CS680 Proposal: Gender Fairness in Open Source Software

Amirreza Shamsolhodaei AMIRREZA.SHAMSOLHODAEI@UWATERLOO.CA
Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada

1. Chosen option and the problem

I have chosen the empirical evaluation option for my project.

Open Source Software (OSS) projects are the result of the collaborative work of developers with diverse backgrounds. Despite the fact that the quality of their contributions should be the sole criteria for assessing their value to the project, as it is discussed in the work by Nadri et al. (2022). studies have demonstrated a correlation between diversity factors, such as gender, and the acceptance or rejection of these contributions. Vasilescu et al. (2015) work suggests that diversity in race and gender is beneficial for the software development industry and studies show that team diversity positively correlates with team output in a positive way. Terrell et al. (2016) highlighted the different acceptance rates of men and women in OSS community in detail and discovered some noteworthy outcomes. They found that gender bias could work both ways, with biases against men occurring in certain situations. Additionally, they identified that factors such as experience and the tendency of women to tackle immediate project issues could contribute to the gender bias observed in OSS.

Looking at the state of AI and its increasing prevalence in our daily lives,Mehrabi et al. (2021) suggests it is important to consider fairness in the design and engineering of artificial intelligence systems. The use of AI in sensitive environments where crucial decisions are made necessitates ensuring that such decisions do not exhibit discriminatory behavior towards specific groups or populations.

With these issues in mind, I plan to use the dataset provided by Zhang et al. (2020) as a base for my project, This is a rich dataset with over 3 million pull requests and 95 features, such as contributor gender, contributor's experience, lines of codes changed, and a number of commits. they have used this dataset in their work Zhang et al. (2023) for the purpose of finding a correlation between the features discussed and the decision that was made for the pull request by the integrator of the pull request.

2. Proposed Machine Learning Techniques

What I will try to do is make contributor gender a dependent variable on some of the features in the dataset and try to find the correlation between them and contributor gender as well as making predict gender based on the features. Machine learning models that I intend to use would be Logistic Regression, Decision Trees, Random Forests, Support vector Machines, and Neural Networks. I plan to use these models to find the features most correlated with the feature and then predict if a pull request was made by a male or female contributor I plan on experimenting with different models and comparing their performance to determine the best approach for this problem. This study will provide insight into the fact that are there any biased toward a specific gender in OSS.

References

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, jul 2021. doi: 10.1145/3457607. URL https://doi.org/10.1145%2F3457607.

Reza Nadri, Gema Rodriguez-Perez, and Meiyappan Nagappan. On the relationship between the developer's perceptible race and ethnicity and the evaluation of contributions in OSS. *IEEE Transactions on Software Engineering*, 48(8):2955–2968, aug 2022. doi: 10.1109/tse.2021.3073773. URL https://doi.org/10.1109%2Ftse.2021.3073773.

Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, and Chris Parnin. Gender bias in open source: Pull request acceptance of women versus men. feb 2016. doi: 10.7287/peerj.preprints.1733v1. URL https://doi.org/10.7287%2Fpeerj.preprints.1733v1.

Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark G.J. van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. Gender and tenure diversity in GitHub teams. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, apr 2015. doi: 10.1145/2702123.2702549. URL https://doi.org/10.1145%2F2702123.2702549.

Xunhui Zhang, Ayushi Rastogi, and Yue Yu. On the shoulders of giants. In *Proceedings of the 17th International Conference on Mining Software Repositories*. ACM, jun 2020. doi: 10.1145/3379597.3387489. URL https://doi.org/10.1145%2F3379597.3387489.

Xunhui Zhang, Yue Yu, Georgios Gousios, and Ayushi Rastogi. Pull request decisions explained: An empirical overview. *IEEE Transactions on Software Engineering*, 49(2):849–871, feb 2023. doi: 10.1109/tse.2022.3165056. URL https://doi.org/10.1109%2Ftse.2022.3165056.