Towards Emotional Awareness in Software Development Teams

Emitza Guzman
Technische Universität München
Garching, Germany
emitza.guzman@mytum.de

Bernd Bruegge
Technische Universität München
Garching, Germany
bruegge@in.tum.de

ABSTRACT

Emotions play an important role in determining work results and how team members collaborate within a project. When working in large, distributed teams, members can lose awareness of the emotional state of the project. We propose an approach to improve emotional awareness in software development teams by means of quantitative emotion summaries. Our approach automatically extracts and summarizes emotions expressed in collaboration artifacts by combining probabilistic topic modeling with lexical sentiment analysis techniques. We applied the approach to 1000 collaboration artifacts produced by three development teams in a three month period. Interviews with the teams' project leaders suggest that the proposed emotion summaries have a good correlation with the emotional state of the project, and could be useful for improving emotional awareness. However, the interviews also indicate that the current state of the summaries is not detailed enough and further improvements are needed.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; D.2.9 [**Management**]: Programming teams

General Terms

Human Factors

Keywords

Sentiment analysis, topic extraction, collaborative software engineering

1. INTRODUCTION

How we feel and how we perceive that others feel, i.e. emotional awareness, plays an important role in collaborative work. Being aware of your own emotions and those of colleagues allows you to adapt your actions accordingly and obtain better results in joint tasks [5]. Software development is a highly collaborative activity in which participants use mailing lists, forums, software code repositories and issue tracking tools, among others, to manage their collaborations [11]. These collaboration artifacts are communication channels where development teams can express their emotional state. With the growing trend of business globalization and distributed teams these forms of communication are taking over more traditional ones, such as face to face meetings and voice [7]. This trend produces a large amount of textual data that can be challenging to analyze and process, making it difficult for team members to stay aware of the emotional state of their project. Previous research has shown that emotions affect productivity, task quality, creativity, group rapport and job satisfaction [3]. We propose an approach for automatically creating summaries with emotional information from collaboration artifacts. From these emotion summaries it could be possible to, for example: identify the change of emotions in a team during different stages of development (e.g. a deadline), or to uncover the general emotion of a team with respect to a specific topic (e.g. a new feature in the software). The emotion summaries we propose could help members of development teams reflect about the emotions in the project, about the factors causing these emotions, and could help them make decisions concerning necessary changes. Our approach uses latent Dirichlet allocation (LDA) [1] for extracting topical information from the artifacts and lexical sentiment analysis [12] for assigning the emotion score. This work presents:

1. A technique for extracting and summarizing emotions in collaboration artifacts.
2. Evidence from interviews that our approach could be useful for improving emotional awareness in development teams.

2. RELATED WORK

Extensive work has studied emotions expressed in twitter messages, question and answer sites, and product reviews, among others. Examples of work in this area are those of Giannopoulos et al. [6], Munson et al. [10] and Kucuktunc et al. [8].

However, few research studying emotions in software development teams exists. McDuff et al. [9] propose a multi-sensor system that is able to detect emotions in the workplace for individual emotion tracking. Their approach is complementary to ours as they do not extract emotions from text, but from sources such as posture, facial expressions and voice.

ESEC/FSE'13, August 18–26, 2013, Saint Petersburg, Russia
Copyright 2013 ACM 978-1-4503-2237-9/13/08...$15.00
http://dx.doi.org/10.1145/2491411.2494578

Also, their approach is meant to improve individual emotional awareness, but not collective emotional awareness. Dullemond et al. [4] study the emotions and topics that developers working in distributed teams express when using a microblogging tool. Their approach differs from ours in that they extract the topics and emotions manually. Furthermore, they are interested in improving collaboration through emotion sharing, but they do not abstract or summarize the emotional information for improving retrospective emotional awareness. The work that is perhaps most similar to ours is the one of De Choudhury and Counts [3]. They propose a methodology based on lexical sentiment analysis for analyzing workplace emotions expressed in social media. They classify emotions into positive and negative and use the most frequent keywords present in the positive and negative texts as emotion summaries. We improve their idea by using LDA to create the emotion summaries and have a more precise depiction of factors affecting the emotional state of software teams. Additionally, we expand their approach by proposing to analyze artifacts that do not make use of social media technologies, such as bug reports and commit messages. Furthermore, our work complements theirs as we provide initial evidence of the perceived usefulness that emotion summaries can have when managing software development teams.

3. METHODOLOGY

The main goal of our approach is to summarize emotions expressed in collaboration artifacts by extracting topics and assigning them an average emotion score. Figure 1 shows our overall approach. We first anonymize the data in the existing artifacts, and remove HTML tags, email headers and file attachments. Before feeding the data to our LDA topic model, we perform three additional preprocessing steps: tokenization, stemming and stopword removal. The topic model outputs a list of topics associated to each artifact. We apply sentiment analysis to the processed data, obtaining an average emotion score for each artifact. We compute a weighted average to combine the results of the sentiment analysis and topic modeling techniques, obtaining a list of topics each assigned to an average emotion score. We consider this associated list a summary. In the following sections we explain the three main aspects of our approach: the input data, sentiment analysis and topic modeling.

3.1 Input Data: Collaboration Artifacts

To accurately model the emotions in a software development team and understand the factors that influence these emotions we propose to analyze different collaboration artifacts. Examples of these artifacts are commit messages in software repositories, bug reports, emails, wikis and twitter messages. We believe that the amount of emotion and detail that developers express depends on the artifact type. For example, an email message from the project's mailing list might have more technical details about an aspect of the project than a twitter message dealing with the same topic. On the other hand, a twitter message might express more positive emotions when dealing with a topic related to a social event than an email message, whereas this topic is not likely to appear in an error report or commit message. We believe that by analyzing a diverse set of collaboration artifacts, the summaries will be richer in their terms and therefore, more explanation about the factors influencing the emotion scores will be possible.

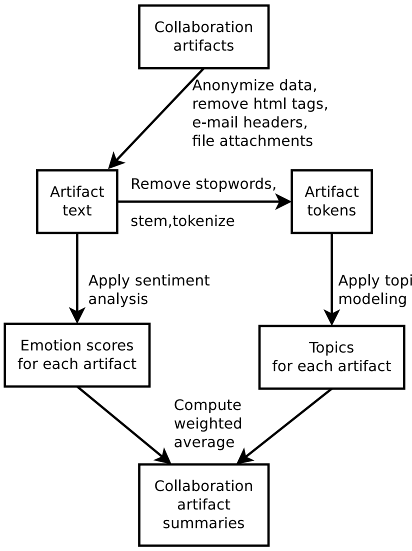
Collaboration artifacts
Anonymize data, remove html tags, e-mail headers, file attachments
Artifact text
Remove stopwords, stem,tokenize
Artifact tokens
Apply sentiment analysis
Apply topic modeling
Emotion scores for each artifact
Topics for each artifact
Compute weighted average
Collaboration artifact summaries

Figure 1: Block diagram of overall approach.

3.2 Sentiment Analysis

Sentiment analysis is the process of assigning a quantitative emotion (positive or negative) to a text snippet [8]. For analyzing emotions in software development teams we use SentiStrength [12], a lexical sentiment extraction tool. SentiStrength divides artifact's text into text snippets. A text snippet can consist of one or more words, or one or more sentences. SentiStrength assigns positive and negative values to these snippets. The idea behind this is that humans can express both positive and negative emotions at the same time, e.g. "I love hating you". SentiStrength assigns positive scores in the $\{+1, +5\}$ range, where $+5$ denotes an extremely positive emotion. Similarly, negative emotions range from $\{-1, -5\}$. SentiStrength assigns fixed scores to tokens in a dictionary where common emoticons are also included. For example, "love" is assigned a score of $[3, -1]$ and "hate" a $[1, -4]$ score. Only words that are present in the dictionary are attributed with an individual score. Modifier words and symbols also alter the score. For example, "absolutely love" is assigned a score of $[4, -1]$. The same score is given to "looove" and "love!!!". The final positive or negative score in a text snippet is computed by adding all positive or negative individual scores. We compute the emotion of an entire artifact by calculating the average of all text snippets in the artifact. Table 1 shows an example of how SentiStrength calculates the positive and negative scores for each snippet and how these results are used to compute the $[+, -]$ emotion average of the artifact. The artifact's total emotion score average is the average of the $[+, -]$ averages.

3.3 Topic Modeling

We use LDA to extract topics from the set of collaboration artifacts and to assign a set of topics to each of the artifacts. Each topic is a set of co-occurring words in the analysed artifacts. An example of a topic could be the set of words *{bug, problem, crash, application}* because these words usually appear in the same type of artifacts. In LDA, artifacts can be associated to more than one topic. For our

Table 1: Example of SentiStrength scores in an email message.

Text snippet	**Word score**	**[+,-] Text snippet score**
You're working well with JIRA and I like that you resolve your tasks.	You're working well with JIRA and I like[2] that you resolve[2] your tasks.	[4,0]
But we're facing one Problem: the reporter of a task also has to close or reopen the task.	But we're facing one Problem[-2] : the reporter of a task also has to close or reopen the task.	[0,-2]
Please do this as soon as possible so that we can have a good overview over the sprint :-)	Please[2] do this as soon as possible so that we can have a good[2] overview over the sprint :-) [1 emoticon]	[6,0]
	[+,-] Emotion score average of artifact	**[3.33,-0.667]**
	Total emotion score average of artifact	**[1.332]**

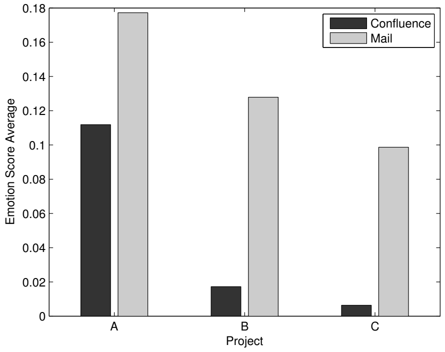

Figure 2: Emotion average score of all projects.

approach we use the Matlab Topic Modeling Toolbox[1]. Let $A = \{a_1, a_2, ..., a_n\}$ be the set of collaboration artifacts to analyze and $T = \{t_1, t_2, ..., t_m\}$ the set of extracted topics. The final output of the LDA computation is the $W_{n \times m}$ matrix where $w_{i,j}$ contains the number of times a word present in artifact a_i is associated with topic t_j.

To create summaries and link the sentiment analysis results with the topic modeling results, we associate the emotion score with the topics by means of a weighted average. That is, for every topic t_j and topic emotion score te_j we have:

$$te_j = \frac{\sum_{i=1}^{n} w_{i,j} \cdot e_i}{\sum_{i=1}^{n} w_{i,j}}$$

where $E = \{e_1, e_2, ..., e_n\}$ denotes the emotion score of each artifact a_i.

4. INITIAL EVALUATION AND RESULTS

Our initial evaluation consisted of two main steps: the application of our approach to collaboration artifacts from software development and interviews with project leaders. In the following sections we describe these two initial evaluation steps.

4.1 Application of Approach on Collaboration Artifacts

We evaluated our approach using text from a mailing list and Confluence[2], a web-based software collaboration tool. In these cases, the collaboration artifacts are emails and web pages, respectively. The analysed artifacts were generated during the iPraktikum lab course of the Technische Universität München in 2012 [2]. During the three month course, students were organised in teams of 5 to 8 students, developing applications for industrial partners. Each team had a project leader, usually a doctoral student, who performed project management activities. We chose three teams for our evaluation, ruling out teams who did not communicate in English, or whose collaboration artifacts were either unavailable or less than 200 for the three month period.

In total, we analysed 857 emails and 143 web pages from Confluence. During the course, the teams had face to face meetings where they worked together and coding problems were discussed. Because of this, both of the chosen artifacts had little to no code snippets or stack traces, and no additional processing had to be done.

Figure 2 presents the average emotion scores of the teams' emails and Confluence pages. Among the teams, mailing lists were mainly used for discussing organizational issues or generic problems, whereas Confluence was mainly used for recording face to face meeting protocols and general knowledge about the project. Therefore, Confluence artifacts tended to have a more neutral language. This explains the values in Figure 2, where the Confluence artifacts of all analyzed teams had a lower average emotion score than their respective emails.

4.2 Interviews with Project Leaders

To evaluate the usefulness of our approach, we interviewed the project leaders of the three teams. During the interviews we showed them examples of texts that were evaluated by SentiStrength. We also presented graphs, similar to those depicted in Figures 2 and 3, demonstrating the emotion score average of their teams' emails and Confluence pages, and the variation of this average during the project. Additionally, we showed them a table containing the topics that were most mentioned in emails and Confluence pages. This table had content similar to the one presented in Table 2. Two of the interviewees were able to correlate positive and negative emotion peaks with the team performance, and overall motivation during the project; as well as with important

[1] http:psiexp.ss.uci.eduresearchprograms_datatoolbox.htm

[2] http://www.atlassian.com/software/confluence

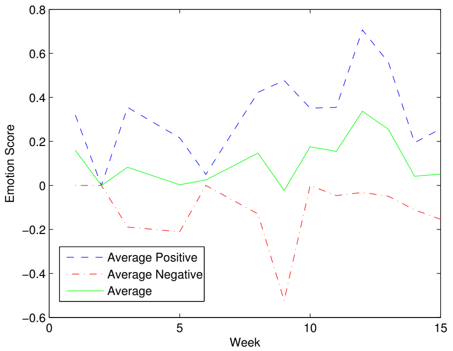

Figure 3: Emotion average fluctuation in the whole project.

Table 2: Example of topics and emotion scores from an evaluated team.

Topic words	Avg. emotion score
code, test, fail, job, user, sent, error, found	0.169
problem, want, good, last, already, implement, status, feedback	0.536
meet, diagram, review, new, creat, add, class	0.422

deadlines. The other project leader could not recall the main events and dates from his project, and was therefore unable to find any correlations. Furthermore, the interviewed project leaders thought that presented graphs would help them react more accordingly, as they would be more aware of their teams emotions. They all agreed that this type of tool would be more useful for large or distributed teams. Additionally, the project leaders mentioned that the approach would be more useful if the summaries were more specific. In the interviews, one of the project leaders expressed fear that developers would change their expression form if they knew that a sentiment analysis approach was being used to analyze the teams' emotions.

5. CONCLUSIONS AND FUTURE WORK

In this work we propose an approach to summarize emotions in collaboration artifacts using lexical sentiment analysis and LDA. We evaluated our approach by applying it to collaboration artifacts produced by three development teams, and interviewed their project leaders. They agreed that the approach could be useful for creating emotional awareness in large or distributed teams, but that finer granularity in the generated summaries was needed. We plan to further refine our approach by applying different summarization techniques and by expanding the lexicon of our sentiment analysis tool. We also plan to test our approach in a larger setting and analyze additional collaboration artifacts. We are also interested in researching how emotional awareness changes the team behavior and plan to analyze possible correlations of development teams' emotions and other contextual factors, such as current tasks or calendar events.

6. ACKNOWLEDGMENTS

This work was partially supported by a scholarship from the Mexican National Council of Science and Technology (Conacyt).

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.

[2] B. Bruegge, S. Krusche, and M. Wagner. Teaching Tornado. In *Proc. of the 8th edition of the Educators' Symposium - EduSymp '12*, pages 5–12, Oct. 2012.

[3] M. De Choudhury and S. Counts. Understanding affect in the workplace via social media. In *Proc. of the 16th Conf. on Computer supported cooperative work - CSCW '13*, pages 303–316, Feb. 2013.

[4] K. Dullemond, B. Van Gameren, and A. S. M.-A. van Deursen. Fixing the 'Out of Sight Out of Mind' Problem: One Year of Mood-Based Microblogging in a Distributed Software Team. In *Proc. of the 10th Int. Conf. on Mining Software Repositories - MSR 2013*, pages 267–276, May 2013.

[5] O. García, J. Favela, and R. Machorro. Emotional Awareness in Collaborative Systems. In *Proc. of the String Processing and Information Retrieval Symposium & International Workshop on Groupware - SPIRE '99*, pages 296–303, Sept. 1999.

[6] G. Giannopoulos, I. Weber, A. Jaimes, and T. Sellis. Diversifying User Comments on News Articles. In *Proc. of the 13th Conf. on Web Information Systems Engineering - WISE'12*, pages 100–113, Nov. 2012.

[7] B. L. Kirkman, R. Benson, P. E. Tesluk, and C. B. Gibson. The Impact of Team Empowerment on Virtual Team Performance: The Moderating Role of Face-to-Face Interaction. *The Academy of Management Journal*, 47(2):175–192, 2004.

[8] O. Kucuktunc, B. B. Cambazoglu, I. Weber, and H. Ferhatosmanoglu. A Large-Scale Sentiment Analysis for Yahoo! Answers. In *Proc. of the 5th Conf. on Web search and data mining - WSDM '12*, pages 633–642, Feb. 2012.

[9] D. McDuff, A. Karlson, A. Kapoor, A. Roseway, and M. Czerwinski. AffectAura: an intelligent system for emotional memory. In *Proceedings of the 20th Conf. on Human Factors in Computing Systems - CHI '12*, pages 849–858, May 2012.

[10] S. A. Munson and P. Resnick. Presenting diverse political opinions. In *Proc. of the 28th Conf. on Human factors in Computing Systems - CHI '10*, pages 1457–1466, Apr. 2010.

[11] M.-A. Storey, C. Treude, A. van Deursen, and L.-T. Cheng. The Impact of Social Media on Software Engineering Practices and Tools. In *Proc. of the Workshop on Future of software engineering research - FoSER '10*, pages 359–364, Nov. 2010.

[12] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment in short strength detection informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, Dec. 2010.