Diversidade de Gênero em Projetos Open Source: um Estudo da Relevância dos Comentários Postados em *Issues* do GitHub

Estela Miranda Batista[1], **Gláucia Braga e Silva**[1], **Thais Regina de Moura Braga Silva**[1]

[1]Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa (UFV)
Florestal - MG - Brazil

{estela.batista,glaucia,thais.braga}@ufv.br

Resumo. *Estudos sobre diversidade de gênero em comunidades de software evidenciam a baixa representatividade feminina e, por consequência, a baixa participação, geralmente medida em função do número de contribuições em código. Este trabalho analisa a participação feminina em termos da relevância dos comentários postados no issue tracking do GitHub utilizando dados de 5 comunidades open source abertas e 5 dedicadas às mulheres. Os resultados mostram que, mesmo com pouca representatividade, na média, a relevância dos comentários feitos por mulheres é similar a dos homens. Em comunidades dedicadas além da representatividade, a relevância se mostra maior.*

1. Introdução

A diversidade de gênero na Tecnologia da Informação (TI) tem sido abordada na literatura, com diversos estudos trazendo evidências sobre os benefícios da inclusão de mulheres no desenvolvimento. Porém, os índices de participação feminina são baixos e o tempo de permanência nos projetos de *software* é curto [Vedres and Vasarhelyi 2019], o que reflete em uma baixa contribuição [Izquierdo et al. 2019, Zacchiroli 2021].

Em geral, a participação das pessoas desenvolvedoras é medida por meio da contabilização das contribuições em códigos nas plataformas de versionamento, como o GitHub [Izquierdo et al. 2019, Zacchiroli 2021]. No entanto, considerando-se a natureza colaborativa e humano-dependente do ambiente de *software*, outros aspectos sociais podem ser investigados em conjunto com as métricas já conhecidas. A diversidade de gênero é o objeto de estudo deste trabalho, que busca avaliar a relevância da participação feminina no âmbito das comunicações ocorridas no ambiente do *issue tracking* do GitHub. O estudo é guiado pelas seguintes questões de pesquisa: **QP1.** Existe diferença na relevância dos comentários postados por homens e mulheres?; **QP2.** Qual a diferença entre o número de comentários postados por homens e mulheres?; e **QP3.** Qual a diferença entre o número de *issues*[1] reportadas por homens e mulheres?

Para responder às questões de pesquisa enumeradas, foram avaliados dados de comunicação extraídos de repositórios de projetos *open source* no GitHub, vinculados tanto a comunidades abertas como a outras dedicadas para mulheres na área de TI. O GitHub foi usado por ser a maior comunidade de código aberto, possuindo mais de 61 milhões de repositórios de *software*[2], além de ser cada vez mais usada entre os pesquisadores como fonte de mineração de dados [Gousios and Spinellis 2017,

[1]Issues representam problemas, bugs, defeitos, sugestões de melhorias, novos requisitos, entre outros.
[2]https://octoverse.github.com/

Saadat et al. 2020]. Os dados foram analisados quali-quantitativamente com base na aplicação da métrica Relevância Temática, proposta para o contexto de *issue tracking* [Neto and Silva 2018], além de outras métricas quantitativas de comunicação, calculadas em função do gênero da pessoa desenvolvedora.

O restante deste trabalho está organizado como segue: a seção 2 traz os trabalhos relacionados; a seção 3 apresenta os materiais e métodos; os resultados obtidos são discutidos na seção 4; e, por fim, as considerações finais são apresentadas na seção 5.

2. Trabalhos Relacionados

No que compete às pesquisas que abordam a baixa participação feminina em projetos de *software*, [Zacchiroli 2021] analisou 1.6 milhões de *commits* feitos num período de 50 anos, e concluiu que 92% do código produzido foi feito por homens. A discrepância também pode ser observada em termos de liderança feminina nos projetos do GitHub, conforme apontaram [Izquierdo et al. 2019]. Os autores analisaram mais de 7000 perfis de usuários na plataforma e observaram que apenas 8% dos líderes de projetos são mulheres. Buscando investigar novos indicadores sobre a participação feminina em projetos *open source*, este trabalho aplica métricas quali-quantitativas sobre os dados de comunicação no *issue tracking* do GitHub, com o intuito de avaliar a qualidade das mensagens postadas em função do gênero da pessoa desenvolvedora.

A análise dos dados de comunicações ocorridas em ambientes de *issue tracking* tem sido abordada na literatura [citação omitida para revisão] [Ortu et al. 2018], mas não foram encontrados trabalhos que investiguem questões de gênero neste contexto.

Considerando a participação feminina no universo *open source*, [Singh 2019] analisou 355 sites de projetos e os resultados apontam que menos de 5% das comunidades tem espaço exclusivos para mulheres. Neste trabalho, foram utilizados dados de ambientes dedicados ao gênero feminino em comparação com comunidades abertas, em que os homens são maioria, de forma a verificar se existem diferenças em termos da qualidade da atuação das mulheres no processo de comunicação.

3. Materiais e Métodos

Para a condução desse estudo de análise dos dados de comunicações no *issue tracking* do GitHub, foram usadas três métricas calculadas em relação ao gênero do autor: a) contabilização de *issues* criadas; b) contabilização de comentários postados; e c) a relevância temática dos comentários.

Para aplicação das métricas de contabilização de *issues* criadas e comentários postados, em função do gênero da pessoa desenvolvedora, os dados foram extraídos da plataforma do GitHub através da sua API Rest[3]. Para adivinhação do gênero das pessoas desenvolvedoras, foi utilizada a ferramenta NamSor[4], que utiliza do nome e sobrenome inseridos, uma vez que não é obrigatório informar o gênero na plataforma [Zolduoarrati and Licorish 2021].

No que diz respeito à avaliação da qualidade dos comentários, foi usada a métrica da Relevância Temática, adaptada por [Neto and Silva 2018] para o ambiente de *issue*

[3]https://docs.github.com/pt/rest

[4]https://namsor.app/

tracking. Essa métrica representa o quanto um texto é relevante em relação ao tema de uma determinada discussão, sendo isso feito através do número de termos relevantes correspondentes entre um comentário e o texto de problematização. Este trabalho realizou adaptações sobre a proposta de [Neto and Silva 2018], substituindo o uso de grafos por uma técnica de cossenos [Medeiros et al. 2014] como forma de comparação da similaridade dos textos. A alteração apresentou melhora significativa no desempenho da aplicação, em especial quando se avaliava um grande conjunto de dados, e não provocou alteração em termos da qualidade dos resultados gerados. A nova fórmula da Relevância Temática, considerando a alteração com o uso da similaridade de cossenos, se tornou a média aritmética entre S_{CI} a semelhança entre o comentário e a *issue*, e S_{CD} a semelhança entre o comentário e a discussão, considerando o título e a descrição da *issue* e os comentários anteriores.

Por fim, no que diz respeito à diversidade das equipes, foi usado o Índice Blau como medida de diversidade de gênero [Biemann and Kearney 2010]. A métrica de diversidade foi usada com dois propósitos: como critério de seleção das comunidades *open source* e para fins de comparação entre os resultados das métricas aplicadas aos dados da comunicação. A Equação 1 apresenta a fórmula para o Índice Blau, que calcula, de um total de N, a porcentagem P de indivíduos em cada categoria i. Neste trabalho, N = 2, visto que as categorias são masculino e feminino. O índice varia entre 0 e 0.5, sendo 0.5 o equilíbrio no número de indivíduos nas categorias.

$$Blau = 1 - \sum_{i=1}^{N} P_i^2 \tag{1}$$

3.1. Seleção e Extração dos Dados

Para este estudo, foram considerados projetos de comunidades abertas e aqueles desenvolvidos no contexto de comunidades específicas para mulheres. Os dados foram extraídos em janeiro de 2022, por meio de uma aplicação na linguagem Python. A implementação utilizou filtros baseados no trabalho de [Neto et al. 2021], divididos em duas categorias: filtros de projetos (FP); e filtros de tópicos (FT), conforme mostra a Tabela 1.

Tabela 1. Filtros de Projetos e Tópicos

[FP1] Possuir, no mínimo, 5 membros	**[FT1]** Possuir, no mínimo, 5 comentários
[FP2] Possuir, no mínimo, 5 *commits*	**[FT2]** Os comentários não podem conter apenas trechos de códigos (um conteúdo textual é requerido)
[FP3] Possuir, no mínimo, 5 *issues* abertas	**[FT3]** Possuir um tempo de abertura mínimo de uma semana
[FP4 Possuir, no mínimo, 5 *issues* fechadas	**[FT4]** *Issues* não podem ter sido reabertas
[FP5] Ter sido criado há pelo menos 6 meses	

Foram extraídos 9151 comentários, de 1275 *issues*, presentes em 28 repositórios de 10 comunidades, sendo 5 comunidades abertas e 5 dedicadas às mulheres. Autores de comentários e *issues* tiveram seu gênero adivinhado, e foi realizado o cálculo da relevância temática de todos os comentários postados. Os dados extraídos foram armazenados em um repositório no GitHub[5]. A Figura 1 apresenta os quantitativos de mulheres e

[5]https://github.com/stardotwav/AnaliseGeneroGitHub

homens, juntamente com os valores do índice Blau das equipes para as 10 comunidades consideradas neste estudo.

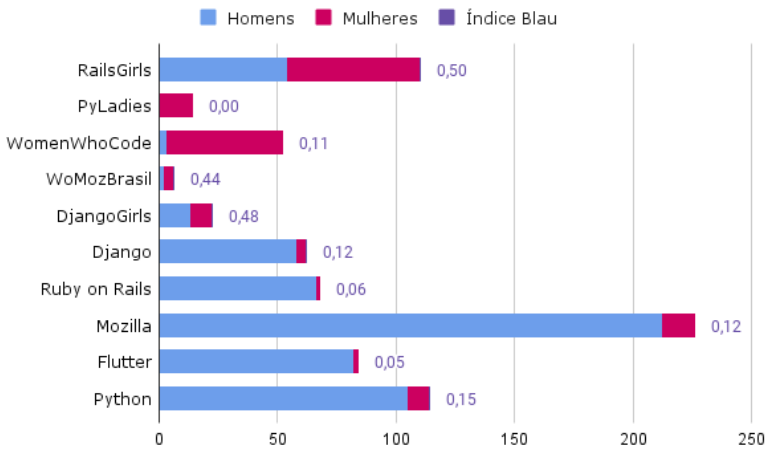


Figura 1. Comunidades, Contagem por Gênero e Índice Blau

As 5 primeiras comunidades apresentadas na Figura 1 são dedicadas para mulheres, o que explica um maior presença feminina e maiores índices de Blau, em alguns casos. No caso da *PyLadies*, o índice Blau é igual a 0.0, o que significa ausência de diversidade, uma vez que essa comunidade se fecha para a presença masculina com o intuito de propiciar um ambiente dedicado e seguro para as mulheres.

4. Resultados e Discussões

Para responder as questões de pesquisa endereçadas neste estudo, os dados foram analisados por meio da ferramenta Jupyter Notebook[6] e os resultados obtidos encontram-se descritos nas próximas subseções.

4.1. QP1. Existe diferença entre a relevância dos comentários postados por homens e mulheres?

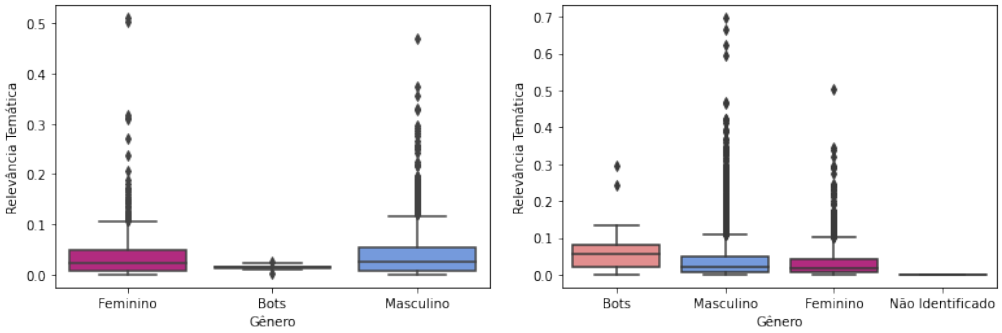
Conforme mostrado no gráfico das Figuras 2 (a)) e 2 (b)), observa-se que o intervalo de dados altera entre os gêneros, segundo o contexto do segmento em análise. Nas comunidades dedicadas (Figura 2 (a)), diferente das abertas (Figura 2 (b)), as mulheres alcançam valores de relevância maiores que os homens.

Além das análises por tipo de comunidade, também foi calculada a relevância média por gênero, para todo o conjunto de dados, resultando em relevâncias similares, sendo 0.03680, para os homens e 0.03394, para as mulheres. Esse resultado nos mostra que os comentários postados por mulheres, apesar de em um número muito menor, são igualmente relevantes para a discussão em torno das *issues*.

4.2. QP2. Qual a diferença entre número de *issues* reportadas por homens e mulheres?

Das 1275 *issues* analisadas, 1071 (84%) foram reportadas por homens, e 204 (16%) por mulheres. Olhando para o recorte das comunidades dedicadas, temos 218 *issues*, das

[6]https://jupyter.org/



(a) Comunidades Dedicadas (b) Comunidades Abertas

Figura 2. Intervalo da Relevância Temática

quais 102 (48%) foram reportadas por mulheres. Já no contexto das comunidades abertas, das 1057 *issues*, apenas 102 ($\sim$ 9%) foram reportadas por mulheres. Os números mostram que as mulheres parecem se sentir mais seguras e confiantes para reportar *issues* no contexto de comunidades dedicadas.

4.3. QP3. Qual a diferença entre número de comentários postados por homens e mulheres?

O gênero dos autores dos comentários foi obtido automaticamente. Em alguns casos foi necessária uma intervenção manual e, apenas uma pessoa desenvolvedora permaneceu, ao final, com gênero indefinido. Além disso, é importante ressaltar que alguns comentários foram postados por *bots*, e, por conta disso, foram considerados em separado nos gráficos da Figura 2. Dos 9151 comentários avaliados, 6897 foram postados por homens, 2059 (22%) por mulheres, 194 por *bots*. Olhando apenas para comunidades dedicadas, tem-se 1923 comentários, sendo 741 (39%) postados por mulheres. Já em comunidades abertas, dos 7447 comentários, apenas 1395 (19%) foram postados por mulheres, o que também evidencia a baixa participação em comparação com os homens neste segmento.

5. Considerações Finais

Com base nos dados analisados, temos que os índices de participação feminina no *issue tracking* de comunidades abertas se mostram baixos, com apenas 19% dos comentários postados e 9% das *issues* reportadas. Em comunidades dedicadas, os números se tornam melhores, chegando a 50% em termos de comentários postados. Porém, a média da relevância para mulheres e homens é similar, sendo que em comunidades abertas as mulheres chegam a ter valores maiores. Visto que os comentários das mulheres se mostram igualmente relevantes nas discussões dos projetos, o desafio é criar um ambiente seguro para elas em qualquer comunidade, usando de ações simples como o estabelecimento de um código de conduta e/ou a presença constante de uma pessoa moderadora, por exemplo.

Como trabalhos futuros, podem ser realizadas entrevistas com mulheres que participam em ambos os tipos de comunidades, para se conhecer as percepções das mesmas quanto aos motivos que levam aos baixos índices obtidos. Além disso, pode-se avaliar a influência dos comentários postados por mulheres na resolução das *issues* e na discussão como um todo. Métricas de estimativa de tempo de resolução de *issues* e análise de sentimentos poderiam ser usadas para guiar tais estudos.

Agradecimentos

Agradecemos ao CNPq pela bolsa de iniciação científica e ao programa MinasCoders da UFV, pelo fomento a pesquisas envolvendo temáticas de gênero em Computação.

Referências

Biemann, T. and Kearney, E. (2010). Size does matter: How varying group sizes in a sample affect the most common measures of group diversity. *Organizational Research Methods*, 13:582–599.

Gousios, G. and Spinellis, D. (2017). Mining software engineering data from github. In *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, pages 501–502.

Izquierdo, D., Huesman, N., Serebrenik, A., and Robles, G. (2019). Openstack gender diversity report. *IEEE Software*, 36:28–33.

Medeiros, D. C., de Queiroz, J. E. R., and Araújo, J. M. F. R. (2014). Análise de funções de similaridade para verificação do conteúdo de mensagens em fóruns de discussão. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, volume 25, page 144.

Neto, L., Silva, G., and Comarela, G. (2021). Estimativa do tempo de resolução de issues no github usando atributos textuais e temporais. In *Brazilian Symposium on Software Engineering*, page 253–262, New York, NY, USA. Association for Computing Machinery.

Neto, L. E. C. and Silva, G. B. e. (2018). Colminer: A tool to support communications management in an issue tracking environment. In *Proceedings of the XIV Brazilian Symposium on Information Systems*, SBSI'18, New York, NY, USA. Association for Computing Machinery.

Ortu, M., Hall, T., Marchesi, M., Tonelli, R., Bowes, D., and Destefanis, G. (2018). Mining communication patterns in software development: A github analysis. In *Proceedings of the 14th International Conference on Predictive Models and Data Analytics in Software Engineering*, PROMISE'18, page 70–79, New York, NY, USA. Association for Computing Machinery.

Saadat, S., Newton, O. B., Sukthankar, G., and Fiore, S. M. (2020). Analyzing the productivity of github teams based on formation phase activity. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 169–176.

Singh, V. (2019). Women-only spaces of open source. In *2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE)*, pages 17–20.

Vedres, B. and Vasarhelyi, O. (2019). Gendered behavior as a disadvantage in open source software development. *EPJ Data Science*, 8.

Zacchiroli, S. (2021). Gender differences in public code contributions: A 50-year perspective. *IEEE Software*, 38(2):45–50.

Zolduoarrati, E. and Licorish, S. A. (2021). On the value of encouraging gender tolerance and inclusiveness in software engineering communities. *Information and Software Technology*, 139:106667.