Avaliação de Ferramentas de Ética no Levantamento de Considerações Éticas de Modelos de Linguagem em Português

Jhessica Silva[1], Alef Ferreira[3], Diego Moreira[1], Gabriel Santos[1], Gustavo Bonil[2], João Gondim[1], Luiz Pereira[1], Helena Maia[1], Nadia Silva[3], Simone Hashiguti[2], Sandra Avila[1], Helio Pedrini[1]

[1]Instituto de Computação [2]Instituto de Estudos da Linguagem
Universidade Estadual de Campinas (UNICAMP)
Campinas – SP – Brasil

[3]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

{jhessica.silva, sandra, helio}@ic.unicamp.br

Abstract. *This paper presents a study using AI Ethics Tools (AIETs) to raise ethical considerations in language models developed for the Portuguese language. The AIETs are intended to help developers, companies, governments, and other interested parties to establish trust, transparency, and responsibility with their technologies. This work briefly discusses whether AIETs can help developers think about their technologies ethically. This study was based on interviews with developers of four language models using the AIETs Harms Modeling and Model Cards. The results suggest that the AIETs serve as a guide in elaborating ethical considerations but require prior knowledge of AI ethics.*

Resumo. *Este artigo apresenta um estudo sobre a utilização de Ferramentas de Ética para IA (AIETs) no levantamento de considerações éticas de modelos de linguagem em português. As AIETs visam ajudar desenvolvedores, empresas, governos e outras partes interessadas a estabelecer confiança, transparência e responsabilidade com suas tecnologias. Busca-se com este artigo discutir brevemente se as AIETs podem auxiliar desenvolvedores a pensarem eticamente seus modelos. Tal estudo foi realizado a partir de entrevistas com desenvolvedores de quatro modelos de linguagem com as AIETs Harms Modeling e Model Cards. Os resultados sugerem que as AIETs servem como guia na elaboração de considerações éticas, mas requerem conhecimentos prévios de ética em IA.*

1. Introdução

Nos últimos anos, no contexto de Inteligência Artificial (IA), muitas pesquisas foram desenvolvidas visando à geração de textos e de conversas realistas através do uso de dados de linguagem natural. Essas pesquisas estão sendo impulsionadas pela disponibilização de grandes modelos de linguagens, como GPT [Radford et al. 2018] e suas variações [Radford et al. 2019, Brown et al. 2020, OpenAI 2023]. Entretanto, apesar dos avanços na área, esses modelos podem gerar informações incorretas e imprecisas, não apresentam citação da fonte utilizada e não possuem informações atualizadas.

Os modelos de linguagem requerem alta quantidade de dados para o treinamento e, apesar desse grande volume de dados, não existem modelos que mitiguem vieses, representem a diversidade, preservem o meio ambiente e combatam a exclusão de grupos

marginalizados [Bender et al. 2021]. Além disso, esses modelos costumam ser treinados em língua inglesa, que não é uma língua pertencente a todas as culturas, e com dados do norte global. Desta forma, representam apenas um conjunto de visões de mundo dessas regiões [Hershcovich et al. 2022], fazendo com que os mesmos não estejam preparados para fazer escolhas éticas de um valor em detrimento de outro [Johnson et al. 2022], visto que os valores variam muito de acordo com a cultura de onde aquela língua está inserida.

No contexto da língua portuguesa, uma língua de baixos recursos computacionais[1], a maioria dos modelos de linguagem multilíngue, além de não apresentar desempenho competitivo [Santos et al. 2023], não é capaz de representar aspectos culturais, históricos e sociais da população onde a língua está inserida [Bender et al. 2021, Johnson et al. 2022], pois são treinados com dados textuais majoritariamente em inglês e com poucos dados em outras línguas. Por outro lado, quando se trata de modelos especialistas no português, embora possuam desempenhos melhores, na grande maioria são treinados com dados textuais traduzidos do inglês. Isso faz com que o modelo, embora direcionado para a população falante do português, replique todos os aspectos do norte global, além de perpetuarem erros e vieses provenientes das traduções automáticas [Santos et al. 2023, Hovy and Prabhumoye 2021].

Esses pontos demostram que avaliar eticamente os modelos de linguagem e apresentar aos usuários finais boas documentações, explicitando os problemas da tecnologia, é de extrema importância. Entretanto, vários modelos atuais estão sendo anunciados apenas acompanhados de "relatórios técnicos", não apresentando, de fato, muitas informações sobre o processo de treinamento ou arquitetura do modelo e muito menos suas considerações éticas. Um dos motivos para que isso ocorra é o fato de que muitos desses modelos estão sendo propostos por profissionais de Computação, sem o apoio de uma equipe multidisciplinar. Infelizmente, muitos destes profissionais da Computação finalizam suas formações básicas sem um conhecimento sólido de disciplinas relacionadas com ética e tecnologias [Brown et al. 2024, Goetze 2023]. Consequentemente, não possuem uma base em como realizar as formulações éticas sobre as tecnologias que propõem. Em contrapartida, atualmente, algumas conferências como FAccT, EMNLP e NeurIPS passaram a solicitar uma seção exclusiva para considerações éticas nas chamadas por artigos, mostrando a importância da reflexão sobre ética na pesquisa na área de tecnologia.

Diante dos pontos apresentados, este artigo visa contribuir para as urgentes discussões em aberto na nossa sociedade sobre ética e tecnologias de IA. Para isso, conduzimos um estudo de caso sobre a utilização de Ferramentas de Ética para IA (AIETs, do inglês *AI Ethics Tools*) no levantamento de considerações éticas de modelos de linguagem desenvolvidos em português. As AIETs são métodos práticos que podem ser aplicados durante o projeto, o desenvolvimento e o uso de uma IA para identificar possíveis problemas éticos que a tecnologia pode gerar e propor maneiras de mitigar esses problemas. Busca-se com este artigo discutir brevemente se as AIETs podem auxiliar desenvolvedores a pensarem e documentarem eticamente suas tecnologias. Tal estudo foi realizado a partir de entrevistas com esses profissionais de quatro modelos de linguagem desenvolvidos para a língua portuguesa.

[1]Embora o português esteja classificado entre as dez principais línguas mundiais em número de falantes nativos, ela é uma língua de baixos recursos do ponto de vista da IA [Kemp and Social 2023], pois possui poucos dados disponíveis para uso nas tecnologias de IA.

2. Metodologia

Para a realização deste estudo, foi feito um extenso levantamento bibliográfico de AIETs na literatura. As 213 AIETs encontradas foram filtradas seguindo critérios de inclusão e exclusão, cujo objetivo era selecionar as que poderiam ser aplicadas em modelos de linguagem prontos para uso e respondidas por seus desenvolvedores. A partir dos critérios aplicados, selecionamos as AIETs *Harms Modeling* [Microsoft 2022] e *Model Cards* [Mitchell et al. 2019]. Ambas AIETs oferecem um questionário que visa auxiliar pesquisadores e desenvolvedores a pensarem eticamente suas IAs, seja identificando ou documentando as questões éticas do modelo analisado. Conduzimos entrevistas orais com desenvolvedores de quatro modelos de linguagem desenvolvidos para a língua portuguesa, utilizando os questionários das AIETs como roteiro para as entrevistas.

Os desenvolvedores convidados para o estudo foram identificados através de busca de artigos ou outras publicações sobre modelos de linguagem em língua portuguesa. Ao todo, foram entrevistados 11 desenvolvedores. Para manter o sigilo dos participantes, os modelos entrevistados não serão mencionados neste artigo. Com base nas entrevistas, os participantes da pesquisa foram convidados a preencherem individualmente um questionário no final do estudo, com o objetivo de identificarmos suas percepções sobre as considerações éticas levantadas por cada uma das AIETs utilizadas.

3. Resultados e Considerações Finais

Os resultados obtidos sugerem que as AIETs aplicadas são eficazes para serem utilizadas como guia na elaboração de considerações éticas gerais sobre modelos de linguagem. Entretanto, notamos que as mesmas são generalistas e não abordam aspectos únicos desses modelos, como expressões idiomáticas. Além disso, as AIETs não auxiliaram no levantamento de potenciais impactos negativos de modelos para a língua portuguesa, como a falta de representatividade de aspectos sociais e culturais das populações falantes do português. Assim, notamos que as AIETs sozinhas não são suficientes no levantamento de considerações éticas sobre os modelos de linguagem e é necessário que os desenvolvedores possuam uma base em ética para levantarem as considerações sobre suas tecnologias.

Os desenvolvedores entrevistados, até o momento das entrevistas, disseram nunca ter avaliado eticamente seus modelos. Esse resultado indica que ainda não é natural para os desenvolvedores adicionarem uma etapa para analisarem eticamente os seus modelos antes de publicá-los, mostrando a necessidade de maior engajamento e divulgação sobre a área de ética em IA. Desta forma, esperamos com este artigo despertar o interesse da comunidade em encontrar maneiras de (i) introduzir o ensinamento de ética em IA nas formações bases dos profissionais da Computação e (ii) encontrar caminhos para incentivar esses profissionais a pensarem eticamente suas tecnologias. Acreditamos que somente assim poderemos tornar o futuro da IA mais transparente, responsável e confiável.

Considerações Éticas. Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Estadual de Campinas sob o número do CAAE: 74643023.8.0000.5404. Para realizar este estudo, seguimos todos os protocolos exigidos pelo Comitê. Este estudo foi realizado por pesquisadores da área de Computação, em colaboração com pesquisadores da área dos Estudos da Linguagem, e carrega todos os vieses desses domínios de conhecimento, influenciando como cada AIET foi interpretada, como as entrevistas foram conduzidas e como os resultados foram analisados. Este resumo expandido faz parte de

um trabalho maior sobre avaliação de AIETs em modelos de linguagem. Por conta do espaço, diversos aspectos e resultados não foram abordados.

Agradecimentos. Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado Arquitetura Cognitiva (Fase 3), DOU 01245.003479/2024-10. D.A.B.M. é parcialmente financiado pela FAPESP 2023/05939-5. A.I.F. e N.S. são parcialmente financiados pelo Centro de Excelência em Inteligência Artificial da UFG. G.O.S é parcialmente financiado pela FAPESP 2024/07969-1. H.P. é parcialmente financiado pelo CNPq 304836/2022-2. S.A. é parcialmente financiada por CNPq 316489/2023-9, FAPESP 2013/08293-7, 2020/09838-0, 2023/12086-9 e Google Award for Inclusion Research 2022. Os autores são gratos aos participantes do estudo.

Referências

[Bender et al. 2021] 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 *ACM FAccT*, pages 610–623.

[Brown et al. 2024] Brown, N., Xie, B., Sarder, E., Fiesler, C., and Wiese, E. S. (2024). Teaching Ethics in Computing: A Systematic Literature Review of ACM Computer Science Education Publications. *ACM TOCE*, 24(1):1–36.

[Brown et al. 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., and et al. (2020). Language Models are Few-Shot Learners. In *NeurIPS*, volume 33, pages 1877–1901.

[Goetze 2023] Goetze, T. S. (2023). Integrating ethics into computer science education: Multi-, inter-, and transdisciplinary approaches. In *ACM SIGCSE*, pages 645–651.

[Hershcovich et al. 2022] Hershcovich, D., Frank, S., Lent, H., and et al. (2022). Challenges and strategies in cross-cultural NLP. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Association for Computational Linguistics (ACL)*, pages 6997–7013.

[Hovy and Prabhumoye 2021] Hovy, D. and Prabhumoye, S. (2021). Five Sources of Bias in Natural Language Processing. *Language and Linguistics Compass*, 15(8):e12432.

[Johnson et al. 2022] Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., and Bertulfo, D. J. (2022). The Ghost in the Machine Has an American Accent: Value Conflict in GPT-3.

[Kemp and Social 2023] Kemp, S. and Social, W. A. (2023). Digital 2023: Global Overview Report.

[Microsoft 2022] Microsoft (2022). Harms Modeling.

[Mitchell et al. 2019] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model Cards for Model Reporting. In *ACM FAccT*, pages 220–229.

[OpenAI 2023] OpenAI (2023). GPT-4 Technical Report.

[Radford et al. 2018] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *OpenAI*.

[Radford et al. 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*.

[Santos et al. 2023] Santos, G., Moreira, D., Ferreira, A., Silva, J., Pereira, L., Bueno, P., Sousa, T., Maia, H., da Silva, N., Colombini, E., Pedrini, H., and Avila, S. (2023). Capivara: Cost-efficient approach for improving multilingual clip performance on low-resource languages. In *MRL-EMNLP*.