PLN e MSR em Repositórios Open Source de Criptomoedas: Um Estudo de Caso no Ethereum

Cleane Sousa Santos[1], Arlino Henrique Magalhães de Araújo[1]

[1]Programa de Pós-Graduação em Ciência da Computação, Universidade Federal do Piauí (UFPI)

cleanesantos@ufpi.edu.br, arlino@ufpi.edu.br

Abstract. *This study proposes an integrated approach combining Natural Language Processing (NLP) and Mining Software Repositories to characterize linguistic and operational patterns in issues from Ethereum's official GitHub repository, investigating their association with short-term Ether (ETH) price variations. The approach combines BERTopic, sentence embeddings with MiniLM, KMeans clustering, and sentiment metrics with VADER and TextBlob. Preliminary results indicate that short and specialized texts limit topic separability, while embedding-based clustering reveals operational profiles associated with message length, authorship profile, and issue resolution time. Sentiment and event-study analyses did not indicate substantive association with short-term ETH returns.*

Resumo. *Este estudo propõe uma abordagem integrada de Processamento de Linguagem Natural (PLN) e Mineração de Repositórios de Software para caracterizar padrões linguísticos e operacionais em issues do repositório oficial da Ethereum no GitHub, investigando sua associação com variações de curto prazo no preço da Ether (ETH). A abordagem combina BERTopic, embeddings de sentenças com MiniLM, clusterização com KMeans e métricas de sentimento com VADER e TextBlob. Os resultados preliminares indicam que textos curtos e especializados limitam a separabilidade temática, enquanto a clusterização por embeddings revela perfis operacionais associados ao tamanho das mensagens, ao perfil de autoria e ao tempo de resolução das issues. As análises de sentimento e de eventos não indicaram associação substantiva com retornos de curto prazo da ETH.*

Informações Gerais. Nível: Mestrado. **Aluno:** Cleane Sousa Santos. **Orientador:** Arlino Henrique Magalhães de Araújo. **Programa:** Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Piauí. **Admissão:** Fevereiro 2025. **Exame de qualificação:** A definir. **Defesa prevista:** Março 2027. **Etapas concluídas:** créditos obrigatórios em disciplinas.

1. Introdução e Motivação

Repositórios *open source* constituem fontes empíricas relevantes para compreender a evolução de sistemas computacionais complexos, pois armazenam decisões arquiteturais, correções, discussões técnicas e interações entre desenvolvedores. Ferramentas como *issues*, *commits* e *pull requests* permitem observar como comunidades distribuídas organizam o fluxo de trabalho em larga escala, sendo amplamente utilizadas em estudos de *Mining Software Repositories* (MSR) para investigar padrões de desenvolvimento e colaboração em projetos de software [Hassan 2008]. Essa perspectiva é particularmente relevante no domínio das criptomoedas, uma vez que plataformas baseadas em *blockchain* dependem de ciclos contínuos de verificação técnica, auditoria pública e evolução colaborativa.

Contudo, rastreadores de *issues* podem formar ecossistemas complexos, com grande volume de dados textuais curtos, técnicos e heterogêneos, compostos por termos de domínio, fragmentos de código e registros operacionais, o que limita a análise manual desses artefatos. Nesse contexto, técnicas de Processamento de Linguagem Natural (PLN) tornam-se relevantes para extrair padrões e conhecimento de forma automatizada [Montgomery et al. 2025]. Modelos baseados em *transformers*, como o CodeBERT [Feng et al. 2020] e abordagens baseadas em *embeddings* de sentenças, como o Sentence-BERT [Reimers and Gurevych 2019], ampliaram as possibilidades de representação semântica de textos técnicos e curtos, como aqueles presentes em *issues*.

Apesar desses avanços, ainda há lacunas na aplicação integrada de PLN e MSR ao contexto de repositórios *blockchain*. Esta pesquisa propõe uma abordagem metodológica para mineração e análise textual de *issues*, aplicada inicialmente ao repositório oficial da Ethereum no GitHub, com o objetivo de caracterizar padrões linguísticos, semânticos e operacionais. Grande parte das investigações sobre *blockchain* concentra-se em fontes externas, como notícias ou redes sociais, enquanto artefatos internos de repositórios permanecem relativamente subexplorados. Entretanto, esses registros documentam parte substancial da evolução técnica das plataformas *blockchain* e constituem evidências diretas das decisões que moldam seu funcionamento.

Como contribuição, busca-se estruturar um processo analítico replicável, passível de reaplicação a repositórios de outras criptomoedas que disponibilizem *issues* e metadados de desenvolvimento. Além disso, a pesquisa avalia a adequação de métodos de PLN ao tratamento de textos técnicos curtos, explicitando suas possibilidades e limitações em artefatos de software.

2. Trabalhos Relacionados

O levantamento sistemático da literatura realizado na etapa de fundamentação desta pesquisa indica que o uso de PLN em criptomoedas permanece fortemente orientado a fontes textuais externas ao processo de desenvolvimento, especialmente redes sociais, previsão de preço e sentimento de mercado. Em contraste, a exploração sistemática de artefatos internos de repositórios, como *issues*, *commits*, *pull requests*, comentários de desenvolvedores e contratos inteligentes, concentra-se em um subconjunto menor e ainda recente de pesquisas.

Entre os trabalhos relacionados, destaca-se a análise de discussões de desenvolvedores da Ethereum com modelos baseados em BERT, voltada a aspectos de sustentabilidade no desenvolvimento *blockchain* [Vaccargiu et al. 2024]. No contexto de detecção

de fraudes, foram propostos o BERT4ETH, um Transformer pré-treinado para representar contas da Ethereum a partir de sequências de transações, e o ZipZap, um *framework* voltado à redução do custo computacional no treinamento de modelos de linguagem em dados transacionais de *blockchain* [Hu et al. 2023, Hu et al. 2024]. Outra vertente aborda a identificação de más práticas em contratos inteligentes por meio de modelos de linguagem, como no SCALM [Li et al. 2025].

Contudo, esses estudos não apresentam uma abordagem integrada centrada em *issues* como artefatos textuais e operacionais de desenvolvimento. Esta pesquisa busca preencher essa lacuna ao combinar conteúdo linguístico, metadados do repositório e séries temporais externas da ETH no contexto de MSR aplicado a repositórios *blockchain*.

3. Problema de Pesquisa

O problema de pesquisa deste trabalho consiste em compreender de que forma dados textuais e operacionais extraídos de repositórios *blockchain* podem ser estruturados, minerados e analisados para revelar padrões associados à dinâmica técnica desses ecossistemas. A questão de pesquisa pode ser formulada da seguinte forma: como técnicas de PLN e MSR podem ser integradas para caracterizar padrões linguísticos, semânticos e operacionais em *issues* de repositórios *blockchain*?

Diferentemente de repositórios de software convencionais, ecossistemas *blockchain* como o Ethereum estão vinculados a um ativo financeiro negociável. Assim, decisões técnicas registradas nas *issues*, como mudanças de protocolo ou eventos de governança, podem compor o contexto informacional acompanhado pela comunidade e pelo mercado, ainda que seus efeitos não sejam necessariamente imediatos ou diretamente observáveis em janelas curtas.

A pesquisa possui aderência à área de Banco de Dados por envolver coleta, estruturação, limpeza, integração, mineração e análise de dados heterogêneos, combinando textos não estruturados, metadados operacionais de repositórios de software e séries temporais externas. Seu diferencial está na integração entre a dimensão textual das *issues*, a dimensão operacional dos metadados do repositório e a dimensão temporal externa representada pelas séries históricas da Ether (ETH).

4. Solução Proposta

A solução proposta consiste em uma abordagem metodológica para mineração e análise textual de repositórios *blockchain*, aplicada inicialmente ao repositório oficial da Ethereum no GitHub. A Ethereum é uma plataforma descentralizada baseada em *blockchain* que suporta contratos inteligentes e aplicações descentralizadas, sendo uma das principais criptomoedas em termos de atividade de desenvolvimento.

A abordagem organiza o processo analítico em etapas integradas, desde a coleta das *issues* até a análise dos padrões textuais, operacionais e temporais. A Figura 1 apresenta a arquitetura geral da solução proposta.

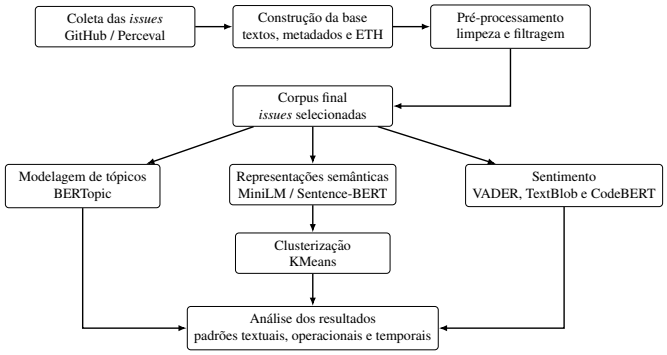

Figura 1. Arquitetura geral da solução proposta para mineração e análise textual de repositórios *blockchain*.

A coleta foi realizada por meio da ferramenta Perceval, componente do ecossistema GrimoireLab, que automatiza a extração estruturada de dados de repositórios de software. Os registros foram convertidos em uma base com atributos textuais e operacionais, incluindo o tempo de resolução, calculado pela diferença entre as datas de abertura e fechamento de cada *issue*.

Após a estruturação dos dados, os textos foram submetidos ao pré-processamento e à geração de *embeddings*, utilizados na modelagem de tópicos e na clusterização. De forma complementar, as *issues* foram integradas a séries históricas da ETH, permitindo calcular retornos em horizontes de dois, três e sete dias após a abertura de cada registro.

5. Metodologia

O pré-processamento textual envolveu normalização dos documentos, remoção de URLs, *hashes* de *commits*, números, blocos extensos de código, padrões de *log* e tokens característicos da Engenharia de Software que não contribuíam semanticamente para a análise. Também foram removidos identificadores de versão, referências a *pull requests*, nomes de funções, variáveis e endereços de memória.

Em seguida, foram aplicados procedimentos de tokenização e lematização com apoio da biblioteca NLTK em Python. Foram eliminadas *stopwords* da língua inglesa e termos específicos recorrentes no ecossistema Ethereum. Para reduzir ruído textual sem comprometer a representatividade semântica, mantiveram-se apenas documentos com comprimento entre dois e cento e quarenta tokens, faixa definida com base na distribuição empírica dos tamanhos do corpus.

Inicialmente, aplicou-se o BERTopic, modelo de modelagem de tópicos baseado em *embeddings*, redução dimensional e ponderação c-TF-IDF [Grootendorst 2022]. Em seguida, utilizou-se o Sentence-BERT MiniLM-L6-v2 para gerar representações vetoriais densas dos textos [Reimers and Gurevych 2019]. Sobre essas representações, realizou-se clusterização não supervisionada com KMeans, algoritmo originalmente proposto por MacQueen [MacQueen 1967], configurado com $k = 6$ grupos.

A qualidade dos agrupamentos foi avaliada por métricas internas: índice Silhouette [Rousseeuw 1987], Calinski-Harabasz [Calinski and Harabasz 1974] e Davies-Bouldin [Davies and Bouldin 1979]. Para caracterizar a valência e a polaridade textual, estimou-se o sentimento por meio do VADER [Hutto and Gilbert 2014] e do TextBlob [Loria 2018]. Também foi calculado um indicador exploratório derivado do Code-

BERT, obtido pela média dos *logits* retornados para cada *issue*, a fim de incluí-lo como variável textual nas análises de correlação e regressão com os retornos financeiros. Esse indicador foi obtido por meio de uma camada de classificação acoplada ao modelo, sem *fine-tuning* específico para a tarefa, o que impede sua interpretação como métrica validada de tecnicidade textual. Por fim, conduziu-se um estudo de eventos com retornos da ETH em D+2, D+3 e D+7.

6. Resultados Preliminares

Após os procedimentos de limpeza, normalização textual e remoção de registros atípicos, o corpus final totalizou aproximadamente 27.135 *issues*. A distribuição inicial do número de tokens evidenciou que a maior parte das *issues* continha textos curtos. Observou-se também elevada proporção de *hapax*, característica típica de vocabulários técnicos altamente especializados, nos quais cada documento tende a utilizar terminologia específica pouco compartilhada com os demais registros.

A Tabela 1 apresenta as estatísticas descritivas antes e depois do pré-processamento. A redução da média de tokens por documento, de 85,34 para 21,62, indica que documentos excessivamente longos, compostos majoritariamente por blocos de código e saídas automatizadas, foram removidos. O valor máximo também foi reduzido de 18.901 para 140 tokens, confirmando a eliminação de registros atípicos que poderiam distorcer os modelos semânticos.

Tabela 1. Estatísticas descritivas do corpus antes e depois do pré-processamento

Métrica / Faixa	Antes	Depois
N documentos	35.555	27.135
Média de tokens	85,34	21,62
Mediana	14	10
Percentil 90%	165	60
Percentil 95%	297	86
Máximo	18.901	140
Faixa 3–10 tokens	33,85%	37,52%
Faixa 11–30 tokens	18,98%	25,03%
Faixa 31–100 tokens	20,24%	19,69%
$>$ 500 tokens	2,56%	0%
Vocabulário total	102.642	45.916
Proporção de *hapax*	5,401	4,901

A aplicação inicial do BERTopic gerou tópicos distribuídos entre rótulos numéricos, sendo o rótulo -1 reservado para documentos que não se encaixam em nenhum tópico definido. Esse comportamento é esperado em corpora técnicos e heterogêneos, nos quais grande parte dos documentos não compartilha termos suficientes para formar grupos coesos. Esse padrão é coerente com o observado por de Groot et al. [de Groot et al. 2022], que demonstram que o HDBSCAN, ao ser aplicado a textos curtos e multidomínio, pode classificar muitos documentos como *outliers*.

Diante desse cenário, o BERTopic foi empregado como etapa exploratória. A clusterização via KMeans sobre *embeddings* semânticos, por sua vez, gerou agrupamentos mais interpretáveis sob a perspectiva operacional. O Cluster 0 concentrou textos extremamente curtos, elevada participação de *committers* e menor tempo médio de resolução,

sugerindo demandas objetivas e de rápida tratativa interna. Em contraste, o Cluster 5 reuniu *issues* mais extensas, baixa participação de desenvolvedores internos e maior tempo médio de resolução, padrão compatível com discussões mais detalhadas e frequentemente iniciadas por usuários externos.

7. Considerações Finais e Próximas Etapas

Os resultados preliminares indicam que o corpus analisado é marcado por predominância de textos curtos, vocabulário técnico especializado e presença de registros operacionais heterogêneos. Esse cenário limita a separabilidade temática por modelos baseados em densidade, como o BERTopic, mas permite identificar perfis operacionais por meio da combinação entre *embeddings* semânticos e clusterização não supervisionada.

No que se refere à integração entre conteúdo textual e dinâmica de mercado, as análises preliminares não identificaram evidências consistentes de associação entre características linguísticas das *issues* e variações de curto prazo no preço da ETH. Embora alguns testes tenham indicado significância estatística pontual em anos específicos, os efeitos observados apresentaram magnitude econômica limitada. Assim, nesta etapa da pesquisa, o principal resultado está na caracterização textual e operacional do repositório, e não na explicação do comportamento de mercado da criptomoeda.

Uma limitação desta etapa está relacionada ao horizonte temporal adotado na análise de retornos da ETH. Embora janelas de D+2, D+3 e D+7 permitam observar associações de curto prazo, discussões técnicas em repositórios *blockchain* podem produzir efeitos em períodos mais longos, especialmente quando associadas a *releases*, *forks*, mudanças de protocolo ou decisões de governança. Por isso, análises futuras deverão considerar agregações por versões, ciclos de lançamento e eventos técnicos relevantes do ecossistema.

Como próximas etapas, pretende-se refinar os critérios de pré-processamento, comparar BERTopic com estratégias alternativas de modelagem de tópicos para textos curtos, e definir baselines e métricas para comparação com o estado da arte, uma vez que a revisão sistemática não identificou estudos que integrem issues, metadados de repositório e séries temporais de preço na mesma abordagem. Também se pretende documentar o pipeline de coleta, pré-processamento e modelagem para garantir a replicabilidade em outros repositórios blockchain.

Uso de IA generativa. Ferramentas de IA generativa foram utilizadas como apoio à revisão linguística e à organização textual do manuscrito. A responsabilidade pelo conteúdo, pela análise dos resultados e pela verificação das referências bibliográficas permanece integralmente com os autores.

Referências

[Calinski and Harabasz 1974] Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27.

[Davies and Bouldin 1979] Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227.

[de Groot et al. 2022] de Groot, M., Aliannejadi, M., and Haas, M. R. (2022). Experiments on generalizability of BERTopic on multi-domain short text. *arXiv preprint arXiv:2212.08459*.

[Feng et al. 2020] Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., and Zhou, M. (2020). Codebert: A pre-trained model for programming and natural languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1536–1547. ACL.

[Grootendorst 2022] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

[Hassan 2008] Hassan, A. E. (2008). The road ahead for mining software repositories. In *Frontiers of Software Maintenance, 2008. FoSM 2008*, pages 48–57. IEEE.

[Hu et al. 2024] Hu, S., Huang, T., Chow, K.-H., Wei, W., Wu, Y., and Liu, L. (2024). Zipzap: Efficient training of language models for large-scale fraud detection on blockchain. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pages 2807–2816. ACM.

[Hu et al. 2023] Hu, S., Zhang, Z., Luo, B., Lu, S., He, B., and Liu, L. (2023). BERT4ETH: A pre-trained transformer for ethereum fraud detection. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*, pages 2189–2197. ACM.

[Hutto and Gilbert 2014] Hutto, C. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM)*, pages 216–225.

[Li et al. 2025] Li, Z., Li, X., Li, W., and Wang, X. (2025). SCALM: Detecting bad practices in smart contracts through llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 470–477.

[Loria 2018] Loria, S. (2018). TextBlob: Simplified text processing.

[MacQueen 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press.

[Montgomery et al. 2025] Montgomery, L., Lüders, C., and Maalej, W. (2025). Mining issue trackers: Concepts and techniques. In *Handbook on Natural Language Processing for Requirements Engineering*, pages 309–336. Springer.

[Reimers and Gurevych 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*.

[Rousseeuw 1987] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

[Vaccargiu et al. 2024] Vaccargiu, M., Aufiero, S., Bartolucci, S., Neykova, R., Tonelli, R., and Destefanis, G. (2024). Sustainability in blockchain development: A BERT-based analysis of ethereum developer discussions. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering (EASE '24)*, pages 381–386, Salerno, Italy. ACM.