Análise temporal do discurso de ódio online e sua correlação com crimes motivados por LGBTfobia

Thaynara Alexandre Cardoso[1], **Ana Cristina Bicharra Garcia**[1]

[1]Programa de Pós-graduação em Informática
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

thaynara.cardoso@edu.unirio.br,cristina.bicharra@uniriotec.br

Abstract. *This study investigates the temporal relationship between public discourse on social media and official records of hate crimes motivated by LGBTphobia in the state of Minas Gerais, Brazil. The research integrates posts collected from Reddit with official incident records provided by the State Secretariat of Justice and Public Security (SEJUSP/MG) over a shared temporal window. Text mining and sentiment polarity analysis are applied to the textual data, followed by training and comparison of five machine learning classifiers (K-Nearest Neighbors, Support Vector Machine, Random Forest, Artificial Neural Networks, and Naïve Bayes) evaluated through accuracy, precision, recall, and F1-score using stratified k-fold cross-validation. As its primary contribution, this work proposes a reproducible analytical pipeline for integrating collaborative online community data with official crime records, advancing the understanding of offline effects of behaviors mediated by collaborative systems. The framework is designed for exploratory temporal analysis, replicability across crime categories and geographic regions, and potential use in informing public security policies and content moderation strategies on digital platforms.*

Resumo. *Este trabalho investiga a relação temporal entre discursos públicos em redes sociais e registros oficiais de crimes motivados por LGBTfobia no estado de Minas Gerais. A pesquisa integra postagens coletadas do Reddit e registros de ocorrências disponibilizados pela SEJUSP/MG em uma mesma janela temporal. Técnicas de mineração de texto e análise de polaridade de sentimentos são aplicadas aos dados textuais, seguidas de treinamento e comparação de cinco algoritmos de aprendizado de máquina (KNN, SVM, Floresta Aleatória, Redes Neurais Artificiais e Naïve Bayes), avaliados por meio de acurácia, precisão, recall e F1-score, com validação cruzada estratificada k-fold. Como contribuição, o trabalho propõe um pipeline analítico reproduzível para integração de dados de comunidades colaborativas online e registros criminais oficiais, avançando a compreensão dos efeitos offline de comportamentos mediados por sistemas colaborativos (CSCW). A metodologia foi concebida com foco em análises temporais exploratórias, replicabilidade para outros tipos de crimes e regiões, e potencial de subsídio a políticas públicas de segurança e moderação de conteúdo em plataformas digitais.*

1. Problema

O crescimento das redes sociais expandiu espaços de livre expressão, mas também facilitou a disseminação de discurso de ódio (DO), caracterizado por hostilidade e

discriminação [Sharma et al. 2024],[Organização das Nações Unidas 2019, p. 2]. Estudos recentes indicam que o discurso de ódio online pode contribuir para a normalização da violência e anteceder ocorrências de crimes motivados por preconceito no mundo offline [Bozhidarova et al. 2023],[Arcila Calderón et al. 2024]. No Brasil, apesar do aumento expressivo de denúncias de crimes motivados por LGBTfobia [Ministério dos Direitos Humanos e da Cidadania 2024], ainda há uma lacuna de estudos que integrem dados de redes sociais e registros oficiais de criminalidade para investigar possíveis correlações temporais e espaciais entre esses fenômenos. Diante desse cenário, este estudo investiga a seguinte questão: qual a correlação temporal e espacial entre manifestações de discurso de ódio online e crimes motivados por LGBTfobia no contexto brasileiro? Como hipóteses exploratórias: H1: aumento de sentimento negativo precede crimes; H2: certas palavras têm maior associação com eventos violentos

2. Objetivos

O objetivo desta pesquisa é analisar padrões de correlação temporal e espacial entre discurso hostil em redes sociais e registros de crimes motivados por LGBTfobia. Este trabalho contribui com (i) evidência empírica sobre efeitos offline de interações em comunidades online; e (ii) um framework metodológico replicável para análise de fenômenos sociais mediados por plataformas digitais.

3. Contexto

Esta pesquisa insere-se no campo de Sistemas Colaborativos ao investigar efeitos offline de interações em comunidades online. Estudos em CSCW indicam que tecnologias computacionais mediam interações sociais e influenciam práticas coletivas [Grudin 1994],[Schmidt and Bannon 1992].

Embora existam pesquisas que relacionam discurso de ódio online a crimes offline [Arcila Calderón et al. 2024], há lacuna no contexto brasileiro, especialmente em relação a crimes motivados por LGBTfobia.

Ao integrar dados de redes sociais e registros oficiais, este estudo busca ampliar a compreensão dessas dinâmicas e apoiar análises de risco social.

4. Motivação

Em 2022, foram registradas 19.128 denúncias de crimes de ódio contra pessoas LGBT+ no Brasil, o maior valor em 7 anos, sendo 52,2% na região Sudeste [Ministério dos Direitos Humanos e da Cidadania 2024]. A incorporação de dados criminais notificados, como os disponibilizados por Secretarias Estaduais de Segurança Pública, permite avaliar se contextos digitais hostis de redes sociais se correlacionam com indicadores concretos de violência, como agressões e homicídios contra a população LGBT+. Essa abordagem contribui para uma compreensão mais ampla das dinâmicas sociais contemporâneas, articulando elementos do discurso online com seus possíveis reflexos [Tran et al. 2024].

5. Solução Proposta

A solução proposta consiste em uma abordagem exploratória integrada de análise temporal e espacial que combina dados de redes sociais e registros oficiais de crimes motivados

por LGBTfobia. A metodologia utiliza técnicas de análise de sentimentos, mineração de texto e aprendizado de máquina para identificar padrões de correlação. O artefato trata-se de um pipeline modular com quatro componentes: (1) pré-processamento: geolocalização, tokenização, anonimização; (2) análise de sentimento: VADER adaptado; (3) modelagem: algoritmos: KNN, SVM, RNA, Random Forest, Naïve Bayes; (4) visualização: séries temporais, mapas de calor, ranking de palavras. O pipeline tem caráter acadêmico e metodológico, com foco em replicabilidade. Seus resultados são direcionados a dois perfis de usuários potenciais, como pesquisadores que desejam aplicar a abordagem a outros estados brasileiros ou outras categorias de crime e gestores de segurança pública interessados em monitorar padrões de risco.

6. Métodos de condução da pesquisa

Esta pesquisa caracteriza-se como um estudo de caso exploratório, com abordagem predominantemente quantitativa [Filippo et al. 2011],[Yin 2005].

A escolha metodológica justifica-se por sua adequação à investigação de fenômenos contemporâneos em contextos reais, nos quais o pesquisador não possui controle sobre as variáveis envolvidas. Neste estudo, analisa-se a possível correlação temporal e espacial entre discurso de ódio online e crimes motivados por LGBTfobia no estado de Minas Gerais, no período de 2023 a 2025. Conforme destacado por [Yin 2005], o método de estudo de caso é particularmente apropriado quando as fronteiras entre o fenômeno e seu contexto não são claramente definidas.

A natureza exploratória da pesquisa decorre da ausência de estudos prévios que integrem, no contexto brasileiro, dados de comportamento em comunidades colaborativas online com registros oficiais de crimes motivados por LGBTfobia. Assim, não se parte de uma hipótese causal rígida, mas de uma questão de pesquisa aberta, característica de investigações exploratórias [Filippo et al. 2011].

A abordagem quantitativa predominante fundamenta-se no uso de técnicas estatísticas e modelos de aprendizado de máquina para análise dos dados. As variáveis consideradas são majoritariamente numéricas, incluindo frequência de postagens, escores de sentimento e número de ocorrências criminais.

Por fim, o estudo se insere no campo de Sistemas Colaborativos ao investigar como tecnologias computacionais mediam interações sociais e como essas mediações se relacionam com fenômenos sociais no mundo offline, conforme discutido por [Grudin 1994]. Especificamente, analisa-se como comportamentos emergentes em comunidades colaborativas online, podem estar associados a dinâmicas sociais observáveis fora do ambiente digital. A pesquisa adapta ao contexto brasileiro o modelo proposto em um estudo no Reino Unido [Jimoh 2023]. Para operacionalizar a coleta, preparação, modelagem e análise dos dados, adota-se o framework CRISP-DM (Cross-Industry Standard Process for Data Mining) [Chapman 2000],[Wirth and Hipp 2000]. A Figura 1 ilustra as etapas

Os dados são obtidos a partir de registros oficiais da SEJUSP/MG e de postagens de comunidades no Reddit do estado de Minas Gerais, considerando a mesma janela temporal. Após o pré-processamento e padronização das variáveis, é aplicada a análise de polaridade de sentimentos.

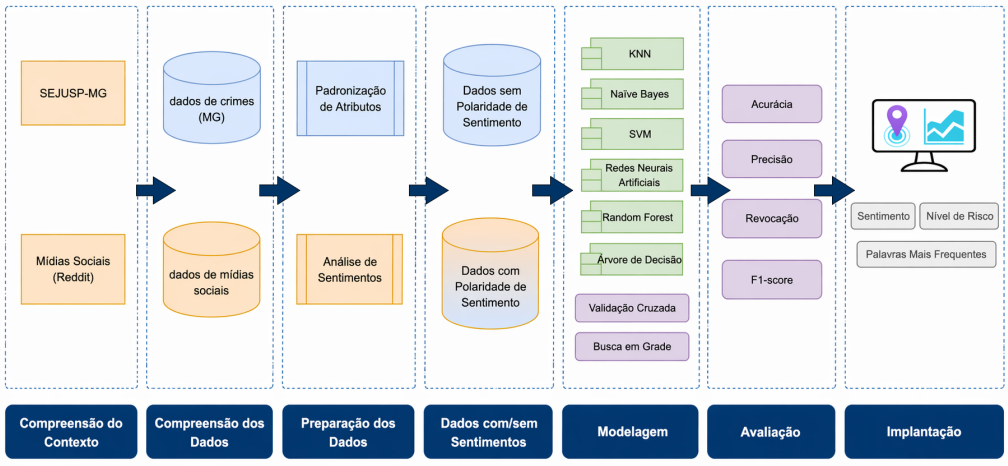

Figura 1. Fases da Metodologia

Na etapa de modelagem, são treinados e comparados cinco algoritmos supervisionados (KNN, SVM, RNA, Floresta Aleatória e Naïve Bayes). A variável-alvo é a categoria de risco do período, derivada dos registros da SEJUSP/MG. As features incluem: volume de postagens com polaridade negativa, score médio de sentimento e período temporal. A validação cruzada estratificada com k-folds é utilizada para garantir a representatividade das classes em cada partição, dado o possível desbalanceamento entre períodos de alto e baixo risco.

Para garantir a reprodutibilidade do estudo, o pipeline analítico, os scripts de coleta e pré-processamento, e os parâmetros de configuração dos modelos serão disponibilizados em repositório (GitHub) após a publicação. Os dados da SEJUSP/MG e os dados do Reddit serão referenciados por meio do link oficial de acesso aberto;

7. Avaliação dos resultados

A avaliação utiliza métricas como acurácia, precisão, recall e F1-score para comparar os modelos. O objetivo é verificar a consistência dos resultados e identificar padrões nos dados.

8. Aspectos Éticos

A pesquisa utiliza dados secundários públicos ou governamentais agregados, dispensando registro CEP/CONEP conforme Resolução CNS 510/2016 [Brasil. Conselho Nacional de Saúde 2016]. Os dados são tratados de forma agregada sem preservação de identificadores de usuário (nomes de usuários ou metadados de perfil). Nenhum dado individual é reportado nos resultados. Em relação aos dados da SEJUSP/MG, os registros não possuem dados nominais de envolvidos. Não há risco de reidentificação. Por fim, destaca-se que os resultados desta pesquisa não devem ser utilizados para identificação ou perfilamento de indivíduos.

Declaração de uso de Inteligência Artificial Generativa (IAG)
Foi utilizado ChatGPT[1] para apoiar na tradução e organização de conceitos deste trabalho.

[1]https://chatgpt.com/

Referências

Arcila Calderón, C., Sánchez Holgado, P., Gómez, J., Barbosa, M., Qi, H., Matilla, A., Amado, P., Guzmán, A., López-Matías, D., and Fernández-Villazala, T. (2024). From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and lgbt communities. *Humanities and Social Sciences Communications*, 11(1):1369.

Bozhidarova, M., Chang, J., Ale-Rasool, A., Liu, Y., Ma, C., Bertozzi, A. L., Brantingham, P. J., Lin, J., and Krishnagopal, S. (2023). Hate speech and hate crimes: a data-driven study of evolving discourse around marginalized groups.

Brasil. Conselho Nacional de Saúde (2016). Resolução nº 510, de 7 de abril de 2016. `https://conselho.saude.gov.br/resolucoes/2016/Reso510.pdf`. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais. Publicada no DOU nº 98, 24 maio 2016.

Chapman, P. (2000). *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS.

Filippo, D., Pimentel, M., and Wainer, J. (2011). *Metodologia de pesquisa científica em sistemas colaborativos*, pages 379–404.

Grudin, J. (1994). Computer-supported cooperative work: History and focus. *Computer*, 27(5):19–26.

Jimoh, F. (2023). *Real Time Crime Prediction Using Social Media*. PhD thesis, University of Salford. Tese de Doutorado.

Ministério dos Direitos Humanos e da Cidadania (2024). Incitação à violência contra a vida na internet lidera violações de direitos humanos com mais de 76 mil casos em cinco anos, aponta observadh. Acesso em: 4 ago. 2025.

Organização das Nações Unidas (2019). Un strategy and plan of action on hate speech. `https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf`. Acesso em: jul. 2025.

Schmidt, K. and Bannon, L. (1992). Taking cscw seriously: Supporting articulation work. *Computer Supported Cooperative Work*, 1:7–40.

Sharma, D., Singh, A., and Singh, V. (2024). Thar: Targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Tran, M., Kashyap, A., Rybak, P., Nakamura, H., Lau, J. H., and Black, A. W. (2024). Harmpot: An annotation framework for evaluating offline harm potential of social media text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics and Language Resources (LREC-COLING)*, pages 6459–6470, Torino, Italy. ELRA.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

Yin, R. K. (2005). *Estudo de caso: planejamento e métodos*. Bookman, Porto Alegre, 3 edition.