Análise de Sentimentos em Discussões de Issues Reabertas do Github

Gláucya Boechat, Joselito Mota Jr, Ivan Machado, Manoel Mendonça

[1]Universidade Federal da Bahia (UFBA)
Campus Ondina – 40.170-110 – Salvador – BA – Brazil

{glaucya.boechat, ivan.machado, manoel.mendonca}@ufba.br,

joselitomota@dcc.ufba.br

Abstract. *The behavior of reopened issues is a perception to be studied to analyze the impact of discussions on the continuity of software project maintenance. Sentiment analysis is presented as a powerful technique to assist such analysis. In this study, we analyzed 12,996 reopened issues, which contained discussions, from 80 Github projects. Based on the analysis of such historical data, we seek to analyze whether a closed issue tends to be reopened from the sentiment analysis of this issue's discussions. The analyzes are performed through the degree of sentiment of the texts of the comments of the issues. The SentiStrength tool, supported by Software Engineering lexicons, were used to classify the degree of polarity of the texts found. The study identified that the polarity of feelings in discussions can directly affect the issue's life cycle, including support for the prediction about reopening issues.*

Resumo. *O comportamento de issues reabertas é uma percepção a ser estudada para analisar o impacto das discussões na continuidade da manutenção de projetos de software. A análise de sentimentos apresenta-se como uma poderosa técnica para auxiliar tal análise. Neste estudo, analisamos 12.996 issues reabertas, contendo discussões, de 80 projetos do Github. Com base na análise dessa massa de dados históricos, buscamos analisar se uma issue fechada tende a ser reaberta a partir da análise de sentimentos das discussões dessa issue. As análises são realizadas através do grau de sentimento dos textos dos comentários das issues. A ferramenta SentiStrength, com suporte aos léxicos da área de Engenharia de Software, foi utilizada para classificar o grau de polaridade dos textos encontrados. O estudo identificou que a polaridade dos sentimentos nas discussões pode afetar diretamente o ciclo de vida da issue, inclusive com suporte à predição sobre a reabertura das issues.*

1. Introdução

Uma atividade fundamental da fase de manutenção do software é compreender porque *issues* fechadas são reabertas [Caglayan et al. 2012]. As *issues* podem ser reabertas depois de serem fechadas devido ao fechamento incorreto, descoberta do problema real da *issue*, etc. As *issues* reabertas podem aumentar o custo de manutenção, degradar a qualidade geral do produto de software, reduzir a confiança dos usuários e trazer trabalho desnecessário para os desenvolvedores [Pan and Mao 2014].

Nesse trabalho, optamos por investigar o comportamento das *issues* e *pull requests* que foram reabertas no GitHub. O GitHub é uma plataforma de hospedagem de código-fonte do sistema de controle de versão git, que contém mais de 36 milhões de usuários e 100 milhões de repositórios [1].

No GitHub, os colaboradores dos repositórios, opcionalmente, podem criar e participar de discussões de *issues* para coletar feedback dos usuários nas discussões sobre o projeto, adição de novas features, bugs e outras tarefas de manutenção. Eles podem ainda participar de discussões de *pull request*, que é um tipo de *issue*, para discutir e revisar alterações realizadas no código-fonte antes de serem incorporadas na *branch* principal [Ortu et al. 2016]. Os comentários dessas discussões das *issues* postados pelos colaboradores não contêm apenas informações técnicas, mas também informações valiosas sobre sentimentos ou emoções [Ortu et al. 2015]. Os sentimentos podem ser classificados como positivo, negativo ou neutro, que estão associados as emoções como felicidade, tristeza, alegria, raiva, dentre outras [Liu 2015]. *A investigação sobre os sentimentos contidos nas issues pode trazer informações ou indicativos que auxiliam no gerenciamento desses repositórios, por exemplo ...* .

O objetivo do trabalho consiste em investigar os sentimentos dos colaboradores durante as discussões das *issues* que ajudam a prever se uma *issue* fechada tem propensão de ser reaberta. As seguintes questões de pesquisa (QP) direcionam esta investigação:

QP1 Existe algum indicativo de que uma *issue* não será reaberta se ela for fechada com um sentimento positivo?

QP2 É possível predizer se uma *issue* será reaberta quando o número de sentimentos negativos adicionados ao número de sentimentos neutros for maior que o número de sentimentos positivos presentes nas suas discussões?

QP3 Uma *issue* com discussões com sentimentos neutros após o fechamento indica que ela será reaberta?

2. Metodologia

Inicialmente, foram selecionados os 90 repositórios listados no desafio *MSR Challenge Dataset* da conferênca *Mining Software Repositories* (MSR)[2] do ano de 2014. Ao validalos, observamos que oito repositórios não possuiam *issues* reabertas com discussões, e dois repositórios não mais encontravam-se disponíveis. Assim, a lista final incluiu os 80 repositórios disponíveis, que englobavam 506.227 *issues* abertas e fechadas, com uma média de aproximadamente 379 mil linhas de código (*LOC*), e média de 2.769 arquivos por repositório. As *issues* foram extraídas entre 4 de Junho à 24 Julho de 2019. Selecionamos apenas as *issues* que foram reabertas e que possuíam discussões, resultando em 12.996 issues no *dataset*. A Figura 1 apresenta um *workflow* com as etapas seguidas neste trabalho. Cada uma das etapas será discutida a seguir.

Os repositórios selecionados foram implementados nas seguintes linguagens: 8 em C, 8 em C++, 8 em C#, 1 em CSS, 3 em HTML, 7 em Java, 8 em JavaScript, 3 em Markdown, 5 em PHP, 10 em Python, 4 em R, 8 em Ruby e 7 em Scala.

[1] Dados de https://github.com/about

[2] http://ghtorrent.org/msr14.html

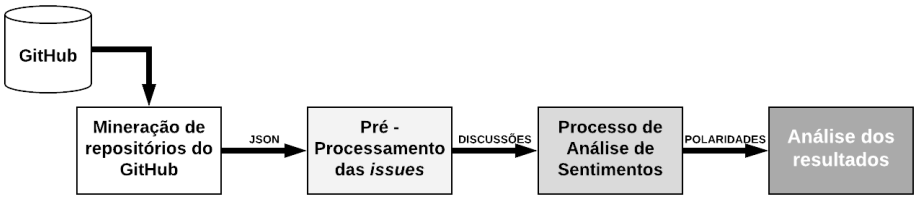

Figura 1. *Workflow* de captura e análise de dados.

Todos os dados dos repositórios utilizados neste projeto encontram-se disponíveis em [Boechat et al. 2019].

2.1. Etapa 1 - Mineração de Repositórios do Github

Os dados foram obtidos a partir de um script de mineração desenvolvido utilizando a biblioteca PyGitHub[3] para extração dos dados. Este script é reponsável por recuperar todas as informações referentes as *issues*: comentários, horário de criação, usuários, reações e outras informações. Processamos todas as respostas das requisições ao GitHub no formato *JavaScript Object Notation* (JSON) e em seguida armazenamos todas as informações em um banco de dados não relacional, MongoDB[4]. O script encontra-se disponível no GitHub[5]. A mineração das *issues* foi realizada em uma estação de trabalho com um processador Core i7-7500U, 8 GB RAM, SSD 240 GB, Sistema Operacional Win 10 x64.

2.2. Etapa 2 - Pré-processamento das *issues*

A etapa de pré-processamento dos dados foi realizada através da limpeza dos textos do título, descrição e comentários das issues. A limpeza dos textos foi feita através do módulo RE[6] do Python, com operações com expressões regulares para excluir trechos indesejados no texto, tais como URLs, código-fonte, trechos de código, erros de compilação, classes, interfaces, imagens, quebra de linhas, excesso de espaços em branco, respostas de comentário, frases de alertas sobre *warning* e exceções.

O pré-processamento considerou o truncamento de palavras, por exemplo *joy** para todas as palavras que começam com *joy*, ao invés de utilizar os processos de lematização e stemização. Não utilizamos a remoção de *stopwords*, pois pode ocorrer de remover palavras que distorcem o verdadeiro sentimento da frase, por exemplo os advérbios de negação ou intensidade podem alterar o sentido da frase [Thelwall et al. 2010].

2.3. Etapa 3 - Processo de Análise de Sentimentos

Durante o processo de análise de sentimentos das discussões das *issues* reabertas, e que tenham pelo menos dois comentários, foi utilizada a versão Java da ferramenta SentiStrength [Thelwall et al. 2010] para classificar as polaridades dos textos. A polaridade pode ser *negativa*, *neutra* ou *positiva*. A SentiStrength é uma ferramenta de análise de sentimentos baseada em um dicionário léxico; esse dicionário é um arquivo léxico de

[3]https://pygithub.readthedocs.io
[4]https://www.mongodb.com/
[5]https://github.com/joselitojunior94/gfetcher
[6]https://docs.python.org/3/library/re.html

emoções, onde palavras com sentimentos negativos estão previamente classificadas com pesos entre -5 e -1, e palavras com sentimentos positivos possuem pesos entre +1 e +5. Utilizamos o dicionário léxico da ferramenta SentiStrength-SE[7] [Islam and Zibran 2017], uma versão desenvolvida para aplicar análise de sentimentos no domínio de Engenharia de Software. A ferramenta SentiStrength divide a sentença em *tokens* e para cada palavra (*token*) que transmite uma emoção é atribuída uma pontuação. Após pontuar todas as palavras, a ferramenta retorna a pontuação máxima dos sentimentos negativos e a pontuação máxima dos sentimentos positivos. O sentimento final do texto é obtido através da soma da pontuação positiva com a pontuação negativa. Para valores menores que zero, o texto é classificado como negativo; para valores iguais a zero, o texto é classificado como neutro; e para valores maiores que zero, o texto é classificado como positivo.

2.4. Etapa 4 - Análise dos Resultados

A ordem cronológica dos eventos de uma *issue* foram consideradas na análise dos resultados. Os eventos são apresentados no diagrama de estado da Figura 2, que exibe os possíveis status da *issue*: "ABERTA" ("OPEN") e "FECHADA" ("CLOSED"), e as transições responsáveis pelas mudanças de status ("FECHAR" e "REABRIR").

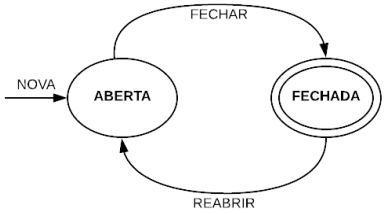

Figura 2. Ciclo de Vida da *issue*.

Durante o ciclo de vida de uma *issue*, os colaboradores do repositório e/ou usuários do GitHub podem colabarorar com o repositório através de comentários nas *issues*. Nos comentários, os colaboradores podem expressar seus sentimentos por meio de textos, emoticons e emojis. Neste estudo, foram analisadas as polaridades dos sentimentos encontradas nos comentários no período de tempo entre os status possíveis da *issue*. A Figura 3 representa o ciclo de vida de uma *issue* reaberta com discussões. Ela apresenta o início da *issue*, no status "Aberta", as discussões dos colaboradores entre a abertura e o fechamento, a mudança de status para "Fechada", as discussões dos colaboradores entre o fechamento e a reabertura da *issue*, a mudança de status para "Reaberta", as discussões dos colaboradores entre a reabertura e o fechamento da *issue*, e por fim a mudança de status para "Fechada".

3. Discussão dos Resultados

Foram analisadas 12.996 *issues* reabertas, que continham discussões, de 80 repositórios do GitHub. A seguir, discutiremos os resultados à luz das QP propostas para este estudo.

QP1 - Existe algum indicativo de que uma *issue* não será reaberta se ela for fechada com um sentimento positivo?

Para calcular os resultados, verificou-se o último sentimento antes do fechamento da *issue*. Assim, caso a *issue* possua pelo menos um fechamento, então é contabilizado.

[7]https://laser.cs.uno.edu/Projects/Projects.html

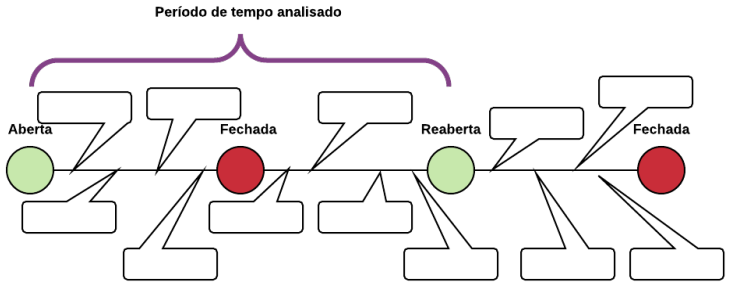

Figura 3. Linha do tempo dos eventos da *issue*.

O ponto central da análise de dados foi encontrar pelo menos um caso durante o tempo de vida de uma *issue* onde existiu fechamento com sentimento positivo, e reabertura logo em seguida. Verificamos que 2.925 (22,51%) *issues* reabertas foram fechadas com sentimento positivo e em seguida foram reabertas. A Figura 4 apresenta a distribuição de issues fechadas com sentimentos positivos com mediana de 14 *issues*. Isso indica que se a *issue* fechar com sentimento positivo, não há garantia de que ela continuará fechada. Entretanto, observa-se uma menor tendência de reaberta dessas *issues*. 10.071 (77,49%) *issues* foram fechadas com sentimento negativo ou neutro, e em seguida foram reabertas.

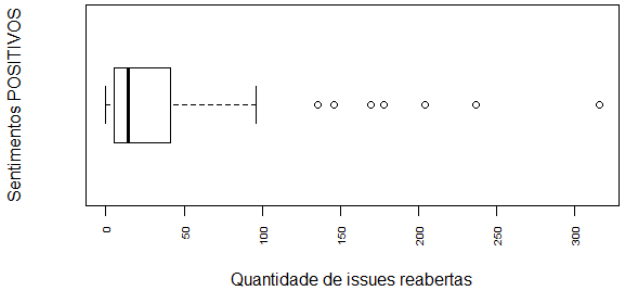

Figura 4. Distribuição de *issues* fechadas com sentimentos positivos.

Dentre os repositórios analisados, apenas o `Impress.js` teve 13 de 18 (72.22%) *issues* reabertas que foram fechadas com sentimento positivo e em seguida foram reabertas. Em contraponto, nenhum caso foi observado nos repositórios `ActionBarSherlock`, `Kestrel` e `Storm` [Boechat et al. 2019].

QP2 - É possível predizer se uma *issue* será reaberta quando o número de sentimentos negativos adicionados ao número de sentimentos neutros for maior que o número de sentimentos positivos presentes nas suas discussões?

Verificamos que 11.722 (90.20%) *issues* foram reabertas com a quantidade de sentimentos negativos ou neutros maior que a quantidade de sentimentos positivos. Encontramos 810 (6,23%) *issues* reabertas com quantidade de sentimentos negativos ou neutros igual a quantidade de sentimentos positivos. Em 464 (3,57%) *issues*, a quantidade de sentimentos positivos foi maior que a quantidade de sentimentos negativos ou neutros. A Figura 5 apresenta a distribuição de *issues* reabertas com mediana de de 57 *issues* para (Neutros+Negativos) > Positivos, em função da polaridade de sentimentos. Nos repositórios `ActionBarSherlock`, `Storm`, `Beanstalkd`, `Kestrel`, `Flockdb`, `Ravendb` e `CCV`, 100% das *issues* foram reabertas com sentimentos negativos ou neutros

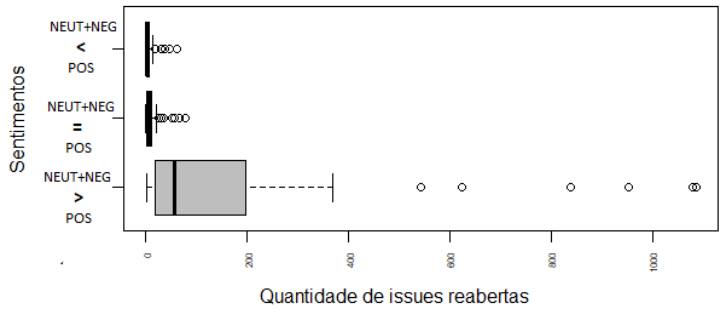

Figura 5. Distribuição de *issues* reabertas em função da polaridade de sentimentos.

maior que a quantidade de sentimentos positivos. Os dados permitiram observar que, se houve uma tendência de discussão negativa ou neutra durante o período de atividade da *issue*, então a propensão de reabertura é maior.

QP3 - Uma *issue* com discussões com sentimentos neutros após o fechamento indica que ela será reaberta?

Encontramos 5.547 (42,68%) *issues* reabertas que possuem comentários com sentimentos neutros entre o fechamento da *issue* e a sua reabertura. A Figura 6 apresenta a distribuição de *issues* que possuem sentimentos neutros entre fechamento e reabertura. 2.079 (16,00%) *issues* reabertas possuem comentários com sentimentos negativos entre o fechamento da *issue* e a sua reabertura. A mediana da Figura 6 foi de 24 *issues*. A *issue* pode ser reaberta quando há sentimentos neutros relacionados, mas existe uma propensão maior de reabertura de *issues* com sentimentos neutros do que com sentimentos negativos. Algumas *issues* observadas foram reabertas sem discussões, portanto não foram contabilizadas.

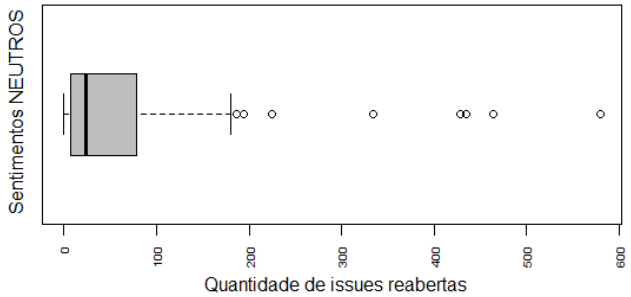

Figura 6. Distribuição de *issues* que possuem sentimentos neutros entre fechamento e reabertura.

4. Ameaças à validade

Validade Interna. A mineração dos dados foi realizada em um período de tempo que começou em 4 de Junho à 24 Julho de 2019. Assim, repositórios ativos podem receber atualizações com novas *issues*, comentários ou mudança de estado (aberta/fechada), alterando a quantidade de dados da nossa base. Como a lista escolhida foi do desafio *MSR Challenge Dataset*, ocorrido em 2014, dentre os repositórios minerados também encontramos alguns que estão arquivados ou são inativos. Entretanto, o cenário utilizado no estudo possui uma quantidade significativa de *issues*.

Validade Externa. Alguns repositórios possuem características que não favorecem o estudo, possuindo poucas *issues*. Resta, portanto, um número menor de *issues* reabertas para serem analisadas. Entretanto, uma menor quantidade propicia uma análise mais criteriosa e assertiva, por serem repositórios verificados e utilizados como referência nos estudos presentes na literatura.

Validade de Construto. Ao passo que havendo uma base de dados com múltiplos domínios distintos, os repositórios possuem tamanhos diferentes entre si. Há uma tendência de alguns repositórios possuírem maior quantidade de dados que os demais. Dessa forma, certos domínios impactam significativamente no resultado. Assim, para encontrar resultados mais homogêneos é preciso expandir a base de dados minerados do GitHub para integrar mais elementos relevantes ao estudo.

5. Trabalhos Relacionados

A pesquisa de [Caglayan et al. 2012] caracterizou os possíveis fatores que podem causar a reabertura de *issues*. Eles identificaram que a rede de proximidade das *issues* e a atividades dos desenvolvedores foram os fatores mais importantes para reabertura das *issues*.

A investigação de [Shihab et al. 2013] identificou que os principais fatores que causam a reabertura de *issues* relacionadas a bugs no Eclipse foram a descrição da *issue*, os textos dos comentários da *issue*, o tempo para solucionar o bug e o componente onde o bug foi encontrado.

Os autores [Zimmermann et al. 2012] categorizaram as principais razões que levam as *issues* relacionadas a defeitos (bugs) serem reabertas com base em um survey realizado com 358 colaboradores da Microsoft. Eles usaram fatores relacionados a processos e organizacionais, incluindo a localização, hábitos de trabalho, além de fatores extraídos dos relatório de *issues* (*issue report*).

O trabalho de [Souza et al. 2015] analisou o projeto Firefox para compreender as relações entre as diferentes formas de rejeição de *patch*. Foi identificado que cerca de 5,7% de todas as *issues* resolvidas do Firefox foram reabertas, o que induziu uma sobrecarga de discussões entre os colaboradores sobre a reabertura de uma *issue*. Também foi identificado que 70% das *issues* fechadas foram reabertas permaturamente devido a equívocos de interpretação dos colaboradores.

6. Considerações Finais

O presente estudo analisou o sentimento de cerca de 13 mil *issues* reabertas (que incluíram cerca de 153 mil comentários) em 80 repositórios de projetos hospedados no GitHub, buscando respostas sobre a previsão de reabertura de *issues* através da análise dos sentimentos presentes nas discussões dos colaboradores. A ferramenta SentiStrength foi fundamental para a classificação do grau de polaridade dos textos encontrados.

O impacto dos sentimentos nas discussões em alguns casos pode afetar ou não diretamente o ciclo de vida da *issue*. Identificado na primeira questão de pesquisa, a reabertura de uma *issue* com sentimento positivo existe, mas ela é menos recorrente do que com outros sentimentos. Também são questionados sobre a tendência da discussão durante a reabertura, que na segunda questão de pesquisa é observado que discussões negativas e neutras podem influenciar diretamente na reabertura de uma *issue*. Durante o

período de fechamento e reabertura de uma *issue*, existe um indicativo maior de reabertura se as discussões forem realizadas com sentimento neutro do que se forem realizadas com sentimento negativo. O estudo mostra a importância da análise de sentimentos para o gerenciamento de repositórios de software. As informações extraídas são indicativos que podem ajudar no gerenciamento do projeto, uma vez que *issues* fechadas podem ser identificados para posterior reabetura.

Como trabalhos futuros, planejamos correlacionar linguagens de programação, tipo de domínio e tipos de colaboradores com a análise de sentimentos. Também existe a necessidade de expandir a base de dados com uma maior amostra de repositórios.

Agradecimentos: O presente trabalho foi realizado com apoio do CNPq e da FAPESB.

Referências

Boechat, G., Júnior, J. M., Machado, I., and Mendonça, M. (2019). Análise de sentimentos em discussões de issues reabertas do github (material suplementar). Zenodo. `http://doi.org/10.5281/zenodo.3376175`.

Caglayan, B., Misirli, A. T., Miranskyy, A., Turhan, B., and Bener, A. (2012). Factors characterizing reopened issues: A case study. In *Proceedings of the 8th Int. Conf. on Predictive Models in Soft. Engineering*, pages 1–10, New York, USA. ACM.

Islam, M. R. and Zibran, M. F. (2017). Leveraging automated sentiment analysis in software engineering. In *14th Int. Conf. on Min. Soft. Repositories(MSR)*, pages 203–214.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. C.U.P.

Ortu, M., Destefanis, G., Adams, B., Murgia, A., Marchesi, M., and Tonelli, R. (2015). The jira repository dataset: Understanding social aspects of software development. In *Proceedings of the 11th Int. Conf. on Predictive Models and Data Analytics in Software Engineering (PROMISE)*, pages 1:1–1:4, New York, NY, USA. ACM.

Ortu, M., Murgia, A., Destefanis, G., Tourani, P., Tonelli, R., Marchesi, M., and Adams, B. (2016). The emotional side of software developers in jira. In *Proceedings of the 13th Int. Conf. on Mining Soft. Repositories(MSR)*, pages 480–483, NY, USA. ACM.

Pan, J. and Mao, X. (2014). An empirical study on interaction factors influencing bug reopenings. In *21st Asia-Pacific Soft. Engineering Conf.*, volume 2, pages 39–42.

Shihab, E., Ihara, A., Kamei, Y., Ibrahim, W. M., Ohira, M., Adams, B., Hassan, A. E., and Matsumoto, K.-i. (2013). Studying re-opened bugs in open source software. *Empirical Software Engineering*, 18(5):1005–1042.

Souza, R. R., Chavez, C. F., and Bittencourt, R. A. (2015). Patch rejection in Firefox: negative reviews, backouts, and issue reopening. *J. of Soft. Eng. Res. and Dev.*, 3(1).

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558.

Zimmermann, T., Nagappan, N., Guo, P. J., and Murphy, B. (2012). Characterizing and predicting which bugs get reopened. In *34th Int. Conf. on Software Engineering (ICSE)*, pages 1074–1083.