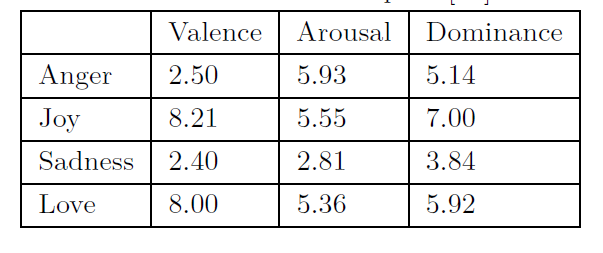
Calculating VAD score

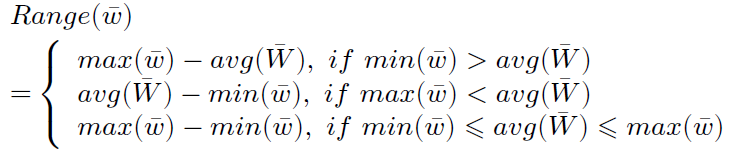
METHOD:

For each report, we extracted title, description and comments, then calculated the VAD measures. We then used R’s tm package tokenizer to extract words, after which we matched the words found to the ones existing in our VAD lexicon (see below). If the word did not match our lexicon, then it was not used as no VAD score could be given to it. This mitigates the problem that issue reports can contain code and other information like stack traces.

Mapping of the words representing discrete emotions to the Valence-Arousal-Dominance space



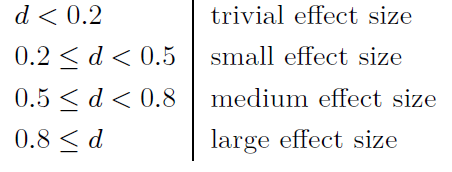
To calculate the corresponding VAD scores for a piece of text (i.e., a list of words ¯ w = [w1,w2, ...,wn]), the Range of the words’ individual VAD scores is computed by taking the two words with the Max and Min Valence, Arousal or Dominance. For the special cases when Max has lower than average value or when Min has higher than average value we set the Max or Min to the average of all words of the lexicon (¯W = [W1,W2, ...,WN], where N is 13,915). Our formula is adapted from the one used by \*SentiStrength [41], which is an industry strength tool to measure positive and negative sentiment. Our avg(¯W ) is similar to zero in \*SentiStrength, i.e. a mid-point and in \*SentiStrength Min values are never allowed to be higher than zero and Max values are never lower than zeros. In more formal terms, our equation is as follows:

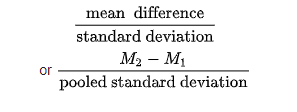


For example, if an issue would contain all the words listed in Table 1, it would receive a Valence score of 5.81 (8.21-2.40, cf. third case in formula). The higher the value, the more extreme the VAD scores are. Overall, our approach is simple and straightforward from a text-mining viewpoint.

\*SentiStrength: http://sentistrength.wlv.ac.uk/

Cohen’s d effect size:





Findings:

* Blocker issues have higher Arousal than Trivial issues, but effect sizes are small (Arousal typically is introduced in a controlled experiment in the form of penalties [46] and rewards [1] that are awarded if a task can be completed within a certain time frame. In the context of issue reports, we believe that the reward for fixing a defect successfully is related to the priority of the issue, with high priority issues (e.g., blocker defects that are showstoppers) boosting a developer’s profile and morale the most.)
* Valence is lowest for Bugs, which supports our expectations (Valence refers to the pleasure or attraction experienced by humans. Typically, humans experience more pleasure in new things)
* To our surprise, high Dominance was associated with high (not low) issue resolution time (Dominance refers to feeling in control of a situation. If a developer writing a comment to an issue report has a firm grip on the situation, this should be reflected in higher Dominance scores)
* From the moment an issue is reported to the time when the issue is closed, Valence and Dominance tend to increase, while Arousal has a small decrease
* The Arousal of Assignees drops as issues are resolved
* The Valence and Dominance of Reporters increase as issues are resolved, yet Arousal remains stable
* For Other commenters’ emotions, Valence and Dominance increase as issues are resolved, while Arousal decreases
* Across VAD metrics, Valence of all comments and the Valence of the issue title have the largest impact on issue resolution time
* We find that metrics that increase Valence decrease Arousal and vice versa

Additional ideas:

Additionally, ideas already found in other papers could be used for improvement, especially in the detection of Arousal. For example, in a small study of a single project with two deadlines [13], the authors find that, as deadlines came closer, more and longer emails were exchanged with higher emotional intensity. Thus, frequent intervals of exchanging messages and reporting issues could be an additional measure of Arousal. Similarly, Arousal due to time pressure could increase the number of spelling mistakes as it is well-known that under time pressure individuals make speed-accuracy tradeoffs [19]. A more focused discussion also suggests an increase in Arousal. This phenomenon has different names.

Finally, we need to point out that the observed effect sizes, for example in RQ1, are small. We think that this only illustrates the general problems in finding emotional responses from text. For example, a highly cited emotion mining paper on Facebook data published in 2014 in the prestigious PNAS journal [20] only reported effect sizes (Cohen’s d) ranging from 0.02 to 0.008, whereas our effect sizes are almost 20 times as high (e.g., 0.389).