Ética para LLMs: o compartilhamento de dados sociolinguísticos

Marta Deysiane Alves Faria Sousa[1,2], **Raquel Meister Ko Freitag**[1,2], **Túlio Sousa de Gois**[3]

[1]Programa de Pós-graduação em Letras - Universidade Federal de Sergipe – UFS

[2]Departamento de Letras Vernáculas – Universidade Federal de Sergipe – UFS

[3]Departamento de Computação – Universidade Federal de Sergipe – UFS

Abstract. *The collection of speech data carried out in Sociolinguistics has the potential to enhance large language models due to its quality and representativeness. In this paper, we examine the ethical considerations associated with the gathering and dissemination of such data. Additionally, we outline strategies for addressing the sensitivity of speech data, as it may facilitate the identification of informants who contributed with their speech.*

Resumo. *Considerando que a compilação de dados de fala feita na área de Sociolinguística pode contribuir para a alimentação de grandes modelos de linguagem, haja vista sua qualidade e representatividade, neste artigo discutimos questões éticas referentes à coleta e compartilhamento desses dados. Apontamos, também, possíveis caminhos para lidar eticamente com a sensibilidade de dados de fala, uma vez que eles podem levar à identificação dos informantes que cederam sua fala.*

1. Introdução

O uso da inteligência artificial (IA) tem se tornado cada vez mais presente na vida da população brasileira. Particularmente, o uso de IA na educação é uma realidade [de Oliveira Figueiredo et al. 2023, Leão et al. 2021] a qual, como educadores, tivemos que nos adaptar rapidamente tentando extrair o melhor delas para tornar o processo de ensino e aprendizagem mais interessante. Contudo, o uso dessas tecnologias em quaisquer áreas do conhecimento perpassa por questões éticas cujo debate no cenário brasileiro ainda é incipiente.

Em 2023, a Confederação de Organizações Europeias de Proteção de Dados reconheceu que a Inteligência Artificial Gerativa (IAG) é uma tecnologia muito nova e que as organizações que cuidam da proteção de dados devem remodelar suas normativas para atender às diferentes demandas que estão surgindo junto com a incorporação desse tipo de tecnologia em diversos espaços. Isso se deve ao fato de que os modelos de IAG são treinados em grandes volumes de dados, em sua maioria mal documentados, codificando, como consequência, viéses, estereótipos, discursos nocivos, além de os reproduzir em suas respostas [Bender et al. 2021].

No Brasil, o Governo Federal elaborou a Proposta de Plano Brasileiro de Inteligência Artificial 2024-2028, que prevê o investimento de 23 bilhões de reais. Sob

a premissa de uma "IA para o bem de todos", são destacadas visões como a centralidade no ser humano, prevenindo desigualdade e viéses, e a transparência e responsabilidade, garantindo a privacidade e a sobreania dos dados. Contudo, apenas 0,45% do orçamento é destinado ao "Apoio ao Processo Regulatório e de Governança da IA", evidenciando ainda mais a necessidade de discussões urgentes sobre a ética na IA no Brasil[1].

A preocupação ética na alimentação de grandes modelos de linguagem (large language models - LLMs) atravessa também o trabalho de linguistas. Como já dito, essas tecnologias precisam de grandes volumes de dados, no caso de chatbots, por exemplo, dados linguísticos são necessários para que as máquinas possam ter um fluxo de fala (ou escrita) similar ao de seres humanos, colocando aos linguistas o desafio de pensar em como proceder para proteger eticamente os dados colhidos. Neste artigo, então, discutimos desde a base de coleta de dados linguísticos ao compartilhamento desses dados para alimentar LLMs.

2. Regulamentação

Nos últimos anos, muito tem-se debatido sobre Ciência Aberta e a centralidade dos dados de pesquisas no fortalecimento do conhecimento produzido pela academia, de tal forma que os periódicos científicos têm demandado dos pesquisadores acesso aos dados que geraram as publicações. Isso se deve ao fato de que a maior transparência na condução, compartilhamento de dados e publicação das pesquisas acarreta não só maior fortalecimento do rigor metodológico e experimental, mas também a confiabilidade na produção científica por parte do público geral [Lyon 2016]. Ademais, no escopo do cenário científico brasileiro, no qual grande parte do financiamento de projetos de pesquisa parte da iniciativa pública, é esperado que os dados coletados sejam de domínio público. Entretanto, existem aspectos concernetes à responsabilização das instituições e dos pesquisadores com os dados fornecidos pelos indivíduos que aceitam participar das pesquisas que devem balizar o compartilhamento desses dados [Freitag 2021, Freitag 2022].

A normativa que trata das pesquisas com seres humanos, no Brasil, é a Resolução Nº 510, de 07 de abril de 2016 do Conselho Nacional de Saúde. Destacamos o fato de que seu artigo 9º apresenta os direitos dos participantes das pesquisas, evidenciando o protagonismo destes na tomada de decisão quanto às informações que podem ser publicadas e tornadas públicas. Ainda, neste documento, encontramos informações acerca da documentação que deve ser entregue aos participantes para torná-los cientes dos possíveis ganhos e danos que possam ser causados em decorrência da pesquisa em curso.

Além da Resolução Nº 510, de 07 de abril de 2016 do Conselho Nacional de Saúde, no Brasil, existe a Lei Geral de Proteção de Dados (Lei 13.709, de 2018, doravante LGPD), inspirada na Regulamento Geral de Proteção de Dados da Europa. Essa lei tem como objetivo assegurar aos cidadãos brasileiros o controle de suas informações pessoais, demandando daqueles que as coletam o pedido de consentimento sobre os usos desses dados bem como a escolha dos usuários sobre as formas de visualização, correção e também exclusão de dados pessoais que circulam na internet.

[1]Proposta de Plano Brasileiro de Inteligência Artificial 2024-2028. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2024/07/plano-brasileiro-de-ia-tera-supercomputador-e-investimento-de-r-23-bilhoes-em-quatro-anos/ia_para_o_bem_de_todos.pdf/view

Ao pensarmos, então, em coleta de dados, principalmente, no caso de pesquisadores da área de Linguística que produzem *corpora* para suas pesquisas, devemos observar essas duas regulamentações para conduzirmos nossos estudos. Assim, os procedimentos éticos de entrada em campo e de compartilhamento de dados devem refletir essas normativas e assegurar aos participantes o controle das informações que eles oferecem.

3. Dados sociolinguísticos

No escopo da Sociolinguística, a compilação de bancos de dados se dá principalmente por meio de entrevistas sociolinguísticas visto que elas se apresentam como o melhor instrumento para se capturar a língua vernacular em grandes volumes (entre uma e duas horas de duração) e com melhor qualidade em termos de gravação [Labov et al. 1981]. Assim, os bancos de dados sociolinguísticos, que já são utilizados no treinamento de ferramentas de PLN [Sousa and Freitag 2024], se tornam produtos com alto valor para a área de IA por possuírem qualidade em termos de gravação, transcrição e quantidade para alimentar LLMs.

É inevitável, então, que os dados de fala coletados preservem as condições da fala natural do indivíduo, tornando-se um paradoxo entre a preservação da identidade dos participantes e o reconhecimento de suas vozes. Para mitigiar os efeitos desse paradoxo, o termo de consentimento e as licenças de uso são fundamentais [Calamai and Frontini 2018, Mello 2021].

3.1. Caminhos a seguir

Primeiramente, o termo de consentimento deve prever que o consentimento para utilização dos dados pode ser interrompido em qualquer estágio da pesquisa e indicar o local em que os dados serão depositados. Ademais, os participantes devem ser informados sobre o possível compartilhamento desses dados, determinando seu tipo circulação e a finalidade desse compartilhamento.

Ademais, visando garantir ainda mais o controle dos dados compartilhados tanto por parte dos pesquisadores quanto dos participantes, é importante a escolha adequada de licenças de uso. Existem dois tipos de licenças para garantir os direitos do autor, a licença GNU General Public License (GNU GPL) e as Creative Commons (CC). A licença GNU GPL é menos restritiva que as CC em termos de reutilização, sendo a única restrição a de que os derivados produzidos com os dados sejam de acesso aberto. Está no bojo das licenças CC respeitar os direitos autorais e conexos.

É no contexto do compartilhamento de dados de fala que uma licença adequada contribui tanto para a atribuição da devida autoria daquele produto intelectual quanto em uma maior transparência das pesquisas na área. Nesse caso, sugerimos "Atribuição-NãoComercial-CompartilhaIgual", uma licença CC que, por meio dela, se protege a autoria e é autorizado o desenvolvimento de novas tecnologias a partir dos dados disponibilizados. Assim, o licenciante (o criador do banco de dados) permite que o licenciado (quem usará o conteúdo) utilize e faça novos trabalhos com os dados gerados, contanto que cite os licenciantes e utilize a mesma licença na criação dos derivados para propósitos não-comerciais.

4. Considerações finais

Neste artigo, abordamos de maneira abrangente a necessidade de se considerar questões éticas associadas à coleta e uso de dados de fala na Sociolinguística para a alimentação de LLMs. Discutimos o dilema de se preservar a identidade dos informantes bem como a utilidade desses dados para IA. Sugerimos o uso de termos de consentimento e licenças de uso como passos importantes para garantir que os dados sejam coletados e compartilhados de maneira ética e responsável. A continuidade das discussões sobre ética na IA no Brasil é urgente, principalmente devido ao seu rápido avanço e assimilação pela sociedade. A implementação efetiva das regulamentações existentes e a promoção de uma cultura de responsabilidade ética serão essenciais para o desenvolvimento sustentável e responsável da IA em nosso país.

References

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Calamai, S. and Frontini, F. (2018). Fair data principles and their application to speech and oral archives. *Journal of new music research*, 47(4):339–354.

de Oliveira Figueiredo, L., Lopes, A. M. Z., Validorio, V. C., and Mussio, S. C. (2023). Desafios e impactos do uso da inteligência artificial na educação. *Educação Online*, 18(44):e18234408–e18234408.

Freitag, R. M. K. (2021). Linguistic repositories as asset: Challenge for sociolinguistic approach in brazil. In *Proceedings of the 1st International Workshop on Digital Language Archives 2021*. University of North Texas.

Freitag, R. M. K. (2022). Sociolinguistic repositories as asset: challenges and difficulties in brazil. *The Electronic Library*, 40(5):607–622.

Labov, W. et al. (1981). Field methods of the project on linguistic change and variation.

Leão, J. C., Leão, J. J. C. C., dos Santos, A. B., Marques, T. M., and Santos, E. M. S. (2021). Inteligência artificial na educação: aplicações do aprendizado de máquina para apoiar a aprendizagem adaptativa. *Revista Multidisciplinar do Vale do Jequitinhonha-ReviVale*, 1(1).

Lyon, L. (2016). Transparency: The emerging third dimension of open science and open data. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, 25(4):153–171.

Mello, H. R. (2021). Trabalhando com dados de fala: a experiência do projeto c-oral-brasil. In Brescancini, C. R., editor, *Projeto VARSUL - Variação Linguística no Sul do Brasil 36 anos*, pages 31–54. Editora Zouk, Porto Alegre, 1 edition.

Sousa, M. D. A. F. and Freitag, R. M. K. (2024). Bancos de dados sociolinguísticos e a ciência aberta: compartilhamento de dados e conhecimentos. *Revista Diálogos*, 12(1):165–187.