

What is the Mood of Your Team? About Adapting Sentiment Analysis Tools to Company-Specific Needs

Jil Klünder¹ and Oliver Karras²

¹ Leibniz University Hannover, Software Engineering Group
jil.kluender@inf.uni-hannover.de

² TIB – Leibniz Information Centre for Science and Technology
oliver.karras@tib.eu

An IFIP SELECT paper for IT-professionals based on the research paper:

Jil Klünder, Julian Horstmann, and Oliver Karras (2020). „Identifying the Mood of a Software Development Team by Analyzing Text-Based Communication in Chats with Machine Learning“. 8th IFIP WG 13.2 International Working Conference, HCSE 2020.

Summary of the original paper

Meetings are one established communication channel in software development teams to effectively coordinate and communicate in order to ensure close collaboration and synchronization among all team members. So far, research has mainly focused on analyzing meetings to determine the reasons for inefficiency and dissatisfying meeting outcomes. In the original paper, we propose to transfer the insights of meeting analysis on text-based communications channels such as chats and emails that are also frequently used in software development teams. In particular, we propose an approach for analyzing interpersonal behavior among team members in text-based communication based on the conversational tone, the familiarity of the sender and receiver of a message, the sender's emotionality, and the appropriateness of the language used. In a first industrial case study, we applied our approach analyzing 1947 messages in a group chat over 5.5 months. We gained an overall picture of the adequacy of textual communication and tendencies in the software development team's mood.

On the subjective perception of communication and the difficulty of using generic approaches Software projects become increasingly complex and are therefore more of a team effort rather than a one-person effort. Well-functioning and effective teamwork requires continuous information exchange so that each team member knows what he or she needs to know. However, teams often deliver reports without adequate information sharing, resulting in insufficiently addressed requirements and dissatisfied customers. The success of teamwork depends heavily on sufficient and adequate communication at both the verbal and the textual levels.

The question of what constitutes sufficient communication can be answered without difficulty: A team communicates sufficiently when everyone receives the relevant information he or she needs for his or her work and the tasks he or she has been given. But how can you define adequate communication? This question is not easy to answer as it depends on several factors. Besides the words used and the message intended in a sentence, personality traits and the characters of the team members influence whether a person perceives a statement as positive, negative, or neutral. Even friendly statements praising a team member can raise different perceptions: Although in most

cases they are meant positively, they can also be meant ironically (praising someone for something that is daily work for him or her).

In addition, the general communication behavior in a team also influences how communication is perceived. For example, a group of team members who have known each other for a very long time (probably also on a private level) may tend to communicate in a more “extreme” way, kidding each other, using a lot of irony and sarcasm, etc. For individual team members, this communication can be perceived negatively, which usually does not reflect the perception of the group.

Nevertheless, analyzing the communication behavior in a team regarding the moods conveyed or the polarity (positive, negative, or neutral) of the statements can help project leaders or managers to observe what is going on in the team. However, these analyses should not be conducted for the communication of individual team members, but the entire team. While it would be interesting to know which team member is in a bad mood, this would be ethically questionable for data protection and privacy reasons.

To provide the necessary anonymity (and to facilitate analysis), several generic tools return the polarity (or in some cases also sentiments) triggered by individual statements. Some of these tools are explicitly developed for the software engineering domain, while others are not. The use of domain-specific tools increases the reliability of the results since domain-specific language and words are considered. For example, in everyday life, the word “kill” conveys a negative sentiment, whereas in the IT domain it is not uncommon to speak about “killing a process”. Nevertheless, these domain-specific tools still do not take into account all the social aspects of the team that need to be observed and analyzed.

The research approach and next steps

Instead of using such a generic tool, we implemented and trained a classifier based on manually labeled data. Overlooking the fact that labeling data requires a lot of time and effort, the manual labeling counteracts the problem of ignoring the personality traits of the team.

To evaluate our approach, we applied it in a case study in a software consulting company with in-house software development. In a software project with 80 developers in one country, we analyzed the communication via Zulip, which is a (group) chat communication tool. We looked at all group communication between Feb 11, 2019, and Jul 24, 2019, resulting in 1947 messages with 7070 sentences. During the data preprocessing, we cleaned the data by removing URLs and markdown commands, anonymizing the messages, applying auto-correction, and splitting the messages into single sentences. Afterward, to construct the classifier, one researcher manually labeled each sentence in the data set as positive, negative, or neutral.

Note that, typically, manual labeling is considered as introducing a bias: As part of quality assurance, the researcher labeled 200 randomly selected sentences once more after one month, resulting in an intrarater agreement of only 66%. That is, in one out of three cases the researcher assigned a different class to a sentence than when it was first labeled. However, this low intrarater agreement also highlights the subjectivity of the perception, resulting in the need to account for this subjectivity when analyzing communication.

Using the manually labeled data, we trained and evaluated the resulting classifier. Overall, the classifier achieved an accuracy of almost 63%, which is not bad but offers potential for improvement.

Results

Applying the classifier to the whole data set shows some interesting insights. In some cases, the human and tool assigned the same class to a sentence. For example “*well done!*” appears to be

positive, whereas *"if this was not your mistake, it has to be fixed as follows"* is neutral. However, the sentences *"Yes, this was my mistake"* and *"had understood you differently this morning"* appear to be negative for the human, but neutral for the tool. Although the tool is trained using the perception of the human, it does not completely reflect his or her perception. So, who is right? Tool or human? Whose sentiment does the result of the tool reflect? Is it the majority of humanity or, more importantly, the majority of the team? How reliable are the results for the team? Are they useful when not reflecting its perception?

Answering these questions requires further research. The generic tools provide first insights on the team mood and help to detect stressful or problematic situations. However, the insights can (and to some extent should) be improved by looking at the personalities of the team members.

Practical Advice and Experiences

In summary, sentiment analysis can help to get an overview of the general mood in the team. In addition, assuming that communication loses professionalism in stressful situations, this type of analysis can also help detecting these stressful situations.

However, generic and pre-trained tools only reflect how the majority of people perceives a sentence or communication, and thus ignore social aspects of the observed team and the context of the software project. Consequently, we propose to use a classifier trained on data labeled by the respective team.

As giving a sufficient number of sentences (a thousand or even more) to the team and asking the team to label them is time-consuming and requires a lot of effort, our future research will focus on how to improve this process, e.g., by asking the team members to label a dozen of the sentences that help in gaining information about whether the team is more optimistic, pessimistic, or neutral. This procedure can help in configuring an existing tool.

Nevertheless, practitioners already benefit from being aware of their team's mood as it helps to improve the collaboration and thus smoothing the way to a successful project.