

Relevância Temática dos Comentários e Diversidade de Gênero em Projetos Open Source

Estela Miranda Batista
Federal University of Viçosa
Florestal, Brasil
estela.batista@ufv.br

Gláucia Braga e Silva
Federal University of Viçosa
Florestal, Brasil
glauucia@ufv.br

RESUMO

Este trabalho analisa a participação feminina em termos da relevância temática dos comentários postados no ambiente de *issue tracking* do GitHub. O trabalho também investiga se existem relações entre a relevância temática e outras métricas ligadas ao desenvolvedor, como reputação, tempo de plataforma e número de *issues* reportadas. Foram analisados dados de 5 comunidades *open source* abertas, e 5 dedicadas às mulheres. Os resultados apontam que, na média, a relevância dos comentários feitos por mulheres é similar à dos homens, se mostrando igualmente capazes, porém demonstram uma representatividade e participação muito baixas, tendo apenas 22% dos comentários postados e 16% das *issues* reportadas.

KEYWORDS

Gênero no GitHub, Comunicações em Issue Tracking, Índice Blau

1 INTRODUÇÃO

Considerando a natureza colaborativa do desenvolvimento de software, as equipes podem ser formadas por pessoas de diversas culturas, idades, nacionalidades, gêneros, entre diversos outros fatores, de forma que aquelas que possuem uma maior diversidade social podem aproveitar de informações mais amplas, experiências e habilidades aprimoradas de resolução de problemas, tornando-se, portanto, mais eficazes [16]. Apesar do fator da diversidade ser benéfico para as equipes de desenvolvimento, existem trabalhos na literatura que apontam a falta dela no desenvolvimento de software, principalmente em questões de gênero, em que mulheres tendem a ter uma permanência menor que a de homens [17]. Apesar da baixa permanência, estudos mostram que as mulheres funcionam como mediadoras dos problemas que surgem [5], sendo peças chave no processo de comunicação.

A baixa participação feminina em projetos de desenvolvimento de software é bem conhecida na literatura e tem sido avaliada, principalmente, em termos na contabilização das contribuições em código e *pull request* aceitas [4, 13]. No entanto, outros aspectos podem ser analisados com o intuito de se investigar a participação feminina na composição de equipes mais diversas, mais inclusivas e, possivelmente, potencialmente mais produtivas [5]. Dentre esses

aspectos sociais, 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 será guiado pelas seguintes questões de pesquisa:

- **QP1.** Qual a diferença de participação de homens e mulheres em termos de *issues* e comentários associados?
- **QP2.** Existe diferença entre a relevância dos comentários postados por homens e mulheres?
- **QP3.** Existe relação entre a relevância dos comentários e a reputação do autor?
- **QP4.** Existe relação entre a relevância dos comentários e o tempo de participação nas plataformas?

Para responder às questões de pesquisa enumeradas, serão avaliados dados de comunicação extraídos de repositórios de projetos *open source* na plataforma do GitHub, em busca de análises comparativas entre a participação de homens e mulheres no contexto do *issue tracking*. O GitHub será usado por ser a maior comunidade de código aberto, possuindo mais de 61 milhões de repositórios de software criados por mais de 16 milhões de desenvolvedores cadastrados na plataforma¹, além de ser cada vez mais usada entre os pesquisadores como fonte de mineração de dados [6, 14]. Dentre os dados extraídos, os mesmos foram classificados em dois tipos de comunidades *open source*, sendo eles comunidades abertas, e comunidades dedicadas à mulheres, de forma a realizar comparações sobre a participação de mulheres e homens em cada um desses ambientes.

Os dados serão analisados com base na aplicação da métrica Relevância Temática, proposta por Neto e Braga [11] para o contexto de discussões em *issue tracking*. A relevância dos comentários será analisada em conjunto com outras métricas de comunicação, sendo elas o número de *issues* criadas e número de comentários postados, em função de outros atributos que caracterizem o perfil do desenvolvedor como gênero, tempo de participação na plataforma e reputação.

Como resultados, serão apresentados indicativos importantes sobre o processo de comunicação dos times e as possíveis diferenças em termos de contribuições em função do gênero do desenvolvedor. Ressalta-se que o termo gênero está sendo tratado neste trabalho como a forma de identificação do usuário na plataforma, sendo masculino ou feminino.

O restante do artigo está organizado da seguinte forma: a seção 2 apresenta informações sobre as Comunicações em *Issue Tracking*, e aborda o conceito de relevância dos comentários. Na seção 3, são apresentados os trabalhos relacionados. A seção 4 traz os materiais

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

POC'22, Março 2022, Florestal, Minas Gerais, Brasil

© 2022 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

¹<https://octoverse.github.com/>

e métodos. Os resultados são discutidos na seção 5. Por fim, as considerações finais são apresentadas na seção 6.

2 COMUNICAÇÕES EM ISSUE TRACKING

Ao longo do desenvolvimento de software surgem diversas questões sobre melhorias e dúvidas apontadas pelos membros dos projetos. Essas questões ficam comumente concentradas em tópicos de discussão, representados por *issues* no sistema de *issue tracking* do GitHub. Segundo Bertran et al. [2], os repositórios de *issue tracking* são repositórios de conhecimento, que concentram grande parte da comunicação e da colaboração dos projetos, além de serem um canal de comunicação em si. Temos então que esses canais geram conhecimento útil e relevante no contexto para a gerência dos projetos.

Análises automatizadas sobre esses dados podem revelar indicadores de qualidade de processo importantes no que diz respeito à participação dos membros do projeto, como, por exemplo, desenvolvedores-chaves, conflitos nas relações, representatividade e participação de novatos e minorias. Para auxiliar nesse processo de análise existem algumas métricas usadas para tal apontamento, como a reputação dos usuários, a contabilização dos anos de participação do usuário na plataforma, e a qualidade dos comentários postados.

2.1 Issue Tracking no GitHub

Nos ambientes de *issue tracking*, as comunicações ocorrem à medida que *issues*, representando *bugs*, sugestões de melhoria e novos requisitos, são reportadas pelos membros dos projetos e/ou por colaboradores externos. Após reportadas, as *issues* podem ser respondidas por qualquer membro da equipe. Ao final da discussão em busca da solução, podem ser gerados novos *commits*, e consequentemente *pull requests* avaliados por gerentes de projeto, ou por algum responsável pelo repositório.

As *issues* são compostas por alguns campos obrigatórios, e outros, que dependendo do projeto não necessitam ser preenchidos. Os campos obrigatórios são: título e descrição. Já os campos não obrigatórios incluem as *labels*, como *bug* ou *question*, caso o projeto possua as mesmas definidas, e a quem aquela *issue* é atribuída. Por fim, tem-se os campos gerados automaticamente pela plataforma, como o autor, data de criação, e o estado, que ao ser criada será *open*, e ao ser fechado por algum membro do projeto será *closed*. Os campos relacionados a uma *issue* podem ser observados na Figura 1².

Os campos *título* e *descrição* contêm a principal fonte de dados textuais, uma vez que representam o conteúdo da discussão realizada na *issue* através dos comentários posteriormente postados, e, por conta disso, serão explorados no contexto deste trabalho.

2.2 Relevância Temática em Issue Tracking

O objetivo de utilizar a relevância temática no contexto do *issue tracking* é identificar se o texto do comentário em questão está relacionado ao contexto do tema proposto na *issue*, com o intuito de se avaliar o impacto do comentário na resolução da *issue*, além de demonstrar a importância de determinados desenvolvedores

na comunidade de um determinado projeto, também é possível investigar se existem correlações entre a relevância temática dos comentários e outros atributos ligados ao desenvolvedor, como gênero, reputação e o tempo de projeto.

A relevância temática é uma métrica capaz de representar o quanto um texto é relevante em relação ao tema de uma determinada discussão [8]. Essa métrica foi proposta por Azevedo [1] para se analisar a relevância de comentários postados em fóruns de discussão educacionais, e adaptada para o ambiente de *issue tracking* por Neto e Braga [11]. O cálculo envolve contabilizar o número de conceitos relevantes usados no texto do comentário que correspondem a conceitos do texto da problematização, que neste caso é dado pela descrição da *issue*, e seu título. Além de contabilizar os conceitos relevantes, o cálculo também considera a frequência dos mesmos no texto e os relacionamentos com outros termos. Assim, quanto mais conceitos relacionados, maior será a relevância do comentário. A Equação 1, adaptada por Neto e Braga [11], apresenta o cálculo da relevância temática.

$$RT = \max\{S_{CI}, S_{CD}, S_{CC}\} \quad (1)$$

Onde RT é a Relevância Temática; S_{CI} , a semelhança entre o comentário e a *issue*; S_{CD} , a semelhança entre o comentário e a discussão, que leva em consideração o título da *issue*, sua descrição e o comentário anterior; e S_{CC} , a semelhança entre o comentário e o comentário anterior, caso exista.

3 TRABALHOS RELACIONADOS

A baixa participação feminina em projetos de desenvolvimento de software tem sido reportada na literatura. Zacchiroli [18] analisou 1.6 milhões de *commits* num período de 50 anos e apresentou que cerca de 92% do código produzido nas plataformas de versionamento de código foi feito por homens. Izquierdo et al. [7] analisaram mais de 7000 perfis de usuários na plataforma do GitHub, em que apenas 7% dos membros do comitê de projeto e 8% dos líderes foram identificados como mulheres.

No que compete à análise dos dados de comunicações em *issue tracking*, diversos trabalhos na literatura têm usado técnicas de mineração de textos sobre o conteúdo textual das *issues*. Ortu et al. [12] analisou os dados de comunicação em projetos de software com o objetivo de avaliar o impacto de determinados fatores no tempo de resolução de uma *issue*. Neto and Silva [11] extraíram conhecimentos sobre os desenvolvedores do projeto, de forma a indicar desenvolvedores chaves no projeto, usando de métricas como a relevância temática dos comentários, o número de *issues* reportadas e o número de comentários postados nas *issues*. Apesar de trazerem investigações pertinentes às comunicações do desenvolvimento de software, os trabalhos apresentados não levam em consideração questões de gênero.

Considerando a participação feminina no universo *open source*, Singh [15] analisou 355 sites de projetos, e os resultados apontam que menos de 5% dessas comunidades possuíam espaço dedicados às mulheres. Neste trabalho, serão utilizados ambientes dedicados para mulheres para fins de comparações com ambientes com baixa diversidade, indicados pelo índice Blau [3], de forma a verificar se as mulheres são mais atuantes em termos de comunicações em ambientes dedicados para elas.

²Adaptado de <https://docs.github.com/pt/issues/tracking-your-work-with-issues/planning-and-tracking-work-for-your-team-or-project>



Figura 1: Exemplo de Issue no GitHub

Buscando apresentar novos indicadores sobre a baixa participação feminina no desenvolvimento de software, este trabalho visa analisar a quantidade e a qualidade do conteúdo das comunicações em *issue tracking*. Para isso, serão usadas as métricas de relevância temática dos comentários, quantidade de *issues* reportadas e comentários postados, reputação do desenvolvedor e os anos de participação na plataforma, sempre calculadas em relação ao gênero do desenvolvedor.

4 MATERIAIS E MÉTODOS

Para analisar os dados de comunicação em *issue tracking* em função do gênero do desenvolvedor, o estudo foi conduzido em 2 etapas: a) Definição das métricas; b) Extração dos dados, conforme detalhado nas próximas subseções.

4.1 Definição e Automatização das Métricas

Para a condução das análises, foram utilizadas cinco métricas gênero-dependentes, chamadas aqui de Métricas de Comunicação (MC).

- MC01. Quantidade de *Issues* Reportadas
- MC02. Quantidade de Comentários Postados
- MC03. Tempo de Participação na Plataforma
- MC04. Reputação dos Desenvolvedores
- MC05. Índice de Diversidade de Gênero das Equipes [3]
- MC06. Relevância Temática dos Comentários

Para obter as medidas para as métricas MC01 e MC04, foram utilizadas APIs e ferramentas da plataforma do GitHub. Já para as métricas MC05 e MC06, foram automatizadas em uma aplicação, desenvolvida no contexto desse trabalho e será explicada ao longo dessa seção.

As métricas MC01 e MC02 foram calculadas de acordo com o gênero do desenvolvedor, de forma que ao final da extração dos dados, usando do Jupyter Notebook³ sobre a base de dados salva em um CSV, em que os dados foram contabilizados através de filtros. Já a métrica MC03 foi calculada de acordo com a data fornecida

pela API GitHub, sendo subtraído o valor do ano de entrada do desenvolvedor pelo ano corrente.

Para a métrica MC04, visto que a plataforma do GitHub não realiza a contabilização automática da reputação dos desenvolvedores, foi utilizada a ferramenta Git Score⁴. Na Figura 2, pode-se observar o campo *reputation*, usado para avaliarmos a reputação dos usuários, sendo importante citar que a métrica leva em consideração para o cálculo da reputação a participação do usuário em projetos presentes na plataforma, a criação de repositórios e *issues*, além de *commits* feitos.



Figura 2: Exemplo de Exibição dos Dados GitScore

No que diz respeito à diversidade de gênero das equipes (MC05), foi usado o Índice Blau como métrica de diversidade [3]. A métrica de diversidade foi usada com o propósito de realizar comparações sobre os resultados das métricas aplicadas aos dados de comunicação para cada um dos segmentos de comunidades avaliadas. A Equação 2 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, consideramos 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.

³<https://jupyter.org/>

⁴<http://www.gitscore.com>

$$Blau = 1 - \sum_{i=1}^N p_i^2 \quad (2)$$

Ao final da extração dos dados de uma determinada comunidade, o nome e o gênero dos desenvolvedores participantes foi salvo em um arquivo CSV. A partir desse arquivo, usando o Jupyter Notebook, foi contabilizado o número de desenvolvedores nos gêneros feminino e masculino. Tais valores foram enviados a uma função simples na linguagem Python, que automatizou a métrica MC05, retornando assim o nível de diversidade daquela comunidade.

Visto que na plataforma de *issue tracking* do GitHub não é obrigatório a identificação do gênero dos desenvolvedores, e as métricas anteriormente apresentadas fazem comparações em relação ao gênero, fez-se necessário "adivinhar" tal informação. Zolduarrati [19] realizou um estudo para qualificar as melhores ferramentas adivinhadoras de gênero. As ferramentas apresentadas foram avaliadas e a ferramenta NamSor⁵ foi escolhida no contexto deste trabalho, devido à sua simplicidade de uso e qualidade dos resultados.

4.2 Relevância Temática dos Comentários

Para a avaliação da qualidade dos comentários, foi desenvolvida uma aplicação automatizada da métrica de relevância temática [11]. Para fins de simplificação e melhoria no desempenho, o cálculo automatizado da métrica foi adaptado no que diz respeito ao cálculo da similaridade entre os termos de um comentário e seu texto referência usando da técnica de cossenos. Na proposta de Neto e Braga e Silva [11], o cálculo da similaridade era realizado através de grafos, com uma dependência externa a uma ferramenta de geração dos mesmos. Essa dependência causava problemas de desempenho em virtude do alto número de requisições feitas à ferramenta externa. A alteração propiciou uma melhora significativa no desempenho da aplicação, sem provocar alterações em termos da qualidade dos resultados gerados. Além disso, a técnica de cossenos foi escolhida devido à simplicidade de aplicação e também por apresentar uma alta precisão em relação à classificação humana, como observado por Medeiros [9].

Ao usar a fórmula da Equação 1, apresentada na Seção 2.2, observou-se que a similaridade do comentário em relação ao seu comentário anterior, S_{CC} , assumia, em diversos momentos, o valor zero no cálculo da similaridade de cossenos. Depois de algumas verificações, foi possível notar que a similaridade do comentário em relação à discussão S_{CD} , e a similaridade do comentário em relação à *issue* S_{CI} , compunham uma relação com a relevância temática. Assim, de forma similar à equação apresentada por Azevedo [1], optou-se pela média aritmética das semelhanças entre o comentário e a *issue* S_{CI} , e entre o comentário e a discussão S_{CD} , como apresentado na Equação 3. Ressalta-se que a discussão corresponde a uma concatenação do título, da descrição e dos comentários anteriores da *issue*.

$$RT = \frac{S_{CI} + S_{CD}}{2} \quad (3)$$

4.2.1 Validação da Adaptação no Cálculo da Métrica. Para validar a alteração e garantir a confiabilidade dos resultados, foi realizado um procedimento de validação, comparando os resultados da versão

⁵<https://namsor.app/>

original [11], versão atualizada e avaliação manual de especialistas. Os grupo de especialistas contou com um total de 12 voluntários, sendo 3 engenheiros de software; 2 especialistas de domínio, ou seja, que participam da *issue* em questão; e 7 desenvolvedores, em vários níveis de conhecimento.

Para a avaliação de especialistas, foi elaborado um questionário que endereçava uma determinada *issue*, sendo solicitado que fosse atribuída uma relevância de 0 a 5 para cada comentário daquela *issue*. Ao todo foram analisados 59 comentários, distribuídos em 12 *issues*. Cada um dos comentários foi analisado por dois especialistas e foi feita uma média com as relevâncias atribuídas pela equipe de desenvolvimento, composta pelo(as) autores(as). A composição da média da relevância dos especialistas com a média da equipe de desenvolvimento foi utilizada como valor real y , enquanto os valores calculados pela abordagem original, e pela abordagem adaptada foram usados como \hat{y} . Esses valores foram usados no cálculo de Erro Absoluto Médio - MAE (Equação 4), que representa a média da diferença absoluta entre os valores reais e os previstos no conjunto de dados.

$$MAE = \frac{1}{N} \sum_{i=1}^N \|y - \hat{y}\| \quad (4)$$

Além do MAE, também foi calculado o Erro Quadrático Médio (MSE), que representa a média quadrada entre os valores reais e os previstos no conjunto de dados, como apresentado na Equação 5.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y - \hat{y})^2 \quad (5)$$

Ao final, observou-se uma melhora nos valores de MSE e MAE, para a versão adaptada, em comparação com a abordagem original [11]. A porcentagem de erro da abordagem original ficou em torno de 21% no MAE e 7% no MSE, enquanto na abordagem adaptada, houve redução do valor de MAE, para 8% e de MSE, para 1%.

4.3 Seleção e Extração dos Dados

Na seleção das bases de dados, foram considerados projetos de comunidades abertas, e comunidades dedicadas à mulheres, de forma a se comparar a participação feminina nesses dois contextos. As comunidades dedicadas foram selecionadas através de uma pesquisa preliminar, em que após a escolha de comunidades abertas, buscou-se comunidades no mesmo segmento que trabalhassem a inclusão de mulheres.

Os dados foram extraídos em janeiro de 2022, por meio de uma aplicação na linguagem Python, utilizando a API Rest do GitHub⁶.

Para a etapa de extração dos dados, foram utilizados filtros baseados no trabalho de Neto et al. [10], divididos em duas categorias: Filtros de Projetos (FPs) e Filtros de *Issues* (FIs). Para selecionar os projetos de interesse, foram definidos cinco filtros FPs:

- **FP1.** Possuir, no mínimo, 5 membros.
- **FP2.** Possuir, no mínimo, 5 *commits*.
- **FP3.** Possuir, no mínimo, 5 *issues* abertas.
- **FP4.** Possuir, no mínimo, 5 *issues* fechadas.
- **FP5.** Ter sido criado a pelo menos 6 meses.

A seleção das *issues* abrangeu a aplicação dos seguinte filtros FIs:

⁶<https://docs.github.com/pt/rest>

- **FI1.** Possuir no mínimo, 5 comentários.
- **FI2.** As comentários não podem conter apenas trechos de códigos, sendo assim necessário textos.
- **FI3.** Possuir um tempo de abertura mínimo de uma semana.
- **FI4.** *Issues* não podem ter sido reabertas.

Ressalta-se que o filtro FP4 foi definido com base no fator de que, as *issues* avaliadas no trabalho estavam com seu *status closed*, de forma que pudéssemos garantir que ao longo da extração dos dados, e da pesquisa não fossem inseridos mais comentários, unido a isso temos a definição do filtro FI4, para trazer mais uma garantia. Já o filtro FP5 foi escolhido de forma a garantir que o projeto estivesse plenamente em funcionamento, e executando as tarefas pertinentes ao sistema de *issue tracking*, unido aos fatores apresentados nos filtros FP2, FP3 e FP4. O filtro FI2 foi inserido com base no fator de que a avaliação dos comentários é feita a partir de seus textos, ou seja, os trechos de código não são inseridos na avaliação da relevância temática, fazendo com que comentários com apenas trechos de código se tornem irrelevantes. Além disso, é importante ressaltar que o filtro FI2 leva em consideração todos os comentários de uma determinada *issue*, de forma que, caso tenhamos um comentário com apenas trechos de códigos, aquela *issue* será desconsiderada.

Ao final, esses dados foram armazenados em um arquivo CSV, que posteriormente foram analisados através do Jupyter Notebook.

5 RESULTADOS E DISCUSSÕES

Após a extração dos dados com aplicação dos filtros, foram obtidos 9151 comentários, de 1275 *issues*, presentes em 28 repositórios de 10 comunidades, sendo 5 comunidades abertas e 5 dedicadas às mulheres. Na Tabela 1, tem-se um sumário dos dados obtidos para cada comunidade, incluindo o número de repositórios, *issues* e comentários. Ressalta-se que as cinco primeiras comunidades são dedicadas às mulheres.

Tabela 1: Visão Geral dos Dados Extraídos por Comunidades

| Comunidade | #Repositórios | #Issues | #Comentários |
|---------------|---------------|---------|--------------|
| RailsGirls | 2 | 5 | 148 |
| PyLadies | 4 | 41 | 360 |
| WomenWhoCode | 3 | 22 | 168 |
| WoMozBrasil | 1 | 5 | 26 |
| DjangoGirls | 2 | 145 | 1014 |
| Django | 6 | 239 | 1985 |
| Ruby on Rails | 4 | 426 | 3079 |
| Mozilla | 2 | 81 | 423 |
| Flutter | 1 | 214 | 1333 |
| Python | 3 | 97 | 644 |

Para cada uma das comunidades, foi calculado o quantitativo de mulheres e homens participantes e também o índice Blau, com o intuito de se conhecer o fator de diversidade em cada uma delas. Os resultados são apresentados na Figura 3, em que é possível observar as discrepâncias em termos de representatividade feminina nas comunidades abertas e, por consequência, o baixo índice de diversidade de gênero.

Conforme esperado, os dados coletados evidenciam baixos índices de diversidade em comunidades abertas e índices altos em

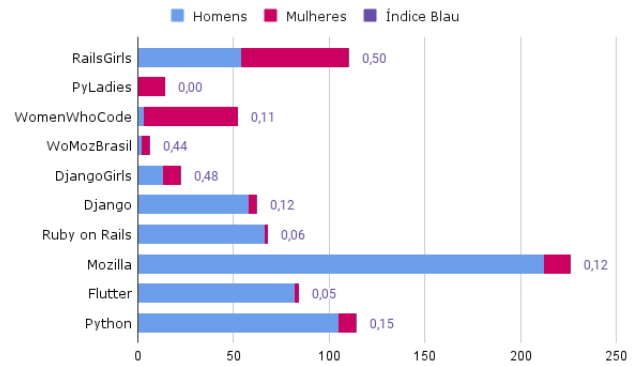


Figura 3: Comunidades, Índice Blau e Quantitativos por Gênero

comunidades dedicadas, já que o maior índice chegou a 0.5, indicando o equilíbrio entre as categorias, na RailsGirls. Importante também destacar que na comunidade PyLadies, apesar do índice Blau ser igual a 0.0 indicar a falta de diversidade, isso ocorre porque a comunidade é fechada à participação masculina, com o intuito de se criar um ambiente seguro para que as mulheres possam desenvolver as habilidades, tanto técnicas, quanto teóricas. Ao se observar a Tabela 1 e a Figura 3, comunidades como o Ruby on Rails tem um alto número de comentários postados, porém, seu nível de diversidade é baixo. Importante também ressaltar, que após uma análise nos dados, verificando o nome e sobrenome dos desenvolvedores da comunidade, foi possível notar que muitas mulheres que participam vieram da comunidade dedicada, RailsGirls, ou seja, já são mulheres mais seguras e empoderadas.

Os demais resultados obtidos com a aplicação das métricas MC01 a MC05, voltadas às questões de pesquisa endereçadas neste estudo, encontram-se descritos nas próximas subseções.

5.1 QP1. Qual a diferença de participação de homens e mulheres em termos de *issues* e comentários associados?

Olhando primeiramente para as *issues*, temos foram extraídas 1275, em que 1071 (84%) foram reportadas por homens, e 204 (16%) por mulheres, como apresentado na Figura 4.

Considerando o recorte das comunidades dedicadas, tem-se 218 *issues* reportadas, das quais 102 (48%) foram reportadas por mulheres. Já em comunidades abertas, foram reportadas 1057 *issues*, sendo que apenas 102 (9%) foram reportadas por mulheres. Tais valores mostram que de acordo com o aumento de mulheres nas comunidades, a porcentagem de participação das mesmas aumenta. Ao contrário dos homens, que possuem valores altos de porcentagens em ambos os tipos de comunidades.

No caso dos comentários, temos que em sua grande maioria, o gênero dos autores foi obtido automaticamente. Em alguns casos foi necessária a intervenção manual, em que apenas um desenvolvedor permaneceu com o gênero indefinido. Além disso, tem-se que alguns comentários foram postados por *bots*, não sendo, assim,

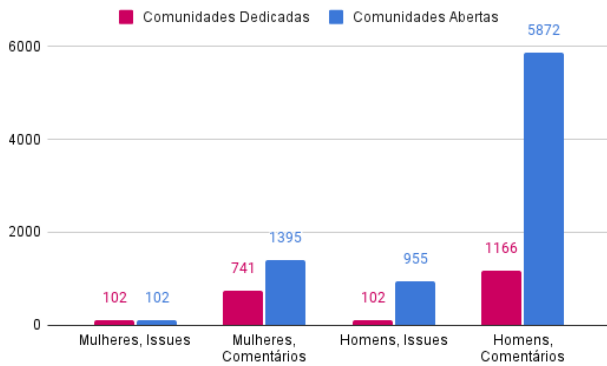


Figura 4: Issues e Comentários em Relação ao Gênero

considerados na Figura 4. A partir disso, dos 9151 comentários avaliados, 6897 foram postados por homens, 2059 (22%) por mulheres e 194 por bots.

No contexto das comunidades dedicadas, dos 1923 comentários, 741 (39%) foram postados por mulheres. E no caso das comunidades abertas, foram postados 7447 comentários, em que apenas 1395 (19%) foram postados por mulheres, o que também evidencia a baixa participação em comparação com os homens neste segmento. Observar-se então que os homens se sentem mais confortáveis para comentar do que as mulheres, até mesmo em comunidades dedicadas.

5.2 QP2. Existe diferença entre a relevância dos comentários postados por homens e mulheres?

Após o cálculo da relevância temática para cada comentário analisado, os resultados são similares na média, sendo 0.03680, para os homens e 0.03394, para as mulheres. Conforme mostrado no gráfico das Figuras 5 e 6, observa-se que o intervalo de dados altera entre os gêneros, segundo o contexto do segmento em análise. Importante também destacar que, nas comunidades dedicadas (Figura 5), diferente das abertas (Figura 6), as mulheres alcançam valores de relevância maiores que os homens.

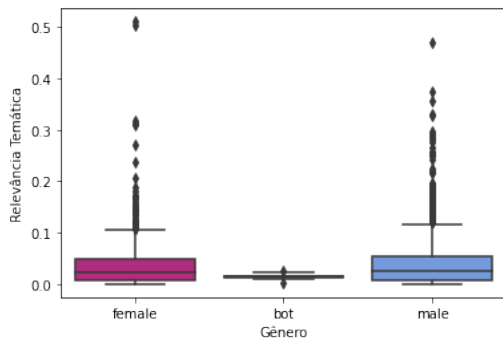


Figura 5: Intervalo de Relevância Temática nas Comunidades Dedicadas

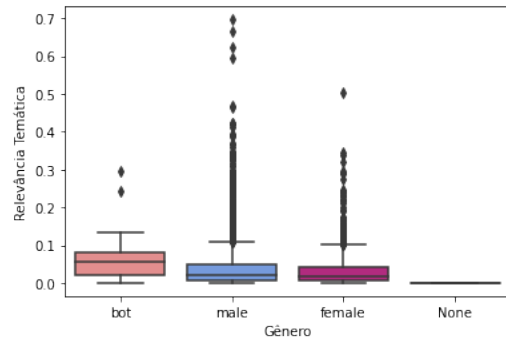


Figura 6: Intervalo de Relevância Temática nas Comunidades Abertas

5.3 QP3. Existe relação entre a relevância dos comentários e a reputação do autor?

Para os dados considerados, obteve-se uma reputação dos desenvolvedores variando entre 0 e 16000. Esses dados foram cruzados com os dados de relevância temática dos comentários postados por esses desenvolvedores, conforme mostrado nos gráficos das Figuras 7 e 8. A partir da análise dos gráficos, não é possível notar uma relação entre os atributos, uma vez que a distribuição dos valores da relevância temática se mostram similares em todos os valores de reputação analisados. Para comprovar esse comportamento nos dados, foi calculado o coeficiente de correlação de Pearson entre os atributos de relevância temática e a reputação dos autores. A partir do coeficiente obtido, -0.045, constatou-se que os atributos não estão correlacionados.

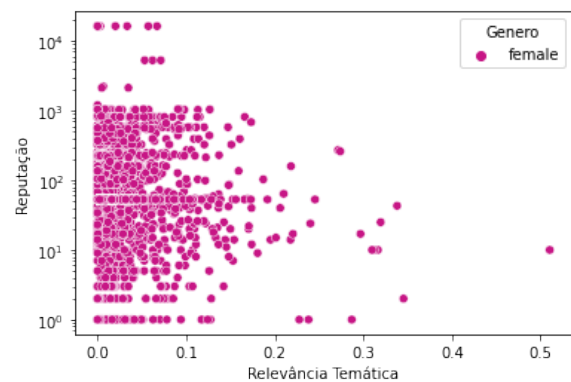


Figura 7: Relevância Temática e Reputação para Mulheres

Nas comunidades dedicadas às mulheres, como apresentado na Figura 9, observam-se alterações nos valores de reputação, com o maior valor presente na faixa de 10^3 , enquanto na base de dados como um todo está em 10^4 . Além disso, é importante ressaltar que o maior valor de reputação é do gênero masculino, e as reputações das mulheres estão em torno de 10^2 e 10^3 .

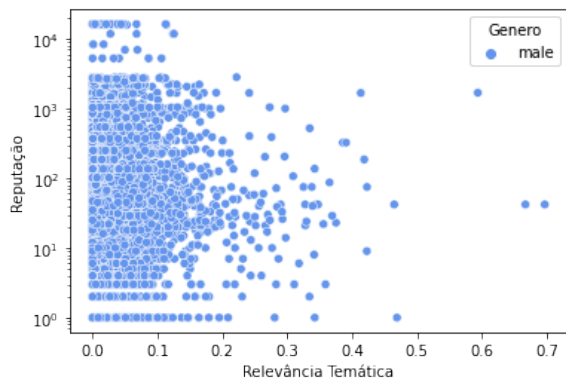


Figura 8: Relevância Temática e Reputação para Homens

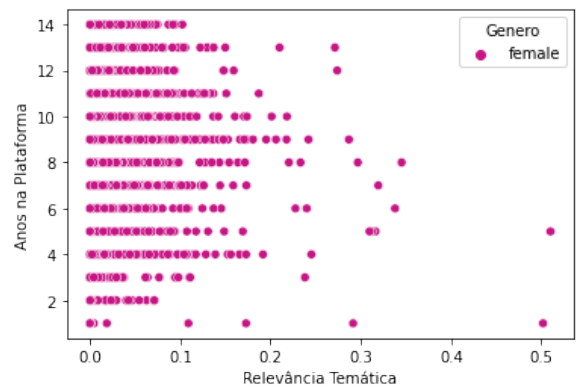


Figura 10: Relevância Temática e Anos na Plataforma para Mulheres

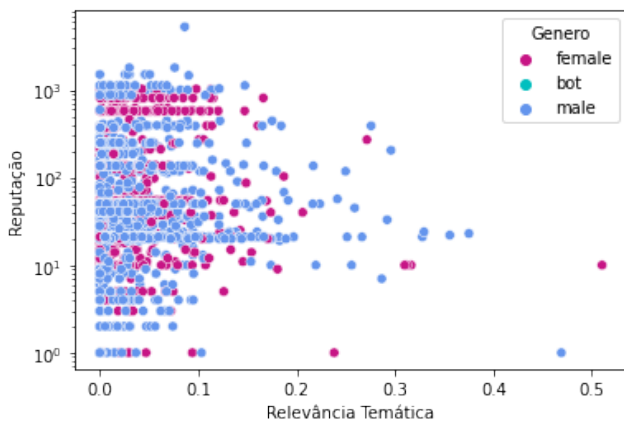


Figura 9: Relevância Temática e Reputação em Comunidades Dedicadas

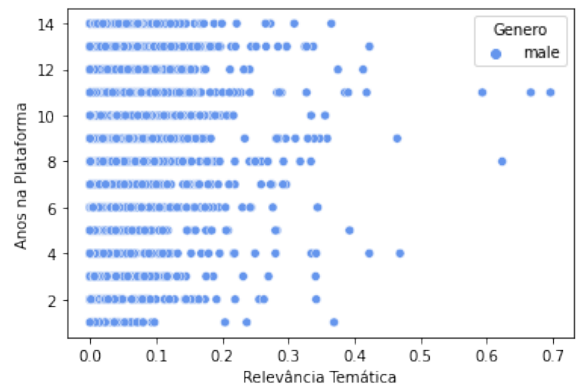


Figura 11: Relevância Temática e Anos na Plataforma para Homens

5.4 QP4. Existe relação entre a relevância dos comentários e o tempo de participação na plataforma?

As Figuras 10 e 11 apresentam a relação entre a relevância dos comentários e o tempo de participação na plataforma por gênero. De forma similar à QP3, os atributos não parecem estar relacionados, uma vez que a distribuição dos valores de relevância temática se encontra similar em todos os anos apresentados. Para confirmar essa análise, foi realizado o cálculo da correlação de Pearson entre os atributos de relevância temática e os anos de participação da plataforma, sendo obtido o valor de -0.052 , que confirmou a falta de relação entre os atributos.

Um fator importante a se citar sobre a Figura 10 é que ao passar dos anos apresentados, o número de mulheres diminui, comprovando assim questões já apontadas na literatura [17]. Os dados mostram uma maior presença feminina no quarto ano de participação, e uma tendência de baixa a partir do décimo primeiro ano.

Levando em consideração que as mulheres possuem um menor tempo de permanência nos projetos, tem-se um impacto nos valores de sua reputação, isso porque quanto menos tempo de participação,

menor será o número de comentários e *commits* realizados. Essa questão pode ser verificada na Figura 7, de forma que em reputações muito altas, acima de 10^3 , temos uma maioria de homens, e uma maioria feminina em valores igual ou abaixo de 10^2 .

Ao observar-se o recorte das comunidades dedicadas, é possível verificar que o intervalo dos dados varia, como mostrado na Figura 12. Observa-se que a participação na plataforma tem início no segundo ano, com maioria masculina, o que apresentam indicadores sobre um tempo de espera para que mulheres comecem a participar de comunidades *open source*, mesmo que dedicadas. As mulheres começam a participar mais ativamente, a partir do quinto ano, mas o tempo de permanência é menor que o dos homens. Também se observa uma queda da participação feminina, a partir do nono ano.

6 CONSIDERAÇÕES FINAIS

Este trabalho apresentou os resultados de uma análise de dados de comunicação em *issue tracking*, observou-se que, na média, as mulheres postam comentários com relevância similar aqueles postados dos homens. O estudo também reforçou índices discrepantes de participação feminina nas comunidades *open source* confirmando-se também no contexto de *issue tracking*, onde apenas 22% dos

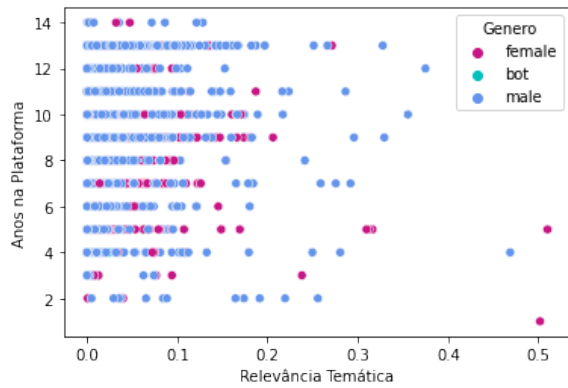


Figura 12: Relevância Temática e Anos na Plataforma em Comunidades Dedicadas à Mulheres

comentários foram postados por mulheres e 16% das *issues* tem autoria feminina. Além disso, olhando para os recortes das comunidades, podemos verificar que as mulheres têm sido mais atuantes em comunidades dedicadas à elas, chegando a 48% das *issues* reportadas. Porém, ao observarmos comunidades abertas elas tem se mostrando com uma baixa participação e representatividade, tendo apenas 9% das *issues* reportadas. Outro ponto importante explorado pelo trabalho é em relação à reputação e ao tempo de participação das mulheres na plataforma, em que elas geralmente estão em um intervalo de 4 anos a 10 anos, enquanto os homens tem seu intervalo variando desde meses até 14 anos.

Como trabalhos futuros, levando-se em consideração questões técnicas sobre a análise de dados, pode-se realizar o cruzamento entre os dados da relevância temática com outros indicadores de participação no desenvolvimento de software, como o número de *commits* realizados, *pull requests* aceitas. Pode-se avaliar se a participação feminina no âmbito das comunicações nos projetos causa algum impacto em termos do tempo de resolução ou fechamento de *issues*. Por fim, com o intuito de complementar a análise quantitativa dos dados, 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.

REFERÊNCIAS

- [1] Breno Fabrício Terra Azevedo. 2011. *Minerafórum : um recurso de apoio para análise qualitativa em fóruns de discussão*. Tese de Doutorado em Informática na Computação. Universidade Federal do Rio Grande do Sul.
- [2] Dane Bertram, Amy Volda, Saul Greenberg, and Robert Walker. 2010. Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (Savannah, Georgia, USA) (CSCW '10)*. ACM, New York, NY, USA, 291–300. <https://doi.org/10.1145/1718918.1718972>
- [3] Torsten Biemann and Eric Kearney. 2010. Size does matter: How varying group sizes in a sample affect the most common measures of group diversity. *Organizational Research Methods* 13 (7 2010), 582–599. Issue 3. <https://doi.org/10.1177/1094428109338875>
- [4] E.D. Canedo, Heloíse Tives, Madianita Bogo, Fabiano Fagundes, and José Siqueira de Cerqueira. 2019. Barriers Faced by Women in Software Development Projects. *Information* 10 (10 2019), 309. <https://doi.org/10.3390/info10100309>
- [5] Gemma Catolino, Fabio Palomba, Damian A. Tamburri, Alexander Serebrenik, and Filomena Ferrucci. 2019. Gender Diversity and Women in Software Teams: How Do They Affect Community Smells?. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, Montreal, QC, Canada, 11–20. <https://doi.org/10.1109/ICSE-SEIS.2019.00010>
- [6] Georgios Gousios and Diomidis Spinellis. 2017. Mining Software Engineering Data from GitHub. In *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*. IEEE, Buenos Aires, Argentina, 501–502. <https://doi.org/10.1109/ICSE-C.2017.164>
- [7] Daniel Izquierdo, Nicole Huesman, Alexander Serebrenik, and Gregorio Robles. 2019. OpenStack Gender Diversity Report. *IEEE Software* 36 (1 2019), 28–33. Issue 1. <https://doi.org/10.1109/MS.2018.2874322>
- [8] Crystiano José Richard MacHado, Alexandre Magno Andrade MacIel, Rodrigo Lins Rodrigues, and Ronaldo Menezes. 2019. An approach for thematic relevance analysis applied to textual contributions in discussion forums. *International Journal of Distance Education Technologies* 17 (2019), 37–51. Issue 3. <https://doi.org/10.4018/IJDET.2019070103>
- [9] Danielle C. Medeiros, José Eustáquio Rangel de Queiroz, and Joseana M. F. R. Araújo. 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)*. SBC, Dourados, Mato Grosso do Sul, Brazil, 144. <https://doi.org/10.5753/cbie.sbie.2014.144>
- [10] Luiz Neto, Gláucia Silva, and Giovanni Comarella. 2021. Estimativa Do Tempo de Resolução de Issues No GitHub Usando Atributos Textuais e Temporais. In *Brazilian Symposium on Software Engineering*. Association for Computing Machinery, New York, NY, USA, 253–262. <https://doi.org/10.1145/3474624.3474647>
- [11] Luiz Eugênio Coelho Neto and Gláucia Braga e Silva. 2018. ColMiner: A Tool to Support Communications Management in an Issue Tracking Environment. In *Proceedings of the XIV Brazilian Symposium on Information Systems (Caxias do Sul, Brazil) (SBSI'18)*. Association for Computing Machinery, New York, NY, USA, Article 50, 8 pages. <https://doi.org/10.1145/3229345.3229398>
- [12] Marco Ortu, Tracy Hall, Michele Marchesi, Roberto Tonelli, David Bowes, and Giuseppe Destefanis. 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 (Oulu, Finland) (PROMISE'18)*. Association for Computing Machinery, New York, NY, USA, 70–79. <https://doi.org/10.1145/3273934.3273943>
- [13] Gema Rodriguez, Reza Nadri, and Meiyappan Nagappan. 2021. Perceived diversity in software engineering: a systematic literature review. *Empirical Software Engineering* 26 (09 2021). <https://doi.org/10.1007/s10664-021-09992-2>
- [14] Samaneh Saadat, Olivia B. Newton, Gita Sukthankar, and Stephen M. Fiore. 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)*. IEEE, Melbourne, Australia, 169–176. <https://doi.org/10.1109/WIAT50758.2020.00027>
- [15] Vandana Singh. 2019. Women-Only Spaces of Open Source. In *2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE)*. IEEE, Montreal, QC, Canada, 17–20. <https://doi.org/10.1109/GE.2019.00010>
- [16] Bogdan Vasilescu, Alexander Serebrenik, and Vladimir Filkov. 2015. A Data Set for Social Diversity Studies of GitHub Teams. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*. IEEE, Florence, Italy, 514–517. <https://doi.org/10.1109/MSR.2015.77>
- [17] Balazs Vedres and Orsolya Vasarhelyi. 2019. Gendered behavior as a disadvantage in open source software development. *EPJ Data Science* 8 (12 2019). Issue 1. <https://doi.org/10.1140/epjds/s13688-019-0202-z>
- [18] Stefano Zacchiroli. 2021. Gender Differences in Public Code Contributions: A 50-Year Perspective. *IEEE Software* 38, 2 (2021), 45–50. <https://doi.org/10.1109/MS.2020.3038765>
- [19] Elijah Zolduoarrati and Sherlock A. Licorish. 2021. On the value of encouraging gender tolerance and inclusiveness in software engineering communities. *Information and Software Technology* 139 (11 2021), 106667. <https://doi.org/10.1016/j.infsof.2021.106667>