SoftSkillMiner: Identificação e Análise de Soft Skills em Desenvolvedores de Software pela sua Pegada Digital

Gustavo Vitor da Silva
Universidade Federal Lavras
São Sebastião do Paraíso
gustavo.silva39@estudante.ufla.br

Johnatan Alves De Oliveira
Universidade Federal Lavras
São Sebastião do Paraíso
johnatan.oliveira@ufla.br

Keywords
SoftSkills, GitHub, Mineração de Repositórios, Detecção de Skills

1 Introdução

A crescente complexidade do desenvolvimento de software, especialmente no contexto de projetos no modo geral, a uma exigência de profissionais com um conjunto diversificado de habilidades técnicas e também interpessoais [2, 3]. Em projetos de desenvolvimento de software e por padrão em projetos de código aberto (OSS), por exemplo, competências como *resolução de problemas*, *organização de tarefas* e *comunicação clara* são vitais para o sucesso do projeto [2]. Contudo, mesmo havendo forte demanda por habilidades técnicas (*hard skills*), recrutadores e gestores valorizam cada vez mais habilidades comportamentais (*soft skills*), pois estas são difíceis de treinar e impactam diretamente a dinâmica de equipe [3, 5].

As *soft skills* – também chamadas de *habilidades interpessoais* – envolvem traços de personalidade, capacidades de interação social, comunicação, adaptabilidade e pensamento crítico [3, 5]. Essas características subjetivas influenciam como os profissionais se relacionam e colaboram, afetando a produtividade coletiva [3, 5]. Portanto, identificar candidatos com *soft skills* adequadas é fundamental para formar equipes de alta performance na engenharia de software [3, 5].

No sentido mais amplo, uma *habilidade* ou *skill* pode ser entendida como a capacidade de um agente realizar tarefas específicas [4]. As *soft skills*, no entanto, abrangem atributos comportamentais como liderança, colaboração e inteligência emocional, que são necessario para harmonizar a interação entre as *hard skills* e potencializam o sucesso em ambientes colaborativos [3]. Estudos apontam que a posse de *soft skills* aumenta significativamente a chance do sucesso individual e coletivo nos projetos de software [3, 5]. Dado esse cenário, o desenvolvimento do **SoftSkillMiner**, uma ferramenta dedicada a extrair indicadores de *soft skills* de desenvolvedores a partir de seu histórico de contribuições em repositórios de código aberto.

O **SoftSkillMiner** opera sobre a *"pegada digital"* deixada pelos programadores no GitHub. Coletamos dados de atividades públicas dos usuários (*commits*, *issues*, *pull requests*, comentários etc.) e com isso podemos extraír métricas comportamentais e de colaboração para a determinação de *softskills* [1]. Em particular, analisamos aspectos como:

- Frequência e consistência de *commits* em projetos, indicando disciplina e compromisso;
- Número de *issues* fechadas e *pull requests* mesclados, refletindo proatividade na resolução de problemas e capacidade de integrar código de outros;
- Participação em discussões (comentários em *issues* e PRs) e contribuição para documentação, evidenciando engajamento e trabalho em equipe;
- Análise de sentimento em comunicações textuais, para inferir traços de comunicação, cortesia e empatia nas interações entre desenvolvedores;
- Tempo médio de resposta a solicitações (*issues*/PRs), sugerindo proatividade e responsabilidade.

Cada métrica acima funciona como um *sinal* mensurável associado a competências comportamentais [2]. Por exemplo, *commits* regulares podem sinalizar disciplina profissional, enquanto análises linguísticas positivas nos comentários podem indicar boa comunicação e colaboração. Baseamo-nos nas abordagens de Thomas et al. [1] e Liang et al. [2], que propuseram usar sinais extraídos do GitHub para detectar habilidades (técnicas e comportamentais) de colaboradores. No **SoftSkillMiner**, haverá a combinação dessas informações em um modelo empírico que atribui uma **pontuação de *soft skills*** a cada desenvolvedor. Essa pontuação sintetiza indicadores de comunicação, trabalho em equipe e outras competências interpessoais, oferecendo um recurso quantitativo para auxiliar recrutadores e coordenadores na identificação de talentos compatíveis com os requisitos para uma posição no projeto de software.

A abordagem foi implementada como um protótipo de minerador de dados, que acessa a API do GitHub para coletar o histórico de contribuições dos desenvolvedores. Em seguida, as métricas são computadas e normalizadas, e técnicas de processamento de linguagem natural (e.g. ferramentas de *sentiment analysis*) avaliam as postagens textuais. Os resultados são apresentados em um *dashboard* que lista as pontuações gerais e detalhadas de *soft skills* por usuário. Espera-se que o **SoftSkillMiner** complemente as práticas tradicionais de seleção de profissionais em engenharia de software, destacando candidatos não apenas pelos conhecimentos técnicos, mas também pelo perfil comportamental indicado em suas contribuições de código aberto e por uma possivel ilustração de seu nivel de *softskills*.

2 Objetivo

Analise, extração, deselvolvimento dos por fóruns comumente utilizados por desenvolvedores (GitHub, StackOverflow et), com objetivo de quantificar de maneira representativa e confiável seu possivel nivel de softskill.

3 Metodologia

Neste trabalho coletaremos e analisaremos a pegada digital de engenheiros de software em fóruns públicos, com foco principal no GitHub. A coleta é realizada por scraping das atividades públicas

(commits, issues, pull requests, comentários e metadados associados) e agregação por usuário para compor o histórico comportamental de cada participante.

A pontuação de *soft skills* segue uma estratégia empírica híbrida: primeiro extraímos sinais mensuráveis (por exemplo, frequência de commits, número de issues resolvidas, PRs mescladas, participação em discussões e métricas derivadas de análise de texto nas interações) e transformamos esses sinais em um score inicial. Em seguida, vinculamos esse score às avaliações obtidas por entrevistas/questionários individuais, procedimento inspirado na abordagem de Zimmerman et al. [1] de modo que cada usuário possui uma pontuação empírica de referência obtida humanamente. O sistema gera automaticamente uma estimativa de *soft skill* por usuário e compete com a avaliação humana; discrepâncias entre as duas são usadas para recalibrar o modelo.

O refinamento do modelo é iterativo e baseado em técnicas estatísticas e de aprendizado: ajustamos pesos dos sinais por regressão regularizada ou modelos de autoregressivos, avaliamos desempenho via validação cruzada e repetimos o ciclo até reduzir a diferença média entre as pontuações automáticas e as empíricas. O objetivo é que a pontuação automática se aproxime cada vez mais da obtida nas entrevistas, transformando medições objetivas da atividade online em um indicador confiável de *soft skills*.

Por fim, todo o processo preserva dados públicos apenas e associa identidades a avaliações humanas somente mediante consentimento explícito dos participantes com a noção de possiveis vieses das autoavaliações e das métricas extraídas.

4 Resultados Esperados

Esperamos que o procedimento descrito gere evidências quantitativas e qualitativas de que é possível inferir indicadores de *soft skills* a partir da pegada digital em plataformas públicas de desenvolvimento, em particular o GitHub. Os resultados previstos incluem:

- Uma **pontuação automática de *soft skills*** por usuário, construída a partir da combinação dos sinais extraídos (commits, issues fechadas, PRs mescladas, participação em discussões e medidas de sentimento).
- Correlação positiva entre a pontuação automática e as **avaliações empíricas** obtidas em entrevistas/questionários; como meta operacional, espera-se um *MAE* médio inferior a 1.0 na escala 0–5 e que ao menos 70% das previsões estejam dentro de uma margem de ±1 ponto da avaliação humana.
- Identificação das **features mais informativas** para cada dimensão de *soft skill* (por exemplo, frequência de commits para disciplina; polaridade/cordialidade nos comentários para comunicação).
- Um **protótipo de dashboard** que apresenta pontuações agregadas e detalhadas por usuário, permitindo inspeção e auditoria dos sinais subjacentes.

Em conjunto, esses resultados devem demonstrar a viabilidade de transformar métricas objetivas de atividade online em indicadores úteis de habilidades comportamentais, ao mesmo tempo em que evidenciam as restrições e os cuidados necessários para interpretação e aplicação desses indicadores.

References

[1] Denae Jenny, Thomas. 2022. Towards Mining OSS Skills from GitHub Activity. *IEEE/ACM International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER), volume=44* (2022).

[2] Haiyi Liang, Jennifer Marlow, and Tianyi Zhang. 2022. Detecting skills in open source software contributors: introducing Disko. In *Proceedings of the 44th International Conference on Software Engineering: New Ideas and Emerging Results*.

[3] Jennifer Marlow and Laura Dabbish. 2013. Exploring the role of social media in open source software communities. In *Proceedings of the Conference on Computer Supported Cooperative Work*.

[4] Paulo Montandon, Igor Steinmacher, and Marco Aurélio Gerosa. 2019. Detecting expertise in OSS communities: Combining data mining and social network analysis. In *Proceedings of the 41st International Conference on Software Engineering*.

[5] Ahmad Rashid and Faheem Ahmed. 2020. Soft skills and software development: A systematic mapping study. *Journal of Systems and Software* 158 (2020).