Replicate an MSR challenge paper

Sentiment Analysis of Commit Comments in GitHub: An Empirical Study
 (MSR'14)

https://github.com/jodokae/cmput663-msr14-replic

Johannes Kaestle and Moein Owhadi Kareshk
University of Alberta

The Motivation

Is there any relation between the programming language and the sentiment of commit messages?

What is the impact of the weekday on the sentiment of commit messages?

Is the sentiment in commit messages related to the geographical distribution of team members?

Are project approval and the sentiment of commit messages related?

Dataset

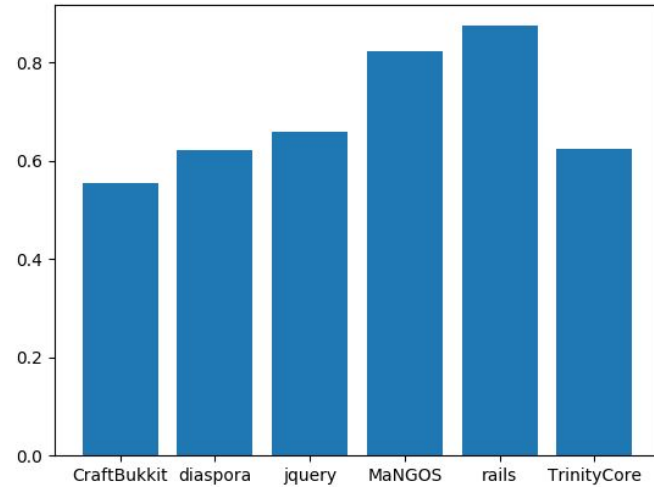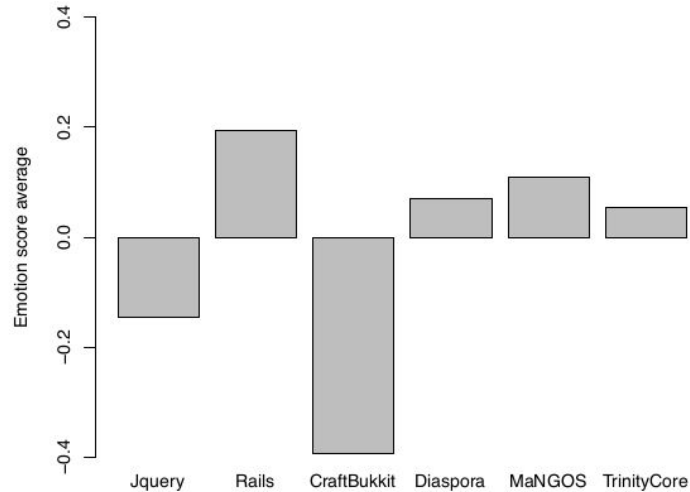90 of the 10-top starred software projects for the 14 top programming languages on Github

In this paper: all projects having more than 200 comments, results in analyzing 29 projects.

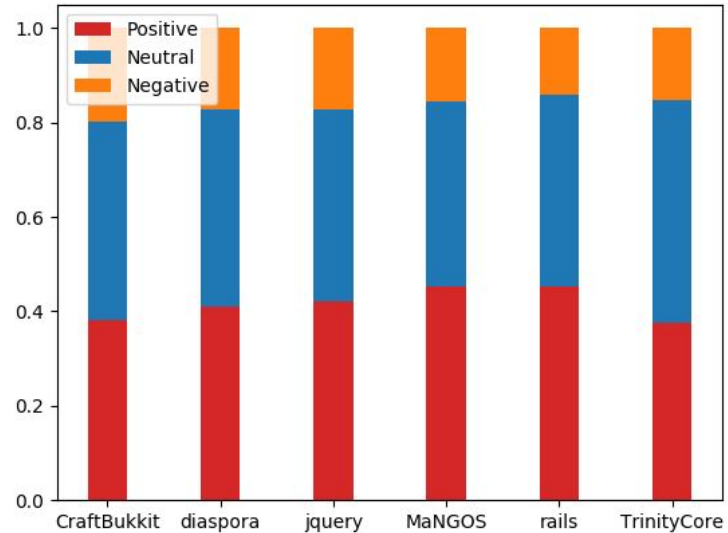60,425 commit comments

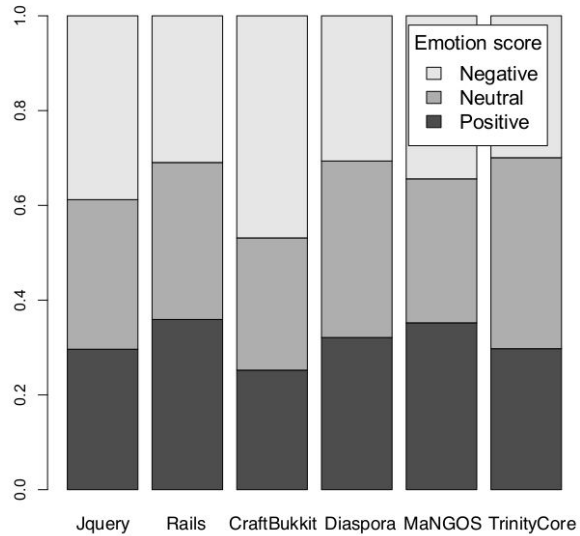Dataset: MSR 2014 Mining Challenge Dataset (http://ghtorrent.org/msr14.html)

Tools for sentiment analysis: SentiStrength (http://sentistrength.wlv.ac.uk/)

1- Emotions in Commit Comments

1- Emotions in Commit Comments

2- Emotions and Programming Language

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

```
+----------+---------+--------+---------------------+
| Language | Commits |  Mean  |     Stand. Dev.     |
+----------+---------+--------+---------------------+
|    C     |    6277 | 0.5417 | 1.6922899082083442  |
|   C++    |   16985 | 0.6725 | 1.7350392064047633  |
|   Java   |    4715 | 0.5599 | 1.7517041522841483  |
|  Python  |    2134 | 0.5553 | 1.7112252969363808  |
|   Ruby   |   15358 | 0.7231 | 1.7686331638947068  |
+----------+---------+--------+---------------------+
```

3- Emotions, Day and Time of the Week

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

```
+--------------+----------+--------+---------------------+
|   Weekday    | Commits  |  Mean  |     Stand. Dev.     |
+--------------+----------+--------+---------------------+
|    Monday    |   9534   | 0.5784 | 1.8016544758809139  |
|   Tuesday    |   9390   | 0.6490 | 1.7271577971644025  |
|  Wednessday  |   9750   | 0.6456 | 1.7291079586526288  |
|   Thursday   |   9564   | 0.6552 | 1.7517612228103145  |
|    Friday    |   9154   | 0.6690 |  1.74051508684218   |
|   Saturday   |   6722   | 0.6617 | 1.7269209060792783  |
|    Sunday    |   6544   | 0.6400 | 1.7342454654302941  |
+--------------+----------+--------+---------------------+
```

3- Emotions, Day and Time of the Week

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

```
+-------------+---------+--------+---------------------+
| Time of Day | Commits |  Mean  |     Stand. Dev.     |
+-------------+---------+--------+---------------------+
|   Morning   |  12750  | 0.6056 | 1.7197171087873917  |
|  Afternoon  |  19859  | 0.6634 | 1.7289907551048684  |
|   Evening   |  16634  | 0.6664 | 1.750382355042099   |
|    Night    |  11415  | 0.6087 | 1.7974757545782019  |
+-------------+---------+--------+---------------------+
```

4- Emotions and Team Distribution

The authors also worked on the greaographocally distribution of team members.

For this analysis, additional information (continent) is rqureed that is not provided by GitHub, GHTorrent, and the authors.

Conclusion

Version differences of dataset and tools

Unclear description of the procedure of sentiment analysis