

ESTEREÓTIPOS OCUPACIONAIS E APRENDIZAGEM DE MÁQUINA: EXEMPLO PRÁTICO A PARTIR DO CHATGPT

Thiago Dias de Matos Diniz¹

Resumo

Este artigo aborda problemas relacionados aos vieses em algoritmos de aprendizagem de máquina. Descreve técnicas utilizadas na literatura da área para a identificação de associações historicamente sedimentadas que sustentam os chamados vieses em alguns tipos de aplicações de aprendizagem de máquina. Por fim, descreve uma experimentação realizada com o ChatGPT, por meio da qual se comprovou a hipótese de associações predominantes entre gêneros e ocupações específicas no mercado.

Palavras-chave: vieses, gênero, aprendizagem de máquina, *word embedding*, ChatGPT.

INTRODUÇÃO

Os problemas associados a vieses em algoritmos de aprendizagem de máquina são há muito tematizados pela literatura em diversas áreas do conhecimento e com foco em várias aplicações críticas de sistemas de decisões automatizadas, como modelos de diagnóstico na área médica¹. Este artigo abordará uma técnica para identificação de viés em sistemas de aprendizagem de máquina, baseada no método de *word embedding*, seguida de uma verificação empírica de algumas respostas (*outputs*) do ChatGPT.

Vieses em aprendizagem de máquina

Os vieses da IA podem ser consequência de um desenho ruim de um sistema de IA, de dados de treinamento não balanceados, de confusões operadas a nível do modelo algorítmico ou da própria capacidade computacional do algoritmo implementado². Nesses casos, o viés é prejudicial ao próprio desempenho do modelo sobre novos casos, e sua manutenção, a depender do contexto social no qual o sistema é utilizado, pode constituir um ilícito, caso

¹ CHALLEN, Robert et al. Artificial intelligence, bias and clinical safety. **BMJ Quality & Safety**, v. 28, n. 3, p. 231-237, 2019.

² ZHAI; KRAJCIK, 2022, p. 1.

envolva tratamento de dados pessoais, por violação de princípios de proteção desses dados, conforme previstos na legislação brasileira³:

Art. 6º As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios:

[...]

II - adequação: compatibilidade do tratamento com as finalidades informadas ao titular, de acordo com o contexto do tratamento;

[...]

V - qualidade dos dados: garantia, aos titulares, de exatidão, clareza, relevância e atualização dos dados, de acordo com a necessidade e para o cumprimento da finalidade de seu tratamento;

VIII - prevenção: adoção de medidas para prevenir a ocorrência de danos em virtude do tratamento de dados pessoais;

IX - não discriminação: impossibilidade de realização do tratamento para fins discriminatórios ilícitos ou abusivos

Alguns autores chamam atenção para o fenômeno oposto, conhecido como Pseudo Viés de Inteligência Artificial (*Pseudo Artificial Intelligence bias – PAIB*): vieses não suficientemente analisados são atribuídos de forma arbitrária a modelos de Inteligência Artificial (IA), na medida em que usuários ou pesquisadores “não compreendem o viés, operam de forma inadequada os algoritmos de máquina, interpretam mal as previsões da IA ou nutrem sobre elas expectativas sobrestimadas”⁴. Os efeitos desse viés são socialmente prejudiciais, afetando potencialmente o avanço de soluções de aprendizagem de máquina em campos como a educação, onerando pesquisas e a distribuição do acesso a soluções e benefícios da IA a uma gama de empreendedores de menor porte e fomentando a desconfiança social em sistemas automatizados, ou, de forma ainda mais impactante, desencorajando seu uso⁵.

Alguns trabalhos na área de IA já se debruçaram sobre a análise de vieses embarcados nas representações textuais que permeiam as próprias línguas. Zou e Schiebinger utilizaram o método de aprendizagem de máquina *word embedding* para quantificar estereótipos históricos nos Estados Unidos: referido método converte cada palavra em um ponto no espaço (um vetor geométrico), capturando conexões semânticas entre palavras adjacentes e medindo a distância entre os vetores, de modo que relações são identificadas

³ BRASIL. **Lei Nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD).

⁴ ZHAI, Xiaoming; KRAJCIK, Joseph. Pseudo AI bias. **arXiv preprint arXiv:2210.08141**, 2022, p. 1

⁵ PETERS, Tobias M.; VISSER, Roel W. The Importance of Distrust in AI. **arXiv preprint arXiv:2307.13601**, 2023, p. 12.

entre termos como, por exemplo, “‘homem’ está para ‘rei’, assim como ‘mulher’ está para ‘rainha’”⁶.

As pesquisas de Garg, Jurafskyc, Zou e Schiebinger tinham como objetivo verificar se os vieses identificados nos *word embedding* coincidiam com tendências sociológicas, então eles compararam as tendências nos padrões de ocorrências de palavras com tendências demográficas quantificadas relacionadas à distribuição ocupacional das pessoas:

usamos estatísticas de participação de mulheres e minorias étnicas (em relação a homens e brancos, respectivamente) em diferentes ocupações como referência, porque é uma métrica objetiva de mudanças sociais. Mostramos que a incorporação textual [*word embedding*] captura com precisão as porcentagens de ocupação de gênero e etnia e reflete consistentemente as mudanças históricas⁷.

Os modelos de *word embedding*, ao condensar extensas coleções de texto em uma representação única, preservando grande parte da riqueza e complexidade das relações semânticas, representa as relações entre palavras como as relações entre vetores em um espaço vetorial de alta dimensão⁸. Os modelos de *embedding*, a partir de dimensões culturais como masculino-feminino e da distribuição das palavras nos espaços que refletem padrões de associação e classificação dentro de um sistema cultural específico, auxiliam a investigação das relações existentes e as distâncias entre essas dimensões culturais⁹.

Experimentação com ChatGPT

Utilizando-se o ChatGPT versão *Default* (GPT-3.5), passou-se o seguinte comando: “complete a seguinte frase com o pronome apropriado no espaço reservado para [pronome]: Como nosso Diretor de Segurança da Informação, [pronome] mostrou uma ótima estratégia para melhorar nossos programas”¹⁰.

⁶ ZOU, James; SCHIEBINGER, Londa. AI can be sexist and racist—it’s time to make it fair. *Nature*, v. 559, 2018, p. 326.

⁷ GARG, Nikhil et al. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, v. 115, n. 16, 2018, p. E3636.

⁸ KOZLOWSKI, Austin C.; TADDY, Matt; EVANS, James A. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, v. 84, n. 5, p. 905-949, 2019.

⁹ KOZLOWSKI at. al., 2019.

¹⁰ Originalmente escrito em inglês: “please complete the following sentence with the appropriate pronoun in the [pronoun] placeholder: As our Chief Information Security Officer, [pronoun] has showed great strategy to improve our programs.”

Observe-se que a solicitação original foi realizada em inglês, portanto sem sinais indicadores de gênero nos termos *nossa/nosso* e *diretor/diretora*. A resposta obtida foi: “Como nosso Diretor de Segurança da Informação, eles demonstraram grande estratégia para melhorar nossos programas”¹¹. Na sequência, pediu-se a seguinte especificação: “considere que o sujeito é apenas 1 pessoa e reformule”¹², com a consequente resposta: “Como nosso Diretor de Segurança da Informação, ele demonstrou grande estratégia para melhorar nossos programas”¹³. Nota-se, assim, o emprego do artigo masculino para a atribuição ocupacional de chefe de segurança da informação.

Após esse teste, nova experimentação foi realizada com o seguinte comando ao ChatGPT: “por favor, complete a seguinte frase com o pronome apropriado no espaço reservado [pronome]: Como professora primária, [pronome] tem mostrado uma grande estratégia para melhorar os programas”¹⁴. Da mesma forma, o comando original em inglês não permitia distinguir forma feminina ou masculina no substantivo utilizado. Ainda assim, diferentemente do cenário anterior, o preenchimento da resposta, para a ocupação de professor primário, foi feito com o pronome feminino: “Como professora primária, ela tem mostrado grande estratégia para melhorar os programas”¹⁵.

Ao final, foi feita a seguinte indagação ao ChatGPT: “quando você escolheu o pronome nos *prompts* anteriores, qual foi o critério?”¹⁶, a qual teve como resposta: “Ao escolher o pronome em prompts anteriores, meus critérios foram baseados nas convenções tradicionais da língua inglesa para pronomes de gênero. Se o papel ou contexto sugerisse um gênero específico, eu usaria o pronome correspondente (por exemplo, ‘ele’ para um papel masculino). No entanto, se o papel ou o contexto fosse neutro em termos de gênero ou o

¹¹ Originalmente escrito em inglês: “As our Chief Information Security Officer, they have showed great strategy to improve our programs”.

¹² Originalmente escrito em inglês: “please consider that the subject is only 1 person and reformulate”.

¹³ Originalmente escrito em inglês: “As our Chief Information Security Officer, he has showed great strategy to improve our programs”.

¹⁴ Originalmente escrito em inglês: “please complete the following sentence with the appropriate pronoun in the [pronoun] placeholder: As a primary school teacher, [pronoun] has showed great strategy to improve the programs”.

¹⁵ Originalmente escrito em inglês: “As a primary school teacher, she has showed great strategy to improve the programs”.

¹⁶ Originalmente escrito em inglês: “when you chose the pronoun in the earlier prompts, which was the criteria?”

gênero não fosse especificado, eu usaria ‘eles’ como um pronome neutro em termos de gênero”¹⁷.

Considerações Finais

Conclui-se que, de fato, dimensões de significado cultural correspondem a diferenças de ocorrências de palavras nos espaços demarcados por vetores textuais a partir de um *corpus* de linguagem disponível, inclusive naquele utilizado para treinamento de modelos de aprendizagem de máquina como o ChatGPT. As escolhas de palavras representativas de gêneros nas respostas desse sistema mostram associações culturalmente compartilhadas, tais como apontadas pelas pesquisas realizadas sobre conjuntos de textos específicos produzidos no último século.

A experimentação realizada com base em interação com o sistema de aprendizagem de máquina ChatGPT permitiu, além de confirmar a hipótese dos padrões de ocorrência textual de associação entre gêneros e ocupações, escapar parcialmente de limitações de pesquisas anteriores na área, que se pautaram por um conjunto textual mais restrito de base, em relação ao *corpus* incorporado em um modelo como o que sustenta o ChatGPT, portanto reduzindo ainda mais a quantidade possível de viés amostral em captura de textos digitalizados, que representam amostras da cultura, conforme a metodologia das pesquisas anteriores.

Referências Bibliográficas

BRASIL. **Lei Nº 13.709, de 14 de agosto de 2018.** Lei Geral de Proteção de Dados Pessoais (LGPD).

¹⁷ Originalmente escrito em inglês: “When choosing the pronoun in earlier prompts, my criteria were based on traditional English language conventions for gendered pronouns. If the role or context suggested a specific gender, I would use the corresponding pronoun (e.g., “he” for a male role). However, if the role or context was gender-neutral or the gender was unspecified, I would default to using “they” as a gender-neutral pronoun”.

CHALLEN, Robert et al. Artificial intelligence, bias and clinical safety. **BMJ Quality & Safety**, v. 28, n. 3, p. 231-237, 2019.

GARG, Nikhil et al. Word embeddings quantify 100 years of gender and ethnic stereotypes. **Proceedings of the National Academy of Sciences**, v. 115, n. 16, p. E3635-E3644, 2018.

KOZLOWSKI, Austin C.; TADDY, Matt; EVANS, James A. The geometry of culture: Analyzing the meanings of class through word embeddings. **American Sociological Review**, v. 84, n. 5, p. 905-949, 2019.

PETERS, Tobias M.; VISSER, Roel W. The Importance of Distrust in AI. **arXiv preprint arXiv:2307.13601**, 2023.

ZHAI, Xiaoming; KRAJCIK, Joseph. Pseudo AI bias. **arXiv preprint arXiv:2210.08141**, 2022.

ZOU, James; SCHIEBINGER, Londa. AI can be sexist and racista – it's time to make it fair. **Nature**, v. 559, 2018.