Sentiment analysis of commit comments in GitHub: an empirical study

Authors:

Emitza Guzman
David Azocar
Yang Li

Conference:

MSR 2014

Presented by:

Andrew Berg
Raidel Hernandez
Tyler Kelly

Overview

- Research Question
- Sentiment Analysis
- Results
- Additional Research

Research Questions

1. Are emotions in commit comments related to the programming language in which a project is developed?
2. Are emotions in commit comments related to the day of the week or time in which the commits were written?
3. Are emotions in commit comments related to the team geographical distribution?
4. Are emotions in commit comments related to project approval?

SentiStrength & Sentiment Analysis

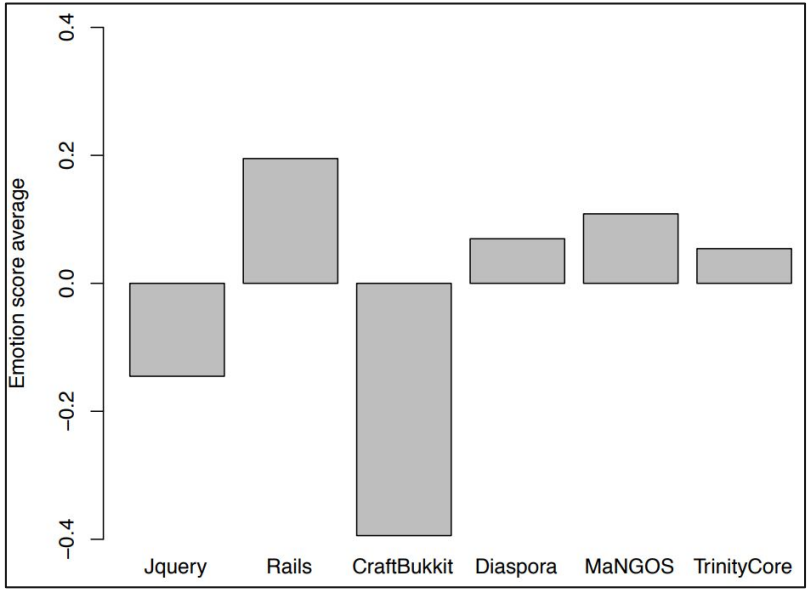
- Assigns a quantitative integer mood value to a text snippet (-5 to 5)
- (1,5] positive, (-5,1] negative, [-1,1] neutral
- SentiStrength designed by University of Wolverhampton, written in Java for social media short text sentiment scoring
- Designed for non-political short messages in social media and claims human level accuracy for this application

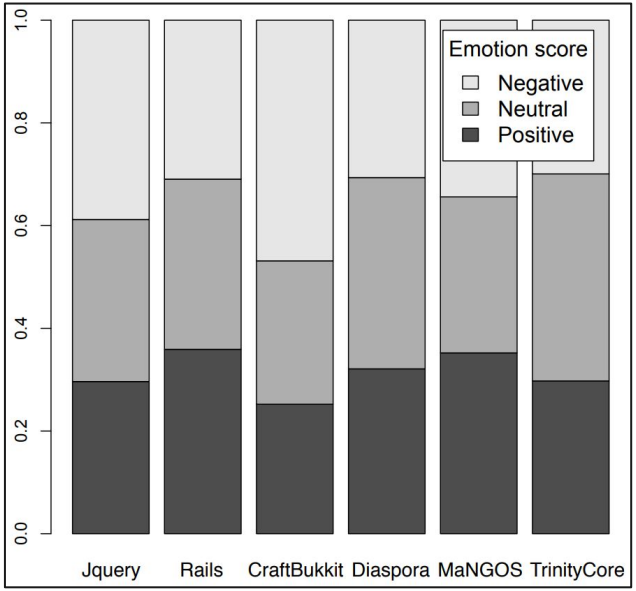
Methodology Used for Analysis

- MySQL MSR 2014 challenge data dump
- 29 projects with only analyzing projects that had more than 200 commit comments in the GitHub repo
- MySQL database queried then stored the results of the commit comment sentiment analysis in a SQLite database
- This data then queried to replicate the results of the paper and answer the research questions presented earlier

Results: Commit Comments

Average emotion score per project, proportion of emotion types





Results: Programming Language

Average emotion score grouped by programming languages

Language	Commits	Mean	Stand. Dev.
C	6257	0.023	1.716
C++	16930	0.017	1.725
Java	*4713*	***-0.144***	*1.736*
Python	2128	-0.018	1.711
Ruby	15257	0.002	1.714

Results: Day and Time of the Week

Average emotion score grouped by weekday, time of the day commited

Weekday	Commits	Mean	Stand. Dev.
Monday	*9517*	***-0.043***	*1.732*
Tuesday	9319	0.005	1.712
Wednesday	9730	0.008	1.716
Thursday	9538	0.001	1.728
Friday	9076	-0.016	1.739
Saturday	6701	-0.027	1.688
Sunday	6544	0.022	1.717

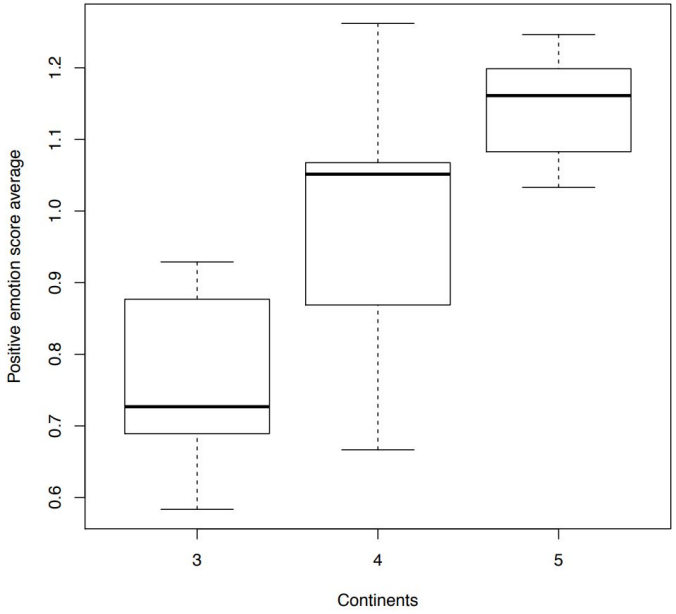
78%

22%

Time of Day	Commits	Mean	Stand. Dev.
Morning	12714	0.001	1.730
Afternoon	*19809*	***0.004***	*1.717*
Evening	*16584*	***-0.023***	*1.721*
Night	11318	-0.016	1.713

Results: Team Distribution

Average emotion score grouped by continent distribution, positive emotion



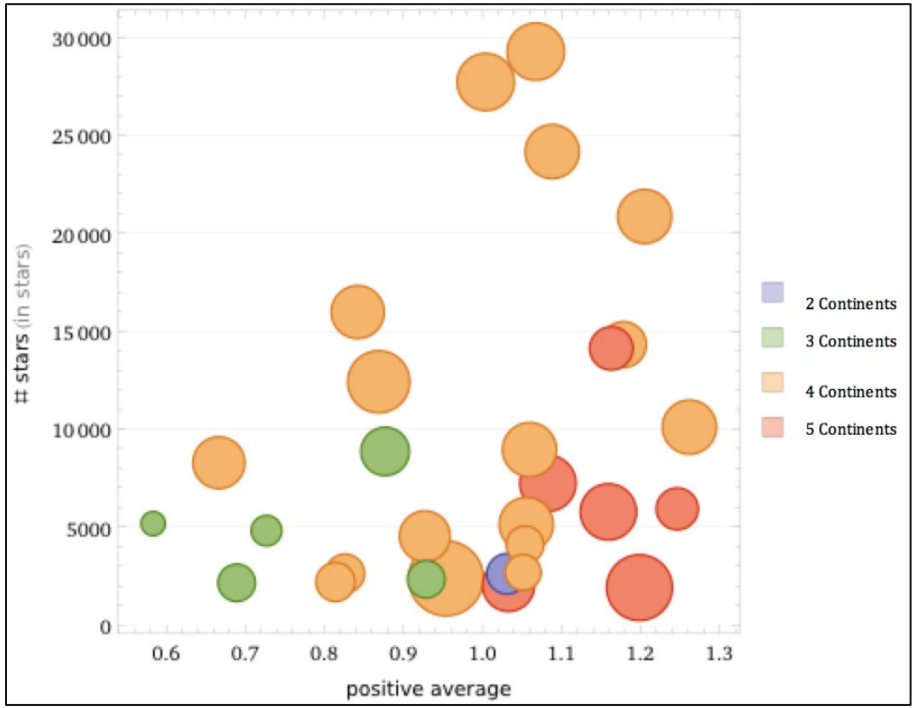
Continents	Mean	Stand. Dev.
2	1.031	--
3	*0.761*	0.141
4	*0.996*	0.157
5	*1.148*	0.078

Results: Project Approval

Relationship between:

- Continent distribution
- Stars (project)

Positive weak correlation



Additional Research

- Provide a link between a committer's sentiment and overall code quality
- Do angry people make bad code?
  - Are XXXX developers prone to being angry/happy, poor/high quality developers?

Additional Implementation

- Examine each committer's public GitHub Repositories
- Determine code quality for user's most used language repositories
  - Code / comment ratio
  - **Linters**
  - Compilation warnings / errors

Additional Results

- Database will have have following schema

| committer | language | avg_sentiment | avg_code_quality_for_language |

- Use pearson correlation coefficient to determine relation

PEARSON

Sentiment analysis of commit comments in GitHub: an empirical study

Authors:

Emitza Guzman
David Azocar
Yang Li

Conference:

MSR 2014

Presented by:

Andrew Berg
Raidel Hernandez
Tyler Kelly