

More Than React: Investigating The Role of Emoji Reaction in GitHub Pull Requests

Teyon Son, Tao Xiao, Dong Wang, Raula Gaikovina Kula, Takashi Ishio, Kenichi Matsumoto
Nara Institute of Science and Technology, Japan

Email: {son.teyon.sr7, tao.xiao.ts2, wang.dong.vt8, raula-k, ishio, matumoto}@is.naist.jp

Abstract—Context: Open source software development has become more social and collaborative, especially with the rise of social coding platforms like GitHub. Since 2016, GitHub started to support more informal methods such as emoji reactions, with the goal to reduce commenting noise when reviewing any code changes to a repository. Interestingly, preliminary results indicate that emojis do not always reduce commenting noise (i.e., eight out of 20 emoji reactions), providing evidence that developers use emojis with ulterior intentions. From a reviewing context, the extent to which emoji reactions facilitate for a more efficient review process is unknown.

Objective: In this registered report, we introduce the study protocols to investigate ulterior intentions and usages of emoji reactions, apart from reducing commenting noise during the discussions in GitHub pull requests (PRs). As part of the report, we first perform a preliminary analysis to whether emoji reactions can reduce commenting noise in PRs and then introduce the execution plan for the study.

Method: We will use a mixed-methods approach in this study, i.e., quantitative and qualitative, with three hypotheses to test.

I. INTRODUCTION

In the past few years, open source software development has become more social and collaborative. Known as social coding, open source development promotes formal and informal collaboration by empowering the exchange of knowledge between developers [9]. GitHub, one of the most popular social coding platforms, attracts more than 72 million developers collaborating across 233 million repositories.¹ Every day, thousands of people engage in conversations about code, design, bugs, and new ideas on GitHub. To promote collaboration, GitHub implements a vast number of social features (i.e., follow, fork, and stars).

Since 2016, GitHub introduced a new social function called “reaction” for developers to quickly express their feelings in issue reports and PRs. Especially for discussing a Pull Requests (PRs), we find that²:

“In many cases, especially on popular projects, the result is a long thread full of emoji and not much content, which makes it difficult to have a discussion. With reactions, you can now reduce the noise in these threads” - GitHub

In the context of code review, we assume that a thread full of emoji may also contribute to the existing forms of confusion for reviewers during the code review process. For instance,

Ebert et al. [10] pointed out that confusion delays the merge decision decreases review quality, and results in additional discussions. Hirao et al. [16] found that patches can receive both positive and negative scores due to the disagreement between reviewers, which leads to conflicts in the review process.

The Figure 1 depicts two typical cases where the emoji reactions occur. Figure 1(a) shows the case where the reaction does reduce unnecessary commenting in the thread, hence may lead to less confusion and conflicts. The example illustrates how Author B reduces the commenting by simply reacting with a quick expression of approval through THUMBS UP 👍. In contrast, as shown in Figure 1(b), there exists a case where the emoji usage has an ulterior intention and does not reduce comments in the discussion thread. In detail, Contributor D uses three positive emoji reactions (THUMBS UP 👍, HOORAY 🎉, and HEART ❤️) to represent the appreciation to this PR. Then later goes on to provide detailed comments on the PR. We posit that the intention of the emoji was to express appreciation for the PR, and did not reduce the amount of commenting in the thread discussions. As part of our preliminary study, we also found other cases where the emoji did not always reduce the commenting in the discussion. Under a closer manual inspection of 20 emoji reactions, we find that there are eight cases where the emoji reactions did not reduce commenting noise.

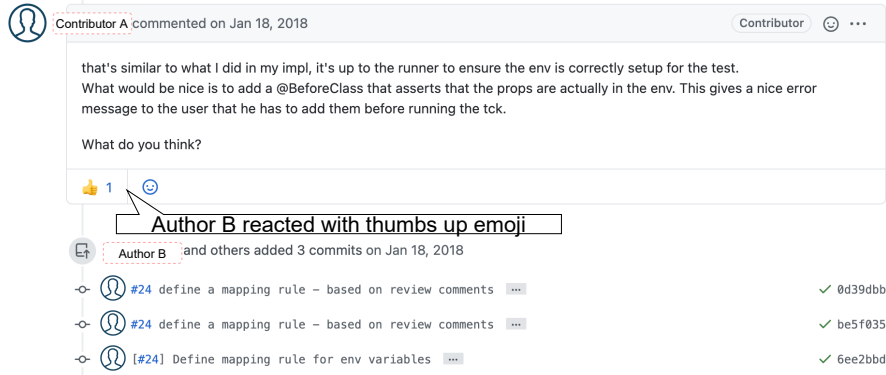
Therefore, in this registered report, we present our study protocol to investigate ulterior intentions and usages of emoji reactions, apart from reducing commenting noise during the discussions. Specifically, we would like to (i) investigate the effect of emoji reaction related factors on the pull request process (i.e., review time), (ii) investigate whether the first time pull request is more likely to receive reactions, (iii) analyze the relationship between the reaction and the intention of comment, and (iv) explore consistency between sentiments of an emoji reaction and the sentiment of the comment. To enable other researchers to extend our study, we plan to make the study data publicly available.

II. PRELIMINARY STUDY

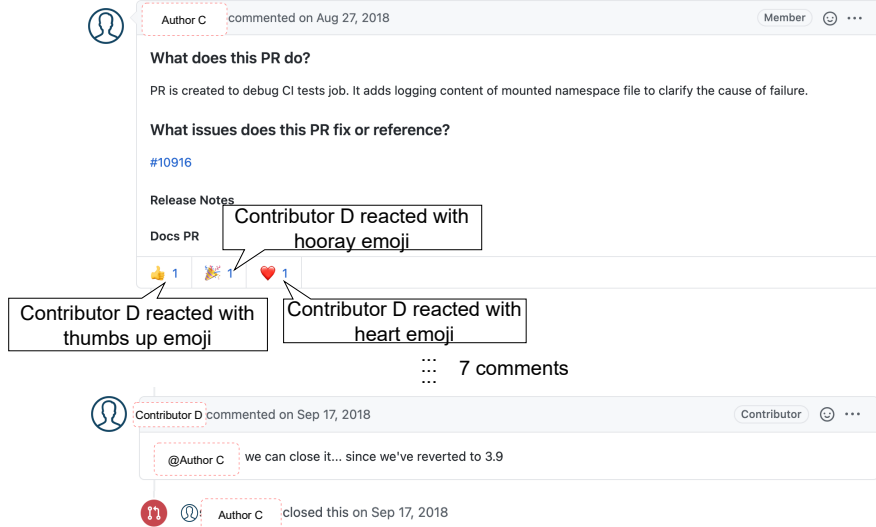
The goal of our preliminary study is to explore the extent to which emoji reaction usage reduces commenting noise of a PR. We selected the Eclipse projects as our case study since it is a mature thriving open source project with a number of contributors that actively submit and merge PRs.

¹<https://github.com/search>

²<https://tinyurl.com/3rpdr6dp>



(a) Example of emoji reaction reduce commenting noise.



(b) Example of emoji reaction does not reduce commenting noise.

Fig. 1. Examples of emoji reactions used in GitHub.

TABLE I
PRELIMINARY DATASET SUMMARY STATISTICS

	# Repositories	# PR	# PR Comments
With reactions	203 (48%)	6,867 (8%)	9,256 (4%)
Without reactions	217 (52%)	76,202 (92%)	249,354 (96%)
Total	420 (100%)	83,069(100%)	258,610 (100%)

A. Data Collection

We collected a list of 683 Eclipse repositories by using the official API of GitHub [2]. Since we focus on the reactions that are used in PRs, we excluded the repositories that do not have any PRs. We obtained 420 repositories that have 83,069 PRs. Then, we extracted PRs with reactions using GitHub GraphQL API [1], along with the information of reaction types, reaction time, the developer who posts the reaction. We excluded those reactions or PRs that are posted by bots since we investigate the reactions used by developers. Based on our manual check, we performed the following two exclusions:

(i) exclude developer name by using a regular expression matching('github.app') (ii) exclude the dependabot, a popular bot used to automatically notify developers of dependency upgrades.³ In the end, we obtained 9,256 comments having emoji reactions across 203 Eclipse repositories, as shown in Table I.

B. Data Analysis

With our preliminary dataset, we conducted three exploratory analyses related to emoji reaction usage. First, we investigate the prevalence of reaction usage yearly since this feature was initially introduced in 2016. To do so, we measured the proportion of PRs that have at least one emoji reaction. Second, we investigate what are the common reactions for developers to express during the PR process. To do so, we grouped the eight existing emoji reactions into four categories:

- **Positive** - is the single usage or the combination usage of THUMBS UP 👍, LAUGH 😄, HOORAY 🎉, HEART ❤️, and ROCKET 🚀 reactions.

³<https://dependabot.com/>

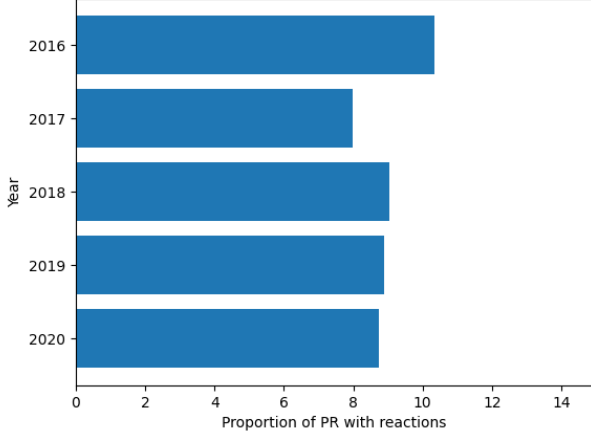


Fig. 2. Proportion of PRs with reactions by year in Eclipse repositories.

- **Negative** - is the single usage or the combination usage of THUMBS DOWN 👎 and CONFUSED 😕.
- **Neutral** - is the usage of EYES 👁️ reaction.
- **Mixed** - is the combination usage of the four categories mentioned above.

For example, if a PR comment is reacted by the THUMBS UP 👍 and the THUMBS DOWN 👎, we then classify this case as *Mixed*. Third, we further investigate whether or not the reaction is used to reduce commenting noise during the review process. To do so, we randomly select 20 PR comment samples from the preliminary dataset and did a manual classification (i.e., reduce commenting noise or not) among the first four authors. After the emoji reactions were posted, if there are no additional comments related to the existing topic by the developers who react, we classify such case as *Reduce Noise*. Otherwise, we classify the case as *Not Reduce Noise*. For example, in Figure 1(a), the Author B reacted with THUMBS UP to the suggestion provided by the Contributor A and added three commits without any additional comments. This case is labeled as *Reduce Noise*.

Positive emoji reactions are widely used in PRs. Two preliminary results are summarized. First, we find that around 8% to 10% of PRs have at least one reaction in Eclipse repositories between 2016 and 2020. Figure 2 shows the proportion of PRs with reactions yearly. Second, we observe that most of the reactions in PR comments are *Positive* (i.e., the single or combination usage of 👍, 😊, 🎉, ❤️, and 🚀), accounting for 98.1%. Upon a further inspection, among these *Positive* reactions, the single usage of 👍 almost reached 86.5%. On the other hand, the usage of *Negative*, *Neutral*, and *Mixed* only accounts for 1.84%. Table II shows the distribution of the sentiment of emoji reaction usage, indicating that the positive reaction is the most prevalent.

Emoji reactions do not always reduce the commenting noise. Table III shows the frequency of samples where whether reactions reduce commenting noise from our manual classifi-

TABLE II
DISTRIBUTION OF SENTIMENTS OF EMOJI REACTIONS.

Emoji reaction sentiments	# PR Comments
Positive	9,084 (98.10%)
Negative	67 (0.72%)
Neutral	74 (0.79%)
Mixed	31 (0.33%)
Total	9,256 (100%)

TABLE III
THE FREQUENCY COUNT OF MANUAL SAMPLES WHERE WHETHER THE EMOJI REACTIONS REDUCED THE COMMENTING NOISE.

Category	Count
Reducing commenting noise	12
Not Reducing commenting noise	8

cation. The table shows that eight samples (40%) are classified as *Not Reduce Noise*. In these samples, after the emoji reactions were posted, developers post additional comments to express and discuss issues on the PR.

Summary: Preliminary results show that around 8% to 10% of PRs have reactions in Eclipse repositories. We find cases where the emoji did not always reduce commenting noise in the discussion. Under a closer manual inspection of 20 emoji reactions, we find that there are eight cases where the emoji reactions did not reduce commenting noise.

III. STUDY PROTOCOLS

In this section, we present the design of our study. This section consists of our research questions with their motivations.

A. Research Questions

Inspired by the motivating examples and the preliminary study, we formulate four research questions to guide our study:

- **RQ1: Does the emoji reaction used in the review discussion correlate with review time?**

Prior studies [3, 19] have widely analyzed the impact of technical and non-technical factors on the review process (e.g., review outcome, review time). However, little is known about whether or not the emoji reaction can be correlated with review time. It is possible that emoji reaction may shorten the review time, as it could reduce the noise during the review discussions. Thus, our motivation for the first research question is to explore the correlation between the emoji reaction used in the review discussion and review time.

- **RQ2: Does a PR submitted by a first-time contributor receive more emoji reactions?**

As shown in Figure 1(b), we find that the emoji reaction might be used to express appreciation for submitting a PR. Our motivation for this research question is to understand if contributors that have never submitted to the

project before receive more emoji reactions. Furthermore, answering this research question will provide insights into a potent ulterior motive for a emoji reaction.

Our assumption is that:

H1: PRs submitted by first-time contributor receive more emoji reactions. Existing contributors express positive feelings to attract newcomers to the project.

• **RQ3: What is the relationship between the intention of comments and their emoji reaction?**

Our preliminary study findings show that emoji reactions do not always reduce the commenting noise. Hence, our motivation for the third research question is to explore the relationship between the intention of comments and their reactions.

Our assumption is that:

H2: Most emojis are uniformly distributed across the different intentions. Specific intentions may explain the ulterior purpose of reacting with an emoji reaction.

• **RQ4: Is emoji reaction consistent with comment sentiment?**

We found that specific sentiments of the emoji (i.e., THUMBS UP 👍) are widely used in PRs from our preliminary study. Our motivation for this research question is to investigate whether there is any inconsistency between sentiments of the comments and sentiments of the emoji reactions. Furthermore, we plan to manually check the reasons why inconsistency happened. We believe answering RQ4 would help newcomers better understand the emoji usage in the PR discussion.

Our assumption is that:

H3: The sentiment of emoji reactions are uniformly distributed across the same comment sentiments. Specific sentiments may explain the ulterior purpose of reacting with an emoji. This may be useful in understanding what information is needed in code review.

IV. DATA COLLECTION

To generalize the results of the study, we plan to expand on our dataset from active software development repositories shared by Hata et al. [13]. Each repository in this dataset has more than 500 commits and at least 100 commits during the most active two-year period. In total, this dataset contains 25,925 repositories from seven languages (i.e., C, C++, Java, JavaScript, Python, PHP, and Ruby). We will use the GraphQL API [1] to obtain PRs created before March 13rd 2016 where GraphQL was introduced. The whole dataset will be used for all four research questions.

V. EXECUTION PLAN

In this section, we present the execution plan of our experiment. We will use a mixed method consisting of both quantitative and qualitative analysis to answer our research questions.

A. Research Method for RQ1:

For the first research question, we plan to use a quantitative method. To investigate the effect of emoji reaction related factors on the pull request process (i.e., review time), we plan to perform a statistical analysis using a non-linear regression model. This model allows us to capture the relationship between the independent variable and the dependent variable. The goal of our statistical analysis is not to predict the review time but to understand the associations between the emoji reaction and the review time.

For the independent variables, similar to the prior studies [19, 23], we will select the following confounding factors as our independent variables:

- *PR size*: The total numbers of added and deleted lines of code changed by a PR.
- *Change file size*: The number of files what were changed by a PR.
- *Purpose*: The purpose of a PR, i.e., bug, document, feature.
- *# Comments*: The total number of comments in a PR discussion thread.
- *# Author Comments*: The total number of comments by the author in a PR discussion thread.
- *# Reviewer Comments*: The total number of comments by the reviewers in a PR discussion thread.
- *Patch author experience*: The number of prior PRs that were submitted by the PR author.
- *Reviewers*: The number of developers who posted a comment to a review discussion.
- *Commit size*: The number of commits in a PR.

Since we investigate the effect of the emoji reaction, we plan to compute additional independent variables that are related to emoji reaction:

- *With emoji reaction*: Whether or not a PR includes any emoji reaction (binary).
- *The number of emoji reactions*: The count of emoji reaction in a PR.

For the dependent variable (i.e., review time), we measure the time interval in hours from the time when the first comment was posted until the time when the last comment was posted. For the model construction, we will adopt the steps that are similar to the prior studies, including (i) Estimating budget for degrees of freedom, (ii) Normality adjustment, (iii) Correlation and redundancy analysis, (iv) Allocating degrees of freedom, and (v) Fitting statistical models.

(a) Analysis Plan: We will analyze the constructed regression models in the following three steps: (i) Assessing model stability. To evaluate the performance of our models, we will report the adjusted R^2 [12]. We will also use the bootstrap validation approach to estimate the optimism of the adjusted R^2 . (ii) Estimating the power of explanatory variables. Similar to prior work [23], we plan to test the significant correlation of independent variables with p-value and employ Wald statistics to measure the impact of each independent variable. (iii) Examining relationship. Finally, we will examine and plot

the direction of the relationship between each independent variable (i.e., especially emoji reaction related variables) and the dependent variable.

B. Research Method for RQ2:

For RQ2, we plan to use a quantitative method. To do so, we will construct two groups of pull requests to compare against: first-time contributors and non-first time contributors (control group). For the first-time contributor group, we will identify all pull requests that are submitted by first-time contributors from our dataset. For the non-first time contributor group, to construct a balanced control group, we will randomly select the equal number of pull requests that are submitted by non-first time contributors. We will then divide the pull requests into ones having emoji reactions and the other ones without emoji reactions, respectively.

(a) **Analysis Plan:** We will present a pivot chart to show the frequency of pull requests having emoji reactions or without emoji reactions by first-time contributors and non-first time contributors. The plot x-axis will represent two groups of first-time contributors and non-first time contributors. Furthermore, each group will be divided into two parts: pull requests with emoji reaction and pull reactions without emoji reactions. The plot y-axis will represent the frequency count of pull requests.

(b) **Significant Testing:** To select a suitable statistical test, we will adopt the Shapiro-Wilk test with $\alpha = 0.05$. In the case when the p-value is greater than 0.05, we will perform a two-tailed independent t-test with $\alpha = 0.05$. Otherwise, we will adopt a two-tailed Mann Whitney U test [21] with $\alpha = 0.05$ to validate.

In addition, we will investigate the effect size. In case when the data is normally distributed, we will use Hedges g effect size [14]. Effect size is analyzed as follows: (1) $|d| < 0.2$ as Negligible, (2) $0.2 \leq |d| < 0.5$ as Small, (3) $0.5 \leq |d| < 0.8$ as Medium, or (4) $0.8 \leq |d|$ as Large. If the data are not normally distributed, we will apply Cliff's delta (Romano et al, 2006) to measure effect size. Effect size is analyzed as follows: (1) $|\delta| < 0.147$ as Negligible, (2) $0.147 \leq |\delta| < 0.33$ as Small, (3) $0.33 \leq |\delta| < 0.474$ as Medium, or (4) $0.474 \leq |\delta|$ as Large.

C. Research Method for RQ3:

For RQ3, we plan to use a quantitative method to classify the intentions of the comments. To categorize the intentions of the comments, we will use a taxonomy of intention proposed by Huang [17]. They manually categorized 5,408 sentences from issue reports of four projects in GitHub to generalize the linguistic pattern for category identification.

The taxonomy of intention category is described below:

- **Information Giving (IG):** Share knowledge and experience with other people, or inform other people about new plans/updates (e.g., "The typeahead from Bootstrap v2 was removed.").
- **Information Seeking (IS):** Attempt to obtain information or help from other people (e.g., "Are there any developers working on it?").

- **Feature Request (FR):** Require to improve existing features or implement new features (e.g., "Please add a titled panel component to Twitter Bootstrap.").
- **Solution Proposal (SP):** Share possible solutions for discovered problems (e.g., "I fixed this for UI Kit using the following CSS.").
- **Problem Discovery (PD):** Report bugs, or describe unexpected behaviors (e.g., "the firstletter issue was causing a crash.").
- **Aspect Evaluation (AE):** Express opinions or evaluations on a specific aspect (e.g., "I think BS3's new theme looks good, it's a little flat style.").
- **Meaningless (ML):** Sentences with little meaning or importance (e.g., "Thanks for the feedback!").

To facilitate the automation, they proposed a convolution neural network based classifier with high accuracy. For RQ3, we will use this classifier to automatically label the intention of the comments. To evaluate the robustness of this classifier in our dataset, we will first use the proposed classifier to automatically classify the intentions of the randomly sampled 30 comments. Then, we will manually check whether the labeled intentions of these 30 comments are correct or not. The result of this sanity check will be presented as a percentage of the false positive, under 10% being considerable.

(a) **Analysis Plan:** To analyze the relationship between the intention of comments and their emoji reaction, we will use the association rule mining technique. To show the diversity of different intentions from the classification, we will draw a histogram plot. To show the results of the relationship, a table will be drawn with descriptive statistics, including the criteria support and confidence.

(b) **Significant Testing:** To inspect whether or not the classified intentions of comments are normally distributed, we will adopt the Shapiro-Wilk test with $\alpha = 0.05$, which is widely used for the normality test. In addition, to inspect whether the intentions of comments are significantly different, we will use Kruskal-Wallis non-parametric statistical test [4].

D. Research Method for RQ4:

For RQ4, we plan to use a qualitative and quantitative method. First of all, we will use a quantitative method to investigate whether there is any inconsistency between sentiments of emoji reaction and sentiments of the comments. We determine the sentiments of the emoji based on the definition we discussed in section 2. Hence, Emoji sentiment can be categorized into the following types: Positive, Negative, Neutral, and Mixed. To extract the sentiment of the first responses, we plan to use SentiStrength-SE [18], the state-of-the-art sentiment analysis tool for software engineering text. Similar to the tool we plan to use for RQ3, the input is reacted comments, and the output is the sentiment score of the given comment. The sentiment score varies from -5 (very Negative) to 5 (very Positive). Based on the above definition, we consider it as inconsistent if the sentiments of the comment are different from the sentiments of the emoji.

Then, we will conduct qualitative analysis to explore the possible reasons for inconsistency between sentiments of emoji and sentiments of the comments. To do so, we will apply the open coding approach [5] to classify the reasons for inconsistency. To discover as a complete list of reasons as possible, we strive for *theoretical saturation* [11]. Similar to prior work [15], we set our saturation criterion to 50, i.e., we continue to code randomly selected comments until no new reasons have been discovered for 50 consecutive comments. Furthermore, we perform the kappa agreement score [22] to evaluate the classification quality. Similar to Hata et al. [13], the agreement of the coding guide will be performed using a kappa agreement. Kappa result is interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. The agreement scores larger than 0.81 (i.e., almost perfect) are considered for the manual analysis. Based on the prior experience, we estimate the size of the samples to range from 200–300 samples.

(a) Analysis Plan: Similar to RQ3, we will depict a histogram plot to show the distribution of emoji types by sentiments of the comment. To show the results of inconsistency reasons, we will draw a histogram plot to show the frequency. As part of our result presentation, we will paste the real examples during the analysis to describe the reason taxonomy.

(b) Significant Testing: Similar to RQ3, we will use Shapiro-Wilk test to inspect whether or not the sentiments of emoji usage are normally distributed and use Kruskal-Wallis non-parametric statistical test to validate the significant difference.

VI. IMPLICATIONS

We summarize our implications with the following take-away messages for the key stakeholders:

- **Researchers:** Answering RQ1 will help researchers understand the impact of emoji reactions, hence may contribute to existing knowledge on the code review process. As informal communication such as the usage of emojis becomes prevalent, it is a need to understand its role in keeping an efficient code review process. We believe that emojis may also help remove toxic and other forms of anti-patterns [8] in the code review process. In terms of the intention of the emoji reactions, our study will complement all related works on emoji usage [6, 7].
- **Contributors:** In terms of practitioners, we envision our study to assist projects to attract and maintain existing and potential contributors to the project. Especially for RQ2, the results of the study may provide some insights into how to attract newcomers and also how to provide a more friendly and welcoming environment. Furthermore, answering RQ3 and RQ4 will provide some insights for contributors into understanding the consistency of emoji reactions. As emoji usage becomes popular [20], the results of the study should provide guidelines on how to

represent the common intentions when an emoji reaction is warranted.

VII. THREATS TO VALIDITY

We identified three key threats to our study. First, the Eclipse projects that were used in our preliminary study may not be representative of all types of GitHub projects. To increase generalizability, we will extend our study to include a sample of random GitHub projects [13]. Our second threat is concerning the qualitative aspect of the study, as this is bias to human error in the classification. This is because the interpretation of emoji usage may not be trivial. To mitigate this, we employ the Kappa method to have multiple co-authors for agreement of each code. For instance, if positive emoji is used in an ironic context, it does not mean positive. This usage of emojis may influence our results. Third, our quantitative analysis and data collection may include some false positives, such as bot reactions and comments. Currently, we manually exclude these bots for the preliminary study. To mitigate this, we plan to carefully identify and systematically remove bots based on official documentation.

REFERENCES

- [1] Github graphql api v4. URL <https://docs.github.com/en/graphql>.
- [2] Github rest api. URL <https://docs.github.com/en/rest>.
- [3] Olga Baysal, Oleksii Kononenko, Reid Holmes, and Michael W. Godfrey. Investigating Technical and Non-technical Factors Influencing Modern Code Review. *Empirical Software Engineering*, page 932–959, 2016.
- [4] NORMAN BRESLOW. A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika*, pages 579–594, 1970.
- [5] Kathy Charmaz. *Constructing Grounded Theory*. SAGE, 2014.
- [6] Zhenpeng Chen, Yanbin Cao, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. Sentimoji: an emoji-powered learning approach for sentiment analysis in software engineering. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 841–852, 2019.
- [7] Zhenpeng Chen, Yanbin Cao, Huihan Yao, Xuan Lu, Xin Peng, Hong Mei, and Xuanzhe Liu. Emoji-powered sentiment and emotion detection from software developers’ communication data. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 30(2):1–48, 2021.
- [8] Moataz Chouchen, Ali Ouni, Raula Gaikovina Kula, Dong Wang, Patanamon Thongtanunam, Mohamed Wiem Mkaouer, and Kenichi Matsumoto. Anti-patterns in modern code review: Symptoms and prevalence. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 531–535, 2021.

- [9] Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social coding in github: transparency and collaboration in an open software repository. In Proceedings of the ACM 2012 conference on computer supported cooperative work, pages 1277–1286, 2012.
- [10] F. Ebert, F. Castor, N. Novielli, and A. Serebrenik. Confusion in code reviews: Reasons, impacts, and coping strategies. In 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER), pages 49–60, 2019.
- [11] Kathleen M Eisenhardt. Building theories from case study research. Academy of management review, 1989.
- [12] T. Hastie, R. Tibshirani, and J.H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.
- [13] Hideaki Hata, Christoph Treude, Raula Gaikovina Kula, and Takashi Ishio. 9.6 million links in source code comments: Purpose, evolution, and decay. In Proceedings of the 41st International Conference on Software Engineering, ICSE '19, page 1211–1221. IEEE Press, 2019.
- [14] Larry V. Hedges and I. Olkin. Statistical Methods for Meta-Analysis. Academic Press, 1985.
- [15] Toshiki Hirao, Shane McIntosh, Akinori Ihara, and Kenichi Matsumoto. The review linkage graph for code review analytics: a recovery approach and empirical study. In Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pages 578–589, 2019.
- [16] Toshiki Hirao, Shane McIntosh, Akinori Ihara, and
- [21] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, Kenichi Matsumoto. Code reviews with divergent review scores: An empirical study of the openstack and qt communities. IEEE Transactions on Software Engineering, 2020.
- [17] Qiao Huang, Xin Xia, David Lo, and Gail C Murphy. Automating intention mining. IEEE Transactions on Software Engineering, 46(10):1098–1119, 2018.
- [18] M. R. Islam and M. Zibran. Sentistrength-se: Exploiting domain specificity for improved sentiment analysis in software engineering text. J. Syst. Softw., 145:125–146, 2018.
- [19] O. Kononenko, T. Rose, O. Baysal, M. Godfrey, D. Theisen, and B. de Water. Studying pull request merges: A case study of shopify’s active merchant. In 2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Practice Track (ICSE-SEIP), pages 124–133, 2018.
- [20] Renee Li, Pavithra Pandurangan, Hana Frluckaj, and Laura Dabbish. Code of conduct conversations in open source software projects on github. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1):1–31, apr 2021.
- 18(1):50 – 60, 1947.
- [22] Anthony Viera and Joanne Garrett. Understanding interobserver agreement: The kappa statistic. Family medicine, pages 360–3, 06 2005.
- [23] Dong Wang, Tao Xiao, Patanamon Thongtanunam, Raula Gaikovina Kula, and Kenichi Matsumoto. Understanding shared links and their intentions to meet information needs in modern code review. In The Journal of Empirical Software Engineering (EMSE), volume 26, page to appear, 2021. doi: <https://doi.org/10.1007/s10664-021-09997-x>.