Emotional Dashboard: a Non-Intrusive Approach to Monitor Software Developers' Emotions and Personality Traits

Leo Silva[1,*], Marília Castro[2], Miriam Silva[2], Milena Santos[2], Uirá Kulesza[3],
Margarida Lima[2], and Henrique Madeira[1]

[1]Centre of Informatics and Systems, University of Coimbra, Polo II, Pinhal de Marrocos, 3030-290, Coimbra, Portugal
[2]Psychology and Educational Sciences, University of Coimbra, Colégio Novo Street, 3001-802, Coimbra, Portugal
[3]Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte, 59072-970, Natal, Brazil
leo.moreira@me.com, castro.mariliag@gmail.com, bernardino.miriam@gmail.com, milena.nestor@hotmail.com,
uirakulesza@gmail.com, mplima@fpce.uc.pt, henrique@dei.uc.pt
*corresponding author

***Abstract*—Developers' emotions are crucial elements that influence the overall job satisfaction of software engineers, including motivation, productivity, and quality of the work, affecting the software development lifecycle. Existing approaches to assess and monitor developers' emotions, such as facial expressions, self-assessed surveys, and biometric sensors, imply considerable intrusiveness on developers' routines and tend to be used only during limited periods. This paper proposes a new non-intrusive and automatable tool (Emotional Dashboard) to assess, monitor, and visualize software developers' emotions during long periods, providing team leaders and project managers with an overview of teams' and software developers' emotional statuses. The idea is to use posts shared by developers on social media to assess their emotions' polarity and visualize the emotional situation on a dashboard, allowing the identification of potentially abnormal emotional periods that may affect the software development. A first evaluation of the tool's accuracy, done by comparing the emotion polarity (negative, positive, or neutral) of posts done by our tool with the manual classification of a set of posts done by three psychologists, has shown an accuracy of 77%. The tool is available for analysis at this link: https://emotional-dashboard.herokuapp.com.**

Keywords*—*Dashboard, software engineering, software quality, software productivity, sentiment analysis, social media.

I. Introduction

The study of the role of emotions in work and professional contexts is a multidisciplinary endeavor involving organizational psychologists, process management, and field domain specialists. Many research studies have established different facets of the impact of human factors on professional work in general [1] and on people's performance at work [2]. For instance, Frost [3] states that unhappy employees tend to be disconnected from their work, which can lead to low productivity and low quality of work, while Diener et al. [4] found that positive emotions influence key variables for workplace success.

Software development is an excellent example of a human-intensive intellectual activity where human factors play a significant role [5]. Modern software development approaches rely on social and communicative processes [6], especially in large-scale software projects, where human factors play a key role [7]. For instance, a developer's negative emotional state maybe not be strong enough to indicate unfavorable project progress. Still, a continuous negative emotion over a few weeks is likely to imply some project problem [8]. The human-intensive nature of the tasks involved in the software development process increases the importance of emotions and personality traits for the effectiveness of the software development process.

This paper proposes an approach to assess, monitor, and visualize developers' emotional states over long periods without causing explicit intrusion or disturbance in software development activities. There are existing approaches to assess developers' emotions, such as self-assessed surveys [9], facial expressions analysis [10], and biometric sensors attached to the developers' body [11], among others. However, These approaches imply some degree of intrusiveness and disturbance in regular software development activities. They are not effectively used in practice for long periods, as required by real-world software projects.

Our approach uses posts shared by developers on social media to assess their emotional polarities (inside and outside working periods) through sentiment analysis techniques [12]. We have developed a prototype tool called Emotional Dashboard to display the emotional information of software developers. The tool aims to provide software managers/scrum masters/team leaders with valuable data to identify potential abnormal periods of negative/positive sentiments of developers that may affect the quality of the software developed and developers' productivity.

Publicly available information on social media offers a rich source of information that can be used to monitor emotions in a non-intrusive way. Additionally, the use of social media is highly disseminated among software developers, and the required sentiment analysis techniques and tools needed to extract people's emotions from social media posts are mature techniques and readily available [12]. These features make the proposed approach fully automatable and applicable with minimal effort in real projects since all the required elements are effectively available.

Compared to other emotional dashboard approaches in software engineering [13] [8], our approach has several ad-

2693-9177/22/$31.00 ©2022 IEEE
DOI 10.1109/QRS57517.2022.00045

vantages. First, we used open-context social media as a data source, allowing data gathering outside the working period. Second, we employ personality traits information to ponder the sentiment analysis results. Our approach also provides a diverse set of visualizations to be used by software managers. Last, our approach is an end-to-end solution, from data collection to data visualization to software managers.

We organize the rest of this paper as follows: section II briefly introduces basic concepts involved in this work; section III presents the components of our approach; section IV shows an experimental study to provide a first approach evaluation, providing details on the utilization of sentiment analysis tools and presenting the dashboard visualizations; section V discusses the preliminary results and the approach application; section VI presents the related work; and section VII provide our conclusions and outlines future research directions.

II. Background

Our approach uses concepts of two psychology topics (emotions and personality traits of an individual), as well as methods and techniques established in the area of text sentiment analysis. This subsection briefly summarizes the relevant concepts.

A. Emotions

Emotions are states of mind raised through external or internal stimuli [14]. More broadly, emotions are intense feelings that are directed to someone or something [15]. Psychologists have proposed many theories and models to classify human emotions [16]. We can mention The Plutchik Wheel [14] and the different basic emotions sets, such as those proposed by Ekman and Friesen [17] and by Cowen and Keltner [18]. In our work, we deal with emotions in a broad way, grouping them into *positive* and *negative* categories. Thus, these emotions become mood states instead of isolating a particular one [19]. We also consider the nonemotional state, i.e., the *neutral* state.

B. Personality Traits

Psychologists consider *personality* as a person's unique long-term pattern of thinking, emotions, and behavior. Psychology also characterizes personalities in terms of traits, which are relatively enduring characteristics that influence our behavior across many situations. Introversion, friendliness, conscientiousness, honesty, and helpfulness are examples of personality traits that are important because they help explain consistencies in behavior [20]. It also refers to the special combinations of talents, values, hopes, loves, hates, and habits that make each person unique [21] [22].

There are several models and theories to classify personality traits. In this work, we use the Five-Factor model (known as Big Five) [23] (also mentioned by the acronym OCEAN), one of the most accepted and used models to trace personality traits [24]. According to McCrae and Costa [23], most human personality traits can be reduced to five large dimensions despite language or culture. A person could score low or high on each dimension or factor.

The five general factors of Big Five are *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. *Openness* refers to a person's intellect or imagination. This factor means a person's creativity and desire to adapt to and explore new things. People with low scores in this factor might be considered pragmatic and driven by conventional methods. *Conscientiousness* implies an individual's desire to pursue aims and do the tasks involved correctly. This is the factor for assessing one's diligence, efficiency, and organizational ability. A high score in this personality trait implies self-discipline, whereas a low score implies spontaneous behavior and even low reliability. *Extraversion* factor measures if an individual is more open to external interactions or prefers being discreet. Extroverted people are perceived as people with high energy levels, whereas introverts prefer more time alone and less stimulation. *Agreeableness* trait reflects social harmony or lack thereof. These people are considered benevolent, trusting, helpful, and willing to compromise for the greater good. On the contrary, disagreeable individuals are prone to be selfish, skeptical, and unfriendly. *Neuroticism* refers to a person's ability to feel negative emotions such as anxiety and anger. This trait measures a person's emotional instability from calm to overwhelmed.

C. Text Sentiment Analysis

There are two main approaches for text sentiment analysis: unsupervised lexicon-based and supervised machine learning-based techniques. Machine-learning approaches are potentially more effective, but they have the disadvantage of needing a sizeable training corpus to develop a classifier. When a training corpus is unavailable (normally when looking for a large training corpus), the alternative is to use an existing sentiment lexicon-based to perform sentiment analysis.

In this paper, we employed the following lexicon-based techniques, all of them created or adapted to the Brazilian Portuguese language (the native language of the developers that have participated in the evaluation of the tool):

- *SentiStrength* [25]: a well-known sentiment analysis method that uses a lexical dictionary labeled by humans enhanced by machine learning. This method used an expanded version of the LIWC dictionary, adding new characteristics for the context of social media;
- *Sentilex-PT* [26]: is a sentiment lexicon specifically designed for the sentiment and opinion analysis about human entities in texts written in Portuguese, consisting of 7,014 lemmas and 82,347 inflected forms;
- *Linguistic Inquiry and Word Count (LIWC)* [27]: aims to analyze texts to detect emotional, social, cognitive words and standard linguistic dimensions of texts. Although LIWC has several metrics, we employed only those related to emotional polarities in this study.

III. The emotional dashboard concept

The proposed tool to assess, monitor, and visualize software developers' emotions during long periods comprises four primary modules: Data Collection, Data Pre-processing, Sentiment Analysis, and Visualization, as shown in Figure 1.

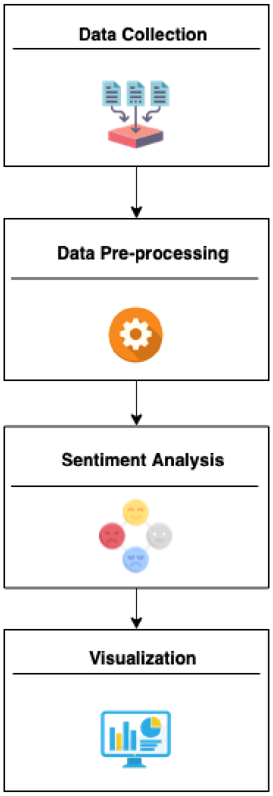


Figure 1. Emotional Dashboard Workflow.

The *Data Collection* collects Twitter public posts and manages a two-step survey to assess developers' personality traits. The first step aims to collect demographic information, and the second comprises answering an instrument (i.e., a set of calibrated questions) to assess personality traits (Big Five Inventory).

The tool periodically collects the developers' tweets (e.g., daily) and performs *Data Pre-processing* to remove stop words, URLs, mentions to other Twitter users, and tweets containing multimedia attachments (videos, photos, and animated gifs). Then, the tool performs *Sentiment Analysis* on the cleaned text (task 3) using an ensemble of dictionary lexicons to classify the tweets regarding their polarities into three categories: *negative*, *neutral*, and *positive*. In this module, we use ground truth established by psychologists. The next step is to use all information gathered and produced to build a set of visualizations in the Emotional Dashboard prototype tool.

It is worth mentioning that the techniques used in the current implementation of the Emotional Dashboard tool do not limit the proposed approach, as we can add other techniques in future versions, especially in the Sentiment Analysis component.

IV. Experimental Study and Preliminary Results

This section presents the first utilization of the Emotional Dashboard with real data from sixteen Twitter users, all software developers. The goal is to demonstrate the use of the dashboard and evaluate the accuracy of the assessment of the emotions polarity (i.e., negative, positive, or neutral) obtained from the posts on Twitter.

We have invited a total of 45 Twitter users (selected through searching in the ”Programming (Technology)” topic of Twitter) according to the following criteria: i) had an open profile, with explicit location and direct message enabled; ii) had at least one tweet per day during the study period; and iii) have posts mainly in the Brazilian Portuguese language (to avoid the added complexity of using more than one language). Sixteen of them voluntarily and anonymously agreed to participate in the experiment.

TABLE I
Participants' characteristics.

		Quantity (%)
Gender	***Male***	**10 (62.5)**
	Female	6 (37.5)
Age	*Less than 20*	1 (6.25)
	21-30	**9 (56.25)**
	31-40	6 (37.5)
Experience in Software Development	*1-3 years*	3 (18.8)
	3-5 years	3 (18.8)
	5-7 years	1 (6.3)
	7-10 years	3 (18.8)
	More than 10 years	**6 (37.5)**
Schooling	*High School*	2 (12.5)
	Higher Education	**14 (87.5)**

Table I displays participants' main characteristics, with bold values highlighting the highest values. Most participants (9/16) were young adults with more than ten years of software development experience (6/16) and with higher education (14/16). The table also shows that the group of participants includes more males (62.5%) than females (37.5%), which reflects the current imbalanced gender situation in the software industry.

A. Data Collection

We ask participants to complete a survey with demographic data such as gender, age, and experience working with software development. In this survey, participants read documents related to ethics and privacy and signed informed consent to participate in the experiment. They also answered the Big Five Inventory to assess their personality traits.

We collected 91,632 (mean $= 5,727$; std $= 5,256$) participants' tweets within 36 months, from March 31, 2018, to March 31, 2021. These tweets cover any topics that authors posted without textual content restrictions applied by our approach. To collect solely tweets written in Portuguese, we set a parameter on the Twitter API, choosing the ”pt” value. Nevertheless, the API has a limitation on the adopted endpoint: it can return only the 3,200 most recent tweets, retweets, replies, and quote tweets posted by the user. In some cases, this limitation can be enough to collect a specific user's entire history of tweets. However, in other cases, it is not. To overcome this and retrieve more tweets than the established limit, we employ a crawler that visits the web page of the user's Twitter timeline and extracts the needed information.

TABLE II
UNSUPERVISED LEXICON-BASED CLASSIFICATION METRICS FOR THE THREE SENTIMENT LEXICONS AND ENSEMBLES.

Lexicon	Positive			Negative			Accuracy	Macro F1-Score
	Precision	*Recall*	*F1-Score*	*Precision*	*Recall*	*F1-Score*		
SS	0.82432	0.84138	0.83276	0.62295	0.59375	0.60800	0.76555	0.72038
SS + SL	0.79012	0.82051	0.80503	0.69565	0.65306	0.67368	0.75591	0.73936
SS + SL + LI	0.80357	0.83851	0.82067	**0.69767**	0.64516	0.67039	0.76772	**0.74553**
SS + LI	0.81176	0.85714	0.83384	0.65151	0.57333	0.60993	0.76695	0.72188
SL	0.76577	0.74561	0.75556	0.67416	**0.69767**	**0.68571**	0.72500	0.72063
SL + LI	0.79861	0.78231	0.79038	0.64835	0.67045	0.65922	0.74043	0.72480
LI	**0.82482**	**0.86923**	**0.84644**	0.60465	0.52000	0.55914	**0.77222**	0.70279
Mean	0.80271	0.82210	0.81210	0.65642	0.62192	0.63801	0.75625	0.72505
Std	0.01919	0.04058	0.02896	0.03261	0.05750	0.04302	0.01610	0.01293

B. Data Pre-processing

We evaluated only text-based tweets, removing those containing multimedia attachments (images, videos, and gifs). These multimedia files could interfere with the tweet's sentiment classification. Thus, we performed data preparation and cleaning steps. Among data cleaning operations for each tweet, we performed retweets exclusion and retweet handles removal, composed of ”RT” and a user citation: ”@user.” Similarly, we removed the user citation in an original tweet (those that are not retweets). Lastly, the next step was to remove stop words, repeated letters, and URL links from tweets. The final dataset used in the experiment consisted of **79,029** (mean = $4,939$; std = $3,421$) original and text-based tweets.

C. Sentiment Analysis

The analysis consisted of a manual analysis for polarity classification of a sample of the posts by a group of psychologists (used as ground truth) and an automated analysis of the whole dataset through three existing lexicon-based sentiment analysis methods.

We invited three psychologists (*evaluators*) to collaborate in the experiment by manually classifying the polarity of a sample of tweets to establish ground truth. Evaluators independently analyzed a randomly generated sample of 35 posts for each participant, totaling 560 anonymous tweets. The goal of this analysis was to manually classify each tweet regarding its polarity in the scale: *negative*, *neutral*, and *positive*. The psychologists used the following criteria:

- Considering a tweet as positive if the review expresses a positive sentiment after analyzing all the characteristic terms of the language;
- Considering a tweet as negative if the review expresses a negative sentiment after analyzing all the characteristic terms of the language;
- Considering a tweet as neutral if the review expresses a neutral sentiment, i.e., without any positive or negative terms characteristic of the language;
- Considering a tweet as neutral if it presents factual information;
- A tweet containing both negative and positive sentiments is considered negative or positive based on the sentiment most relevantly presented;
- Considering positive and negative emphasis, such as emojis, punctuation, and capital letters;
- Analyzing only the tweet text;
- Do not access external links.

Using the same criteria, we also asked the participants to classify their tweets in the same sample to provide additional important control on the quality of the classification of the posts.

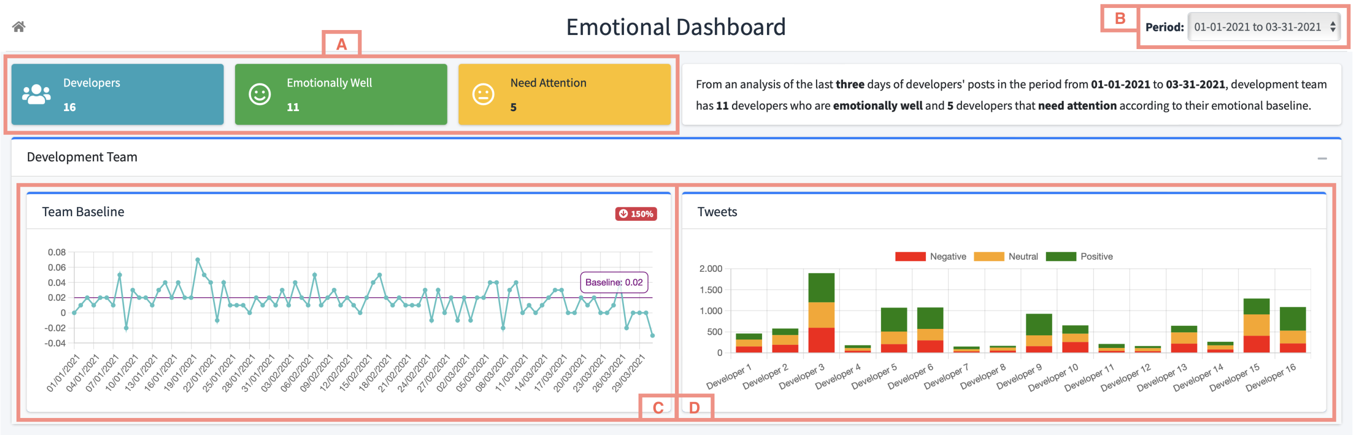
Previous work [28] [29] [30] [31] [32] employed an annotation process using Shaver framework [33]. These previous works used non-experts as evaluators, such as computer science students [30] [32] and IT professionals [32]. As suggested by [29], we provided clear guidelines for tweets' manual classification by our evaluators and participants. We achieved a Cohen's Kappa index ($\kappa = 0.710$) similar to those achieved by [30] ($\kappa = 0.740$) and [31] ($\kappa = 0.740$). Differently from previous work, we used experts to perform manual classification and confirmed the reliability of this process with the tweets' authors: the participants.

This preliminary result suggests that psychologists could label a sample of developer's tweets with a polarity in a real software development environment. This labeling can be further compared with the results of lexicons or ensemble inference to evaluate their accuracy, F1-Score, and other metrics.

Another element of our approach is the lexicons used to perform sentiment analysis over tweets. As mentioned, we used three lexicons: SentiStrength, Sentilex-PT, and Linguistic Inquiry and Word Count (LIWC). We executed these lexicons separately and in ensembles in a total of 7 combinations. After each execution, we store each tweet's generated scores and polarities in a database. We kept the classification scale adopted by lexicons and by previous works in sentiment polarity classification consisting of three categories: *negative*, *neutral*, or *positive*.

In addition to the lexicon analysis, we also analyzed the polarity of emojis in every tweet through the ranking elaborated by Novak et al. [34]. Thus, the analysis of each post includes the text itself and the emojis, resulting in the final polarity classification of each post as positive, negative, or neutral.

In this experiment, we employ the score from -1 to $+1$ to indicate a negative or positive post and use threshold values of -0.05 and $+0.05$ to define the polarity categories, as proposed



From an analysis of the last **three** days of developers' posts in the period from **01-01-2021** to **03-31-2021**, development team has **11** developers who are **emotionally well** and **5** developers that **need attention** according to their emotional baseline.

Figure 2. Emotional Dashboard: team view.

by Hutto and Gilbert [35]. *Negative* messages have scores below the negative threshold, *positive* messages have scores above the positive threshold, and the remainder are *neutral* messages.

Table II presents the preliminary results of the automated classification obtained using lexicons and ensembles of lexicons. Due to paper size restrictions and readability reasons, we present preliminary results with Macro F1-Score greater than 0.70. We calculate precision, recall, and F-Score using the manual classification done by the psychologists as ground truth. We represent the lexicons in Table II using these acronyms: SentiStrength (SS), Sentilex-PT (SL), and LIWC (LI).

Our preliminary results show that all lexicons and ensembles performed better on classifying positive than negative tweets, scoring an F1-Score mean of 0.812, precision mean of 0.803, and recall mean of 0.822. For negative tweets, they scored an F1-Score mean of 0.638, precision mean of 0.656, and recall mean of 0.622. The evaluation of these lexicons available in the literature also reported differences between negative and positive performances [36] [37] [38] [39].

The best preliminary results are spread among several lexicons and ensembles, as observed in the bold blue metrics in Table II. LI achieved the best precision score (0.825), recall (0.869), and F1-Score (0.846) for positive posts. For negative tweets, the best results were concentrated in SL (recall of 0.698; F1-Score of 0.686) and an ensemble of SS, SL, and LI (precision of 0.698). In general, the LI lexicon provided better quality in inferring positive polarities, while SL achieved the best results for negative tweets. However, considering the best accuracy and macro F1-Score, LI (accuracy = 0.772; macro F1-Score = 0.703) and the ensemble SS + SL + LI (accuracy = 0.768; macro F1-Score = 0.745) achieved the best results.

D. Visualization

The Emotional Dashboard has a set of visualizations in a web application to support software managers/scrum masters/team leaders' decisions to improve the software development process, mainly on software quality and developers' productivity. The visualizations offer team and individual feedback on emotional polarities over a long period, besides showing each developer's personality traits and a set of actions that managers could take. Figure 2 shows the team view of the prototype dashboard. The dashboard can be explored at: https://emotional-dashboard.herokuapp.com.

This figure presents a summary visualization of a software development team, considering the sixteen participants mentioned earlier at the beginning of this section. In this figure, region *A* shows the number of developers at the left in a cyan color. The green block in the center of region *A* exhibits the number of developers considered *emotionally well*, and the orange block on the right presents the number of developers considered that *need attention*.

The criteria to identify developers that may *need attention* is the following: a developer is in the *need attention* state if the mean of tweets' polarities in the last three consecutive days of analysis inside the established period was lower than the mean of the established period. Otherwise, the developer is considered *emotionally well*.

Figure 2 shows the period *B* of the analysis presented in the dashboard. The user can change the period by selecting another group of three months to be displayed. The chosen period applies as a filter to all dashboard visualizations.

Region *C* shows the team's emotional baseline. The line chart presents the mean of all developers' emotional baseline (in terms of polarity) for the period. The straight purple line shows the team baseline in the three months, and each dot in the cyan line represents the mean of the team baseline for a day. The top right position of the chart exhibits an indicator badge meaning the variation of the team baseline in the last three days of the selected period. In the example of Figure 2, the mean of the team is lower than the baseline at 150%.

The region *D* of Figure 2 shows a bar chart with the total of tweets of the period for each developer. This chart exhibits the total of tweets of each category, demonstrating the developers' activity on Twitter.

The web page of team' visualizations offers an overview of each software developer's polarity means within the same

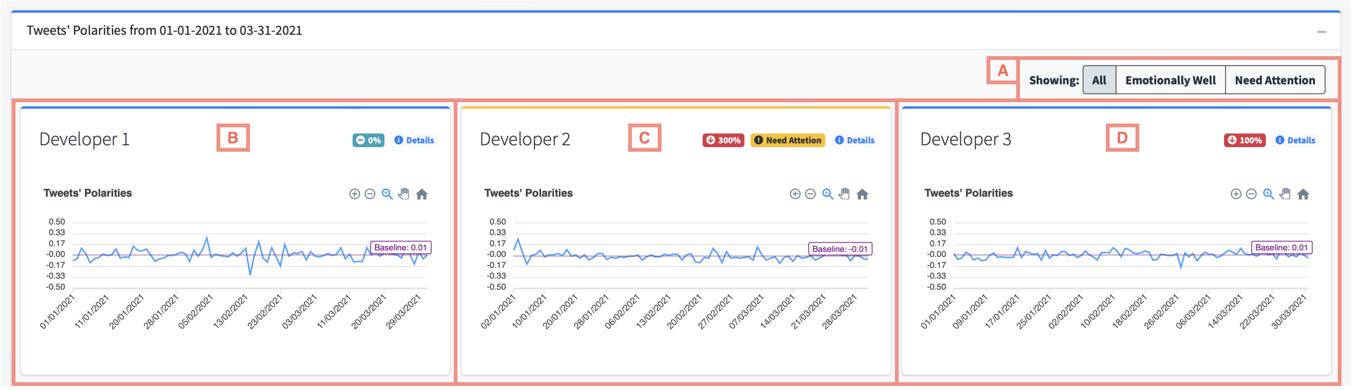


Figure 3. Emotional Dashboard: developers' overview.

period, as shown in Figure 3. This figure shows each team developer on a card with a line chart. The user can filter these cards by *all*, *emotionally well*, and *need attention*, as shown in region *A* of Figure 3.

To explain the visual components of each card, we will look at the first three software developers of the sixteen participants, shown in regions *A*, *B*, and *C*. The three cards show a line chart with the developers' polarity baseline for the period (the three months indicated in Figure 2, region *B*) drawn in purple.

Developer 1 (Figure 3, region *B*) is an example of a developer with no variation between the polarity baseline for the period and the baseline for the last three days, as indicated in the cyan badge. That means the polarity mean for the last three days was equal to the baseline mean in the three months.

Developer 2 (region *C*) is an example of a developer with the *need attention* state. This developer posted tweets with emotional polarity equal to or below the baseline in the last three days, triggering an alert indicated by the *need attention* badge. Furthermore, the mean of the polarity dropped 300% (as displayed in the red badge), always comparing the baseline for the last three days and the polarity baseline for the period.

The region *D* in Figure 3 shows an example of a developer (Developer 3) that had a decrease in the polarity mean (red badge) but posted at least one tweet with a polarity equal to or higher than the baseline. In this case, the developer did not receive the *need attention* yellow badge.

The dashboard shows a card similar to the ones in Figure 3 for each developer, with their baselines, badges, and a line chart with the mean of the tweets' polarities for each day in the period. For every card, the dashboard user can click on the *Details* badge to see detailed information for that developer.

Figure 4 shows a detailed view of a developer. Region *A* presents data on developer's experience in software development, developer's actual project, and how long the developer has been working on that project. This view shows the developer activity on Twitter, displaying the total of tweets posted during the entire period and the distribution of these tweets into the categories *negative*, *neutral*, *positive*.

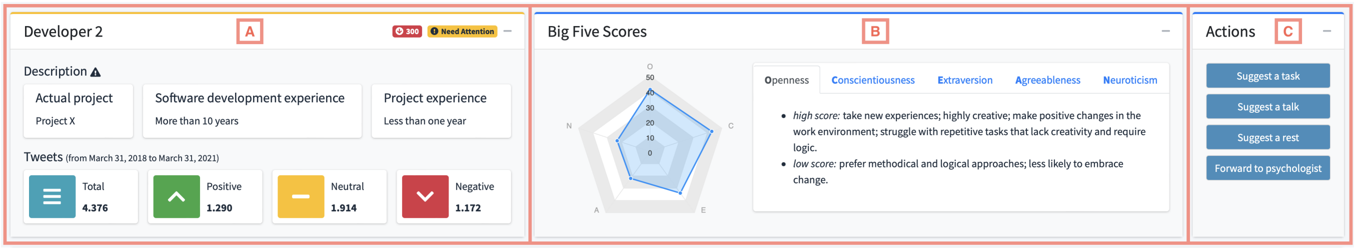
The importance of region *B* is that it shows the developer' personality traits scores in the Big Five Factor model. The left side exhibits a radar chart with the scores on each factor using the acronym OCEAN for the Five-Factor model. The right side of this region displays a tabbed panel with a brief description of persons' characteristics on scoring high or low in each factor.

With our approach, software managers/scrum masters/team leaders can visualize and monitor the development team's emotional state and take action. Region *C* suggests a set of actions, notifying both developer, manager, and psychologist (if necessary) by email:

- *Suggest a task*: this action suggests the developer engage in a simple and quick development task. Developers in the *need attention* state should be more likely to make mistakes in complex tasks;
- *Suggest a talk*: by taking this action, the manager suggests to the developer a simple conversation to understand that *need attetion* state. This conversation could be the beginning of solving a specific problem, mainly if the problem is in the working environment;
- *Suggest a rest*: in cases related to high-stress load, it may be interesting to suggest that the developer rest for a day or a shift. This action may depend on the manager's monitoring over time;
- *Forward to psychologist*: despite all the attempts and actions taken, there are cases in which the manager may realize that it is necessary to refer the developer to a professional in the field of psychology.

The last approach's visualization is a scatter chart that shows the tweets and their polarities during the selected period (in our example, from 01-01-2012 to 03-31-2021, as indicated in region *A*), as shown in Figure 5. The straight purple line shows the developer baseline, and each dot signifies one or more tweets, on a given date, for a given polarity score. For instance, if two tweets were posted on the same day and had equal polarity scores, they would be represented as one single dot. A tooltip will show how many tweets each dot denotes. Green dots mean *positive* tweets, orange dots represent *neutral* ones, and red dots mean *negative* tweets.

The scatter chart in Figure 5 contains several highlighted areas divided into two groups: light red and light green. The



Description

Actual project	Software development experience	Project experience
Project X	More than 10 years	Less than one year

Tweets (from March 31, 2018 to March 31, 2021)

Total	Positive	Neutral	Negative
4.376	1.290	1.914	1.172

Openness · Conscientiousness · Extraversion · Agreeableness · Neuroticism

- *high score:* take new experiences; highly creative; make positive changes in the work environment; struggle with repetitive tasks that lack creativity and require logic.
- *low score:* prefer methodical and logical approaches; less likely to embrace change.

Figure 4. Emotional Dashboard: developer detail summary, personality traits, and actions.

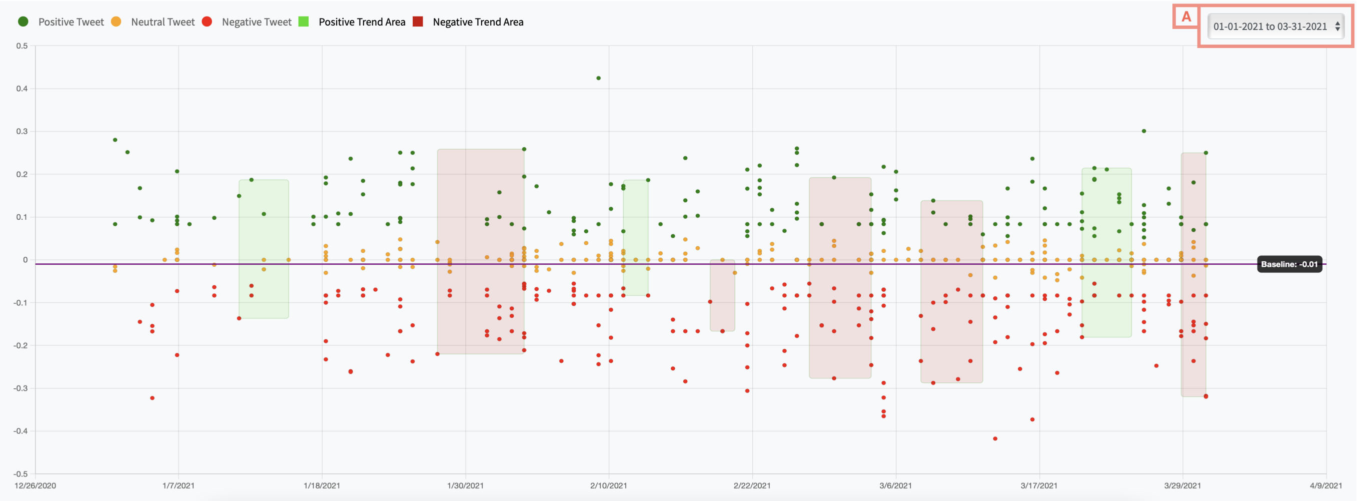


Figure 5. Emotional Dashboard: detailed developer's tweets.

light red rectangle highlights developers' days with a negative trend regarding tweets' polarities, meaning that the developer has a mean of tweets' polarity below the baseline for three consecutive days, at least. The opposite occurs for positive trends: at least three straight days of a mean of tweets' polarity above the baseline means a light green positive trends area. In the figure example, the developer has five negative and three positive trends in the selected three months period.

V. Discussion

The preliminary results suggest that the approach followed in the Emotional Dashboard tool is feasible and accurate enough to constitute a simple and non-intrusive method to assess, monitor, and visualize the polarity of developers' emotions over long periods.

Emotions have an impact on problem-solving [40]. As problem-solving tasks are the basis of the software development ecosystem, it is worth measuring and monitoring to understand those emotions. If successfully measured and visualized, a manager could aid developers by knowing and handling negative emotions over time, reducing the impact of those emotions on the software development process, mainly software quality and developers' productivity.

According to Landis and Koch [41], our preliminary result of Kappa index of 0.710 for emotionally loaded (negative or positive) [42] posts indicates a substantial strength of agreement between analysis from psychologists and participants, as mentioned in section IV-C. This result suggests that psychologists could accurately label a sample of developer's tweets with a polarity in a real software development environment and then be used as ground truth to select an ensemble of lexicons to perform automatic and long-period sentiment analysis over developers' tweets.

We employed ensemble learning techniques that generally produce more accurate polarity predictions by combining different classifiers [43]. We use a post hoc combination of three lexicons (plus emojis polarity analysis elaborated by Novak et al. [34]) to create the ensembles. For each tweet, we run each one of the lexicons to calculate their individual polarity score. For each ensemble, the tweet's polarity score is defined by the mean of the polarity scores of each lexicon involved in the ensemble. Indeed, using this strategy, we achieved a very acceptable accuracy of 0.745 and then analyzed all tweets in the dataset and defined their polarities.

An important aspect of the approach followed in the Emotional Dashboard is that we use public data from an open-context social media platform (Twitter). Using this platform, we aimed to analyze posts on a non-work-related platform to infer emotions regarding any subject posted by developers, allowing the approach to measure, monitor, and display posts produced outside the work period in a dashboard. It is worth mentioning that, although the proposed tool uses software developers as participants, it can be easily and with a minimum effort applied to any professional worker and implanted in any professional environment.

Software managers can analyze the team view (see Figure 2) to get instant feedback on how many developers they should pay attention to and eventually take action. The team view also shows the team's sentimental state (in terms of polarity) across time. Managers could link the variations shown in region *C* of Figure 2 with project outcomes. For instance, in our example, the development team is about five days with polarity mean below the team baseline, which indicates a clear persistent negative variation that may influence software development aspects, mainly software quality, and team productivity.

The cards, as displayed in Figure 3, help the software managers/scrum masters/team leaders to identify the developers with persistent negative states (below their baseline for the last three days). It is possible to perceive the severity of the variation in a simple and objective colored badge and navigate to see detailed information about the developers. Along with the team's view in Figure 2, managers have valuable information to compare with the progress of the project, check in detail the sentimental state of each developer, and then make decisions.

To our knowledge, our proposed dashboard is the first that includes personality traits data as crucial visual information to evaluate, interpret, and ponder the variations of tweets' polarities. The developer's personality traits and tweets polarities may also support managers in distributing tasks.

The Big-Five factor scores presented in Figure 4 shows that *Developer 2* scored high on *Openness* and *Conscientiousness*. These scores mean that *Developer 2* appears to be cautious in managing its social media profiles [44] and, simultaneously, tends to have more extensive networks [45]. Knowing the standard behavior of developers through their Big-Five Factor scores can help dashboard users to ponder the emotional scores and to support software managers/scrum masters/team leaders' decisions over each developer.

One of the main visualizations of the prototype dashboard is Figure 5, where the manager can observe detailed monitoring of developer's emotional polarities over time. In this figure, we can clearly perceive the high variability of negative and positive tweets. Furthermore, certain days or even several weeks present predominantly positive or negative emotions polarity, as highlighted in light red and light green boxes.

Developer 2 interleaved periods of intense activity with less activity on Twitter, and the sentiment polarity of posts changed considerably between positive and negative. Our approach allows managers to take a closer look into small periods (days or weeks) and identify short periods of polarity (positive or negative) predominance.

This kind of dashboard showing the sentiment polarity of all software development team members can provide helpful information for project managers and software team members. Additionally, the approach accurately identifies abrupt changes in developers' sentiment polarity (*need attention* badge), indicating that some important event may have occurred to certain developers or even they need professional help.

A. Challenges and Limitations

Although the preliminary results of the proposed approach are quite promising, it is important to discuss possible difficulties. The reduced number of participants engaged in the study is a limitation of the sentiment analysis' evaluation. However, we consider that the number of sixteen software developers is enough to draw reasonably reliable conclusions, especially considering the vast number of posts (79,029) and three years considered for the participants' social media activity. Another difficulty is that our approach assumes that the developers have frequent post activity to generate enough data for the analysis. However, this does not seem to be a strong limitation. In fact, all the sixteen developers that participated in our study had quite intensive and regular social media activity.

The manual evaluation of the polarity of a sample of tweets by the evaluators is another element of concern. The evaluators reported a few difficulties in classifying some posts, mainly related to the context of tweets with content about software engineering. However, the self-classification of the same tweets performed by the participants strongly agrees with the classification performed by the psychologists (Cohen's Kappa Index of 0.710), indicating that the polarity annotation used as ground truth is reliable.

A clear challenge to our approach is to evaluate the prototype dashboard in a real software development environment. We conducted the study with sixteen real software developers. We developed the dashboard based on a dataset of their tweets, but we did not evaluate the proposed dashboard in their companies, which is one of our future works.

B. Ethics and Privacy

Using social media as a data source may raise privacy and ethical issues. We ask participants to read documents about data protection and informed consent statements and agree with the terms and conditions to enroll in the study. The documents informed them about what data we need, how we collect it, what we will do, and how we guarantee anonymity and privacy. Afterward, participants authorize us to collect these data and perform the study. We anticipate that the real utilization of the proposed approach in software companies should go for a similar informed consent process.

Social media companies store data for long periods, and much of this data is searchable. We must ensure anonymity, and protecting the participants' identity is a key issue when dealing with post data. As a final remark, we have asked for approval from the Research Ethics and Deontology Committee of the Faculty of Psychology of our University. The committee unanimously approved our experiment.

VI. Related Work

Our approach proposes an instrument to identify potentially abnormal periods of negative or positive sentiments that may affect software development, mainly software quality and developers' productivity. Currently available methods to assess developers' emotions include self-assessed surveys [9], facial expressions analysis [10], and biometric sensors and wearable devices used by developers' body [11]. These methods imply some intrusiveness and disturbance in regular software development activities.

Concerning lexicons, the language of the posts does not play a relevant role in our approach, as the results and conclusions are not dependent on the language. Since we used lexicons for the Brazilian Portuguese language, it is important to review relevant related work that used such lexicons, even if they are not used to evaluate software developers' posts over long periods.

Souza and Vieira [39] evaluated models of negation and pre-processing techniques using the lexicons Sentilex-PT and OpLexicon [46]. Using tweets written in the Brazilian Portuguese language, the authors obtained a 0.55 of F1-Score on classifying positive tweets and 0.45 for negative. Also considering a different context and approach, Ruiz et al. [47] proposed the LexReLi, a context-sensitive lexicon approach for analyzing book reviews in Brazilian Portuguese. The authors created an ensemble that involves Sentilex-PT, OpLexicon, and LIWC lexicons, achieving an accuracy of 77.98%. The UniLex lexicon uses a context-dependent database of 14,084 tweets for its evaluation [48]. This approach can not be employed to assess software developers' sentiments because its creation is context-dependent and produces inaccurate results for use in the real software development environment.

Our approach uses domain-independent lexicons to perform sentiment analysis which acceptable accuracy (0.768) over Twitter posts written originally in the Brazilian Portuguese language (translation steps are not reliable [37]). We aim to analyze posts on the common social media platform to infer emotions regarding any subject posted by developers at any time, including posts outside the working period.

Another key element of our approach is the dashboard. Vivian et al. [13] work present a dashboard tool that extracts and communicates team role distribution and team emotion information. Their dashboard comprises team and individual views, using line and radar charts as visualizations. Another closely related previous study was made by Neupane et al. [8], proposing an approach named EmoD for supporting emotion awareness in software development. Their tool can automatically collect project teams' communication records, identify their emotions and intensities, model them into time series, and provide data management. Our work differentiates from Vivian et al. [13] and Neupane et al. [8] by i) using open-context social media as a data source; ii) creating a ground truth with specialists, iii) using an ensemble of lexicons and emoji classification to determine the tweets' polarities, iv) using personality traits information, and v) providing a diverse set of visualizations to be used by software managers.

Built for Marketing and Social Media context, emotionVis [49] is a text inference tool that automatically detects emotional dimensions from the text. The tool can detect six core emotions: joy, empowerment, excitement, fear, anger, and sadness, and it can also extract overall emotions such as aroused/calm and positive/negative reflected in a text. A social media manager can compare the emotionality of posts made by the organization with the comment chain generated in reaction to the published post. Our tool is different. We aim to improve software development, especially software quality and developers' productivity, by highlighting and warning software managers of developers' emotional states to help them to identify polarity variations and take further actions.

VII. Conclusion and Future Work

The study of human factors and their impact on software engineering has gained increasing attention. Sentiments influence software developers' motivation, with an unavoidable impact on the software development process. Self-assessed surveys, facial expressions analysis, and sensors attached to the user's body assess individual's psychological aspects that impose a non-negligible degree of intrusiveness, limiting their utilization in real software development setups.

We propose a new tool that uses information from social media to assess, monitor, and visualize software developers' sentiment polarity in a non-intrusive way. This approach can be used for long periods and does not cause any intrusiveness or disturbance in software developers, which means that it can be easily and with minimum effort applied to real-world software development setups. Our tool offers a dashboard with a set of visualizations that are extremely useful to software managers/scrum masters/team leaders to be aware of the team and individual emotional polarities over a long period, besides showing each developer's personality traits. With this information, managers can take action and make decisions to improve the software development process, mainly on software quality and developers' productivity.

Our preliminary results suggest that the approach is feasible and accurate enough to constitute a simple and non-intrusive method that could be useful in real-world software development. Indeed, we have been working on this approach, planning to conduct qualitative research with software managers in real software development teams to evaluate the feasibility of adopting the dashboard in their software development environments.

VIII. Acknowledgment

This work was supported in part by the Grant CISU-CUID/CEC/00326/2020, funded in part by the European Social Fund, through the Regional Operational Program Centro 2020.

References

[1] Howard M. Weiss and Russell Cropanzano. Affective Events Theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. *Research in Organizational Behavior*, 18(1):1–74, 1996.

[2] Teresa M Amabile, Sigal G Barsade, Jennifer S Mueller, and Barry M Staw. Affect and creativity at work. *Administrative science quarterly*, 50(3):367–403, 2005.

[3] Lars Glasø and Tina Løkke Vie. Toxic emotions at work. *Scandinavian Journal of Organizational Psychology*, 2(1):13–16, 2009.

[4] Ed Diener, Stuti Thapa, and Louis Tay. Positive emotions at work. *Annual Review of Organizational Psychology and Organizational Behavior*, 7:451–477, 2020.

[5] Nicole Novielli, Fabio Calefato, and Filippo Lanubile. The challenges of sentiment detection in the social programmer ecosystem. *7th International Workshop on Social Software Engineering, SSE 2015 - Proceedings*, pages 33–40, 2015.

[6] Margaret-Anne Storey. The evolution of the social programmer. In *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories*, pages 140–140, 2012.

[7] Robert Feldt, Lefteris Angelis, Richard Torkar, and Maria Samuelsson. Links between the personalities, views and attitudes of software engineers. *Information and Software Technology*, 52(6):611–624, 2010.

[8] Krishna Neupane, Kabo Cheung, and Yi Wang. EmoD: An End-to-End Approach for Investigating Emotion Dynamics in Software Development. *Proc. - 2019 IEEE International Conference on Software Maintenance and Evolution, ICSME 2019*, pages 252–256, 2019.

[9] Daniel Graziotin, Xiaofeng Wang, and Pekka Abrahamsson. Are happy developers more productive? The correlation of affective states of software developers and their self-assessed productivity. *Lecture Notes in Computer Science*, 7983 LNCS(Profes):50–64, 2013.

[10] A. Kolakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wrobel. Emotion recognition and its application in software engineering. *2013 6th Int. Conf. on Human System Interactions, HSI 2013*, pages 532–539, 2013.

[11] Renan Vinicius Aranha, Cleber Gimenez Correa, and Fatima L.S. Nunes. Adapting software with Affective Computing: a systematic review. *IEEE Transactions on Affective Computing*, 2019.

[12] Zulfadzli Drus and Haliyana Khalid. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714, 2019.

[13] R. Vivian, H. Tarmazdi, K. Falkner, N. Falkner, and C. Szabo. The Development of a Dashboard Tool for Visualising Online Teamwork Discussions. *Proc. - Int. Conf. on Software Eng.*, 2:380–388, 2015.

[14] Robert Plutchik and Henry Kellerman. EMOTION: Theory, Research, and Experience. *Theories of Emotion*, page ii, 1980.

[15] Nico H Frijda. Moods, emotion episodes, and emotions. *Handbook of emotions*, pages 381–403, 1993.

[16] Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[17] P. Ekman and W. V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[18] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909, 2017.

[19] SP Robbins, TA Judge, and T Campbell. Emotions and moods. organizational behaviour, 2010.

[20] I Clement. Introduction to Psychology. *Psychosocial Foundation of Nursing*, pages 123–123, 2010. https://doi.org/10.5005/jp/books/11375$_6doi : 10.5005/jp/books/11375_6$.

[21] Jerry M Burger. *Personality: Theory and research*. Wadsworth Publishing Company, 1986.

[22] Walter Mischel. Toward an integrative science of the person. *Annu. Rev. Psychol.*, 55:1–22, 2004.

[23] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[24] Anderson S. Barroso, Jamille S.Madureira Da Silva, Thiago D.S. Souza, Bryanne S.De A. Cezario, Michel S. Soares, and Rogerio P.C. Do Nascimento. Relationship between personality traits and software quality Big Five model vs. object-oriented software metrics. *ICEIS 2017 - Proceedings of the 19th International Conference on Enterprise Information Systems*, 3(Iceis):63–74, 2017.

[25] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[26] Paula Carvalho and Mário J. Silva. SentiLex-PT: Principais características e potencialidades. *Oslo Studies in Language*, 7(1), 2015.

[27] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of LIWC2015. *Austin, TX: University of Texas at Austin*, pages 1–22, 2015.

[28] Alessandro Murgia, Parastou Tourani, Bram Adams, and Marco Ortu. Do developers feel emotions? An exploratory analysis of emotions in software artifacts. *11th Working Conference on Mining Software Repositories, MSR 2014 - Proceedings*, (May):262–271, 2014.

[29] N. Novielli, D. Girardi, and F. Lanubile. A benchmark study on sentiment analysis for software engineering research. *Proc. - International Conference on Software Engineering*, pages 364–375, 2018.

[30] F. Calefato, G. Iaffaldano, F. Lanubile, and B. Vasilescu. On developers' personality in large-scale distributed projects: The case of the apache ecosystem. *Proc. - Int. Conf. on Software Engineering*, pages 92–101, 2018.

[31] Nicole Novielli, Fabio Calefato, Davide Dongiovanni, Daniela Girardi, and Filippo Lanubile. Can We Use SE-specific Sentiment Analysis Tools in a Cross-Platform Setting? *Proc. - 2020 IEEE/ACM 17th Int. Conf. on Mining Software Repositories, MSR 2020*, pages 158–168, 2020. http://arxiv.org/abs/2004.00300 arXiv:2004.00300.

[32] M. Herrmann, M. Obaidi, L. Chazette, and Jil Klünder. On the subjectivity of emotions in software projects: How reliable are pre-labeled data sets for sentiment analysis? *The Journal of Systems & Software*, 193:111448, 2022.

[33] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

[34] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of Emojis. *PLOS ONE*, 10(12):e0144296, dec 2015.

[35] C.J. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media*, page 18, 2014.

[36] Julio Reis, Pollyanna Gonçalves, Matheus Araújo, Adriano César Pereira, and Fabricio Benevenuto. Uma Abordagem Multilíngue para Análise de Sentimentos. 2020.

[37] Douglas Cirqueira, Antonio Jacob, Fábio Lobato, Adamo Lima de Santana, and Márcia Pinheiro. Performance evaluation of sentiment analysis methods for Brazilian Portuguese. In *Lecture Notes in Business Information Processing*, volume 263, pages 245–251. 2017.

[38] Pedro Paulo Balage Filho, Thiago Pardo, and Sandra Aluísio. An Evaluation of the Brazilian Portuguese LIWC Dictionary for Sentiment Analysis. *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 215–219, 2013.

[39] Marlo Souza and Renata Vieira. Sentiment Analysis on Twitter Data for Portuguese Language. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7243 LNAI, pages 241–247. 2012.

[40] Sebastian C Müller and Thomas 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*, volume 1, pages 688–699. IEEE, 2015.

[41] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[42] Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32, 2018.

[43] J. Ko, H. W. Kwon, H. S. Kim, K. Lee, and M. Y. Choi. Model for Twitter dynamics: Public attention and time series of tweeting. *Physica A: Statistical Mechanics and its Applications*, 404:142–149, 2014.

[44] Yair Amichai-Hamburger and Gideon Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289–1295, 2010. Online Interactivity: Role of Technology in Behavior Change.

[45] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski, and J. Crowcroft. The personality of popular facebook users. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 955–964, 2012.

[46] Marlo Souza, Renata Vieira, Rove Chishman, and Isa Mara Alves. Construction of a Portuguese Opinion Lexicon from multiple resources. *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, pages 59–66, 2011.

[47] Mateus T. Machado, Thiago A.S. Pardo, and Evandro Eduardo Seron Ruiz. Creating a Portuguese Context Sensitive Lexicon for Sentiment Analysis. *Lecture Notes in Computer Science*, (i):335–344, 2018.

[48] Karine França de Souza, Moisés Henrique Ramos Pereira, and Daniel Hasan Dalip. UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro. *Abakós*, 5(2):79, 2017.

[49] Chris Zimmerman, Mari Klara Stein, Daniel Hardt, Christian Danielsen, and Ravi Vatrapu. EmotionVis: Designing an emotion text inference tool for visual analytics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9661 LNCS(July 2018):238–244, 2016.