Predicting Usefulness of Code Review Comments using Textual Features and Developer Experience

Abstract: Although peer code review is widely adopted in both commercial and open source development, existing studies suggest that such code reviews often contain a significant amount of non-useful review comments. Unfortunately, to date, no tools or techniques exist that can provide automatic support in improving those non-useful comments. In this paper, we first report a comparative study between useful and nonuseful review comments where we contrast between them using their textual characteristics and reviewers' experience. Then, based on the findings from the study, we develop RevHelper, a prediction model that can help the developers improve their code review comments through automatic prediction of their usefulness during review submission. Comparative study using 1,116 review comments suggested that useful comments share more vocabularies with the changed code, contain salient items like relevant code elements, and their reviewers are generally more experienced. Experiments using 1,482 review comments report that our model can predict comment usefulness with 66% prediction accuracy which is promising. Comparison with three variants of a baseline model using a case study validates our empirical findings and demonstrates the potential of our model.

Comparative Study

Code Review Comments: The comments below are used for our comparative study between useful and non-useful comments

- CS (256)
- SM (281)
- MS (288)
- SR (291)
- All useful comments (618)
- All non-useful comments (498)

Tools/Items for Replication:

- Stop words
- Python keywords
- Readability Ease calculator library
- Regex for question identification: `"[?]($|\s)"` and sentence identification: `"[?!.]($|\s)"`
- Regex for code element identification: `"`{3}[\S+\s+]+`{3}|`[\S+]+`|`[\S+\s+]+`"`
- GitHub API librarry
- Git Bash

Experimental Data

Review Comments for Evaluation:

- All useful comments (618)
- All non-useful comments (498)

Review Comments for Validation & Case study:

- CS (81)
- SM (99)
- MS (99)
- SR (87)
- All useful comments (262)
- All non-useful comments (104)

Prediction Model for Comment Usefulness

- Feature Set
- Evaluation dataset
- Validation dataset
- RevHelper Model (Training & Testing)
- RevHelper Model (Validation)