Inteligência Artificial Sustentável baseado em Engenharia de Dados, Aprendizado de Máquina e Transferência de Conhecimento para Processamento de Linguagem Natural

Washington Cunha[1], **Leonardo Rocha**[2], **Marcos A. Gonçalves**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais

[2]Departamento de Ciência da Computação – Universidade Federal de São João del Rei

{washingtoncunha,mgoncalv}@dcc.ufmg.br, lcrocha@ufsj.edu.br

Resumo. *Grandes Modelos de Linguagem (GMLs), baseados em técnicas de Inteligência Artificial, têm transformado o Processamento de Linguagem Natural (PLN), sendo referência em tarefas como classificação de texto, análise de sentimentos, sumarização e perguntas-e-respostas. No entanto, sua construção e adaptação exigem alto custo computacional, demandando infraestrutura especializada e grande consumo energético, o que acarreta impactos ambientais negativos, como a emissão de CO_2. O modelo atual adotado pelos grandes players – baseado na "Lei do Mais" (mais dados, mais hardware, mais energia) – é insustentável e pouco viável para países com recursos limitados, como o Brasil, dificultando a competitividade internacional. Neste tutorial, propomos uma alternativa a essa abordagem dominante, focando em soluções inovadoras baseadas em engenharia de dados e técnicas de IA avançada. O objetivo é aumentar a eficiência dos modelos, reduzindo os custos computacionais e o consumo energético, contribuindo para um desenvolvimento mais sustentável e acessível.*

1. Introdução

Nas últimas décadas, a Web transformou radicalmente o papel dos usuários, que passaram de simples consumidores a produtores ativos de conteúdo. Esse processo resultou em um crescimento exponencial do volume de dados disponíveis online – sobretudo nas redes sociais –, tornando cada vez mais desafiadora a tarefa de localizar informações específicas com precisão e eficiência. Nesse cenário, técnicas de Processamento de Linguagem Natural (PLN) têm se mostrado essenciais, principalmente para o tratamento de dados textuais, ainda predominantes na Web. Aplicações como Classificação Automática de Documentos (CAT) e Análise de Sentimento (AS) exemplificam a relevância dessas técnicas para transformar grandes volumes de texto em informação útil e estruturada.

O PLN avançou significativamente na última década, impulsionado por modelos de Aprendizado Profundo, especialmente os baseados em arquiteturas *Transformers*, como RoBERTa e BART [Cunha et al. 2023a], e mais recentemente pelos Grandes Modelos de Linguagem (GMLs), como GPT e LLama [Cunha et al. 2025b]. Esses modelos representam o estado da arte em diversas tarefas, como recuperação de informação, ranqueamento, dentre outros. No entanto, seu desempenho excepcional vem acompanhado de custos elevados – não apenas computacionais e financeiros, mas também ambientais.

Tipo de Tutorial: Tutorial inédito – Avançado – Apresentação em Português. **Público Alvo:** Pesquisadores, profissionais e desenvolvedores de IA interessados em PLN, especialmente atuantes em contextos com recursos computacionais limitados e em busca de soluções mais eficientes, sustentáveis e acessíveis.

Diante desse cenário, predominam: (1) o aumento da complexidade dos modelos, com mais camadas e parâmetros, e (2) o uso de quantidades cada vez maiores e mais diversificadas de dados de treinamento. Ambas, entretanto, resultam em uma escalada dos custos e levantam sérias questões éticas, como o uso de dados pessoais ou protegidos por direitos autorais. Esse paradigma, amplamente adotado por grandes corporações, tem sido informalmente denominado de "Lei do Mais": mais dados, mais hardware, mais energia. Embora eficaz para atores com grandes recursos – como empresas nos EUA ou na China –, trata-se de uma estratégia insustentável e pouco acessível a países como o Brasil, que enfrentam limitações estruturais, financeiras e humanas.

Diante desse contexto, este tutorial propõe uma alternativa concreta à "Lei do Mais", voltada à construção de uma Inteligência Artificial Sustentável, baseada em princípios de eficiência, acessibilidade e responsabilidade ambiental, ao qual consideramos primordial para a comunidade nacional no desenvolvimento de soluções de baixo custo – financeiro, computacional e ambiental – voltadas à criação e ajuste-fino de Grandes Modelos de Linguagem para tarefas de PLN e Recuperação de Informação.

Especificamente, nossa proposta se fundamenta em duas frentes principais: (1) Engenharia de dados, com ênfase em técnicas de pré-processamento [Siino et al. 2024] e seleção de instâncias [Cunha et al. 2023a, Pasin et al. 2024], voltadas à melhoria da qualidade dos dados e à redução do volume necessário para o treinamento dos modelos; e (2) Estratégias de Aprendizado de Máquina e Transferência de Conhecimento, como: (i) compressão de modelos[Nardini et al. 2023], que visa reduzir a complexidade e o tamanho das redes neurais profundas; e (ii) aprendizado ativo[Bianco et al. 2023], que busca otimizar o processo de treinamento com o uso eficiente de exemplos anotados por humanos (*human-in-the-loop*), minimizando o esforço manual e o custo computacional. Essas estratégias buscam tornar o desenvolvimento de modelos mais acessível para contextos com recursos limitados, ao mesmo tempo em que promovem práticas mais éticas e ambientalmente responsáveis. Acreditamos que essa abordagem representa um caminho promissor para fortalecer a atuação da comunidade brasileira de IA, ampliando sua competitividade e capacidade de inovação em um cenário global cada vez mais exigente.

2. Sumário do tutorial e descrição dos tópicos

Evolução: De Métodos Tradicionais a SLMs e LLMs Nas últimas décadas, a PLN consolidou-se como um conjunto de técnicas fundamentais no enfrentamento da sobrecarga informacional da Web [Cunha et al. 2021]. Inicialmente, abordagens tradicionais de PLN baseavam-se em algoritmos supervisionados, como SVMs e Random Forests, que operavam sobre representações simples do texto, como TF-IDF. Embora eficientes, esses métodos exigiam forte engenharia de atributos e apresentavam limitações na captação de relações semânticas profundas. O surgimento dos *Small Language Models* (SLMs), como RoBERTa e BART, marcou um avanço significativo, ao integrar aprendizado profundo à tarefa de classificação textual com desempenho muito superior. Mais recentemente, os *Large Language Models* (LLMs), como GPT e LLama, elevaram o estado da arte ao utilizarem massivos volumes de dados e parâmetros, alcançando resultados impressionantes em uma ampla gama de tarefas de PLN. Este módulo abordará essa trajetória evolutiva, discutindo vantagens, limitações e implicações dessas diferentes gerações de modelos. Além disso, nesse módulo, iremos apresentar implementações práticas de modelos a serem disponibilizados à audiência.

Compromisso entre Efetividade e Custo: Avaliando o Custo-Benefício dos Modelos Apesar de os LLMs apresentarem ganhos de efetividade em diversas tarefas, como sumarização e classificação automática de textos, esses avanços nem sempre justificam os custos envolvidos. Estudos recentes indicam que, embora os LLMs superem os métodos tradicionais e os SLMs, os ganhos sobre os SLMs são frequentemente modestos – com melhora de até 4,9% - e acompanhados de um aumento expressivo no custo computacional, chegando a ser até 590 vezes maior que o de métodos tradicionais. Este módulo discutirá em profundidade esse trade-off entre efetividade e custo, considerando aspectos como tempo, consumo energético e emissão de carbono. Serão apresentados cenários práticos nos quais o uso de LLMs, SLMs ou métodos tradicionais se mostra mais adequado, com base nas demandas específicas de desempenho e restrições de recursos. Além disso, nesse módulo, iremos apresentar uma abordagem prática para mensurar o compromisso entre efetividade e custo (monetário e emissão de gases estufa).

A "Lei do Mais": Um Paradigma Insustentável na IA A atual corrida por resultados cada vez mais expressivos em tarefas de PLN tem levado grandes corporações a adotarem estratégias baseadas na chamada "Lei do Mais": mais dados, mais parâmetros, mais hardware e mais energia. Essa abordagem dominante foca no escalonamento de modelos e volumes de dados, frequentemente desconsiderando questões éticas e ambientais, como o uso de dados pessoais ou protegidos por direitos autorais e o impacto energético crescente. Embora eficaz para atores com recursos quase ilimitados – como empresas nos EUA e na China –, trata-se de um modelo economicamente inviável e ambientalmente insustentável para realidades como a brasileira. Este módulo analisará os limites e riscos dessa lógica de expansão contínua e discutirá por que ela representa um caminho pouco promissor para países com restrições estruturais e orçamentárias, incentivando a reflexão sobre modelos alternativos mais acessíveis e responsáveis.

Alternativas à Lei do Mais: Eficiência e Inovação com Recursos Limitados Diante da insustentabilidade do modelo hegemônico baseado na "Lei do Mais", surgem alternativas mais alinhadas com realidades de menor capacidade computacional e orçamentária. Embora abordagens como o desenvolvimento de novos modelos de deep learning ou o uso de hardwares especializados (e.g., TPUs e GPUs) tragam avanços, seu alto custo ainda os torna inacessíveis para muitos contextos. Este módulo propõe um caminho mais viável: a engenharia de dados aliada a técnicas avançadas de Aprendizado de Máquina e Transferência de Conhecimento. Estratégias como seleção de instâncias, aprendizado ativo, poda de modelos (pruning) e destilação de conhecimento (distillation) ganham destaque por reduzirem a complexidade computacional sem comprometer significativamente a performance. Esses métodos, foco central do tutorial, serão explorados como ferramentas-chave para viabilizar o desenvolvimento de modelos eficientes, éticos e sustentáveis - tornando possível a inovação em IA mesmo em ambientes com recursos limitados, como pequenas e médias empresas e grupos de pesquisa.

Técnicas de Otimização e Adaptação: Aprendizado Ativo, Poda, Destilação, Quantização e Métodos LoRA/QLoRA Diversas técnicas de otimização e adaptação têm sido fundamentais para viabilizar o uso de LLMs, incluindo aprendizado ativo, que reduz o esforço de anotação ao selecionar exemplos mais informativos; poda de modelos, que remove conexões menos relevantes para diminuir a complexidade sem perda significativa de desempenho; destilação de conhecimento, que transfere o aprendizado de mode-

los grandes para menores mantendo boa performance; e quantização, que reduz a precisão dos pesos para acelerar a inferência e economizar memória. Além disso, métodos recentes como LoRA e QLoRA possibilitam o ajuste-fino eficiente de LLMs em ambientes com recursos limitados. Este módulo abordará como essas estratégias podem tornar o uso e a personalização de LLMs mais viáveis, sustentáveis e alinhados à realidade da pesquisa.

Seleção de Instâncias: Otimizando Dados para Modelos Sustentáveis A seleção de instâncias desponta como uma estratégia promissora para tornar o desenvolvimento de LLMs mais acessível, eficiente e sustentável. Diferente do paradigma da "Lei do Mais", essa abordagem visa reduzir o volume de dados de treinamento de forma inteligente, mantendo - e em muitos casos até melhorando - a eficácia dos modelos por meio da remoção de ruídos e redundâncias. Além de contribuir diretamente para a diminuição do tempo de processamento e do consumo energético, a seleção de instâncias também pode ampliar a explicabilidade dos modelos gerados, tornando-os mais interpretáveis e confiáveis. Este módulo apresentará os fundamentos e aplicações práticas dessa técnica, destacando seu papel central no ajuste-fino de GMLs e sua relevância como pilar da engenharia de dados voltada à construção de uma IA mais sustentável e alinhada à realidade brasileira. Iremos apresentar também um pacote completo de abordagens de seleção de instâncias à ser compartilhado com a audiência.

Resultados Alcançados: Redução de Dados com Efetividade e Sustentabilidade As técnicas de seleção de instâncias desenvolvidas recentemente, como E2SC [Cunha et al. 2023b] e biO-IS [Cunha et al. 2025a], demonstraram resultados expressivos e promissores rumo a uma IA mais eficiente e sustentável. Em avaliações comparativas com 13 métodos de referência do estado da arte em Classificação Automática de Textos (ATC), utilizando 22 conjuntos de dados e grandes modelos linguagem (como BERT, RoBERTa e LLaMA), as propostas conseguiram reduzir os conjuntos de treinamento em até 60% sem comprometer a eficácia. A técnica biO-IS, em particular, superou todas as abordagens anteriores ao eliminar redundâncias e ruídos, alcançando acelerações médias de 1,67x (e máximas de até 2,46x) e estabelecendo-se como o novo estado da arte em seleção de instâncias para PLN. Esses resultados comprovam que é possível treinar modelos avançados com dados mais representativos, e não necessariamente maiores volumes, promovendo economia de recursos e redução das emissões de carbono - um passo concreto em direção a uma IA mais verde, acessível e eficiente. Por fim, compartilharemos um *benchmark* que compreende códigos, documentação, resultados, datasets e partições, permitindo comparação direta e avanços adicionais pela comunidade.

3. Interesse e potencial atração do público alvo

Este tutorial é relevante porque oferece alternativas práticas e sustentáveis ao atual modelo dominante em IA, que é muitas vezes inacessível para pesquisadores e profissionais com recursos limitados. Ao focar em técnicas de Engenharia de Dados, Aprendizado de Máquina e Transferência de Conhecimento como seleção de instâncias, compressão de modelos e aprendizado ativo, o tutorial capacita na prática o público a desenvolver soluções eficazes com menor custo computacional, financeiro e ambiental. Além disso, apresenta métodos estado da arte já validados experimentalmente, demonstrando que é possível inovar e competir globalmente mesmo fora dos grandes centros tecnológicos.

4. Breve Curriculum Vitae dos Proponentes

Washington Cunha é doutor em Ciência da Computação pela UFMG (2024), com mestrado pela mesma instituição (2019, Menção Honrosa no CTDBD-SBBD). Atua em Recuperação de Informação, Aprendizado de Máquina e PLN, com publicações em conferências e periódicos de destaque, como ACM SIGIR e ACM CSUR. É professor substituto e pós-doutorando no DCC-UFMG em projeto sobre IA Generativa na Saúde e pesquisador associado do INCT-TILD-IAR, onde desenvolve pesquisas em IA sustentável.

Leonardo Chaves Dutra da Rocha é Professor Titular da Universidade Federal de São João Del Rei, com formação em Ciência da Computação pela UFMG (graduação, mestrado, doutorado), incluindo pós-doutorado na Ohio State University e visita técnica na University of Connecticut. Atua nas áreas de inteligência artificial, PLN, mineração de dados, banco de dados e recuperação de informação, com mais de 240 publicações. Recebeu diversos prêmios, incluindo o Prêmio Capes de Tese (2024), e coordena projetos financiados pelo CNPq e FAPEMIG. É bolsista de produtividade em pesquisa (PQ-2).

Marcos André Gonçalves possui graduação em Ciência da Computação pela Universidade Federal do Ceara (1995), mestrado pela Universidade Estadual de Campinas (1997), doutorado pela Virginia Polytechnic Institute and State University (Virginia Tech) e pós-doutorado pela Universidade Federal de Minas Gerais (2006) e pela Politécnico di Torino (IT, 2024). Atua em recuperação de informação, aprendizado de máquina e PLN. Recebeu diversos prêmios, incluindo dois Prêmios CAPES de Tese (2024 e 2020) e distinções por melhores artigos. Foi Membro Afiliado da ABC, é bolsista de produtividade do CNPq (nível 1-A), ex-membro da CEX/FAPEMIG e atual coordenador do INCT-TILD-IAR.

Referências

Bianco, G. D., Duarte, D., and Gonçalves, M. A. (2023). Reducing the user labeling effort in effective high recall tasks by fine-tuning active learning. *IIS*, 61(2):453–472.

Cunha, W. et al. (2023a). A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. *ACM Comput. Surv.*

Cunha, W., França, C., Fonseca, G., Rocha, L., and Gonçalves, M. A. (2023b). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *ACM SIGIR*, pages 665–674.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., et al. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *IP&M*.

Cunha, W., Moreo Fernández, A., Esuli, A., Sebastiani, F., Rocha, L., and Gonçalves, M. A. (2025a). A noise-oriented and redundancy-aware instance selection framework. *ACM TOIS*, 43(2):1–33.

Cunha, W., Rocha, L., and Gonçalves, M. A. (2025b). A thorough benchmark of automatic text classification: From traditional approaches to large language models. *arXiv preprint arXiv:2504.01930*.

Nardini, F. M., Rulli, C., Trani, S., and Venturini, R. (2023). Neural network compression using binarization and few full-precision weights. *arXiv preprint arXiv:2306.08960*.

Pasin, A., Cunha, W., Goncalves, M., and Ferro, N. (2024). A quantum annealing instance selection approach for efficient and effective transformer fine-tuning. In *ACM ICTIR*.

Siino, M., Tinnirello, I., and La Cascia, M. (2024). Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers. *Inf. Sys.*, 121:102342.

* Esse trabalho foi financiado por CNPq, CAPES, INCT-TILD-IAR, AWS, FAPEMIG, Google, Finep e FAPESP.