Aumento de Dados utilizando LLMs

Matéria: Tópicos Avançados em Biologia Computacional
Autor: Matheus Muniz Damasco
Professora: Barbara Quintela

Projeto Viva sem Tabaco

O projeto Viva sem Tabaco é uma iniciativa promissora que busca apoiar e orientar indivíduos que desejam parar de fumar. Desenvolvido por especialistas da Universidade Federal de Juiz de Fora (UFJF), em parceria com o Centro de Referência em Pesquisa, Intervenção e Avaliação em Álcool e Outras Drogas (CREPEIA) e o Departamento de Ciência da Computação (DCC), o projeto oferece uma plataforma virtual disponível para toda a população [1]. A plataforma do Viva sem Tabaco tem como objetivo fornecer informações e ferramentas práticas para auxiliar no processo de cessação do tabagismo.

Natural Language Processing (NLP)

Processamento de Linguagem Natural (NLP) é uma área da Inteligência Artificial que capacita os computadores a compreender, interpretar e gerar textos em linguagem humana. Baseando-se em técnicas de Machine Learning e Deep Learning, NLP analisa grandes volumes de dados textuais para identificar padrões, extrair informações e realizar tarefas como tradução, classificação e análise de sentimentos [3]. Originalmente fundamentado em regras e heurísticas, evoluiu para métodos estatísticos e neurais capazes de lidar com dados não estruturados de forma eficiente [4]. Essa evolução viabiliza aplicações inovadoras em diversas áreas, desde comunicação e educação até diagnósticos e intervenções em saúde.

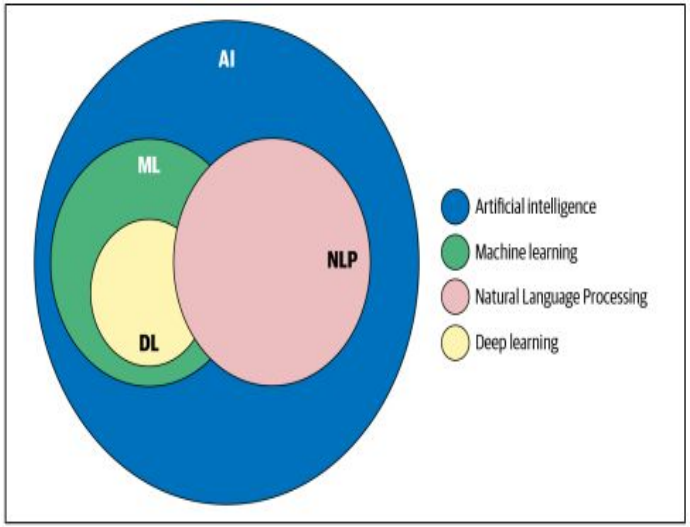

Figura 1: Como NLP, ML e DL estão relacionados [2].

Large Language Models (LLMs)

Os Modelos de Linguagem de Grande Escala (LLMs) são um tipo de modelo de Inteligência Artificial criado para entender e gerar texto. Esses modelos são treinados em grandes volumes de dados da internet, aprendendo padrões sobre como as palavras e frases são comumente usadas juntas. Quando alimentado com uma nova entrada de texto, um LLM tentará prever ou gerar a continuação mais provável desse texto com base no que aprendeu durante o treinamento [5].

Benefícios dos LLMs no Campo de Saúde Mental

Os LLMs oferecem múltiplos benefícios para o campo da Saúde Mental por exemplo, sua capacidade eficiente de recuperação e sumarização de informações a partir de extensos conjuntos de dados de diversas fontes (por exemplo, registros eletrônicos de saúde, interações em dispositivos móveis, plataformas de mídia social, etc. [6]) pode fornecer aos clínicos insights sobre comportamentos e experiências dos pacientes. Isso pode auxiliar em estratégias de intervenção precoce e planos de tratamento personalizados. Além disso, por meio de interações conversacionais, eles têm o potencial de oferecer um meio relacionável e acessível para que os indivíduos articulem seus estados emocionais e experiências pessoais [7]. Isso auxilia tanto na auto expressão do usuário quanto potencialmente aprimora os processos terapêuticos clínicos, caso possam ser validados com sucesso e integrados aos fluxos de trabalho clínicos [8].

Desafios dos LLMs no Campo de Saúde Mental

Apesar dos benefícios que os LLMs podem apresentar ainda existem desafios e problemas a serem levados em conta como a falta de conjuntos de dados especializados e multilíngues anotados por especialistas, preocupações quanto à precisão e confiabilidade do conteúdo gerado, desafios na interpretabilidade devido à natureza "caixa preta" dos LLMs, problemas de privacidade dos dados e o potencial de uma dependência excessiva dos LLMs tanto por parte dos médicos quanto dos pacientes, o que poderia comprometer as práticas médicas tradicionais [9]. Como podemos observar os LLMs não devem ser considerados substitutos aos serviços profissionais de saúde mental.

A Escassez dos Dados Públicos de Texto

No artigo ‘Will we run out of data? Limits of LLM scaling based on human-generated data’, estimou-se que o estoque de textos públicos humanos de alta qualidade atingirá seu ponto médio de esgotamento em 2028, com alta probabilidade de esgotamento completo até 2032 [10].

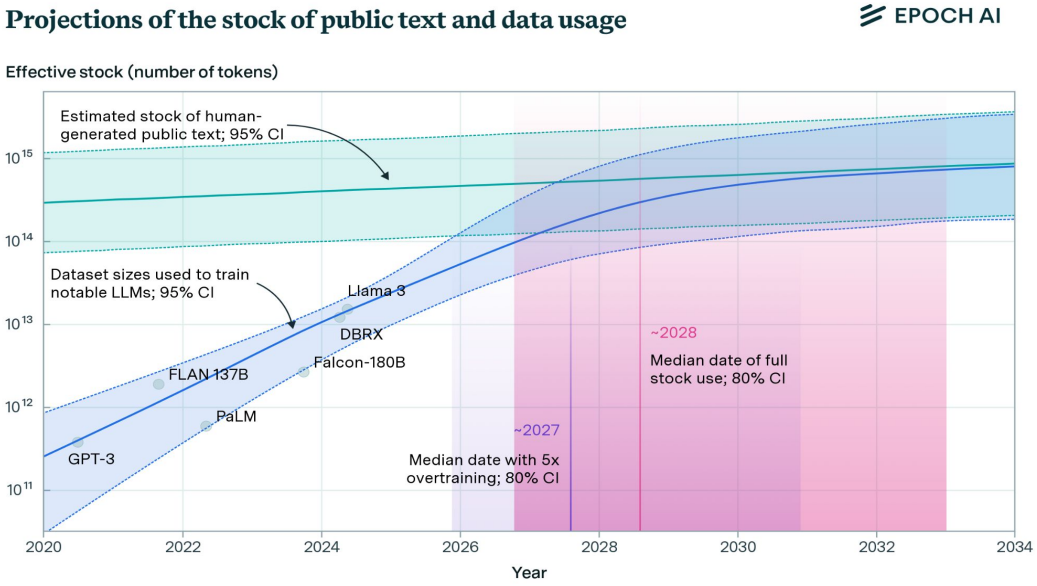

Figura 2: Projeção de estoque efetivo de texto público gerado por humanos e tamanhos de conjuntos de dados usados para treinar LLMs notáveis. Pontos individuais representam tamanhos de conjuntos de dados de modelos notáveis específicos. A projeção de tamanho de conjunto de dados é uma mistura de uma extrapolação de tendências históricas e uma projeção baseada em computação que assume que os modelos são treinados de forma otimizada para computação [10].

Aumento de Dados

O aumento de dados é o processo de gerar artificialmente novos dados a partir de dados existentes, principalmente para treinar novos modelos de Inteligência Artificial (IA). Os modelos de IA exigem conjuntos de dados grandes e variados para o treinamento inicial, mas o fornecimento de conjuntos de dados reais suficientemente diversos pode ser um desafio devido aos data silos, regulamentações e outras limitações. O aumento de dados aumenta artificialmente o conjunto de dados fazendo pequenas alterações nos dados originais. As soluções de inteligência artificial generativa (IA) agora estão sendo usadas para aumento rápido e de alta qualidade de dados em vários setores [11].

Aumento de Dados usando LLMs

Baseado no artigo ‘AugGPT: Leveraging ChatGPT for Text Data Augmentation’ [12], meu estudo propõe investigar o uso de diferentes Grandes Modelos de Linguagem (LLMs) como GEMMA, LLAMA e outros, para o aumento de dados textuais. O objetivo é comparar o desempenho de diversos LLMs para a tarefa de aumento de dados, analisando métricas como Cosine Similarity [13] e TransRate [14] para avaliar a fidelidade (ou seja, se as amostras de dados geradas estão próximas das amostras originais) e a compactação (ou seja, se as amostras de cada classe são compactadas o suficiente para boa discriminação) dos dados aumentados. Caso os resultados se mostrem eficazes, esse método poderá ser utilizado posteriormente na minha dissertação de Mestrado, com foco na aumento de dados do Programa Viva sem Tabaco.

Referências

[1] Gomide, H.P.: Desenvolvimento e avaliação de uma intervenção para tabagismo mediada por internet. Ph.D. thesis, UNIVERSIDADE FEDERAL DE JUIZ DE FORA (2014)

[2] Vajjala, S., Majumder, B. P., Gupta, A., & Surana, H. (2020). Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media. ISBN: 978-1492054054.

[3] Jurafsky, D. & Martin, J. H. (2020). Speech and Language Processing (3ª ed.). Prentice Hall.

[4] Manning, C. D. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

[5] O Que São Large Language Models (LLMs)? - https://blog.dsacademy.com.br/o-que-sao-large-language-models-llms/

[6] Choudhury, M. D., Pendse, S. R. & Kumar, N.Benefits and harms of large language models in digital mental health (2023).2311.14693.

[7] Ma, Z., Mei, Y. & Su, Z.Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support (2023).2307.15810.

[8] Hua, Y., Liu, F., Yang, K., Li, Z., Sheu, Y.-h., Zhou, P., Moran, L. V., Ananiadou, S., & Beam, A. (2024). Large Language Models in Mental Health Care: a Scoping Review. arXiv preprint arXiv:2401.02984. Disponível em https://arxiv.org/abs/2401.02984

[9] Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large Language Model for Mental Health: A Systematic Review. arXiv preprint arXiv:2403.15401. Disponível em: https://arxiv.org/pdf/2403.15401

[10] Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. arXiv:2211.04325v2 [cs.LG]. Disponível em: https://arxiv.org/html/2211.04325v2

Referências

[11] Amazon Web Services. (s.d.). O que é aumento de dados?. Disponível em: https://aws.amazon.com/pt/what-is/data-augmentation/

[12] Zhang, S., Li, X., Chen, X., Zhou, J., & Zhang, Y. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. arXiv. https://arxiv.org/pdf/2302.13007

[13] Wikipedia contributors. (2025, March 17). Cosine similarity. Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Cosine_similarity

[14] Huang, L.-K., Huang, J., Rong, Y., Yang, Q., & Wei, Y. (2022). Frustratingly easy transferability estimation. In International Conference on Machine Learning (pp. 9201–9225). PMLR.