Identificação de Perfis em Múltiplas Redes Sociais

Mariana Barreto[1], **Sérgio Lifschitz**[1]

[1]Departamento de Informática – (PUC-Rio) - Rio de Janeiro - RJ

marianabarreto@aluno.puc-rio.br,

sergio@inf.puc-rio.br

Resumo. *Este trabalho apresenta o desenvolvimento de uma ferramenta web projetada para identificar perfis de usuários em várias plataformas de redes sociais, como Twitter, Facebook e Instagram, empregando exclusivamente medidas de similaridade de texto. O estudo se concentra na utilização da distância de Levenshtein, uma métrica conhecida por medir a diferença entre sequências de caracteres, para avaliar a semelhança entre usernames e nomes de perfil, buscando determinar a eficácia dessa abordagem em produzir resultados precisos na identificação de perfis correspondentes entre diferentes plataformas. A pesquisa conduz uma análise do desempenho e da precisão dessa métrica, explorando e implementando estratégias para aprimorar esses aspectos. Além disso, descreve o processo de criação da ferramenta web, destacando como ela facilita a construção de um banco de dados que associa perfis digitais a indivíduos reais. A aplicação da distância de Levenshtein permite uma identificação mais eficiente e rápida de conexões entre perfis de diferentes redes sociais, potencializando o reconhecimento e a análise de presenças online de usuários em múltiplos ambientes digitais.*

1. Introdução

A ascensão das redes sociais transformou a forma de se comunicar, compartilhar informações e até mesmo de se apresentar ao mundo. Plataformas como Facebook, Twitter, Instagram, LinkedIn, entre outras, oferecem janelas únicas para as vidas pessoais, profissionais e criativas dos usuários. No entanto, essa fragmentação da presença digital em múltiplas redes sociais apresenta desafios significativos, especialmente no que diz respeito à análise de dados e à gestão de identidade online.

Comumente, as análises em redes sociais são realizadas de forma isolada para cada plataforma, ignorando que os perfis em diferentes redes podem representar uma única pessoa. Essa abordagem fragmentada não apenas limita a compreensão do comportamento e das preferências dos usuários, mas também ignora a complexidade de suas identidades digitais. Além disso, a ausência de uma ferramenta centralizada para identificar e associar perfis de um mesmo indivíduo em diversas plataformas dificulta significativamente a construção de um entendimento holístico da presença online de uma pessoa. Este cenário apresenta um desafio tanto para análises que buscam insights precisos quanto para a gestão de identidade e reputação online.

Diante desses desafios, diversas técnicas têm sido exploradas para identificar e associar perfis de um mesmo usuário em múltiplas redes sociais, visando superar a limitação

de análises baseadas em uma única plataforma. No entanto, até o momento, essas abordagens têm sido aplicadas de maneira ad hoc, sem uma solução generalizada que considere tanto a eficácia quanto a eficiência do processo.

O presente projeto propõe uma nova abordagem para abordar esse problema de forma mais sistemática. Ao invés de tratar perfis em diferentes redes sociais como representantes de diferentes indivíduos, nosso objetivo é desenvolver uma metodologia generalizada que reconheça a unidade da identidade digital através das plataformas. Para isso, propomos a criação de uma ferramenta web destinada a construir e gerenciar um banco de dados que associe de maneira eficiente perfis a pessoas. Essa ferramenta visa facilitar o processo de determinar quais perfis pertencem a uma determinada pessoa, permitindo análises multirredes mais integradas e profundas.

2. Trabalhos Relacionados

Trabalhos sobre a identificação de perfis em redes sociais têm muitos estudos significativos (e.g. [Vosecky et al. 2009] e [Li et al. 2018]). O trabalho de [Vosecky et al. 2009] estabeleceu as bases para a pesquisa científica, propondo uma abordagem que calcula a similaridade entre vetores de atributos genéricos, somando um valor global com pesos predefinidos. No entanto, essa técnica não abrange as nuances de novas redes sociais e atributos emergentes, como *usernames*, além de não considerar o desempenho do cálculo da similaridade de todos os atributos, especialmente em buscasem grandes volumes.

Os autores em [Li et al. 2018] focaram na análise de *usernames* e nomes de exibição, confirmando que esses atributos podem identificar perfis de um mesmo usuário em múltiplas redes sociais. No entanto, técnicas de aprendizado de máquina (*machine learning* ou ML) (e.g. [Shu et al. 2017], [Li et al. 2023], [Zafarani and Liu 2013], e [Zhang et al. 2014]) também foram exploradas. Apesar de sua ampla adoção, constatou-se que para nosso caso específico essas técnicas de ML não seriam satisfatórias devido aos altos custos de treinamento e latência na entrega dos resultados.

Motivados por essas limitações, decidimos explorar técnicas diretas de similaridade, conforme analisado por [Sackers et al. 2017]. Essas técnicas, embora menos complexas, demonstraram ser viáveis e preferíveis em certos contextos, oferecendo uma alternativa eficaz e de fácil implementação. Acreditamos que as metodologias desenvolvidas neste trabalho contribuem significativamente para o campo, oferecendo uma perspectiva complementar às abordagens baseadas em ML e servindo como base valiosa para futuras pesquisas que integrem técnicas de ML e não-ML.

3. Método proposto

Neste trabalho exploramos dinâmicas e tendências em redes sociais selecionadas com base na disponibilidade e relevância dos dados. Optamos pelo Twitter, Facebook e Instagram devido à acessibilidade de suas APIs e sua notoriedade entre usuários influentes como políticos e artistas. O Reddit foi descartado devido à dificuldade na obtenção de dados e à predominância de perfis anônimos [Gagnon 2013]. O TikTok também foi excluído devido à insuficiência de dados coletados até então.

Inicialmente, utilizamos a API acadêmica do Twitter para acessar informações detalhadas dos perfis dos usuários. Após mudanças nas políticas de acesso em maio de

2023, recorremos aos dados disponíveis na base da ferramenta eTC[1]. Para Facebook e Instagram, utilizamos a API do CrowdTangle, reconhecida pela Meta, que proporcionou um conjunto relevante de perfis (105,575 do Facebook e 48,864 do Instagram).

3.1. Escolha do algoritmo de cálculo das distâncias entre os textos

O processo de escolha do método para calcular a distância entre textos envolveu uma análise detalhada de diferentes algoritmos, cada um com suas peculiaridades e aplicações. Entre os métodos considerados, destacam-se a distância de Levenshtein, a semelhança de Jaccard e a similaridade de Cosseno. Em relação à complexidade, todos possuem a mesma e, por isso, esse fator não influenciou na escolha.

A distância de Levenshtein mede o número mínimo de operações (e.g. substituições) necessárias para transformar uma string em outra. É particularmente útil para aplicações que exigem a correção de erros de digitação ou a comparação de textos com pequenas variações. Por exemplo, a distância entre as palavras SOCIAL e FOBIA é 3, pois basta substituir o C pelo B, o S pelo F e remover o L no final. Já a semelhança de Jaccard compara membros de conjuntos para determinar a similaridade e diversidade entre eles. Aplica-se ao dividir o tamanho da interseção pelo tamanho da união dos conjuntos de itens. É frequentemente usada para comparar documentos com base na presença (ou ausência) de palavras específicas. Por fim, a similaridade de cosseno calcula a similaridade entre dois vetores de um espaço vetorial, medindo o cosseno do ângulo entre eles. É amplamente utilizado em sistemas de recomendação e processamento de linguagem natural, sendo particularmente eficaz para comparar documentos em termos de conteúdo.

A distância de Levenshtein, mais especificamente sua forma mais genérica, a *Minimum Edit Distance*, foi identificada como a mais apropriada para fazer nossos cálculos. Nós a escolhemos pois ela apresenta uma vantagem única para o contexto de nomes de perfis e *usernames*: considerar a ordem em que os caracteres estão posicionados.

3.2. Análise do desempenho para o cálculo das distâncias

Após a escolha do algoritmo, a segunda etapa foi implementar a análise sem nenhuma alteração, para visualizar o desempenho do método sem nenhuma alteração. De início, a eficiência logo se tornou uma questão, já que para fazer tais comparações, a complexidade é O(n^2), em que n é a quantidade de perfis. Como estamos testando apenas com a base do Twitter, que possui um pouco mais de 2,5 milhões de perfis, esse número ultrapassaria 10^{12} comparações.

Foi constatado que é necessário saber, portanto, o número de usuários razoável para esse tipo de análise. Por isso, foi simulado um teste de desempenho criando a matriz do cálculo da distância para uma quantidade diferente de usuários (10, 100, 1 000 e 10000). O tempo encontrado para 10000 usuários era um pouco maior do que uma hora, o que para uma análise automática é um tempo viável ainda que não seja imediato. No entanto, as buscas em redes sociais podem chegar a milhões de usuários e, de acordo com essa função, a projeção nesse caso seria de cerca de 3 anos. Sendo assim, a análise do cálculo de distância não poderia ser feita automaticamente para um número de usuários elevado. Mesmo que seja inviável executar o algoritmo para uma busca genérica com

[1]https://etc.biobd.inf.puc-rio.br/

n usuários, a análise continua relevante ao buscar possíveis perfis a partir de um nome específico. Nesse cenário, a complexidade passa a ser O(n).

3.3. Levenshtein Modificado para Aumentar a Precisão do Resultado

Para avaliar a precisão do algoritmo, utilizamos um conjunto de dados contendo indicações de acertos e erros em que fosse possível indicar quando o algoritmo escolheu corretamente ou não o par de nome de perfil e *username*. Utilizamos o próprio conjunto de dados de perfis do X, pois essa rede social possui tanto o atributo com o nome de perfil quanto o *username*. Dessa forma, os dados de uma única rede social puderam ter sido utilizados como um *golden conjunto de dados*.

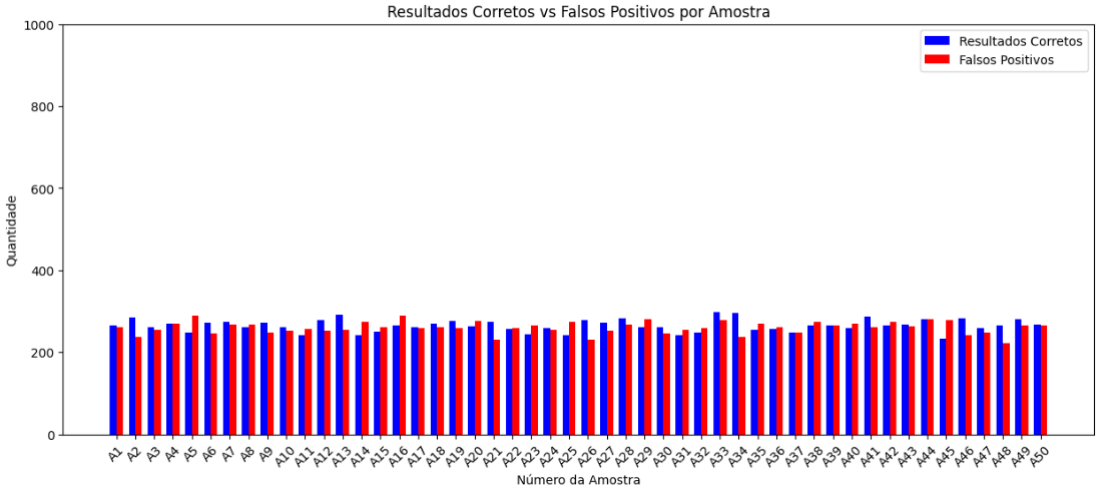

Figura 1. Resultados corretos e falsos positivos por amostra sem transformação

A análise da Figura 1 revela que, considerando as 50 amostras, a média é de 265 resultados corretos e 260 falsos positivos, com o desvio padrão de aproximadamente 14 para cada. O algoritmo exibe um desempenho consistente, embora modesto. A proximidade das médias de acertos e erros, junto ao baixo desvio padrão, indica uma precisão uniforme mas também sugere que sua capacidade de distinção é ligeiramente superior à de uma escolha aleatória. Essa observação sublinha que, apesar de sua consistência, o algoritmo ainda necessita de aprimoramentos para alcançar uma confiabilidade efetiva.

Foram propostas três modificações na distância de Levenshtein para melhorar os resultados. Primeiro, aumentamos o custo da substituição na *Minimum Edit Distance*, partindo da hipótese de que usuários de redes sociais tendem mais a suprimir ou adicionar caracteres ao nome do que a substituir (e.g., Mariana Porto Barreto para maripbarreto, José Silva para josesilva92). Segundo, aplicamos uma função para padronizar nomes de perfis para se assemelharem mais a *usernames*, realizando alterações como transformar todos os caracteres em minúsculo, substituir espaços por ”_”e remover outros símbolos não aceitos na criação de *usernames*. Por fim, para reduzir falsos positivos, determinamos uma variação máxima aceitável na distância de Levenshtein entre o nome de perfil e o *username*, proporcional ao número de caracteres do *username*. Assim, quanto menor a distância, menor a variação permitida.

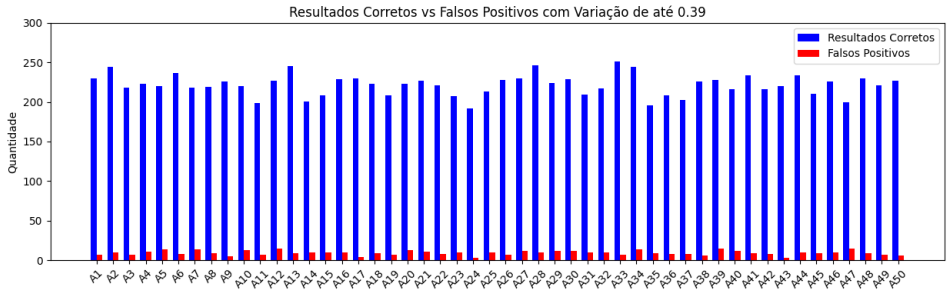

Figura 2. Figura com a quantidade de resultados corretos e falsos positivos por amostra após o aumento do custo de substituição, transformação de nome de perfil para username e variação de 0,39

A Figura 2 apresenta o resultado encontrado após todas as transformações mencionadas. Houve uma redução significativa na proporção de resultados corretos para faltos positivos que ficou em cerca de 5%. Já a quantidade média de resultados corretos encontrada foi de 221,62, o que é apenas 16,37% menor do que o resultado sem as modificações. Por mais que a quantidade de resultados corretos tenha diminuído, isto é compensado devido à redução da quantidade de falsos positivos.

4. Implementação da Interface

Foi desenvolvida também uma aplicação web para construir coletivamente um conjunto de dados de nomes e seus perfis associados. É possível cadastrar os perfis de uma determinada pessoa e, caso não saiba, pesquisar os possíveis perfis desta pessoa nas redes, utilizando nosso algoritmo de Levenshtein modificado.

Figura 3. Página de resultados encontrados que utiliza o algoritmo modificado da distância de Levenshtein para retornar possíveis perfis associados ao nome pesquisado

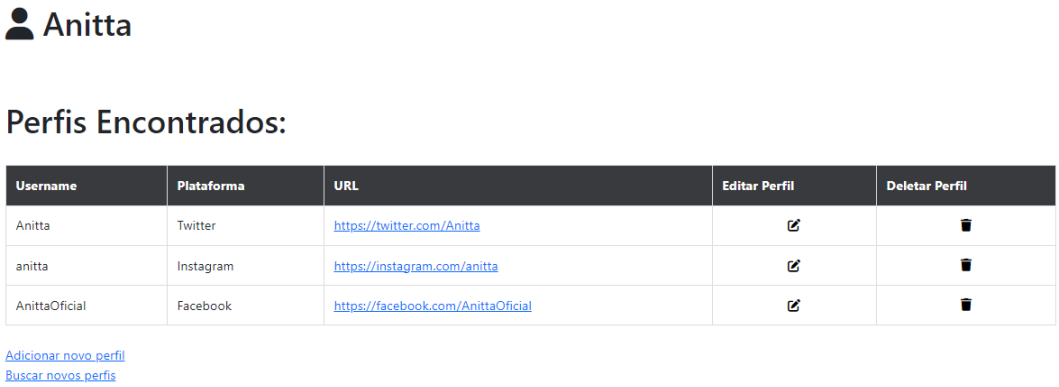

Figura 4. Página com os perfis encontrados caso a pessoa pesquisada já tenha sido cadastrada previamente

As Figuras 3 e 4 apresentam os dois caminhos possíveis após pesquisar um nome na página inicial. O primeiro seria no caso do nome não ter sido cadastrado previamente. É feita uma busca utilizando o nosso algoritmo em todas as redes sociais procurando os principais resultados de *usernames* ou nomes de perfis com o nome digitado. Já o segundo caminho ocorre quando o nome pesquisado corresponde a uma pessoa já cadastrada previamente. Nesse caso, o usuário é redirecionado para a página da pessoa com os perfis encontrados. Há a possibilidade também de cadastrar novos perfis, como ilustrado na página similar à da Figura 3.

5. Conclusões

Este trabalho apresenta um avanço na identificação de perfis em diversas redes sociais, focando na eficácia e eficiência do processo. Utilizando a distância de Levenshtein como métrica de similaridade textual e implementando técnicas para otimizar seu cálculo, demonstramos a viabilidade de obter bons resultados apenas com similaridades textuais. Além disso, desenvolvemos uma ferramenta web genérica, capaz de operar em redes sociais como Twitter, Facebook e Instagram, destacando a flexibilidade e aplicabilidade do método. Esta ferramenta não só facilita a consulta de perfis em diferentes plataformas, como também contribui para a construção de um valioso conjunto de dados para análises multirredes.

Os resultados obtidos com a distância de Levenshtein validam nossa abordagem, mostrando que uma métrica simples pode alcançar precisão satisfatória na identificação de perfis. Este sucesso inicial sugere um grande potencial para o refinamento e expansão das técnicas. No entanto, enfrentamos desafios no desempenho ao lidar com análises multirredes em escala massiva, evidenciando limitações de processamento com milhões de resultados. Isso ressalta a necessidade de pesquisas futuras focadas na melhoria do desempenho e escalabilidade das soluções propostas. Em conclusão, este estudo representa um passo importante para uma compreensão mais profunda e uma abordagem mais eficiente na identificação de perfis em redes sociais, incentivando a continuidade da pesquisa para superar os desafios de desempenho e explorar plenamente o potencial das técnicas de identificação de perfis multirredes.

Referências

Gagnon, T. (2013). The disinhibition of reddit users. *Adele Richardson's Spring*.

Li, Y., Peng, Y., Zhang, Z., Yin, H., and Xu, Q. (2018). Matching user accounts across social networks based on username and display name. *World Wide Web*, 22(3):1075–1097.

Li, Z., Lin, D., and Li, P. (2023). Across online social network user identification based on usernames. In Jiang, X., editor, *Machine Learning and Intelligent Communication*, pages 117–127, Cham. Springer Nature Switzerland.

Sackers, M., de Vries, A. P., and de Boer, M. H. (2017). A comparison of string distance metrics on usernames for cross-platform identification.

Shu, K., Wang, S., Tang, J., Zafarani, R., and Liu, H. (2017). User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter*, 18(2):5–17.

Vosecky, J., Hong, D., and Shen, V. Y. (2009). User identification across multiple social networks. In *2009 First International Conference on Networked Digital Technologies*. IEEE.

Zafarani, R. and Liu, H. (2013). Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 41–49.

Zhang, H., Kan, M.-Y., Liu, Y., and Ma, S. (2014). Online social network profile linkage. In *Information Retrieval Technology: 10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014. Proceedings 10*, pages 197–208. Springer.