“Mais com Menos”- Processamento de Linguagem Natural Inteligente e Sustentável baseado em Engenharia de Dados e Inteligência Artificial Avançada

Marcos André Gonçalves¹, Leonardo Rocha², Washington Cunha¹, Guilherme Dal Bianco³

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

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

³Universidade Federal da Fronteira Sul

mgoncalv@dcc.ufmg.br, lcrocha@ufsj.edu.br

washingtoncunha@dcc.ufmg.br, guilherme.dalbianco@uffs.edu.br

Resumo. *Grandes Modelos de Linguagem (GMLs), baseados em técnicas de Inteligência Artificial (IA), tem revolucionado o Processamento de Linguagem Natural (PLN), sendo considerados o estado-da-arte em diversas tarefas práticas de PLN tais como classificação de texto, análise de sentimentos, sumarização de textos, e sistemas de perguntas-e-respostas. No entanto, tanto a construção desses modelos pré-treinados, quanto a sua adaptação a tarefas específicas (i.e. ajuste fino), demandam um custo computacional elevadíssimo, exigindo hardware e infraestrutura energética especializados, com impactos ambientais negativos em termos de emissão de CO₂. Mais ainda, a infraestrutura necessária para trabalhar nesses desafios no modelo hoje utilizado pelos big players, o qual denominamos de “Lei do Mais” (mais dados, mais hardware, mais energia) é não apenas insustentável, mas também desinteressante do ponto de vista de competitividade nacional – não temos os recursos financeiros e humanos para competir em pé de igualdade. Nesse contexto, nossa proposta de desafio é confrontar a “Lei do Mais” por meio do desenvolvimento de soluções inovadoras baseadas em engenharia de dados e IA Avançada que aumentem a efetividade dos modelos enquanto reduzem os custos computacionais e o consumo energético, com impactos ambientais positivos.*

Abstract. *Large Language Models (LLMs), based on Artificial Intelligence (AI) techniques, have revolutionized Natural Language Processing (NLP) and are considered state-of-the-art for many practical NLP tasks such as text classification, sentiment analysis, text summarization, and question-answering systems. However, constructing and adapting these pre-trained models to specific tasks (i.e., fine-tuning) requires extremely high computational costs, demanding specialized hardware and energy infrastructure, with negative environmental impacts regarding CO₂ emissions. Furthermore, the infrastructure needed to address these challenges under the current paradigm used by big players – which we refer to as the “More is Better Law” (more data, more hardware, more energy) – is unsustainable and unappealing from a national competitiveness perspective. We lack the financial and human resources to compete on equal footing. In this context, our challenge proposal is to confront the “More is Better Law” by*

developing innovative solutions based on data engineering and Advanced AI that enhance model effectiveness while reducing computational costs and energy consumption, yielding positive environmental impacts.

Desafio

Desenvolver novas soluções *sustentáveis* que reduzam os *altos custos* computacionais, financeiros e de consumo energético, e consequente o impacto ambiental, associados à criação e ao ajuste-fino de Grandes Modelos de Linguagem (GMLs) utilizados em tarefas de PLN por meio de soluções avançadas de Engenharia de Dados e de Inteligência Artificial. Este desafio conecta 3 dos grandes temas da chamada da SBC: (1) IA; (2) Ciência dos Dados; e (3) Computação Sustentável. Potencialmente conecta também o tema (4) Computação Quântica pois existem soluções preliminares para este desafio (e.g. seleção de instâncias) implementadas por nosso grupo com esse paradigma computacional [7].

Argumentação

A Web permitiu que os usuários desempenhassem um papel crucial, não apenas no consumo, mas também na criação ativa de conteúdo. Isso vem resultando em volumes cada vez maiores de dados disponibilizados na Web (incluindo as Redes Sociais), dificultando a tarefa de encontrar de forma eficaz informações específicas. Organizar automática e adequadamente essas informações é portanto uma tarefa fundamental. Particularmente, no caso de dados textuais, ainda o principal tipo de dado encontrado na Web e nas redes sociais, tarefas de Processamento de Linguagem Natural (PLN), tais como Classificação Automática de Documentos e Análise de Sentimento tem se mostrado essenciais.

Técnicas de PLN testemunharam um avanço significativo na última década com propostas de estratégias inspiradas em arquiteturas de Redes Neurais (Aprendizado Profundo), incluindo as arquiteturas *Transformers* de 1ª e 2ª geração, tais como *RoBERTa* e *BART* [3, 5], e mais recentemente os Grandes Modelos de Linguagem (GMLs) (*Large Language Models* - LLMs) como *GPT* e *LLama* [10]. Os GMLs constituem o atual estado-da-arte em PLN, tendo alcançado resultados notáveis em diversas tarefas afim, incluindo recuperação de informação, ranqueamento (*ranking*), e perguntas-e-respostas (Q&A). Todavia, para alcançar esse notável desempenho, essas estratégias necessitam de grandes quantidades de dados no estágio de treinamento, muitas das vezes ruidosos e enviesados, impactando significativamente o custo computacional.

De fato, uma questão fundamental relacionada ao treinamento ou adaptação para tarefas específicas (i.e. ajuste-fino) de um modelo de aprendizagem profunda são os altos custos computacionais e de infraestrutura energética que, por sua vez, geram um grande impacto ambiental em termos de emissão de carbono. A quantidade de energia e tempo necessários para realizar o ajuste dos parâmetros para otimizar a eficácia dos modelos pode variar dependendo de diversos fatores, incluindo: (i) o tamanho e a complexidade do modelo; (ii) o hardware especializado utilizado, como GPU e TPU (consumo energético); e (iii) a quantidade de dados. Como a maior parte da eletricidade mundial é gerada a partir de combustíveis fósseis [9] o processo acima pode ser considerado diretamente responsável pela liberação de dióxido de carbono no meio ambiente. Nesse contexto, Patterson et al. [13] destacam que a fase de pré-treinamento do GPT-3 consumiu 1.287 MW/h. Esse consumo de energia resultou diretamente na emissão de pelo menos 552 toneladas de CO₂, equivalente ao consumo de um carro por cerca de 2,1 milhões de quilômetros.

Na busca constante por soluções mais efetivas, duas alternativas se destacam: (1) tornar os modelos mais complexos (e.g., mais parâmetros) ou (2) aumentar a quantidade e diversidade dos dados de treinamento, esbarrando por vezes em questões éticas, tais como o uso de dados pessoais ou com direitos preservados (*copyrighted*). O comum às duas alternativas é que ambas implicam no aumento de custos computacionais, energéticos e ambientais. Atualmente, esse é o modelo mais utilizado pelo *big players*, o qual denominamos de “Lei do Mais”: mais dados, mais hardware, mais energia. Trata-se de um modelo que não é apenas insustentável, mas também desinteressante do ponto de vista de competitividade do Brasil a nível internacional, uma vez que não dispomos dos mesmos recursos financeiros e humanos que nações como os EUA e China, tão pouco das grandes companhias de TI do Vale do Silício. Em suma, o cenário que ilustramos envolve volumes de dados em crescente expansão, requisitos constantes de retreinamento, orçamentos multi-milionários e modelos de custo computacional e de energia de alta demanda.

A alternativa que advogamos em nossa proposta de Desafio vai na contra-mão da “Lei do Mais”. De fato propomos como Grande Desafio para a comunidade nacional o investimento no desenvolvimento de soluções *sustentáveis* e de *baixo custo* (*financeiro, computacional e ambiental*) para a criação e ajuste-fino de Grandes Modelos de Linguagem utilizados em tarefas de Processamento de Linguagem Natural e de Recuperação de Informação. Particularmente, advogamos soluções baseadas em (1) engenharia de dados, incluindo pré-processamento [2, 14] e seleção de instâncias [3, 4, 7, 6], destinadas a melhorar a qualidade e reduzir o volume dos dados de treinamento; e (2) estratégias de IA Avançadas, tais como: (i) compressão de modelos [12] aplicadas para reduzir a complexidade dos modelos de aprendizado profundo e (ii) aprendizado ativo [1], que também foca na redução de treinamento com “humanos-no-circuito”(humans-in-the-loop).

Direcionamentos

Para enfrentar o desafio proposto, apontamos como direcionamento para a comunidade nacional o desenvolvimento de novas soluções de engenharia de dados, e.g. seleção de instâncias [4], aplicadas na construção ou ajuste-fino de GMLs que sejam capazes de otimizar *três restrições simultaneamente*: (i) reduzir o volume de dados de treinamento; (ii) manter (ou mesmo melhorar) a eficácia removendo ruídos e redundâncias do treinamento; e (iii) reduzir o tempo total de aplicação de um modelo de ponta a ponta (que inclui desde as etapas tradicionais de pré-processamento até a etapa de treinamento do modelo) e, conseqüentemente, a emissão total de CO₂. Em situações específicas podemos incluir um quarto requisito, que seria: (iv) permitir que os modelos sejam mais explicáveis.

Generalizando, é fato que basicamente todas as tarefas de PLN fundamentadas em processos de aprendizado de máquina supervisionado e IA enfrentam as questões acima. Dessa forma, soluções para o desafio deverão também beneficiar diversas outras áreas, tais como sistemas de busca e recuperação da informação, sistemas de perguntas-e-respostas e sistemas de recuperação baseados em aprendizado de ranqueamento.

Do ponto de vista de técnicas de inteligência artificial avançadas, destacamos três abordagens principais para o desafio: duas relacionadas a compressão de modelos [12]: (i) quantização, um conjunto de técnicas que reduz a precisão numérica dos pesos dos modelos, geralmente convertendo valores em ponto flutuante para inteiros de 8 ou 4 bits; e (ii) poda (*pruning*), um conjunto de técnicas que elimina partes menos relevantes

do modelo, como neurônios e suas conexões, funcionando como uma generalização do *dropout*. Ambas as abordagens resultam na redução do tamanho dos modelos, otimizando seu desempenho e eficiência. A terceira alternativa que acreditamos que tem bastante potencial no contexto de reduzir a “Lei do Mais” é o Aprendizado Ativo (AA) que no contexto supervisionado busca selecionar um conjunto mínimo de instâncias não-rotuladas que, quando rotuladas pelo usuário, produzem um modelo com eficácia igual (ou até mesmo superior) se comparado a um conjunto completo de dados [1]. Diferentemente da SI, o AA foca em dados não rotulados e utiliza conhecimento especializado do usuário para a criação de conjuntos de treinamento mínimos para o ajuste-fino de modelos.

Questões Objetivas

1 - Por que o desafio proposto é relevante para o Brasil? Um exemplo prático imposto pelo uso de GMLs é o de empresas e grupos de pesquisa com restrições orçamentárias, especialmente em países em desenvolvimento. Diferente de nações como os EUA e a China, ou de grandes companhias do Vale do Silício, que dispõem de vastos recursos financeiros e humanos, no Brasil enfrentamos limitações significativas em ambas as dimensões. Isso afeta a capacidade de aplicar esses modelos em larga escala. Além disso, é comum a necessidade de realizar milhares de experimentos para alcançar avanços científicos ou inovações práticas. Portanto, este projeto tem o potencial de expandir significativamente a utilização desses modelos, tornando-os mais acessíveis, ao mesmo tempo em que promove uma abordagem mais sustentável e computacionalmente eficiente.

2 - Por que é viável considerar a proposta um desafio? A redução dos custos de treinamento e ajuste de GMLs é uma questão de relevância global considerando seu impacto ambiental. Com o compromisso político adequado e a alocação de recursos necessários, o projeto é realizável dentro do período de dez anos, visto a tendência mundial tanto na busca pela redução de custos quanto no aumento da adoção dessas tecnologias.

Nosso grupo vem trabalhando em algumas das direções apontadas alcançando relevantes resultados que demonstram a viabilidade de enfrentar este desafio com tecnologia nacional. Destaca-se a tese de doutorado de Felipe Viegas, vencedora do Prêmio Capes 2024. A tese propõe um modelo simples, barato e criativo de representações de texto – Cluwords – que aproveita a eficiência e a interpretabilidade das representações (matriciais) tradicionais baseadas em frequências de palavras, ao mesmo tempo que explora as capacidades semânticas de modelos modernos baseados em *embeddings* de palavra. Esse modelo se mostrou tão ou mais eficaz que modelos do estado-da-arte baseado em Transformers, muito mais caros e complexos, em tarefas como Modelagem de Tópicos e Análise de Sentimentos [11, 15, 17, 16, 18]. Outro exemplo notável é a tese de doutorado de Washington Cunha [8], focada em *seleção de instâncias*, permitindo treinamento de modelos com conjuntos mínimos de dados, mantendo a efetividade. Mais especificamente, propomos e orquestramos novas etapas de pré-processamento baseadas em engenharia de dados aplicadas à PNL, oferecendo resultados tão eficazes quanto o estado-da-arte com um tempo de treinamento até 6,1 vezes menor. Esses exemplos demonstram a viabilidade de enfrentar o desafio, com fortes evidências de que o pré-processamento dos dados pode ser mais importantes que trabalhar em modelos de IA para alcançar resultados mais eficazes a menor custo.

3 - Quem deverá estar envolvido na sua solução? Nossa proposta convida toda a

comunidade científica brasileira a estar diretamente envolvida, com destaque especial para Computação. Pesquisadores, engenheiros e profissionais dessa área fornecerão as soluções tecnológicas essenciais para o desenvolvimento do projeto, desde a otimização de algoritmos até a criação de infraestruturas computacionais mais sustentáveis. Além disso, colaborações com outras áreas do conhecimento além da computação, como Matemática e Ciências Ambientais, são necessárias para integrar aspectos como eficiência energética e impacto ambiental nas soluções propostas. Além disso, a colaboração entre universidades, centros de pesquisa, empresas e órgãos governamentais será essencial para que a solução proposta se concretize e para que o Brasil se destaque no cenário global, não só pela adoção de novas tecnologias, mas também por sua capacidade de propor soluções sustentáveis para os desafios da computação de alta performance.

4 - Quais as iniciativas paralelas em outros Países? Diversos países ao redor do mundo já estão investindo em iniciativas para reduzir os custos e o impacto ambiental do treinamento de GMLs. Nos Estados Unidos, empresas como OpenAI, Google e Microsoft estão desenvolvendo técnicas de otimização que visam tanto a redução do consumo de energia quanto a melhoria na eficiência dos modelos. A criação de metodologias mais sustentáveis e a adoção de energias renováveis em data centers são algumas das medidas adotadas para mitigar os efeitos das emissões de carbono. Na Europa, países como Itália, Alemanha e França estão à frente em pesquisas que buscam alternativas mais verdes para o desenvolvimento de IA. Projetos como FAIR AI¹ incluem novos métodos de pré-processamento sustentáveis (*Green-aware AI*), oferecendo suporte ao desenvolvimento de IA de forma mais eficiente e com menor impacto ambiental.

Breve Curriculum Vitae dos Proponentes

M. A. Gonçalves é graduado pela UFC (1995), mestre pela UNICAMP (1997) e doutor pela Virginia Tech (2004), em Ciência da Computação. É professor titular da UFMG, pesquisador 1-B do CNPq e ex-membro afiliado da Academia Brasileira de Ciências.

L. Rocha é doutor em Ciência da Computação pela UFMG (2009), professor associado do Dept. de Ciência da Computação da UFSJ e Pesquisador 2 do CNPq.

W. Cunha é doutor em Ciência da Computação pela UFMG (2024) onde também obteve seu título de Mestre (2019) com Menção Honrosa no CTDBD-SBBD, possuindo publicações em importantes conferências e periódicos (e.g. *ACM SIGIR* e *ACM CSUR*).

G. Dal Bianco é doutor em C. Computação pela UFRGS (2014), e é professor adjunto na Univ. Fed. da Fronteira Sul (UFFS, 2015). Atualmente é pós-doutorando na UFMG.

Agradecimentos

Este trabalho contou com apoio do CNPq, CAPES, INCT-TILD-IAR, FAPEMIG, AWS, Google, NVIDIA, CIIASaúde, and FAPESP.

¹<https://fondazione-fair.it/en/>

Referências

- [1] G. D. Bianco, D. Duarte, and M. A. Gonçalves. Reducing the user labeling effort in effective high recall tasks by fine-tuning active learning. *IIS*, 61(2):453–472, 2023.
- [2] W. Cunha, S. Canuto, F. Viegas, T. Salles, C. Gomes, V. Mangaravite, E. Resende, T. Rosa, M. Gonçalves, and L. Rocha. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *IP&M*, 2020.
- [3] W. Cunha et al. A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. *ACM Comput. Surv.*, 2023.
- [4] W. Cunha, C. França, G. Fonseca, L. Rocha, and M. A. Gonçalves. An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *ACM SIGIR*, pages 665–674, 2023.
- [5] W. Cunha, V. Mangaravite, C. Gomes, S. Canuto, C. Nascimento, F. Viegas, C. França, W. S. Martins, J. M. Almeida, et al. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *IP&M*, 58(3):102481, 2021.
- [6] W. Cunha, A. Moreo Fernández, A. Esuli, F. Sebastiani, L. Rocha, and M. A. Gonçalves. A noise-oriented and redundancy-aware instance selection framework. *ACM Trans. Inf. Syst.*, 43(2), Jan. 2025.
- [7] W. Cunha, A. Pasin, M. Goncalves, and N. Ferro. A quantum annealing instance selection approach for efficient and effective transformer fine-tuning. In *ACM ICTIR*, 2024.
- [8] W. L. M. da Cunha. *A comprehensive exploitation of instance selection methods for automatic text classification*. PhD thesis, Aug. 2024. Available at <http://hdl.handle.net/1843/76441>.
- [9] B. Dudley et al. Bp statistical review of world energy 2016. *British Petroleum Statistical Review of World Energy*, Bplc. editor, Pureprint Group Limited, UK, 2019.
- [10] P. Liang et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification.
- [11] W. Luiz, F. Viegas, R. Alencar, F. Mourão, T. Salles, D. Carvalho, M. A. Gonçalves, and L. Rocha. A feature-oriented sentiment rating for mobile app reviews. In *WWW'18*, pages 1909–1918, 2018.
- [12] F. M. Nardini, C. Rulli, S. Trani, and R. Venturini. Neural network compression using binarization and few full-precision weights. *arXiv preprint arXiv:2306.08960*, 2023.
- [13] D. Patterson et al. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.
- [14] M. Siino, I. Tinnirello, and M. La Cascia. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers. *Inf. Sys.*, 121:102342, 2024.
- [15] F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, and M. A. Gonçalves. Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling. In *Proc. of the 12th ACM Int. Conf. on Web Search and Data Mining, WSDM '19*, page 753–761, 2019.
- [16] F. Viegas, S. D. Canuto, W. Cunha, C. França, C. M. V. de Andrade, G. Fonseca, A. Machado, L. Rocha, and M. A. Gonçalves. Pipelining semantic expansion and noise filtering for sentiment analysis of short documents - clusent method. *J. Interact. Syst.*, 15(1):561–575, 2024.
- [17] F. Viegas, W. Cunha, C. Gomes, A. Pereira, L. Rocha, and M. Goncalves. CluHTM - semantic hierarchical topic modeling based on CluWords. In *ACL*, pages 8138–8150, 2020.
- [18] F. Viegas, A. Pereira, W. Cunha, C. França, C. Andrade, E. Tuler, L. Rocha, and M. A. Gonçalves. Exploiting contextual embeddings in hierarchical topic modeling and investigating the limits of the current evaluation metrics. *Computational Linguistics*, pages 1–41, 03 2025.