recognized that linking emotions and personality can shed light on both: (i) personality may influence behavior indirectly via its influence on emotions and (ii) study of emotions too is enhanced through establishing its link to personality. In our SLR we have focused on emotions.

Finally, it is worth mentioning that there are a few SLRs about emotions in other areas such as marketing [57], healthcare [58] and music [59,60]. Therefore, the main contribution of our SLR is to identify and to understand how emotions have been investigated in the domain of SE and among software practitioners.

3 Methodology

This study was carried out following the guidelines given by Kitchenham and Charters [61], with one exception. Based on a data extraction form, we extracted data by qualitatively coding the selected articles as most of the papers contained only qualitative statements and little numerical data. In consequence, we adapted a SLR protocol to define the plan for the review. In this section, we present the research goal and questions, search strategy, filtering strategy, and data extraction

and synthesis. Additionally, we present the selected studies used as data sources and discuss their quality assessment. Figure 2 shows an overview of the research process used in our SLR.

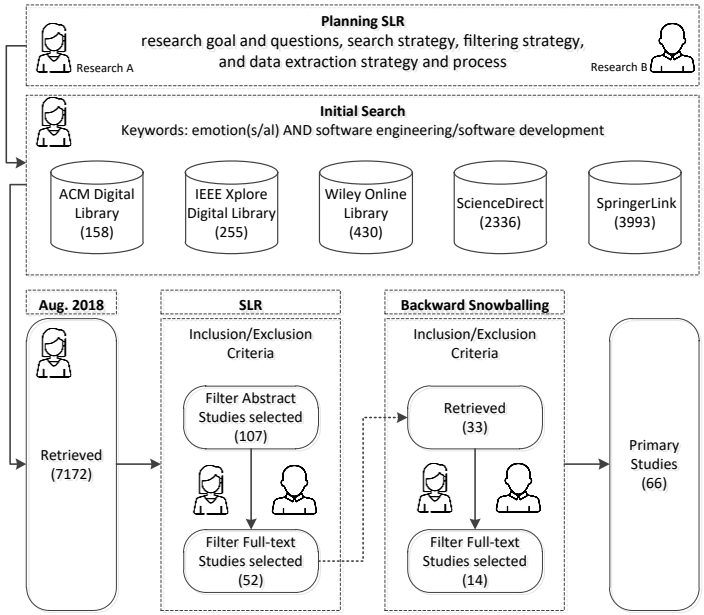

Figure 2. An overview of the research process used in this SLR.

3.1 Research goal and questions

The goal of this paper is to systematically review and synthesize the state of the art in the related area, to get an explicit view of software developers' emotions research including the recent trends and directions in this field to identify research gaps for future study. Therefore, this SLR is performed with the following three specific objectives in mind. First, we would like to understand the emotional facet of software developers through an empirical research in this area. Second, we would like to investigate and find out the most reported methods for emotion measurement. Finally, we would like to see if there is a growing interest in the field or not. Based on the mentioned goals, we raise the following review questions (RQs) grouped under two categories:

Group 1. Trends of primary studies:

- RQ1. What is the trend of studies related to developers' emotions that have been reported in major bibliographic databases? Knowing the types of publication, source and affiliation types of authors of the primary studies will enable us and the readers to get a high-level view of the research landscape.
- RQ2. What type of research methods are used in the studies? It is valuable to know and differentiate the types of research methods and the rigor used in different empirical studies so that readers can benefit from them.
- RQ3. What is the citation landscape of the studies in this area? Characterizing how the primary studies are cited by other papers gives us an idea of their impact and popularity.

The highly-cited papers in the area make it easy to pinpoint the most influential research in terms of subjects and time period.

Group 2. Specific to the domain (emotions in software engineering):

- RQ4. What are the software developers' emotions addressed or investigated that have been reported in the studies? Answering this question would provide practical and useful findings for practitioners and researchers alike. We were expecting to find discrete approaches and dimensional approaches.
- RQ5. How software developers' emotions are measured? Answering this question would provide practical and useful findings for practitioners. We were expecting to find self-report measures, physiological measures or behavioral measures as well as tools proposed and datasets freely available for download.

3.2 Search strategy

Given the exploratory nature of much of the research in this field, in order to identify the relevant studies, the two key terms used were: "emotion'' and "software engineering". The first one is related to the subject of the study. Two inflected forms of that term were also included "emotions" and "emotional". However, we recognize that some authors could overlook the term "emotion" and use only "mood", or "feeling", or "affect" when they study emotions. But, we also believe that it would be unexpected because our SLR is based on scientific literature and the controversy about the use of those terms is well-known in this research field and hence the term "emotion" should be mentioned in such literature. The second one was incorporated to exclude studies related to other fields that are not software engineering. In addition, an alternative search term was "software development". Boolean operators were used to construct the final search string. The "OR" operator was used to concatenate the related terms and "AND" to concatenate the two major key terms. As a result, the final search string applied to locate primary studies from databases was *("emotion'' OR "emotions" OR "emotional") AND ("software engineering" OR "software development")*.

However, it is worth noting that a trial search was conducted with a trial search string including a set of basic emotions —*"anger" OR "disgust" OR "sadness" OR "joy" OR "fear" OR "surprise"*— as another alternative search set of terms. From the five databases consulted, only ScienceDirect showed a growth in the amount of results retrieved. Thus, this database returned 6538 versus 2336 studies with this extended search string. Bearing in mind that it would demand us a lot of time and effort, we analyzed one percent of these studies. It was observed that *"surprise"* and "*fear*" are commonly used in the text (5447 studies), although they are not linked with the emotion research —e.g., "*... the fear of losing market share …"* or "*such a procedure is not surprise neutral*"—. Taking all these results together, we focus on the previous search string (without the list of basic emotions). Authors believe, it does not reduce the importance of the findings, but gives opportunity for further investigations.

3.2.1 Primary search process

The primary search phase for the relevant studies was performed on five major databases. Figure 2 shows the lists of databases consulted, all of them of scientific nature. Moreover, those databases have been suggested too by Kitchenham and Charters [61] and constituted the search engine typically used in Computer Science and Software Engineering SLR studies. The search string was successfully executed on all databases by the first author. The final search string was executed in August 2018 on each of the online databases. Moreover, the time period for the studies selected was not limited but, when available, we filtered results only to "Computer Science". The searches return a total number of 7172 results (see Figure 2).

The first author removed duplicated and totally unrelated papers from the results based on their title, keywords, and abstract. As a result, irrelevant studies were removed based on the inclusion/exclusion criteria (see more details in section 3.3). When those texts did not provide enough information to decide, other parts of the study was considered —i.e. conclusions and discussion if necessary— and the decision was made based on the inclusion/exclusion criteria. However, if the doubt remained, the paper was included in the relevant group, leaving the possibility to discard the paper during the next stage when the full texts of the papers were studied. At this point, the study published by Thomas Shaw [62] in 2004 was excluded because it was a work in progress (a little more than two pages), and no continuation of the project is currently known. However, we recognized that it deserves special mention due to the fact that he was one of the first to explore software developers' emotions. Until then, the focus was mainly on job outcomes such as turnover, burnout, and satisfaction [63]. Finally, a total number of 107 papers passed the criteria.

Next, a careful process was performed in order to ensure the accuracy and reliability of the final study selection. First, a reference manager was used by the first author to collect full-text versions of the papers. Second, texts from the full papers of all included studies (107) were imported into a data analysis tool. Third, each co-author accurately read each paper independently through the data analysis tool and made the decision whether or not to include it as a relevant paper based on the criteria discussed above. Then, the co-authors compared the relevant papers, and each of the differences was discussed after re-reading and re-analyzing the paper in order to find a shared attribution. Some of the studies were excluded because they did not include any empirical evidence, although, the abstracts had led us to think so. For instance, the term "case study" can mean a study of a real-world case, but in some papers it referred to proposals not used in real-life. Furthermore, the consistency and the validity of this primary study selection process are supported by the high value (81.47%) of Krippendorff's alpha (α) calculated after the data extraction process (see section 3.5). The mathematical formula of this alpha for "binary or dichotomous data, two observers, no missing data" is provided in [64]. The values produced by Krippendorff's alpha are between 1 (perfect agreement), 0 (units statistically unrelated) and -1 (perfect disagreement). Ideal values go around 80% to conclude that the validation of the inter-reliability process is acceptable.

Out of the 107 articles, 52 passed our exclusion criteria and, therefore, they were included in the data analysis process. Once the filtering strategy was applied and a paper was selected for inclusion, we used backward snowballing as Figure 2 shown.

3.2.2 Strategy for secondary search process

In the guidelines [61], it is recommended that the reference lists from the identified articles in the previous searches should be considered as well, in order to identify further relevant articles through the reference lists of the articles found using the search strings. Consequently, the secondary search phase included a backward snowballing process by screening of all the studies listed in the references section of the selected primary studies. This process helped us to find other relevant studies pertinent to our SLR as suggested by Jalali and Wohlin [65]. After analyzing all of the abstracts based on inclusion and exclusion criteria, there were 33 studies selected and after a full text analysis the final sample resulted in 14 primary studies (see Figure 2). That is, the same search process described in the previous section. Two examples of the papers found during snowballing are described in what follows. [PS65] was found by backward snowballing of [PS64] while [PS66] was found by backward snowballing of [PS57]. A similar backward snowballing process was carried out based on the previous related works [17,18] reported in section 2.3, however, no additional study was found.

Despite the effort made, there is a risk that some papers have been missed. Therefore, although this study cannot guarantee completeness, we believe that it can still be trusted to give a good overview of the relevant empirical studies on this field.

3.3 Selection criteria

The study selection criteria aim at identifying those studies that provided direct empirical evidence about our research questions [61]. Therefore, the selection of studies was conducted by applying a set of inclusion and exclusion criteria. The inclusion criterion was as follows:

- Domain: the main domain must be within software engineering, and the study should be focused on software developer's emotions. Thus, the study subjects should be related to the people in the software industry such as students or professionals.
- Method: empirical studies that use quantitative, qualitative, and mixed methods such as case studies or experiments, whether observed in the field or in a laboratory or classroom. In particular, sentiment analysis tools for SE that have been empirically evaluated are pertinent.
- Type: the type of study could be an article, conference paper, magazine article, or a book chapter.
- Language: studies written in English language only.

The exclusion criteria were:

- Method: non-empirical studies such as secondary studies and review studies, or studies that include the authors' personal views or assumptions without supporting data.

- Type: studies appear as work-in-progress, posters, or short papers containing less than 4 pages
- The full-text of the study cannot be accessed.

3.4 Quality assessment

Despite the fact that there is no agreed upon definition for the "quality of study", this assessment followed the quality checklist suggested by Kitchenham and Charters [61], and Kitchenham and Brereton [66]. A similar approach was used in an earlier SLR about personality [20]. The information focuses on biasing and validity issues related with the various phases of empirical studies, as well as the minimum information required to establish credibility. The quality questionnaire along with the phases is as follows:

- Design
    - Is there a clear statement of the aims of the research?
    - Are the emotions studied by empirically measured data?
    - Are the measures used in the study fully defined?
- Conduct
    - Does the paper provide relevant data related to the research topics?
- Analysis
    - How adequately is the research results documented? For example, participants or observational units.
- Conclusions
    - Does the study allow answering the research questions?

Each primary study was assessed for quality at the same time as the data extraction process was conducted. Each question was independently answered by each co-author according to the scores proposed by Kitchenham and Brereton [66]. The scores are as follows: Yes = 1, Partially = 0.5 and No = 0 points. In consequence, the maximum quality score for a primary study is 6. However, the scoring could provide only limited evidence of the actual value of the methodology. For instance, if the objective of the study was to undertake preliminary results, it could score well on the questionnaire, although, overall it could only be said to provide very limited evidence of the actual value of the methodology. Hence, we include a research question in our SLR about the research methods used in the primary studies. Additionally, we computed the Krippendorff's alpha (α) [64] for "nominal data, two observers, no missing data" in order to check the quality assessment results and demonstrate the process consistency. In our analysis, we found that the value of Krippendorff's alpha was 75.33%, so, we can conclude that data are interpreted in a similar and acceptable way among co-authors, since the value of alpha is around 80%. Finally, the discrepancies among the evaluations were discussed and a consensus was reached.

3.5 Data extraction strategy and process

Each author worked independently to extract data from all primary studies, guided by an extraction form. The full reference was gathered as provided by the libraries. This contains information about authors' names, title, conference/journal, year of publication, number of pages, keywords and abstract. As this information was automatically extracted, no inconsistencies in the extraction were found. The bibliographic details for all primary studies are available in Appendix A. Furthermore, the data extracted from all the primary studies is available as an archived open data [21]. In order to provide a structured approach for the review, we designed a data extraction form by considering the research questions as well as the quality questionnaire. Such a form was set according to the guidelines provided by Kitchenham and Charters [61]. We believe most of the attributes are self-explanatory (see the data extraction form fields in [21]), except for those addressed for emotions (RQ 4) and their measurement (RQ 5) which are explained in the background (sub-sections 2.1 and 2.2).

After reading the abstract, the first author obtained the full-text versions and kept them in a reference manager (Zotero 5.0.59). The 140 selected studies from primary (107) and secondary (33) search processes were imported from the reference manager tool to the data analysis tool (Nvivo 12.0). Each study was opened in turn and carefully read through NVivo by each co-author independently. When a study was classified as relevant, the full text was coded according to their content and the data extraction form. We measured how the co-authors agreed with each other on the primary study selection process by computing Krippendorff's alpha (α) for "binary or dichotomous data, two observers, no missing data" [64]. The inter-rater agreement test showed an agreement of 81.47% between the co-authors. This Krippendorff's alpha was calculated based on the data extraction form without the inclusion of the information that was automatically extracted and the quality questionnaire. Therefore, we can conclude that the validation of the inter-reliability process is acceptable because, as mentioned before, ideal values of Krippendorff's alpha go around 80%. Any differences in the extracted data were discussed in order to reach a consensus. This data analysis tool was useful in several ways. First, it allowed the key themes to emerge from the data. Secondly, we were able to gain a holistic understanding of the evidence base, as described in the next section. Lastly, the coding themes helped us to identify which papers would contribute to answer the research questions.

Once the data extraction process was completed, a separate soft copy of the extracted data was exported to a spreadsheet application (Microsoft Excel) for further analysis and an identity code serial number (i.e. data item ID) was formulated and assigned to it. Consequently, the primary studies are referred in the rest of this paper in that form: [PS01] - [PS66], including the appendixes. To ensure quality and validity of our results, a final meeting between co-authors was held to perform cross-checking of the extracted data.

4 Results

The first important result of our SLR is that only 66 publications of 7172 met our selection criteria which represent empirical studies on software developer's emotions. The extracted data from

these studies include both qualitative data (e.g. emotions or the measurement of emotions) and quantitative data (e.g. number of subjects involved or quality level). This section presents our findings grouped by research question and depicted using descriptive statistical tools (e.g. tables, pivot chart, simple bar chart, multiple bar chart, and pie chart).

4.1 Group 1: Trends of primary studies.

This section addresses research questions RQ1 to RQ3.

4.1.1 What is the trend of studies related to emotions in SE that has been reported in major bibliographic databases? (RQ1)

Given that a research paper abstract usually demonstrates the content of the research paper and should contain important keywords, an overview of the topics covered in the papers analyzed was obtained from their abstracts by generating a frequency table. Table 1 shows the frequency of these words. It was generated using NVivo 12 ("Word Frequency" query). As one can see, keywords such as analysis, sentiment, positive, negative, issue and comments are among the most popular besides the obvious ones related to software development, developers, emotion(s), engineering, emotional, team(s), and project(s).

Table 1. Popularity of the topics shown by the word frequencies of all paper abstracts.

Word	#	Word	#	Word	#	Word	#
software	55	sentiment	21	issue	13	role	11
results	34	positive	19	information	13	emotion	10
study	32	teams	18	factors	13	affect	10
analysis	30	project	17	projects	12	individual	10
development	28	negative	17	comments	12	developer	9
developers	27	data	15	tools	12	performance	8
emotions	27	work	15	communication	11	quality	8
engineering	26	team	15	studies	11	tasks	8
emotional	22	human	14	empirical	11	social	8

The total number of 66 primary studies that were included in this SLR (see list of studies in Appendix A), are distributed within different publishers as follows: 26 ACM [PS01, PS03–PS09, PS11–PS26, PS34, and PS36], 18 IEEE [PS10, PS27–PS33, PS35, and PS37–PS45], 8 Springer [PS46–PS49, PS55–PS57, and PS63], 10 Elsevier [PS50–PS54, PS58–PS62], 1 Wiley [PS64], and 3 Others [PS02, PS65, and PS66]. The last group corresponds to three journals found during the backward snowballing process. Next, we report a summary on the trends of the primary studies, based on the following aspects:

- Number of studies by publication type
- Number of studies over time by publisher (growth of attention in this area)

- Number of studies by affiliation types of the study authors
- Number of studies by quality level

The pie chart in the left of Figure 3 shows the number of primary studies by publication type divided into four categories. There were 27 conference papers (41%), 22 journal papers (33%), 14 workshop papers (21%) and 3 symposium papers (5%). Most of the studies (51 out of 66) are journal articles and conference papers.

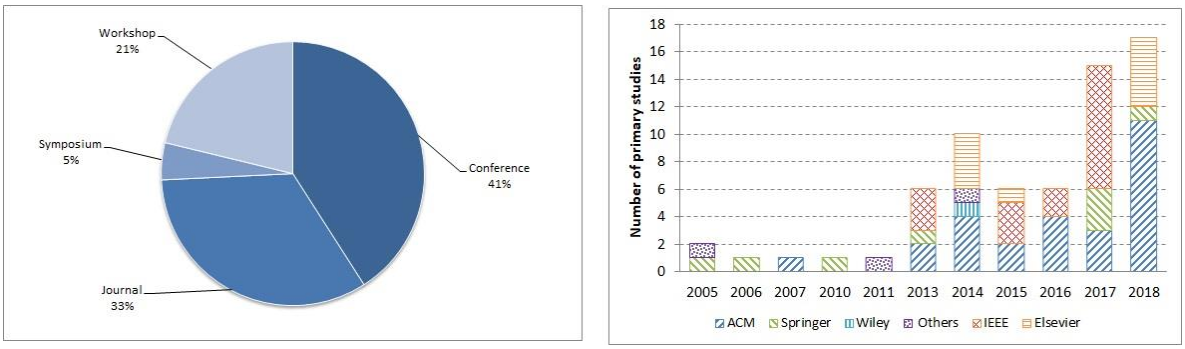

Figure 3. Proportions of studies by publication type (left) and Number of studies over time per publisher (right).

The bar chart in the right of Figure 3 (as a stack chart) shows the number of primary studies published by year. The oldest studies are from 2005 [PS46] (Springer) and [PS65] (Others), and the latest ones from 2018 [PS16]–[PS26] (ACM), [PS58]–[PS62] (Elsevier), and [PS63] (Springer). Just over 82% (54/66) of the primary studies were published after 2014. This shows that empirical research focusing on affective states is novel, despite the fact that human factors in SE have been acknowledged and researched since the 1970s, and in particular, research focusing on personality is much more recent, with the vast majority of the studies developed since 2002 [20]. As it was expected, the bar chart suggests a growing interest on the topic. Although we found sparse empirical evidence in this topic until 2011, researchers seem to have started to study the role of affective states in SE around the year 2013. However, their contributions on this topic were presented and discussed in diverse conferences and workshops. For instance, 10 primary studies were identified in 2014, out of which, 3 of them came from the International Conference on Mining Software Repositories (MSR) while 6 of them came from different journals. Moreover, the fact that our SLR is focused on empirical studies has probably reflected in the low amount of primary studies on the early days. An alternative explanation may be related to the workshop SEmotion. It was launched back in 2016 and its aim is to create an international forum for researchers and practitioners interested in the role of affect in SE. In fact, 35% of the primary studies were presented there in the last three years, denoting the increasing attention of researchers on the topic. The peak years in terms of the number of papers were years 2017 and 2018 in which 15 and 17 papers were published, respectively.

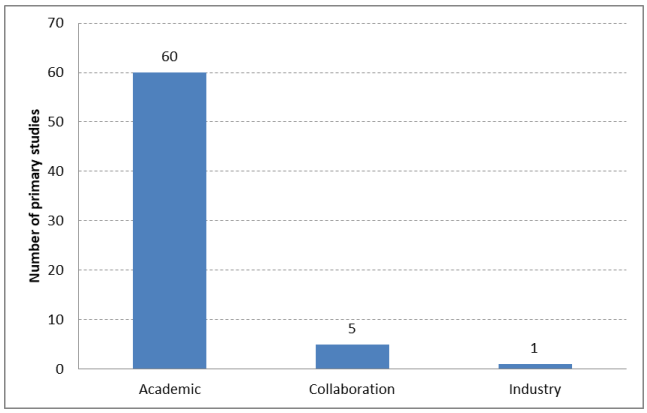

Figure 4. Number of studies by affiliation types of the study authors

Figure 4 depicts a bar chart detailing the number of primary studies by affiliation types of the study authors, classified as follows: (i) academic when a study is published solely by academic authors, (ii) industrial when a study is published solely by industry authors or (iii) collaborative when a mix of academic and industry authors had worked on it. As a result, 60 primary studies written by academic authors were the majority (91%). There were five collaborative works [PS17, PS30, PS33, PS46, and PS58], and only one industry paper [PS58]. Although the result is not surprising because, we focused on academic databases, and this fact reveals the need for more industry-academic collaborations in this area.

After assessing the selected studies based on the quality criteria detailed in section 3.4, a quality level was calculated, taking into account the range of values from 0 (poor) to 6 (very good). Studies equal to or below 3 should be excluded due to their low quality. However, there was no study having such a low quality score. Most of the studies (55 out of 66, 83%) were classified as "very good" quality based on the quality assessment while the remaining 11 papers ([PS15], [PS17], [PS21], [PS27], [PS31], [PS33], [PS34], [PS37], [PS40], [PS42], [PS43]) were ranked as "good" quality.

4.1.2 What type of research methods are used in the studies? (RQ2)

The rationale behind this question is to identify the research methods used in the primary studies, as well as the type of participants in each one. Table 2 shows an overview of the research methods from these perspectives grouped by the data source.

Table 2. Number of studies by research methods (left) and number of subjects in each study (right)

Data Source	Method Research	Number of Primary Studies					Number of Participants				
		Subjects of investigation			FREQ.		Subjects of investigation			FREQ.	
		Both	Student	Professionals	n	f (%)	Both	Student	Professionals	n	f (%)
Dataset	Case study		2	10	**12**	**18.18**		26		**26**	**0.98**
	Experiment		2	26	**28**	**42.42**		19	474	**493**	**18.55**
Other Sources	Case study			6	**6**	**9.09**			123	**123**	**4.63**
	Ethnography			1	**1**	**1.52**			5	**5**	**0.19**
	Experiment	5	7	2	**14**	**21.21**	231	700	98	**1029**	**38.71**
	Survey		1	3	**4**	**6.06**		72	590	**662**	**24.91**

	Survey/Eth.			1	**1**	**1.52**			320	**320**	**12.04**
FREQ.	**n**	**5**	**12**	**49**	**66**	**100.00**	**231**	**817**	**1610**	**2658**	**100.00**
	f (%)	**7.58**	**18.18**	**74.24**	**100.00**		**8.69**	**30.74**	**60.57**	**100.00**	
FREQ denotes frequency while *n* represents absolute frequency and *f* represents relative frequency											

The left of Table 2 presents the number of primary studies. As one can see, 60.60%, namely 18.18% plus 42.42% (40/66) of the studies used datasets but there were only 13 datasets publicly available on the Web (see Appendix C). In fact, there were three links that did not work at the moment of verification, Dec 2018. The remaining 39.40% of primary studies used other data sources such as self-assessment, biometrics, peripherals and interviews (see details in Figure 6). From this first perspective, 74.24% of the studies used professionals as subjects of investigation while 18.18% used students and 7.58% used both professionals and students. However, it is worth noting that 37 out of 40 studies that used datasets did not mention the number of participants. Therefore, on the right of the Table 2, the number of participants came from three studies: [PS58] (474 professionals), [PS31] and [PS34] (26 and 19 students, respectively). These three studies reported 19.53% of all participants while the remaining 80.47% were reported from studies that used other data sources. From this second perspective, 60.57% of participants were professionals, 30.74% were students and 8.69% were mixed groups of type of participants (professionals and students).

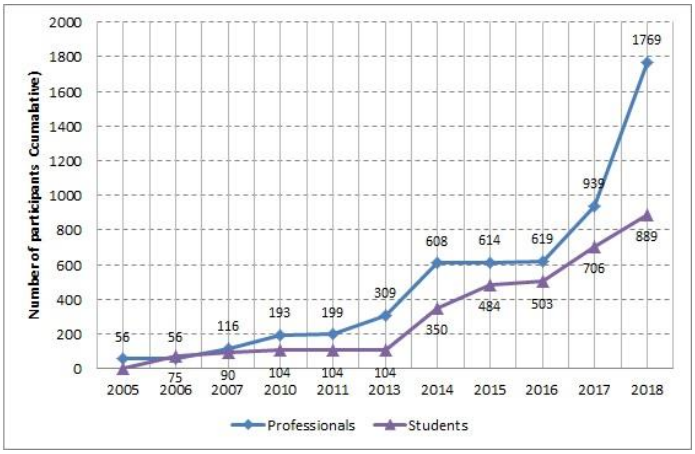

Figure 5. Cumulative number of participants in the studies by year

In a global context, the distribution between types of participants in the studies is not balanced: 67% (1769/2658) are professionals, while only 33% (889/2658) are students (see Figure 5). With regards to professionals, we found one study [PS58] using a dataset, which involved artifacts from 474 IBM Jazz practitioners, and two large surveys [PS44], [PS60]. In [PS44], their authors posted links of the survey on Reddit groups, Quora and in Computer Science Facebook groups; they also emailed it to software development mailing lists. As a result, 311 software developers answered the questionnaire. The authors also conducted an observational study (ethnographic) with 9 professional software developers to investigate the feasibility of predicting fatigue from interaction history. Therefore, we included 320 participants for [PS44]. In the other survey [PS60], the authors extracted a set of developer contacts from the GitHub Archive. Although 2220

individuals participated (7% response rate), only 1318 provided valid data for the open questions on causes and consequences of happiness and unhappiness. A total of 317 subjects provided answers with regards to what happens when developers are happy and unhappy while developing software. Thus, [PS60] is the most large-scale quantitative and qualitative survey of software developers on this topic, and the complete results are archived as open data[1]. When the data from those three studies (1111 professionals) is removed from the total amount, the distribution between types of participants in the studies is more balanced: 57% (889/1547) use students and 43% (658/1547) use professionals.

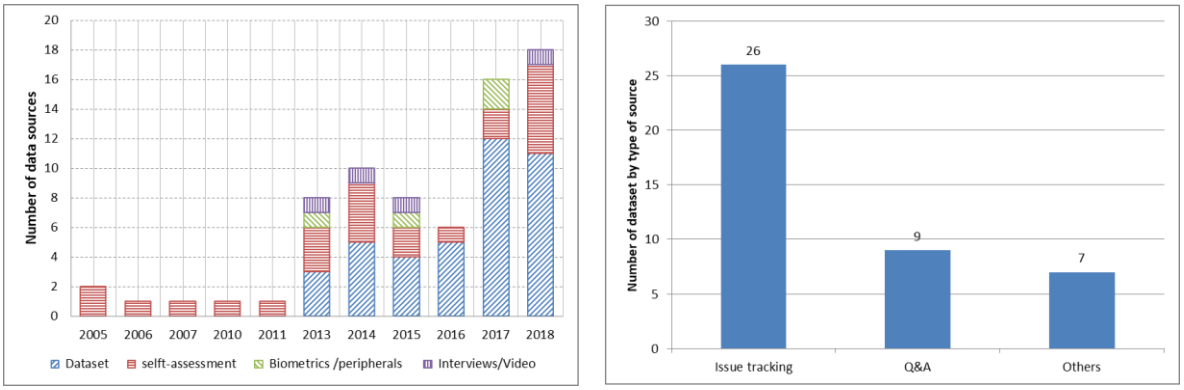

Figure 6. Number of data sources over time (left) and datasets by type of source (right).

On the left of Figure 6 the number of data sources per year is expressed. As mentioned before, a large portion of studies (60%) is using datasets which may indicate not only the interest in the topic but also the difficulty in using data collected by means like biometrics, peripherals (i.e. mouse/keyboard) and interviews. The second most used category is self-assessment (24 of 66, 36%) which is composed entirely of questionnaires, with the exception of the ethnographic study [PS33] that use a notes template. The right of Figure 6 presents the number of datasets by type of source (see details in Appendix B). The majority of datasets (26 of 42, 61.9%) were extracted from issue tracking tools (GitHub, Jira, Bugzilla, SourceForge and bug report from Eclipse, Android and JBoss), followed by Q&A sites (9 of 42, 21.4%) (Stack Overflow, Piazza, Serebro), and others types of tools (7 of 42, 16.7%) such as code review (Gerrit), content collaboration tools (Confluence) and microbloging (Twitter). However, it is worth noting that two studies [PS26], [PS56] used two different sources (Q&A and issue tracking), so that there are 42 sources for 40 studies.

In the last five years, a growing interest in qualitative research methods is revealed by the use of a coding strategy. In fact, the research design of 37.87% of the primary studies (25/66) adopted this approach. Hence there were people (115) involved as coders/raters to categorize emotions. The majority of those studies (19/25) used datasets and 99 coders/raters were involved. The remaining studies used other data sources ([PS29], [PS30], [PS33], [PS52] [PS60] and [PS62]) and 16 coders were involved.

[1] https://figshare.com/collections/Online_appendix_the_happiness_of_software_developers/3355707

4.1.3 What is the citation landscape of the primary studies? (RQ3)

To characterize how the primary studies are cited by other papers, we extracted the citation data from Google Scholar on Dec. 18, 2018. We performed the citation analysis in a similar way to a recent study about the top-100 highly-cited SE papers in SE [67] in which the authors proposed two metrics: (1) the absolute number of citations to each paper, and (2) normalized citations (i.e. average number of citations per year).

Figure 7 depicts the citation landscape of the 66 primary studies from both perspectives: on the left of the figure, the first metric and on the right of the figure the second one. The average values for the two metric values were 24.46 and 7.86, respectively. It means that the papers in this area are reasonably cited. Another interesting point is that more recent papers have higher citations in terms of normalized citations. Furthermore, 9% of the papers (6/66) had no citations at all, but it is worth noting that 2 of them were published in 2017 (2/15) while the remaining in 2018 (6/17). That all gives us an idea of the impact and popularity of the primary studies in this SLR.

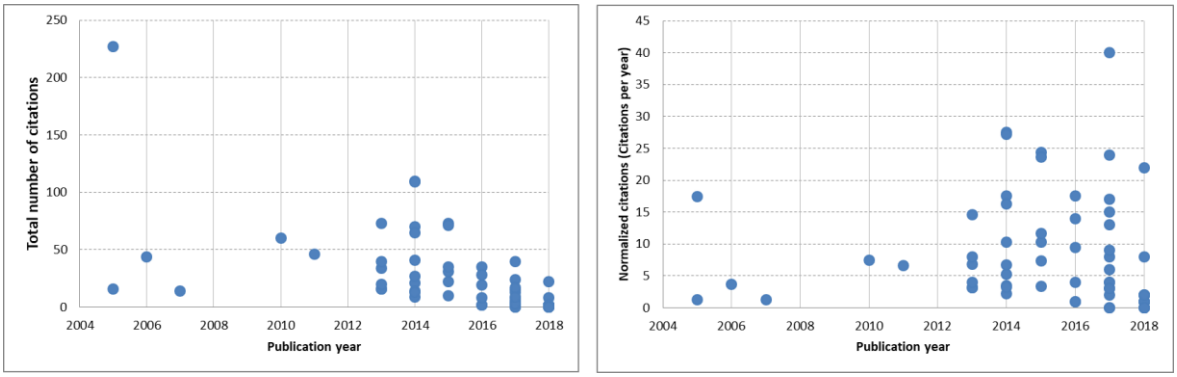

Figure 7. Citation analysis of the primary studies.

To complement this view, it is necessary to look at the average number of authors per year. As Figure 8 shows, the trend for the average number of authors per year is around three which is consistent with the previous finding of Garousi and Fernandes [67] in the study of highly-cited SE papers.

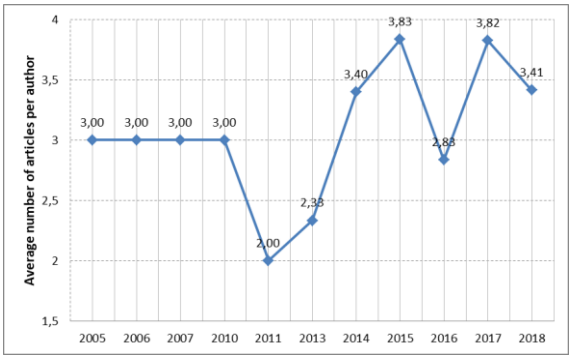

Figure 8. Average number of authors for articles per year.

Finally, citation rank can lead to pinpoint the most influential research. Thus, the top-10 papers by the absolute number of citations are shown in Table 3. We put the most recent paper first when there is a draw in the ranking (number of citations), i.e. the papers in positions 4 and 5 in Table 3. The column "annual average" shows the normalized citations. Regarding the papers' titles, one can see both old and recent papers in which various topics are represented. The top-10 papers are also a mixture of different approaches to study emotions in SE. Items 2, 3, 5, 6 and 7 are using datasets from repositories while item 4 is mainly using biometrics and the remaining items are based on software developers' self-reporting. Taking into account their high impact, readers such as new researchers and graduate students are encouraged to read and benefit from them.

Table 3. The five most cited papers based on the number of citations.

#	Study ID	Title	Year	Cited by	Annual average (%)
1	PS65	The effect of music listening on work performance	2005	227	17.5
2	PS05	Do developers feel emotions? an exploratory analysis of emotions in software artifacts	2014	110	27.5
3	PS06	Sentiment analysis of commit comments in GitHub: an empirical study	2014	109	27.3
4	PS30	Stuck and Frustrated or in Flow and Happy: Sensing Developers' Emotions and Progress	2015	73	24.3
5	PS03	Towards emotional awareness in software development teams	2013	73	14.6
6	PS32	Are Bullies More Productive? Empirical Study of Affectiveness vs. Issue Fixing Time	2015	71	23.7
7	PS07	Security and emotion: sentiment analysis of security discussions on GitHub	2014	70	17.5
8	PS66	Happy software developers solve problems better: psychological measurements in empirical software engineering	2014	65	16.3
9	PS48	Do moods affect programmers' debug performance?	2010	60	7.5
10	PS02	Using the Affect Grid to Measure Emotions in Software Requirements Engineering	2011	46	6.6

4.2 Group 2: Specific to the domain (emotions in SE).

This section addresses research questions RQ4 to RQ5.

4.2.1 What are the emotions addressed or investigated that have been reported in the studies? (RQ4)

Given that there is not a commonly agreed-upon classification regarding emotions, we distinguished that most of the primary studies are focused on following one of the next two types of approaches: (i) identifying the basic emotions as Ekman and Davidson proposed [68], or a subset of them. (ii) identifying the sentiment polarity in text as positive, negative or neutral.

By reviewing the 66 primary studies, we found that more than one emotion was explored in 23 of the studies, while forty discrete emotions were identified in 35 of them. It means that 12 studies

investigated only one emotion while the majority (21/35, 60%) investigated up to four discrete emotions. Moreover, although [PS37] is the study that most emotions explore (20/40, 50%), it does not seem to be the most relevant one, since its main aim is the demonstration that the use of freeform drawing helps distributed teams to enhance individual positive emotions.

Figure 9 shows the absolute frequency of the most reported emotions, i.e. number of times a discrete emotion has been observed to occur (see details in Appendix D). The relative frequency was calculated by dividing the absolute frequency by the total number of discrete emotions reported. Therefore, the whole distribution of those emotions is positive 38.82% (59/152) versus 53.95% (82/152) of *negative*. The remaining 7.24% (11/152) is made up of *surprise* (5.26%), *anticipation* (1.32%), and *interest* (0.66%) which depending on the context could be either positive or negative. It is not surprising if we take into account that *joy, anger, fear, sadness, and surprise* are common in four well-known models of emotions as Figure 1 shows.

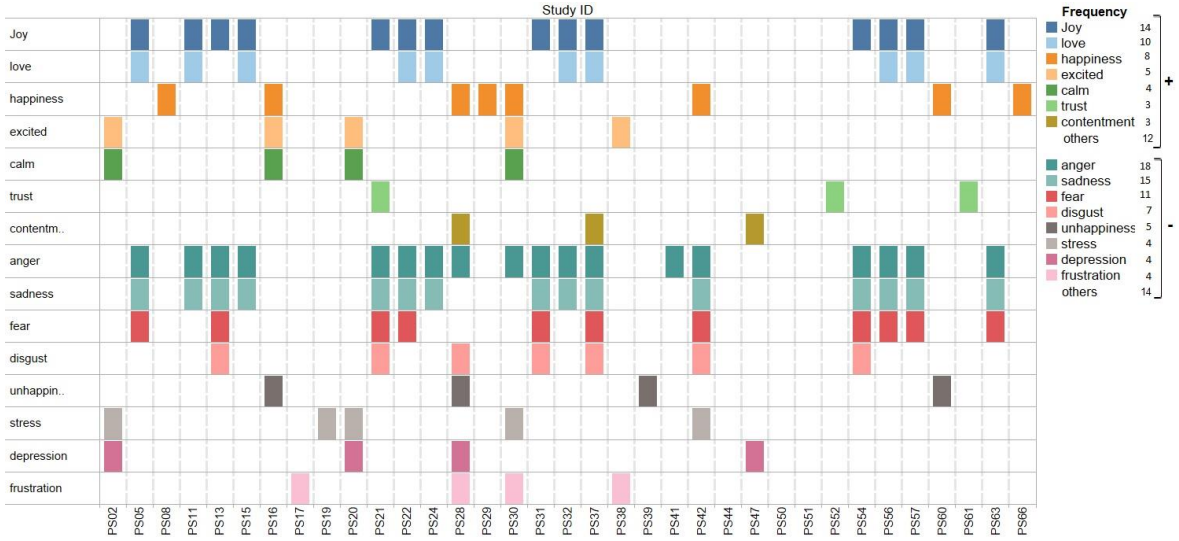

Figure 9. The most frequent emotions in the primary studies.

On the upper part of Figure 9, the positive emotions sum 30.92% (47/152) of *joy, love, happiness, excited, calm/relaxation, trust and contentment*. The remaining 7.89% (12/152) is distributed among eleven emotions (*admiration, amusement, compassion, enthusiasm, empathy, enjoying, in control, optimistic, pleased, pleasure, and relief*) addressed in five primary studies ([PS16], [PS28], [PS37], [PS47], [PS50]). On the lower part of the Figure 9, the negative emotions are made up 44.74% (68/152) of *anger, sadness, fear, disgust, unhappiness, stress, depression, and frustration*. The remaining 9.21% (14/152) is distributed among eleven emotions (*contempt, disappointed, pride, annoyance, anxiety, controlled, fatigue, guilt, hate, regret, and shame*) considered in eight primary studies ([PS16], [PS28], [PS30], [PS37], [PS42], [PS44], [PS47], [PS51]). One interesting point in these findings is that the number of negative emotions outweighs the number of positive ones which is consistent with the models in Figure 1. Finally, *sadness, anger, fear and joy* seem to be the most reported emotions.

Regarding the dimensional approach, we found fifty-one primary studies addressing one of the following two kinds of dimensions: (i) Valence, Arousal and Dominance (VAD), and (ii) positive, negative or neutral (see details in Appendix D). However, two of them [PS24], [PS30] investigated both dimensions. Figure 10 also depicts the absolute frequency of those dimensional approaches. Therefore, 35.85% (19/53) of the studies claim to use a positive, negative and neutral approach while 39.62% (21/53) of the studies state to apply a positive and negative approach. Moreover, one study [PS18] is focused on a positive polarity. On the other hand, 7.55% (4/53) claim to use a VAD approach while 13.21% (7/53) state to apply valence and arousal dimension. In addition, one study [PS40] is only focused on arousal. In fact, 20 of those studies are also investigating at least one discrete emotion as Figure 10 shows.

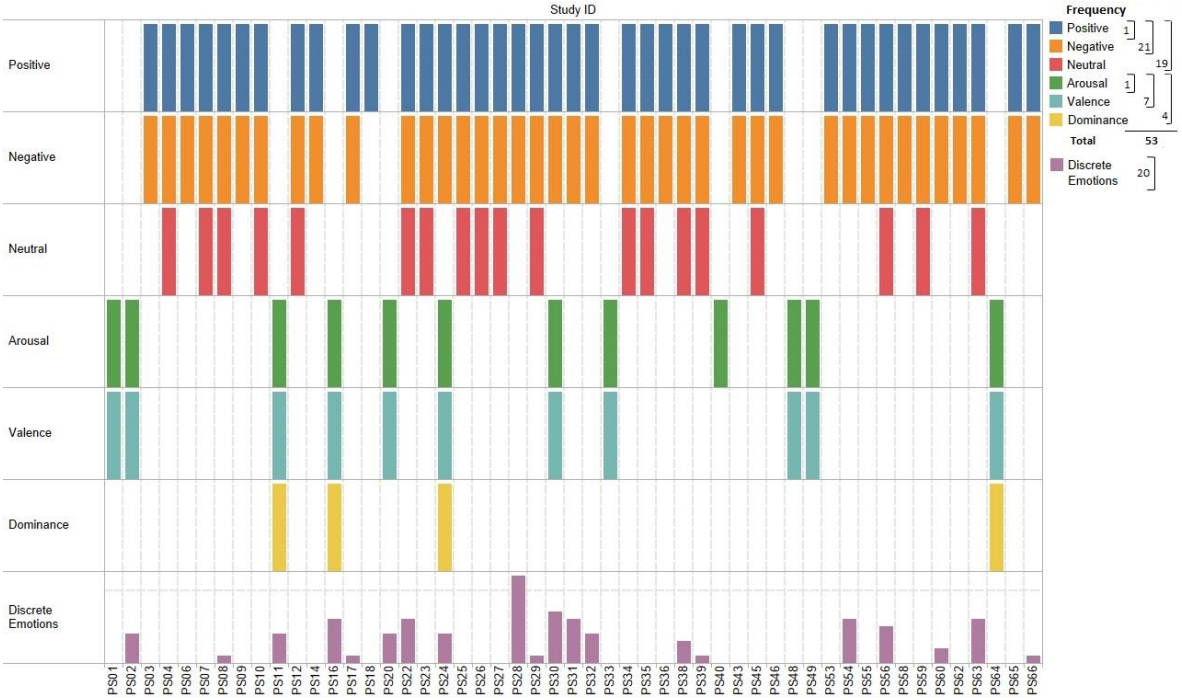

Figure 10. Dimensional approach by primary study.

4.2.2 How software developers' emotions are measured? (RQ5)

After a review of the primary studies, one can see that any given study can be related to more than one approach so that the sum of the approaches in the chart of Figure 11 is greater than the number of primary studies (66) (see details in Appendix E). The approaches were categorized using a content analysis approach. Until the end of the review period (year 2018), out of the 66 primary studies, 40 of them used a dataset but just 37 (56%) presented sentiment analysis. Those 37 studies use alternatively a lexicon-based approach, machine learning or both. The manual annotation of emotions was applied on 15 of those studies as well. There are also other three studies that did the exploration of affect by labeling of emotions present on issue comments [PS05] and forum posts (serebro [PS34] and stack overflow [PS22]). In particular, [PS05] is based on issue reports from Apache's Jira-based repository (software artifacts). The authors of this study adopted Parrott's framework as a reference for emotions (*love, joy, anger, sadness, fear, and*

surprise) to conduct a manual annotation. In turn, [PS22] released a dataset of 4,800 questions, answers, and comments from Stack Overflow, manually annotated with emotion labels using Shaver's framework, i.e. *love, joy, anger, sadness, fear, and surprise*. Final gold labels were assigned using majority agreement among three coders. Moreover, it is worth noting that [PS05] and [PS22] built two large datasets which support research on emotion awareness in software development. In fact, the first one has been already used in [PS15] to understand how developers' sentiments and emotions evolved over time, during the development process, and has been considered in the development of DEVA [PS20]. Furthermore, a mapping between the second dataset [PS22] and their positive, negative, and neutral polarity has been used for training a sentiment analysis tool called Senti4SD [PS63].

Moreover, 13 studies are based on a machine learning approach ([PS04], [PS14], [PS23], [PS26], [PS30], [PS32], [PS38], [PS39], [PS41], [PS42], [PS45], [PS57], [PS63]). For example, [PS30] used biometric measures as input and a decision tree classifier, while [PS42] collected team members' facial expressions as input and returned set of emotions for each face, as well as the bounding box for the face using Microsoft Face API. [PS57] demonstrates the feasibility of a machine learning classifier to identify issue comments containing *gratitude, joy* and *sadness*. Moreover, the authors confirmed their previous findings of [PS05]: (i) issue comments which do express emotions, in particular *gratitude, joy and sadness*, and (ii) the more context is provided about an issue report, the more human raters start to doubt and nuance their interpretation. Similar results were also found in [PS13] when the authors classified primary emotions (*sadness, anger, joy, disgust, and fear*) from a Stack Overflow dataset using API AlchemyLanguage Emotion Analysis.

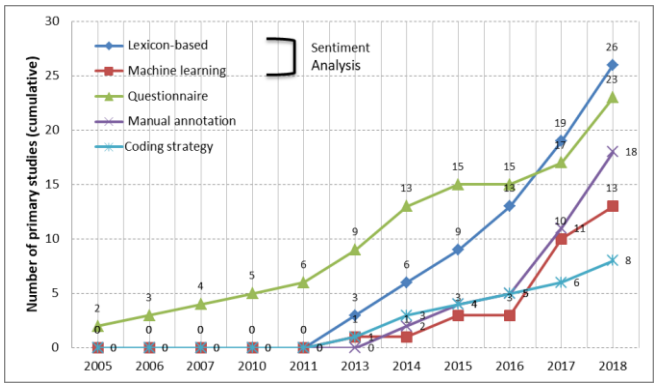

Figure 11. Cumulative trend of studies by type of assessment of emotions.

In what follows we focused our attention on the specific sentiment analysis tools, dictionaries and other approaches such as SAM, PANAS, and SPANE used to measure software developers' emotion.

— Sentiment analysis tools

Table 4 shows a list of sentiment analysis tools used per primary paper. The column "approach" denotes the emotional approach reported in the study. Sentistrengh is the most used lexicon–

based approach ([PS03], [PS04], [PS06], [PS08], [PS10], [PS12], [PS24], [PS27], [PS32], [PS35], [PS36], [PS43], [PS55]). For instance, in 2013, the relation between emotions and activity of its contributors in Gentoo project was studied by [PS04]. The authors found that it is the emotional intensity which defines activity, rather than its polarity in terms of positive or negative emotions. [PS03] used SentiStrength to compute the emotion of an entire artifact and latent Dirichlet allocation (LDA) to extract topics from a set of collaboration artifacts and to assign a set of topics to each of the artifacts. Based on that approach, Guzman in [PS27] proposed a visualization prototype which includes general and detailed views of the topics and emotions expressed in software project collaboration artifacts. Such an approach was evaluated by interviewing the project leaders, who agreed that it could be useful for creating emotional awareness in large or distributed teams, but that finer granularity in the generated summaries is needed. In 2014, [PS08] evaluated the usage of SentiStrength to identify distress or *happiness* in a development team. The results show that (i) user and developer mailing lists carry both positive and negative sentiment and have a slightly different focus, while (ii) work is needed to customize automatic sentiment analysis techniques to the domain of SE, since they lack precision when facing technical terms. In addition, [PS06] confirmed the importance of not only considering the average emotion score of a whole document, in that case, GitHub commits. The positive and negative average emotion scores, as well as the distribution of positive, negative and neutral documents should also be taken into consideration to get a deeper understanding of the emotional content, as averages tend to opaque this information. Other tools, less used, are Stanford coreNLP ([PS14], [PS26]), Syuzhet R package ([PS18], [PS21]) and API AlchemyLanguage ([PS13]). Besides, we included Danescu et al.'s tool [69] due to the interconnectedness of emotions with (im)politeness [70].

Table 4. Tools used for sentiment analysis.

	Measures	Approach	#	Study ID
1	SentiStrength	$P,N^{\pm}$ $P,N,N^{\pm n}$	13	[PS03]$^{\pm}$, [PS06]$^{\pm}$, [PS24]$^{\pm}$, [PS32]$^{\pm}$, [PS36]$^{\pm}$, [PS43]$^{\pm}$, [PS55]$^{\pm}$, [PS04]$^{\pm n}$, [PS08]$^{\pm n}$, [PS10]$^{\pm n}$, [PS12]$^{\pm n}$, [PS27]$^{\pm n}$, [PS35]$^{\pm n}$
2	Stanford coreNLP	$P,N^{\pm}$ $P,N,N^{\pm n}$	2	[PS14] $^{\pm}$, [PS26]$^{\pm n}$
3	Syuzhet R package	P^{+} discrete^{e}	2	[PS18]$^{+}$, [PS21]^{e}
4	API AlchemyLanguage	discrete^{e}	1	[PS13] ^{e}
5	Danescu et al.'s tool [69]	$P,N^{\pm}$	5	[PS24]$^{\pm}$, [PS25]$^{\pm}$, [PS32]$^{\pm}$, [PS36]$^{\pm}$, [PS55]$^{\pm}$

In particular, we found six sentiment analysis tools specific for SE domain: SentiStrength-SE [PS59], DEVA [PS20], SentiCR [PS45], Senti4SD [PS63], SentiSW [PS23] and MEME [PS21]. Each one of them claims that their empirical evaluations demonstrate advantage over other tools in the field.

SentiStrength-SE [PS59] is a tool developed for improved sentiment analysis in texts. It is designed to be used in the SE domain and it reuses the lexical approach of SentiStrength. The empirical comparisons with the three popular domain independent tools/toolkits (NLTK, Stanford NLP,

SentiStrength) suggest that SentiStrength-SE is significantly superior to its domain independent counterparts in detecting emotions within software engineering textual contents. DEVA [PS20] applies a dictionary-based lexical approach specifically designed for operation on software engineering text. The tool also includes a set of heuristics to increase accuracy as well as capturing emotional states such as *excitement, stress, depression, and relaxation* through the detection of both arousal and valence. For capturing *arousal*, the authors constructed a new arousal dictionary for DEVA by combining the SEA (Software Engineering Arousal) dictionary with the ANEW (Affective Norms for English Words) dictionary. For empirical evaluation of DEVA, a ground-truth dataset was manually annotated by three human raters and a baseline tool was implemented. From the comparisons, DEVA was found to be superior to both the baseline and TensiStrength. MEME [PS21] —a Method for EMotion Extraction— was built using functions from Syuzhet R package and the NRC Lexicon. The Syuzhet R Package identifies eight classes of emotions as suggested by Plutchik's wheel of emotions: *joy and sadness, trust and disgust, fear and anger, surprise and anticipation*. This dimensional framework of emotions is balanced with 4 positive and 4 negative emotions. The evaluation results, suggest a better performance of MEME in contrast to Syuzhet R package. SentiCR [PS45] is a supervised learning based sentiment analysis tool used for code review comments. The authors built a sentiment oracle by manually labeling a set of selected review comments and evaluated seven popular sentiment analysis tools (five lexicon-based Afinn, NLTK with Hu and Liu opinion lexicon, SentiStrength, USent, NLTK (Vader) and two supervised learning based tools (TextBlog and Vivekn), using the oracle. SentiSW [PS23] is an entity-level sentiment analysis tool, specific for SE domain, implementing a supervised machine learning method to perform sentiment classification. When SentiSW was compared to SentiStrength-SE and SentiStrength, results demonstrate the advantages of SentiSW. Finally, Senti4SD [PS63] is a classifier trained to support sentiment analysis in developers' communication channels. With respect to SentiStrength, Senti4SD reduces the misclassifications of neutral and positive posts as emotionally negative.

In contrast, a recent study [PS25] published in 2018 gave negative results when the authors aimed to build a software library recommender exploiting developers' opinions mined from Stack Overflow. In consequence, the authors carried out an investigation of the accuracy of sentiment analysis tools (SentiStrength, NLTK, Stack Overflow, Stanford CoreNLP, SentiStrength-SE and Stanford CoreNLP SO) to identify the sentiment of SE related texts. The findings revealed that although, in particular, Stack Overflow is not really a place where emotions run high since developers discuss technicalities there, there is an intrinsic problem with the accuracy of current state of the art sentiment analysis tools, given that this field is not mature enough yet. In 2017, a previous study [PS56] revealed negative results as well. Such a result clearly highlighted that the well-known sentiment analysis tools (Sentistrength, NLTK, Alchemy, Stanford NLP sentiment analyser) do not agree with the manual labeling of emotions (*love, joy, anger, sadness and fear*) and neither do they agree with each other. Indeed, the authors concluded that such a disagreement can lead to diverging conclusions and that previously published results cannot be replicated when different sentiment analysis tools are used. In consequence, there is a need for sentiment analysis tools specially targeting the SE domain. Furthermore, another study [PS26],

published in 2018, evaluated some existing tools for sentiment analysis (SentiStrength, NLTK, Alchemy, Stanford NLP, Senti4SD, SentiCR) and politeness detection (Danescu et al.'s tool [69]). The outcomes confirmed previous findings [PS56] claiming that "not only the tools have a low agreement with human ratings on sentiment and politeness, human raters also have a low agreement among themselves". The authors also remarked that it demonstrates the need for standardized coding schemes for the human coders in order to build an oracle and then perform customized training on the tools to perform reliable affect analysis in the software engineering domain.

— Dictionaries

Table 5 shows a list of 7 dictionaries that have been used in the primary studies. The column "approach" denotes the dimensional approach reported in the study. For instance, [PS54] studied development issues of nine GitHub projects by using the Wordnet-affect lexicon to classify words within the six basic emotions identified by Ekman and Davidson [68] (*sadness, joy, anger, fear, disgust, and surprise*). As a result, although, both polarity and emotional analysis are applicable, the emotional analysis seems to be more suitable to this kind of corpus. In the academic context, a dashboard tool for visualizing online teamwork discussions was proposed in [PS31]. To extract individual emotions, the contributions are matched against the NRC Word Emotion Lexicon so that the dashboard extracts and communicates team role distribution and team emotion information in real-time. Eight basic emotions (*anger, anticipation, disgust, fear, joy, sadness, surprise, and trust)* are mined from the member contributions to the discussion. Furthermore, we found two dictionaries explicitly designed for SE domain. The Software Engineering Arousal lexicon (SEA) [PS40] was specifically designed to address the problem of detecting emotional *arousal* in the software developer ecosystem. The authors included seed words potentially indicative of arousal from different sources such as NASA TLX, Russell's circumplex model of affect, and words from a text analysis application called Linguistic Inquiry and Word Count (LIWC) about *anxiety*, time, and achievement. [PS14] proposes an emotion words–based dictionary for verifying bug reports' textual emotion based on positive and negative terms. Such an approach aims to predict bug severity to reduce developers' efforts by implementing a new algorithm EWD-Multinomial. To do so, the authors modified a well-known machine learning algorithm called Naïve Bayes multinomial. The new algorithm outperforms the others when it was compared with the baselines, including Naïve Bayes multinomial and a Lamkanfi study, for open source projects such as Eclipse, Android, and JBoss.

Table 5. Dictionaries.

	Measures	Approach	#	Study ID
1	Warriner et al. [71]	$A^{\emptyset}$, $VAD^{\dagger}$	3	[PS11]†, [PS24]†, [PS40]$^{\emptyset}$
2	LIWC	$A^{\emptyset}$ $P,N^{\pm}$	3	[PS40]$^{\emptyset}$, [PS41], [PS58]$^{\pm}$
3	ANEW	VA^{*} $P,N,N^{\pm n}$	2	[PS20]*, [PS59]$^{\pm n}$
4	NLTK		2	[PS07], [PS45]
5	Wordnet Affect label	$P,N^{\pm}$	2	[PS32], [PS54]$^{e\pm}$

		discrete[e]		
6	NRC Word Emotion Lexicon	discrete[e]	2	[PS21][e], [PS31]
7	SentiWordNet	P,N[±]	1	[PS14][±]

— Others approaches

Table 6 presents other approaches for measurement of emotions. The following four approaches have been used only once in the set of primary studies: Geneva Emotion Wheel, Job Emotions Scale (JES), the wellbeing questionnaire, and Act4teamsLight. Furthermore, Parrot's framework and Shaver's framework have been explicitly used to map the emotions. With regard to questionnaires, 5 of 23 studies developed or adopted multi-item scales from prior studies for the measurement of *stress* [PS19], *fatigue* [PS44], *collective empathy* [PS50], *pride* [PS51] and *trust* [PS61]. For example, perceived trustworthiness was measured by adapting a previously validated survey instrument developed by Johnson-George and Swap. The other 17 studies applied recognized approaches to assess the affect. Table 6 shows that SAM is the most used questionnaire, followed by PANAS, Russell circumplex model of affect and SPANE. For example, [PS16] studied the effects of automated competency evaluation on software engineers' emotions and motivation by implementing a web-based platform that includes the SAM and the Intrinsic Motivation Inventory (IMI). The findings show that automation has a positive impact on both emotions and motivation of the employees, and no disadvantages were identified. Likewise, the results of the studies carried out in [PS49] revealed significant average correlations between mood measurement and personalized regression models based on keyboard and mouse interaction data. It is worth noting that participants in one of these studies worked on a programming task while listening to high or low arousing background music.

Table 6. Other approaches for measurement of emotions.

	Measures	**Approach**	**#**	**Study ID**
1	SAM	VA[*], VAD[†]	5	[PS01][*], [PS16][†], [PS48][*], [PS49][*], [PS64][†]
2	PANAS	P,N	4	[PS09], [PS46], [PS62], [PS65]
3	SPANE	P,N	2	[PS60], [PS66]
4	Russell's circumplex model of affect	A[Ø], VA[*], VAD[†]	4	[PS02][*], [PS30][*], [PS33][*], [PS40][Ø]
5	Parrot's framework	P,N	3	[PS05], [PS32], [PS57]
6	Shaver's framework	P,N	3	[PS22], [PS41], [PS63]
7	Others*	discrete[e]	4	[PS37] Geneva Emotion Wheel[e] [PS28] Job Emotions Scale (JES) [e] [PS47] The wellbeing questionnaire [e] [72] [PS17] Act4teamsLight

5 Discussion

This SLR found 66 primary studies and their results have given us a useful insight into the state of the art of the software developers' emotions research. In general, most of the studies were found

as of reasonably good quality according to the quality criteria used in our SLR. In this section, we discuss the implications of our SLR findings (section 5.1) and the detail threats to the validity of our findings (section 5.2).

5.1 Implications

Software developers are, like other knowledge workers, capable skilled professionals. Beyond this, Capretz [73] stated that "people sometimes struggle to remember that we are dealing with creatures of logic and emotions, not just ones and zeros". In other words, developers are also humans and prone to human sensitivities [74]. Thus, this SLR focused on a better understanding of software developers' emotions through a holistic view of SE research.

The orientation toward emotion in the primary studies is as follows: 23% (15/66) of the studies chose discrete emotions (anger, fear, etc.) while 47% (31/66) chose a dimensional approach and 30% (20/66) indicated they used both approaches. By analyzing all those studies that chose both approaches, in addition to those that had chosen only the discrete choice (a total of 53% of those), we found which emotions (out of a list of 40) are the most researched. There were more negative emotions —*anger* (12%), *sadness* (10%), *fear* (7%), and *disgust* (5%)— than positive ones —*joy* (9%), *love* (7%), and *happiness* (5%)—, and *surprise* (5%) that can take both positive and negative meaning.

Comparing these findings to a recent survey involving 250 emotion researchers and carried out by Ekman [75] who aimed to evaluate the status of emotion research, one can see: (i) The trend of the orientation toward emotion in both fields was similar with regards to the discrete emotions (23% vs. 18%). However, dimensions were much more researched in SE (47% vs. 16%), maybe due to the better exploitation of sentiment analysis tools by software engineering researchers. Finally, although, almost a third reported both views in SE field (30% vs. 55%), the majority of emotion researchers chose this approach, namely, discrete emotion and dimensions. It means that most emotion researchers find both a discrete and a dimensional view of emotions useful in their research [56], as suggested by Wundt in 1896, more than 100 years ago. In contrast, SE empirical research is more focused on the dimensional approach. (ii) Among emotion researchers, there was high agreement on five emotions (all of which were described by both Darwin and Wundt) that they consider or think, should be considered as the most basic about emotions: *anger* (91%), *fear* (90%), *disgust* (86%), *sadness* (80%), and *happiness* (76%). In comparison with our SLR, there was high agreement about the negative emotions, but there was low agreement about *joy, love* and *happiness*. By analyzing the survey [75], one can see that Ekman proposed happiness as a category of emotion related to *joy* so that *joy* is not part of the list of discrete emotions. Moreover, *shame, surprise,* and *embarrassment* were endorsed by 40%–50% emotion researchers while *surprise* and *shame* were less studied in selected studies in our SLR and *embarrassment* not at all. Other emotions, currently under study by various emotion researchers drew substantially less support: *guilt* (37%), *contempt* (34%), *love* (32%), *awe* (31%), *pain* (28%), *envy* (28%), *compassion* (20%), *pride* (9%), and *gratitude* (6%). Although most of these emotions were part of the forty emotions in our SLR, it seems that *awe*, *pain*, and *envy* are not relevant in the SE context. (iii) There was also

agreement on the circumplex and positive-negative as the most basic about emotions. However, emotion researchers also found another useful dimension, approach-avoidance, which has not even been mentioned in the primary studies of our SLR.

According to Ekman [75], the agreement about the evidence for universals in emotional signals and the evidence for five emotions is robust. In the SE field, the empirical evidence is limited but provides coherent and consistent results. From our findings, the need for further research on happiness is obvious and supported by Graziotin et al. [76] who made a call for software engineering researchers to take (un)happiness into account in their studies. Furthermore, there was no agreement 20 years ago about whether moods differ from emotion, but today most, emotion researchers agree that moods are related to emotions as well as personality and psychopathology are related in some way to specific emotions [75]. Therefore, software engineering researchers should take care not to use interchangeably the term “emotions” and the terms “moods”, “personality” and “emotional disorders”. In terms of emotion measurement, sentiment analysis tools on one hand, and SAM and PANAS, on the other hand, were mostly adopted in the primary studies of our SLR. While research in the field of sentiment analysis has received a growing attention over the last years —SentiStrength-SE [PS59], DEVA [PS20], SentiCR [PS45], Senti4SD [PS63], SentiSW [PS23] and MEME [PS21]—, the emerging fields of multi-class emotion recognition, which entails classifying text into one or more categories of emotion such as joy, love, sadness and anger—[PS32] and [PS41]—, have remained underexplored. In general, emotion detection is a challenging problem that includes the related tasks of sentiment analysis and emotion detection. Although both tasks suffer from the subtleties that the implicit nature of language holds, the second one is more complicated due to the greater number of emotions and the innate similarities among different emotions. In fact, several theories of emotion have been proposed by psychologists over the years, each detailing a slightly different set of emotions. Hence, even human annotators often find hard to distinguish emotions and, as a result, there is low agreement among themselves and sentiment analysis tools specific for SE domain [PS25], [PS56], [PS26]. So far, high accuracy of sentiment analysis tools has been difficult to achieve.

The last five years of research has been productive, but as this SLR revealed, there are still many aspects of emotion that deserve further empirical research to reduce the disagreements that still persist. Perhaps most important, robust evidence is needed in order to support the role of emotions in software development process. Here some questions arise. Firstly, research design should consider a bigger set of emotions than a more reduced one or, in another case, a right balance between negative and positive emotions. Although studies looking closely at a single emotion can provide valuable information, research on the emotional timeline and the relationships among software developers’ emotions is needed to move toward a mapping of their effect in terms of performance, productivity, quality and wellbeing. The question is also whether we need further research in more realistic or typical contexts, rather than in controlled, laboratory settings. Concerning the measurement of emotions, it is very large, very diverse, and very complex field [50]. There is a very broad range of approaches and each one has its advantages and its own limitations. For instance, capturing physiological signals not only requires special devices but also

is far more labor-intensive and costly than gathering self-reports. However, self-report questionnaire data are solicited while physiological and facial scaling data are not necessarily solicited, but the participant is usually aware of the data collection process, which is sometimes intrusive and cumbersome. Therefore, another key question is how to choose a pertinent approach to measure emotions in the SE context. Apart from that, the extent to which cognitive effects of emotions can be differentiated is an area that needs further research. In particular, it is worth questioning to what extent software developers' emotions are useful in assisting them in making practical choices. Finally, cross-cultural research is also necessary to assess the extent of universality versus culture-specificity in SE.

5.2 Limitations and threats to validity

The results of our SLR might have some limitations with regard to the underlying research method. In what follows, the threats of validity are discussed in the context of four main types of threats of validity based on a standard checklist adopted from [77]: construct , internal, conclusion, and external.

Construct validity is related to the degree in which an investigation measures what it claims to be measuring. A threat to construct validity comes from the lack of empirical evidence in the primary studies. In consequence, we aimed to identify as many relevant primary studies as were possible using two key terms: "emotion'' and "software engineering". However, we recognize that the first key term implies a first limitation because some software engineering researchers could overlook the term "emotion" and use only "mood", or "feeling", or "affect" when they study emotions. To reduce this threat, although other major affective phenomena ("feelings", "mood" or "affects") were not included in the search process, they were not included as exclusion criteria during the filter process. Moreover, there might be a selection bias due to having chosen five academic search engines, but, those databases are commonly used in existing SLRs (e.g. [20], [78]) and their selection is also based on Kitchenham and Charters [61]. In addition, a backward snowballing process of the selected primary studies was done to ensure that all relevant references had been included.

Internal validity is the extent to which a causal conclusion based on a study is warranted. It is determined by the degree to which a study minimizes systematic errors. A threat to internal validity in this study lies in bias on data extraction because it may result in inaccuracy of the extracted data, and thus affecting the analysis of the primary studies. To minimize this threat, we adapted a protocol from well-established guidelines [61]. In particular, we designed a data extraction form that includes the research questions and the quality questionnaire. With regard to the evaluation of quality level of primary studies, it was considerably subjective but we gained a broader perspective by reviewing the research methods of each study. Moreover, given that a paper may be retrieved from more than one database, to avoid misleading, we have checked and removed the duplicates based on their publishers. The maturity of the field is another factor that can affect internal validity, however, we believe that more than 10 years, since the first empirical

studies were identified in the literature on software developers' emotions [63], is enough time to review this research field in a systematic way.

Conclusion validity of a review study deals with reaching appropriate conclusions through rigorous and repeatable treatment. The traceability between the data extracted and the conclusions was strengthened through the use of a data analysis tool, which helped us to minimize the possibility of missing evidence, however, human errors may have occurred. Therefore, we provide a replication package as archived open data [21] which not only gives the possibility to others to check our work but also, allow them to expand or improve it. To reduce researcher biases both authors were actively involved in this research and the final decision to include a study depended on the agreement of them. In fact, the Krippendorff alpha was calculated and showed a high agreement. All primary studies were analyzed and the data was reviewed, extracted and synthesized by the two authors. The discrepancies were discussed by both the authors and resolved by consensus.

External validity is the degree to what extent the results can be generalized to other contexts. The limitation into academic search engines represents an academic research so that studies that are published as (non-academic) books and grey literature (such as technical reports, white papers, work in progress) were not included in this study. Although we recognized that additional relevant published studies may have been overlooked, we believe that despite that limitation our SLR gives a significant contribution by itself and this review can be extended in future. We also limited ourselves to publications written in English so that relevant studies in other languages are missed out, but it is expected a relatively small effect because English is the most common language on this research context. Furthermore, given that the software developers' emotions research is a specific research field in empirical software engineering, with its particularities and peculiarities, we cannot claim that our results can be generalized beyond this field, but we believe that the value of our SLR should not be undermined.

6 Conclusions

The number of primary studies included in the SLR was 66. This fact indicates that a limited empirical research has been done on software developers' emotions to this date. We found that the publication years of these primary studies ranged from 2005 to 2018. The aim was to identify the software developers' emotions research and provide an overview of the state of the art in this topic to benefit the readers (both practitioners and researchers) in providing the most comprehensive and holistic view of the field.

The findings of the SLR helped to identify 40 discrete emotions in the studies. The most frequent emotions were: *anger, fear, disgust, sadness, joy, love and happiness*. 22.74% of all primary studies used only a discrete approach. We also found two different dimensional approaches in 46.96% of studies: VAD —valence, arousal and dominance— and polarity —positive, negative and neutral. Here, a clear gap is that the approach-avoidance dimension so far was not used in the

primary studies of our SLR. The remaining 30.30% used both approaches, that is, discrete and dimensional. Most of the studies used software practitioners and datasets from industrial context as subjects. Although they may not be representative of the whole software industry, the evidence is enough to support that software developers not only feel emotions, but also display them in artifacts and communications during their daily work.

The findings also showed that there is not a common agreed approach to measure emotions in SE. Sentiment analysis seems to be the most popular, although, a reliable sentiment analysis tool in the SE domain was not found. It was observed that some primary studies included diverse self-reported instruments. SAM was the most used questionnaire, followed by PANAS, Russell circumplex model of affect and SPANE. Moreover, only little research has focused on developers' emotions and the use of biometric sensors —electroencephalography (EEG), eye-related measures, skin- and heart-related signal) or peripheral devices (keyboard and mouse).

As one can see, there are a lot of opportunities for future empirical research in this area. The most obvious is a replication of previous studies and provide evidence to help the area mature. One of the most important benefits of replication studies centers around the possibility of arriving at negative results [54]. In particular, sentiment analysis accuracy would be interesting to research further, because they seemed to have a large impact on emotion awareness. Once reliable sentiment analysis tools are developed, they have to be well enough validated so that they can be adopted with confidence by software practitioners. In addition, given that the sentiment analysis tools were evaluated only on English datasets, it would be interesting to experiment with data from different languages in future work. In addition, future improvements in physiological measures or behavioral measures will likely bring new ethical challenges.

Future work exploring the nature of the relationship between personality and discrete emotions could lend useful insights as well. Regarding basic emotions, they could be studied to deepen the understanding of their effect and give more insights about how to apply the measurements. However, the question posed by Ekman [75] remains true in SE domain: will compelling evidence for more than just five emotions (*anger, fear, disgust, sadness, and happiness*) be forthcoming in the coming years, or is that all that can be empirically established?

Our future work includes using the findings of this SLR. Thus, authors propose a deeper analysis and comparison between the primary studies, with particular emphasis on understanding the effect of emotions on the software development process expressed in terms such as performance, productivity, quality, and wellbeing.

Appendix A. List of primary studies included in the SLR

#	Primary Study	Authors	Year	Title
1	PS01	Iftikhar Ahmed Khan, Robert M. Hierons, Willem-Paul Brinkman	2007	Mood independent programming
2	PS02	R. Colomo-Palacios and C. Casado-	2011	Using the Affect Grid to Measure Emotions

		Lumbreras		in Software Requirements Engineering
3	PS03	Emitza Guzman, Bernd Bruegge	2013	Towards emotional awareness in software development teams
4	PS04	David Garcia, Marcelo Serrano Zanetti, Frank Schweitzer	2013	The Role of Emotions in Contributors Activity: A Case Study on the GENTOO Community
5	PS05	Alessandro Murgia, Parastou Tourani, Bram Adams, Marco Ortu	2014	Do developers feel emotions? an exploratory analysis of emotions in software artifacts
6	PS06	Emitza Guzman, David Azócar, Yang Li	2014	Sentiment analysis of commit comments in GitHub: an empirical study
7	PS07	Daniel Pletea, Bogdan Vasilescu, Alexander Serebrenik	2014	Security and emotion: sentiment analysis of security discussions on GitHub
8	PS08	Parastou Tourani, Yujuan Jiang, Bram Adams	2014	Monitoring sentiment in open source mailing lists: exploratory study on the apache ecosystem
9	PS09	Kurt Schneider Leibniz , Olga Liskin Leibniz ,Hilko Paulsen , Simone Kauffeld	2015	Media, Mood, and Meetings: Related to Project Success?
10	PS10	F. Calefato, F. Lanubile, M. C. Marasciulo, and N. Novielli	2015	Mining successful answers in stack overflow
11	PS11	Mika Mantyla , Bram Adams , Giuseppe Destefanis , Daniel Graziotin , Marco Ortu	2016	Mining Valence, Arousal, and Dominance - Possibilities for Detecting Burnout and Productivity?
12	PS12	Vinayak Sinha, Alina Lazar, Bonita Sharif	2016	Analyzing developer sentiment in commit logs
13	PS13	Awdren Fontão, Oswald M. Ekwoge, Rodrigo Santos, Arilo Claudio Dias-Neto	2017	Facing up the primary emotions in Mobile Software Ecosystems from Developer Experience
14	PS14	Geunseok Yang, Seungsuk Baek, Jung-Won Lee, Byungjeong Lee	2017	Analyzing emotion words to predict severity of software bugs: a case study of open source projects
15	PS15	Giuseppe Destefanis, Marco Ortu, Steve Counsell, Stephen Swift, Roberto Tonelli, Michele Marchesi	2017	On the randomness and seasonality of affective metrics for software development
16	PS16	Gul Calikli, Mohammed Al-Eryani, Emil Baldebo, Jennifer Horkofff, Alexander Ask	2018	Effects of automated competency evaluation on software engineers' emotions and motivation: a case study
17	PS17	Nils Prenner, Jil Klünder, Kurt Schneider	2018	Making meeting success measurable by participants' feedback
18	PS18	Karl Werder	2018	The evolution of emotional displays in open source software development teams: an individual growth curve analysis
19	PS19	Miikka Kuutila, Mika V. Mäntylä, Maëlick Claes, Marko Elovainio	2018	Daily questionnaire to assess self-reported well-being during a software development project
20	PS20	Md Rakibul Islam, Minhaz F. Zibran	2018	DEVA: sensing emotions in the valence arousal space in software engineering text
21	PS21	Karl Werder, Sjaak Brinkkemper	2018	MEME: toward a method for emotions extraction from github
22	PS22	Nicole Novielli, Fabio Calefato, Filippo Lanubile	2018	A gold standard for emotion annotation in stack overflow
23	PS23	Jin Ding, Hailong Sun, Xu Wang, Xudong Liu	2018	Entity-level sentiment analysis of issue comments
24	PS24	Giuseppe Destefanis, Marco Ortu, David Bowes, Michele Marchesi, Roberto Tonelli	2018	On measuring affects of github issues' commenters
25	PS25	Nasif Imtiaz, Justin Middleton, Peter Girouard, Emerson Murphy-Hill	2018	Sentiment and politeness analysis tools on developer discussions are unreliable, but so

				are people
26	PS26	Bin Lin, Fiorella Zampetti, Gabriele Bavota, Massimiliano Di Penta, Michele Lanza, Rocco Oliveto	2018	Sentiment analysis for software engineering: how far can we go?
27	PS27	Emitza Guzman	2013	Visualizing emotions in software development projects
28	PS28	Michal R. Wrobel	2013	Emotions in the software development process
29	PS29	Kevin Dullemond , Ben van Gameren , Margaret-Anne Storey , Arie van Deursen	2013	Fixing the 'Out of sight out of mind' problem one year of mood-based microblogging in a distributed software team
30	PS30	Sebastian C. Müller , Thomas Fritz	2015	Stuck and Frustrated or in Flow and Happy: Sensing Developers' Emotions and Progress
31	PS31	Rebecca Vivian , Hamid Tarmazdi , Katrina Falkner , Nickolas Falkner , Claudia Szabo	2015	The Development of a Dashboard Tool for Visualising Online Teamwork Discussions
32	PS32	Marco Ortu , Bram Adams , Giuseppe Destefanis , Parastou Tourani , Michele Marchesi , Roberto Tonelli	2015	Are Bullies More ProductiveAre Bullies More Productive? Empirical Study of Affectiveness vs. Issue Fixing Time
33	PS33	Michal R. Wrobel	2016	Towards the participant observation of emotions in software development teams
34	PS34	Allen Marshall , Rose F. Gamble , Matthew L. Hale	2016	Outcomes of Emotional Content from Agile Team Forum Posts
35	PS35	Md Rakibul Islam , Minhaz F. Zibran	2016	Towards understanding and exploiting developers' emotional variations in software engineering
36	PS36	Jirayus Jiarpakdee, Akinori Ihara, Ken-ichi Matsumoto	2016	Understanding question quality through affective aspect in Q&A site
37	PS37	Mengyao Zhao , Yi Wang , David Redmiles	2017	Using Playful Drawing to Support Affective Expressions and Sharing in Distributed Teams
38	PS38	Grant Williams , Anas Mahmoud	2017	Analyzing, Classifying, and Interpreting Emotions in Software Users' Tweets
39	PS39	Amol Patwardhan	2017	Sentiment Identification for Collaborative, Geographically Dispersed, Cross-Functional Software Development Teams
40	PS40	Mika V. Mäntylä , Nicole Novielli , Filippo Lanubile , Maëlick Claes , Miikka Kuutila	2017	Bootstrapping a Lexicon for Emotional Arousal in Software Engineering
41	PS41	Daviti Gachechiladze , Filippo Lanubile , Nicole Novielli , Alexander Serebrenik	2017	Anger and Its Direction in Collaborative Software Development
42	PS42	Georgios Dafoulas , Cristiano Maia , Almaas Ali , Juan Carlos Augusto , Victor Lopez-Cabrera	2017	Understanding Collaboration in Global Software Engineering (GSE) Teams with the Use of Sensors: Introducing a Multi-sensor Setting for Observing Social and Human Aspects in Project Management
43	PS43	Rodrigo Souza , Bruno Silva	2017	Sentiment Analysis of Travis CI Builds
44	PS44	Saurabh Sarkar , Chris Parnin	2017	Characterizing and Predicting Mental Fatigue during Programming Tasks
45	PS45	Toufique Ahmed, Amiangshu Bosu, Anindya Iqbal, and Shahram Rahimi	2017	SentiCR: a customized sentiment analysis tool for code review interactions
46	PS46	Sharifah Lailee Syed-Abdullah, John Karn, Mike Holcombe, Tony Cowling, Marian Gheorge	2005	The Positive Affect of the XP Methodology
47	PS47	Sharifah Syed-Abdullah , Mike Holcombe , Marian Gheorge	2006	The Impact of an Agile Methodology on the Well Being of Development Teams
48	PS48	I. A. Khan, W.-P. Brinkman, and R. M. Hieron	2010	Do moods affect programmers' debug performance?

49	PS49	I. A. Khan, W.-P. Brinkman, and R. Hierons	2013	Towards estimating computer users' mood from interaction behaviour with keyboard and mouse
50	PS50	Ali E. Akgün , Halit Keskin, A. Yavuz Cebecioglu, Derya Dogan	2014	Antecedents and consequences of collective empathy in software development project teams
51	PS51	Ofira Shmueli, Nava Pliskin, Lior Fink	2014	Explaining over-requirement in software development projects: An experimental investigation of behavioral effects
52	PS52	Ban Al-Ani, Sabrina Marczak, David Redmiles, Rafael Prikladnicki	2014	Facilitating contagion trust through tools in Global Systems Engineering teams
53	PS53	Kangning Wei, Kevin Crowston, Na Lina Li, Robert Heckman	2014	Understanding group maintenance behavior in Free/Libre Open-Source Software projects: The case of Fire and Gaim
54	PS54	Francisco Jurado , Pilar Rodriguez	2015	Sentiment Analysis in monitoring software development processes: An exploratory case study on GitHub's project issues
55	PS55	Marco Ortu , Giuseppe Destefanis , Steve Counsell , Stephen Swift , Roberto Tonelli , Michele Marchesi	2017	How diverse is your team? Investigating gender and nationality diversity in GitHub teams
56	PS56	Robbert Jongeling, Proshanta Sarkar, Subhajit Datta, Alexander Serebrenik	2017	On negative results when using sentiment analysis tools for software engineering research
57	PS57	Alessandro Murgia, Marco Ortu, Parastou Tourani, Bram Adams, and Serge Demeyer	2017	An exploratory qualitative and quantitative analysis of emotions in issue report comments ofopen source systems
58	PS58	Sherlock A. Licorish, Stephen G. MacDonell	2018	Exploring the links between software development task type, team attitudes and task completion performance: Insights from the Jazz repository
59	PS59	Md Rakibul Islam, Minhaz F. Zibran	2018	SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text
60	PS60	Daniel Graziotin , Fabian Fagerholm, Xiaofeng Wang , Pekka Abrahamsson	2018	What happens when software developers are (un)happy
61	PS61	Erik H. Trainer, David F. Redmiles	2018	Bridging the gap between awareness and trust in globally distributed software teams
62	PS62	Kurt Schneider, Jil Klünder, Fabian Kortum, Lisa Handke, Simone Kauffeld	2018	Positive affect through interactions in meetings: The role of proactive and supportive statements
63	PS63	F. Calefato, F. Lanubile, F. Maiorano, N. Novielli	2018	Sentiment Polarity Detection for Software Development
64	PS64	Daniel Graziotin , Xiaofeng Wang, Pekka Abrahamsson	2014	Do feelings matter? On the correlation of affects and the self-assessed productivity in software engineering
65	PS65	T. Lesiuk	2005	The effect of music listening on work performance
66	PS66	Daniel Graziotin, Xiaofeng Wang, Pekka Abrahamsson	2014	Happy software developers solve problems better: psychological measurements in empirical software engineering

Appendix B. Dataset by source

	Sources	**#**	**Study ID**
1	Q&A	9	Piazza [PS31], Serebro [PS34], Stack Overflow [PS10], [PS13], [PS22], [PS26]*, [PS36], [PS56]†, [PS63]

2	Issue tracking	27	JIRA [PS05], [PS11], [PS15], [PS20], [PS26]*, [PS32], [PS40], [PS41], [PS57], [PS59], GitHub [PS06], [PS07], [PS12], [PS18], [PS21], [PS23], [PS24], [PS25], [PS35], [PS43] , [PS54], [PS55] , [PS56] †, Bugzilla [PS04], Bug report from Eclipse, Android and JBoss [PS14], SourceForge [PS53]
3	Others	7	Confluence [PS03], Apache Software Foundation [PS08], Twitter (microbloging)[PS38], Gerrit [PS45], IBM Jazz [PS58], Not available [PS27], [PS39]
*,† denote that the study use two source			

Appendix C. Datasets publicly available on the Web

	Study ID	#	URL
1	[PS12]	1	http://seresl.csis.ysu.edu/MSR16challenge/
2	[PS13], [PS36]	2	http://2015.msrconf.org/challenge.php
3	[PS41]	1	http://goo.gl/2e6mbk
4	[PS56]	1	http://ow.ly/HvC5302N4oK
5	[PS22]	1	https://github.com/collab-uniba/EmotionDatasetMSR18
6	[PS63]	1	https://github.com/collab-uniba/Senti4SD
7	[PS23]	1	https://github.com/Jasmine-DJ-420/SentiSW
8	[PS45]	1	https://github.com/senticr/SentiCR/
9	[PS43]	1	https://gitlab.com/rodrigorgs/msr17-challenge
10	[PS26]	1	https://sentiment-se.github.io/replication.zip
11*	[PS20]	1	https://figshare.com/s/277026f0686f7685b79
12*	[PS11], [PS40]	2	http://openscience.us/repo/social-analysis/socialaspects.html
13*	[PS05], [PS57]	2	http://ansymore.uantwerpen.be/system/files/uploads/artefacts/alessandro/MSR16/archive3.zip
* denotes that the link does not work. Verified: Dec 2018			

Appendix D. Primary studies by emotional approach

Discrete emotions affecting software development team.

+\-	#	Emotion	Frequency		Study ID
			Absolute n	Relative f (%)	
POSITIVE (+)	1	joy	14	9.21	[PS05], [PS11], [PS13], [PS15], [PS21], [PS22], [PS24], [PS31], [PS32], [PS37], [PS54], [PS56], [PS57], [PS63]
	2	love	10	6.58	[PS05], [PS11], [PS15], [PS22], [PS24], [PS32], [PS37], [PS56], [PS57], [PS63]
	3	happiness	8	5.26	[PS08], [PS16], [PS28], [PS29], [PS30], [PS42], [PS60], [PS66]
	4	excited	5	3.29	[PS02], [PS16], [PS20], [PS30], [PS38]
	5	calm/relaxation	4	2.63	[PS02], [PS16], [PS20], [PS30]
	6	trust	3	1.97	[PS21], [PS52], [PS61]
	7	contentment	3	1.97	[PS28], [PS37], [PS47]
	8	others	12	7.89	[PS16] (in control), [PS28], [PS47] (enthusiasm), [PS28] (pleased, optimistic, enjoying), [PS37] (pleasure, amusement, admiration, relief, compassion), [PS50] (empathy)
NEGATIVE (-)	1	anger	18	11.84	[PS05], [PS11], [PS13], [PS15], [PS21], [PS22], [PS24], [PS28], [PS30], [PS31], [PS32], [PS37], [PS41], [PS42], [PS54], [PS56], [PS57], [PS63]
	2	sadness	15	9.87	[PS05], [PS11], [PS13], [PS15], [PS21], [PS22], [PS24], [PS31], [PS32], [PS37], [PS42], [PS54], [PS56], [PS57], [PS63]
	3	fear	11	7.24	[PS05], [PS13], [PS21], [PS22], [PS31], [PS37], [PS42], [PS54], [PS56], [PS57], [PS63]
	4	disgust	7	4.61	[PS13], [PS21], [PS28], [PS31], [PS37], [PS42], [PS54]
	5	stress	5	3.29	[PS02], [PS19], [PS20], [PS30], [PS42]
	6	unhappiness	4	2.63	[PS16], [PS28], [PS39], [PS60]
	7	depression	4	2.63	[PS02], [PS20], [PS28], [PS47]
	8	frustration	4	2.63	[PS17], [PS28], [PS30], [PS38]
	9	others	14	9.21	[PS28, PS37] (disappointed), [PS37, PS42] (contempt), [PS37, PS51] (pride), [PS16] (controlled), [PS30] (annoyance), [PS37] (guilt, hate, shame, regret), [PS44] (fatigue), [PS47] (anxiety)
- / +	1	surprise	8	5.26	[PS05], [PS21], [PS22], [PS31], [PS42], [PS54], [PS57], [PS63]
	2	anticipation	2	1.32	[PS21], [PS38]
	3	interest	1	0.66	[PS37]
TOTAL			**152**	**100.00**	

Dimensional approach affecting software development team.

	#	Emotion	Frequency		Study ID
			Absolute n	Relative f (%)	
POLARITY	1	Neutral, Negative, Positive	19	35.85	[PS04], [PS07], [PS08], [PS10], [PS12], [PS22], [PS23], [PS25], [PS26], [PS27], [PS29], [PS34], [PS35], [PS38], [PS39], [PS45], [PS56], [PS59], [PS63]
	2	Negative, Positive	21	39.62	[PS03], [PS06], [PS09], [PS14], [PS17], [PS24]*, [PS28], [PS30]*, [PS31], [PS32], [PS36], [PS43], [PS46], [PS53], [PS54], [PS55], [PS58], [PS60],

					[PS62], [PS65], [PS66]
	3	Positive	1	1.89	[PS18]
VAD	1	Dominance, Valencia, Arousal	4	7.55	[PS11], [PS16], [PS24]*, [PS64]
	2	Valencia, Arousal	7	13.21	[PS01], [PS02], [PS20], [PS30]*, [PS33], [PS48], [PS49]
	3	Arousal	1	1.89	[PS40]
TOTAL			**53**	**100.00**	
Discrete emotions reported in these studies			20	37.74	[PS02], [PS08], [PS11], [PS16], [PS17], [PS20], [PS22], [PS24]*, [PS28], [PS29], [PS30]*, [PS31], [PS32], [PS38], [PS39], [PS54], [PS56], [PS60], [PS63], [PS66]

Appendix E. Type of assessment of emotions

	Approach	**#**	**Study ID**
1	Lexicon-based	26	[PS03], [PS04], [PS06], [PS07], [PS08], [PS10], [PS11], [PS12], [PS13], [PS18], [PS20], [PS21], [PS24], [PS25], [PS27], [PS31], [PS35], [PS36], [PS38], [PS40], [PS43], [PS54], [PS55], [PS56], [PS58], [PS59]
2	Machine learning	13	[PS04], [PS14], [PS23], [PS26], [PS30], [PS32], [PS38], [PS39], [PS41], [PS42], [PS45], [PS57], [PS63]
3	Questionnaire	23	[PS02], [PS09], [PS16], [PS17], [PS19], [PS28], [PS29], [PS30], [PS37], [PS44], [PS46], [PS47], [PS48], [PS49], [PS50], [PS51], [PS60], [PS61], [PS62], [PS64], [PS65], [PS66]
4	Manual annotation	18	[PS05], [PS08], [PS20], [PS22], [PS23], [PS25], [PS26], [PS31], [PS32], [PS34], [PS38], [PS39], [PS40], [PS41], [PS56], [PS57], [PS58], [PS63]
5	Coding strategy	8	[PS29], [PS30], [PS33], [PS37], [PS52], [PS53], [PS60], [PS62]
6	Other	1	[PS15]
Note: one study can use more than one approach			

References

[1] I. Sommerville, Software Engineering, 9th ed., Addison-Wesley, 2010.

[2] D. Gachechiladze, F. Lanubile, N. Novielli, A. Serebrenik, Anger and Its Direction in Collaborative Software Development, in: 2017 IEEE/ACM 39th International Conference on Software Engineering: New Ideas and Emerging Technologies Results Track (ICSE-NIER), 2017: pp. 11–14. doi:10.1109/ICSE-NIER.2017.18.

[3] M.R. Wrobel, Emotions in the software development process, in: 2013 6th International Conference on Human System Interactions (HSI), 2013: pp. 518–523. doi:10.1109/HSI.2013.6577875.

[4] R. Colomo-Palacios, C. Casado-Lumbreras, P. Soto-Acosta, Á. García-Crespo, Using the Affect Grid to Measure Emotions in Software Requirements Engineering, Journal of Universal Computer Science. 17 (2011) 1281–1298. doi:10.3217/jucs-017-09-1281.

[5] L.F. Capretz, Bringing the Human Factor to Software Engineering, IEEE Software. 31 (2014) 104–104. doi:10.1109/MS.2014.30.

[6] D. Graziotin, X. Wang, P. Abrahamsson, Happy software developers solve problems better: psychological measurements in empirical software engineering, PeerJ. 2 (2014) e289. doi:10.7717/peerj.289.
[7] R. Colomo-Palacios, J.M. Gomez-Berbis, A. Garcia-Crespo, I. Puebla-Sanchez, Social Global Repository: using semantics and social web in software projects, International Journal of Knowledge and Learning. 4 (2008) 452–464. doi:10.1504/IJKL.2008.022063.
[8] R. Colomo-Palacios, C. Casado-Lumbreras, P. Soto-Acosta, S. Misra, F.J. García-Peñalvo, Analyzing Human Resource Management Practices Within the GSD Context, Journal of Global Information Technology Management. 15 (2012) 30–54. doi:10.1080/1097198X.2012.10845617.
[9] P. Lenberg, R. Feldt, L.G. Wallgren, Behavioral software engineering: A definition and systematic literature review, Journal of Systems and Software. 107 (2015) 15–37. doi:10.1016/j.jss.2015.04.084.
[10] D. Graziotin, X. Wang, P. Abrahamsson, The Affect of Software Developers: Common Misconceptions and Measurements, in: 2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering, 2015: pp. 123–124. doi:10.1109/CHASE.2015.23.
[11] A. Murgia, P. Tourani, B. Adams, M. Ortu, Do Developers Feel Emotions? An Exploratory Analysis of Emotions in Software Artifacts, in: Proceedings of the 11th Working Conference on Mining Software Repositories, ACM, New York, NY, USA, 2014: pp. 262–271. doi:10.1145/2597073.2597086.
[12] D. Graziotin, X. Wang, P. Abrahamsson, Do feelings matter? On the correlation of affects and the self-assessed productivity in software engineering, Journal of Software: Evolution and Process. 27 (2015) 467–487. doi:10.1002/smr.1673.
[13] E. Guzman, B. Bruegge, Towards Emotional Awareness in Software Development Teams, in: Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering, ACM, New York, NY, USA, 2013: pp. 671–674. doi:10.1145/2491411.2494578.
[14] M. Ortu, A. Murgia, G. Destefanis, P. Tourani, R. Tonelli, M. Marchesi, B. Adams, The Emotional Side of Software Developers in JIRA, in: 2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR), 2016: pp. 480–483. doi:10.1109/MSR.2016.059.
[15] A. Fountaine, B. Sharif, Emotional Awareness in Software Development: Theory and Measurement, in: 2017 IEEE/ACM 2nd International Workshop on Emotion Awareness in Software Engineering (SEmotion), 2017: pp. 28–31. doi:10.1109/SEmotion.2017.12.
[16] M. Storey, A. Zagalsky, F.F. Filho, L. Singer, D.M. German, How Social and Communication Channels Shape and Challenge a Participatory Culture in Software Development, IEEE Transactions on Software Engineering. 43 (2017) 185–204. doi:10.1109/TSE.2016.2584053.
[17] P. Lenberg, R. Feldt, L.G. Wallgren, Behavioral software engineering: A definition and systematic literature review, Journal of Systems and Software. 107 (2015) 15–37. doi:10.1016/j.jss.2015.04.084.
[18] J.L. Barros-Justo, S. Zapata, N. Martinez-Araujo, Are you sure you are happy?, IEEE Latin America Transactions. 16 (2018) 1213–1218. doi:10.1109/TLA.2018.8362159.
[19] S.S.J.O. Cruz, F.Q.B. da Silva, C.V.F. Monteiro, P. Santos, I. Rossilei, Personality in software engineering: Preliminary findings from a systematic literature review, in: 15th Annual Conference on Evaluation Assessment in Software Engineering (EASE 2011), 2011: pp. 1–10. doi:10.1049/ic.2011.0001.
[20] S. Cruz, F.Q.B. da Silva, L.F. Capretz, Forty years of research on personality in software engineering: A mapping study, Computers in Human Behavior. 46 (2015) 94–113. doi:10.1016/j.chb.2014.12.008.

[21] M. Sánchez-Gordón, R. Colomo-Palacios, Online appendix: the emotional pulse of software engineering, Figshare. (2019). https://doi.org/10.6084/m9.figshare.c.4360214.v5.
[22] M. Cabanac, What is emotion?, Behavioural Processes. 60 (2002) 69–83. doi:10.1016/S0376-6357(02)00078-5.
[23] P.R. Kleinginna, A.M. Kleinginna, A categorized list of emotion definitions, with suggestions for a consensual definition, Motiv Emot. 5 (1981) 345–379. doi:10.1007/BF00992553.
[24] C.D. Batson, L.L. Shaw, K.C. Oleson, Differentiating affect, mood, and emotion: Toward functionally based conceptual distinctions, in: Emotion, Sage Publications, Inc, Thousand Oaks, CA, US, 1992: pp. 294–326.
[25] K.R. Scherer, Psychological models of emotion, in: The Neuropsychology of Emotion, Oxford University Press, New York, NY, US, 2000: pp. 137–162.
[26] E. Shouse, Feeling, Emotion, Affect, M/C Journal. 8 (2005) 1.
[27] P.A. Thoits, The sociology of emotions, Annual Review of Sociology. 15 (1989) 317–342. doi:10.1146/annurev.so.15.080189.001533.
[28] R.W. Picard, Affective computing, 1st pbk. ed., MIT Press, Cambridge, MA, USA, 2000.
[29] K.S. Fleckenstein, Defining Affect in Relation to Cognition: A Response to Susan McLeod, Journal of Advanced Composition. 11 (1991) 447–453.
[30] S. Gordon, The sociology of sentiments and emotion, in: M. Rosenberg, R.H. Turner (Eds.), Social Psychology: Sociological Perspectives, Basic Books, New York, 1981: pp. 562–592.
[31] H.A. Murray, C.D. Morgan, A clinical study of sentiments (I), Genetic Psychology Monographs. 32 (1945) 153–311.
[32] C. Clavel, Z. Callejas, Sentiment Analysis: From Opinion Mining to Human-Agent Interaction, IEEE Transactions on Affective Computing. 7 (2016) 74–93. doi:10.1109/TAFFC.2015.2444846.
[33] A. Yadollahi, A.G. Shahraki, O.R. Zaiane, Current State of Text Sentiment Analysis from Opinion to Emotion Mining, ACM Comput. Surv. 50 (2017) 25:1–25:33. doi:10.1145/3057270.
[34] F. Calefato, F. Lanubile, F. Maiorano, N. Novielli, Sentiment Polarity Detection for Software Development, Empir Software Eng. 23 (2018) 1352–1382. doi:10.1007/s10664-017-9546-9.
[35] A. Mehrabian, Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament, Current Psychology. 14 (1996) 261–292. doi:10.1007/BF02686918.
[36] R. Plutchik, Emotions : a general psychoevolutionary theory, Approaches to Emotion. (1984) 197–219.
[37] R.W. Levenson, Basic Emotion Questions, Emotion Review. 3 (2011) 379–386. doi:10.1177/1754073911410743.
[38] P. Ekman, D. Cordaro, What is Meant by Calling Emotions Basic, Emotion Review. 3 (2011) 364–370. doi:10.1177/1754073911410740.
[39] P. Ekman, Universals and cultural differences in facial expressions of emotion, Nebraska Symposium on Motivation. 19 (1971) 207–283.
[40] P. Ekman, W.V. Friesen, P. Ellsworth, Emotion in the human face: Guidelines for research and an integration of findings, Pergamon Press, Oxford, England, 1972.
[41] P. Ekman, An argument for basic emotions, Cognition and Emotion. 6 (1992) 169–200. doi:10.1080/02699939208411068.
[42] P. Shaver, J. Schwartz, D. Kirson, C. O'Connor, Emotion knowledge: further exploration of a prototype approach, J Pers Soc Psychol. 52 (1987) 1061–1086. doi:http://dx.doi.org/10.1037/0022-3514.52.6.1061.
[43] W.G. Parrott, Emotions in social psychology: Essential readings, Psychology Press, New York, NY, US, 2001.

[44] J. Posner, J.A. Russell, B.S. Peterson, The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Dev Psychopathol. 17 (2005) 715–734. doi:10.1017/S0954579405050340.
[45] D. Watson, L.A. Clark, A. Tellegen, Development and validation of brief measures of positive and negative affect: the PANAS scales, J Pers Soc Psychol. 54 (1988) 1063–1070. doi:http://dx.doi.org/10.1037/0022-3514.54.6.1063.
[46] E. Diener, D. Wirtz, W. Tov, C. Kim-Prieto, D. Choi, S. Oishi, R. Biswas-Diener, New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings, Soc Indic Res. 97 (2010) 143–156. doi:10.1007/s11205-009-9493-y.
[47] M.M. Bradley, P.J. Lang, Measuring emotion: The Self-Assessment Manikin and the semantic differential, Journal of Behavior Therapy and Experimental Psychiatry. 25 (1994) 49–59. doi:10.1016/0005-7916(94)90063-9.
[48] J. Russell, A. Weiss, G. Mendelsohn, The Affect Grid - a Single-Item Scale of Pleasure and Arousal, J. Pers. Soc. Psychol. 57 (1989) 493–502. doi:10.1037//0022-3514.57.3.493.
[49] K.R. Scherer, What are emotions? And how can they be measured?, Social Science Information. 44 (2005) 695–729. doi:10.1177/0539018405058216.
[50] G. Coppin, D. Sander, 1 - Theoretical Approaches to Emotion and Its Measurement, in: H.L. Meiselman (Ed.), Emotion Measurement, Woodhead Publishing, 2016: pp. 3–30. doi:10.1016/B978-0-08-100508-8.00001-1.
[51] Microsoft Docs, Detect faces in an image - Face API - Azure Cognitive Services, (n.d.). https://docs.microsoft.com/en-us/azure/cognitive-services/face/face-api-how-to-topics/howtodetectfacesinimage.
[52] T. Vogt, E. André, N. Bee, EmoVoice — A Framework for Online Recognition of Emotions from Voice, in: E. André, L. Dybkjær, W. Minker, H. Neumann, R. Pieraccini, M. Weber (Eds.), Perception in Multimodal Dialogue Systems, Springer Berlin Heidelberg, 2008: pp. 188–199.
[53] S.C. Müller, T. Fritz, Stuck and Frustrated or in Flow and Happy: Sensing Developers' Emotions and Progress, in: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, 2015: pp. 688–699. doi:10.1109/ICSE.2015.334.
[54] R. Jongeling, P. Sarkar, S. Datta, A. Serebrenik, On negative results when using sentiment analysis tools for software engineering research, Empir Software Eng. 22 (2017) 2543–2584. doi:10.1007/s10664-016-9493-x.
[55] T. Fritz, A. Begel, S.C. Müller, S. Yigit-Elliott, M. Züger, Using Psycho-physiological Measures to Assess Task Difficulty in Software Development, in: Proceedings of the 36th International Conference on Software Engineering, ACM, New York, NY, USA, 2014: pp. 402–413. doi:10.1145/2568225.2568266.
[56] W. Michel, Y. Shoda, R.E. Smith, Introduction to Personality: Towards an Integration, John Wiley, New York, 2004.
[57] S.S. Gaur, H. Herjanto, M. Makkar, Review of emotions research in marketing, 2002–2013, Journal of Retailing and Consumer Services. 21 (2014) 917–923. doi:10.1016/j.jretconser.2014.08.009.
[58] A.J. Wearden, N. Tarrier, C. Barrowclough, T.R. Zastowny, A.A. Rahill, A review of expressed emotion research in health care, Clinical Psychology Review. 20 (2000) 633–666. doi:10.1016/S0272-7358(99)00008-2.
[59] S. Garrido, A systematic review of the studies measuring mood and emotion in response to music., Psychomusicology: Music, Mind, and Brain. 24 (20150406) 316. doi:10.1037/pmu0000072.
[60] E. Schubert, Emotion felt by the listener and expressed by the music: literature review and theoretical perspectives, Front. Psychol. 4 (2013). doi:10.3389/fpsyg.2013.00837.

[61] B. Kitchenham, S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, School of Computer Science and Mathematics, Keele University, 2007.
[62] T. Shaw, The Emotions of Systems Developers: An Empirical Study of Affective Events Theory, in: Proceedings of the 2004 SIGMIS Conference on Computer Personnel Research: Careers, Culture, and Ethics in a Networked Environment, ACM, New York, NY, USA, 2004: pp. 124–126. doi:10.1145/982372.982403.
[63] D. Graziotin, X. Wang, P. Abrahamsson, Software Developers, Moods, Emotions, and Performance, IEEE Software. 31 (2014) 24–27. doi:10.1109/MS.2014.94.
[64] K. Krippendorff, Computing Krippendorff's Alpha-Reliability, 2011.
[65] S. Jalali, C. Wohlin, Systematic literature studies: Database searches vs. backward snowballing, in: Proceedings of the 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, 2012: pp. 29–38. doi:10.1145/2372251.2372257.
[66] B. Kitchenham, P. Brereton, A systematic review of systematic review process research in software engineering, Information and Software Technology. 55 (2013) 2049–2075. doi:10.1016/j.infsof.2013.07.010.
[67] V. Garousi, J.M. Fernandes, Highly-cited papers in software engineering: The top-100, Information and Software Technology. 71 (2016) 108–128. doi:10.1016/j.infsof.2015.11.003.
[68] P. Ekman, R. Davidson, The Nature of Emotion: Fundamental Questions, Series in Affective Science, Oxford University Pr ess, 1995.
[69] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, C. Potts, A computational approach to politeness with application to social factors, in: 2013: p. 10.
[70] A. Langlotz, M.A. Locher, (Im)politeness and Emotion, in: J. Culpeper, M. Haugh, D.Z. Kádár (Eds.), The Palgrave Handbook of Linguistic (Im)Politeness, Palgrave Macmillan UK, London, 2017: pp. 287–322. doi:10.1057/978-1-137-37508-7_12.
[71] A.B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas, Behav Res. 45 (2013) 1191–1207. doi:10.3758/s13428-012-0314-x.
[72] P. Warr, The measurement of well-being and other aspects of mental health, Journal of Occupational Psychology. 63 (1990) 193–210. doi:10.1111/j.2044-8325.1990.tb00521.x.
[73] L.F. Capretz, Bringing the Human Factor to Software Engineering, IEEE Software. 31 (2014) 104–104. doi:10.1109/MS.2014.30.
[74] M. Ortu, G. Destefanis, S. Counsell, S. Swift, R. Tonelli, M. Marchesi, Arsonists or Firefighters? Affectiveness in Agile Software Development, in: H. Sharp, T. Hall (Eds.), Agile Processes, in Software Engineering, and Extreme Programming, Springer International Publishing, 2016: pp. 144–155.
[75] P. Ekman, What Scientists Who Study Emotion Agree About, Perspect Psychol Sci. 11 (2016) 31–34. doi:10.1177/1745691615596992.
[76] D. Graziotin, F. Fagerholm, X. Wang, P. Abrahamsson, What happens when software developers are (un)happy, Journal of Systems and Software. 140 (2018) 32–47. doi:10.1016/j.jss.2018.02.041.
[77] C. Wohlin, P. Runeson, M. Hst, M.C. Ohlsson, B. Regnell, A. Wessln, Experimentation in Software Engineering, Springer Science & Business Media, 2012.
[78] E. Moreno-Campos, M.-L. Sánchez-Gordón, R. Colomo-Palacios, A. de Amescua Seco, Towards Measuring the Impact of the ISO/IEC 29110 Standard: A Systematic Review, in: Proceedings of 21st EuroSPI 2014 Conference, Springer-Verlag, Luxembourg, 2014: pp. 1–12. doi:10.1007/978-3-662-43896-1_1.