O preconceito da sociedade propagado pela Inteligência Artificial

O normal é se pensar que artifícios desenvolvidos em bases matemáticas, como algoritmos, sejam objetivos e completamente imparciais, o que os tornaria mais justos que as decisões tomadas pelo ser humano. Entretanto, não é o que os estudos demonstram. Os preconceitos baseados em estereótipos decididos pelos humanos estão sendo reproduzidos pelas Inteligências Artificiais (IA).

Existem estudos que procuram quantificar e minimizar os viés preconceituosos que possam vir a serem gerados pelas IAs. Os estudos analisados usam Word embedding (WE), que é um conjunto de técnicas nas quais palavras individuais são representadas como vetores num espaço multi-vetorial pré-definido. Cada palavra é mapeada para um vetor, o valor desse vetor é utilizado para comparar com outros, assim vetores de valores semelhantes tendem a ter semânticas semelhantes. São citados também os Testes de Associação Implícita (TAI). O TAI é um projeto multidisciplinar que procura uma forma de entender e medir atitudes, estereótipos, crenças ou algum preconceito que possam influenciar a percepção, julgamento e ações.

Bolukbasi et al. fizeram o estudo, *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. Ele tem como objetivo desenvolver um algoritmo que reduza a conexão preconceituosa entre determinadas palavras, sem perder conexões que são relevantes. Por exemplo tirar a conexão entre o termo recepcionista e mulher e não perder a conexão entre o termo rainha e mulher. Para isso, foi utilizado o word2vec, um grupo de modelos relacionados do Google, que tem mais de 3 milhões de palavras e termos em cerca de 300 dimensões e trabalhadores baseados nos EUA da Amazon Mechanical Turk (AMTurk). Foram selecionadas as palavras mais frequentes com 20 ou menos caracteres. Depois, o pessoal do AMTurk foi dividido em dois grupos, um onde os pesquisadores requisitavam palavras para verificar se eles já tem termos preconceituosos, e outro em que eram solicitadas classificações de palavras ou analogias a partir do material que era apresentado. A partir dessas coletas foram verificadas quais as relações entre termos eram sexistas. Para reduzir conexões preconceituosas, o algoritmo do estudo alterou a vetorização de palavras neutras removendo suas associações com gênero. Por exemplo, enfermeiro(a) passa a ser relacionada igualmente a homem e mulher. Além disso termos que são exclusivamente femininos, por exemplo, câncer de ovário, e masculino, como câncer de próstata, não tiveram seu status alterados, o que mostra que o algoritmo gerado é eficiente.

Aylin Caliskan et al. realizou o estudo *Semantics derived automatically from language corpora contain human-like biases* que fala sobre os preconceitos e estereótipos que são visíveis em algoritmos. Foram reproduzidos estudos das ciências humanas, psicologia e sociologia, que mostram diversos casos de preconceitos, utilizando-se de TAI e WE tendo como base o GloVe, que possui diversas palavras e

frases já vetorizadas e com *scores* utilizando treinamento padrão da ferramenta. De início ele compara os resultados de um TAI onde as pessoas tinham que associar inseto e flores a agradável e desagradável, em ambos os testes o resultado foi o mesmo, flores possuem muito mais chance de serem relacionadas a agradável do que insetos. Para isso eles utilizam a métrica de tamanho efetivo e *p-value*, quanto maior o tamanho efetivo e menor o *p-value* maior semelhança semântica entre os vetores. Isso demonstra a efetividade do modelo, uma vez que sem nenhum conhecimento natural prévio ou ajuda ele conseguiu relacionar de forma análoga ao TAI aplicado anteriormente. Em seguida ele reproduz estudos que demonstram o viés racial, em um deles, ele testa se existe alguma relação com nomes afro-americanos e euro-americanos com os conceitos agradável e desagradável. O resultado foi que ambos, TAI e WE, os nomes euro-americanos foram muito mais relacionados a agradável e em contrapartida os nomes afro-americanos foram muito mais relacionados a desagradável. Além desses testes, outros são feitos como relação de homem/mulher com artes/matemática; carreira/família e em todos, o estereótipo nos modelos se confirma.

O que os estudos analisados sugerem é que as relações preconceituosas geradas nas WE são meros reflexos da sociedade. Para reduzir de maneira efetiva esse tipo de propagação por parte dos algoritmos é importante que se reduzam os preconceitos na sociedade. Outras sugestões são permitir acesso ao público do algoritmo, assim ele pode monitorar e sugerir melhorias no código, além de uma colaboração entre o programador e especialistas em desigualdades.