XXX Congresso de Iniciação Científica UNICAMP

Eficiência como viés algorítmico nas técnicas de aprendizado de máquina: caracterização baseada na produção tecnocientífica da Google

Palavras-Chave: inteligência artificial, sociologia da tecnologia, neoliberalismo.

Autores:

Rafael Gonçalves – FEEC/Unicamp

Pedro P. Ferreira (orientador) – DS/IFCH/Unicamp

Resumo: Sistemas de inteligência artificial e técnicas de aprendizado de máquina têm se tornado parte integrante do funcionamento da sociedade. Dessa forma, o uso generalizado de tais tecnologias traz novas questões a serem pensadas. Uma dessas problemáticas é a do viés em aprendizado de máquina. A partir do estudo de artigos científicos produzidos pela empresa Google, este trabalho objetivou caracterizar o viés algorítmico nas tecnologias de aprendizado de máquina. Concluiu-se que existe um viés no sentido de incentivar resultados ótimos, isto é, uma maximização de acertos ou uma minimização de erros, em detrimento de outras implicações dos resultados. Essa tendência observada coincide com a promoção do máximo desempenho presente na racionalidade neoliberal. Assim, constata-se a existência de uma ressonância entre o funcionamento técnico do aprendizado de máquina e normativo neoliberal no sentido de valorização da noção de eficiência. Esse viés – por negligenciar outros valores possíveis – contribui com a diminuição da diversidade cultural e para o aumento de discriminação social.

INTRODUÇÃO

Laymert Garcia dos Santos (2003) chamará de "virada cibernética" o momento no final do século XX a partir do qual há uma convergência entre os interesses de acumulação capitalista e os desenvolvimentos tecnocientíficos. Para o autor, a noção de informação e principalmente o avanço da empresa capitalista sobre a informação digital e genética marcam uma mudança significativa no escopo do capitalismo contemporâneo, que passa a ser capaz de lucrar sobre a realidade em potencial – nas palavras do autor: "da dimensão virtual da realidade".

Nesse contexto, ganham importância as noções de vigilância e controle. Zuboff (2018) chama de "capitalismo de vigilância" a lógica capitalista emergente que é fortemente relacionada com as técnicas ligadas ao *big data*[1] e da qual a Google seria a primeira e mais importante empresa a funcionar sob. Diante disso, os sistemas de inteligência artificial (IA) passam a ocupar um papel cada vez mais importante no funcionamento da sociedade e, consequentemente, as técnicas de processamento automático de dados denominados "aprendizado de máquina" ganham notoriedade.

Uma das principais questões ligada a essas tecnologias é o *problema do viés*, isto é, a aparição de tendências não intencionais nos resultados de um algoritmo de aprendizado de máquina e que geralmente é vista como problema do ponto de vista ético ou político. Exemplos dessa forma de ação social algorítmica são os casos recentes de tecnologias sexistas, que de alguma forma privilegiam o gênero masculino nos seus resultados (CARRERA, 2020; GONÇALVES; FERREIRA, 2021; PRATES; AVELAR; LAMB,

[1] Embora sejam termos com origens diferentes, inteligência artificial e *big data* estão muito relacionados. Sobre essa relação, conferir (ELISH; BOYD, 2018).

2020), e de tecnologias racistas, que classificam pessoas negras com base em esteriótipos ou falham seletivamente na produção de resultados (ANGWIN et al., 2016; CARRERA, 2020; CRAWFORD; PAGLEN, 2019; SILVA, 2020).

Pasquinelli e Joler (2021) propõem a tipificação do viés em três níveis. O primeiro seria o *viés histórico*: formas estruturais de hierarquias presentes na sociedade anteriormente à intervenção tecnológica, mas que seriam incorporadas, reafirmadas e naturalizadas no modelo de aprendizado de máquina. O segundo nível seria o *viés de base de dados*, a criação de tendências e desbalanços no momento de criação das bases de dados que serão utilizadas pelo modelo. Isso pode acontecer tanto por conta de intervenção humana direta, como no caso dos rotuladores de imagens, mas também por sensores, algoritmos de extração de dados, etc. Nesse sentido, os próprios dados já podem ser resultados de processos algorítmicos (PASQUINELLI; JOLER, 2021, p. 1266). Por fim, há o *viés algorítmico*, uma amplificação das duas formas anteriores de viés que aconteceria propriamente pelo modelo de aprendizado de máquina. Para os autores:

> O problema do viés [algorítmico] surge principalmente devido ao fato de que algoritmos de aprendizado de máquina são um dos [meios] mais eficientes para *compressão de informação*, o que engendra questões de resolução de informação, difração e perda. (PASQUINELLI; JOLER, 2021, p. 1265)

Essa capacidade de compressão é o que seria responsável pelos grandes lucros das empresas de IA, que passam a processar informação mais rápida e eficientemente; mas também é ela que produziria como consequência discriminação social e perda de diversidade cultural (PASQUINELLI; JOLER, 2021, p. 1265).

Diante de tais questões, o objetivo deste trabalho é contribuir para o entendimento mais preciso do funcionamento do viés no aprendizado de máquina. Diferentemente de trabalhos que foquem o viés como resultado de formas históricas de opressão nos meios de onde serão extraídos dados (KOERNER, 2020) ou como desbalanços presentes em bases de dados (CRAWFORD; PAGLEN, 2019), nossa análise focará a produção de viés no algoritmo. Para isso, realizamos a leitura de artigos científicos da área, pois foram considerados por nós como forma privilegiada de descrição do funcionamento de tecnociências emergentes.

MATERIAL E MÉTODOS

O nosso recorte foi um *corpus* inicial de dez artigos científicos de relevância na área de IA produzidos por cientistas da Google. Isso se justifica não só pela posição central e pioneira que a empresa ocupa no uso dessas técnicas no capitalismo contemporâneo (ZUBOFF, 2018), mas também porque a Google é uma das principais produtoras dessas tecnociências[2]. Com isso, a metodologia consistiu no estudo qualitativo de cinco artigos selecionados do *corpus* inicial com ênfase na caracterização da produção e reprodução de valores pelas técnicas ali descritas. Durante a pesquisa, foi dada atenção sobre as noções de neutralidade, objetividade e agência para posterior articulação com uma bibliografia sobre o modo de existência dos objetos técnicos, sua ação social e as relações possíveis com o capitalismo contemporâneo e com o neoliberalismo.

COMPRESSÃO DE INFORMAÇÃO

A principal classe de algoritmos de aprendizado de máquina utilizada atualmente é a das redes neurais profundas. Baseadas originalmente no funcionamento cerebral de animais, sua novidade está na capacidade de representar em vários níveis um dado informacional (LECUN; BENGIO; HINTON, 2015). A diversidade de aplicações possíveis para esses algoritmos faz com que as formas de compressão de informação também sejam bastante variadas. Um modo particularmente notável de representação e compressão de informação são as técnicas de codificação vetorial. Estas consistem em representar um dado qualquer como um ponto num espaço multidimensional (ou ainda, um vetor numérico de *n* elementos, o que dá no mesmo)[3]. Dessa forma, passa a ser possível realizar operações matemáticas com o dado de entrada. Um

[2] Uma pesquisa informal sobre as publicações de um dos principais congressos da área constatou que a Google seria a principal organização produtora de artigos na área de IA em 2020: <https://chuvpilo.medium.com/whos-ahead-in-ai-research-in-2020-2009da5cd799> (acesso: 17/07/2022).

[3] Como exemplo didático, um dado poderia ser representada por um vetor de 4 elementos como: [0,23; 1,01; 0,50; 2,08].

exemplo disso é a técnica de codificação de palavras conhecida como "word2vec" (MIKOLOV et al., 2013):

> De forma surpreendente, essas questões [que visam estabelecer relações entre pares de palavras] podem ser respondidas pela realização de operações algébricas simples com a representação vetorial da palavra. Para descobrir uma palavra que é similar a pequeno [*small*] no mesmo sentido que maior [*biggest*] é similar a grande [*big*], nós podemos simplesmente computar o vetor X = vetor("*biggest"*) - vetor("*big"*) + vetor("*small*"). Então, nós procuramos no espaço vetorial pela palavra mais perto de X medida pela distância de cosseno, e usamos ela como resposta para a questão (nós descartamos as palavras da questão de entrada durante essa busca). Quando os vetores de palavras estão bem treinados, é possível encontrar a palavra correta (palavra menor [*smallest*]) usando esse método. (MIKOLOV et al., 2013, p. 5)

Se do ponto de vista da representação é evidente que a possibilidade de operar matematicamente sobre dados que não são originalmente numéricos é de grande valor, também do ponto de vista da compressão a representação vetorial é vantajosa. Esse tipo de representação diminui a complexidade computacional em relação às técnicas precedentes (MIKOLOV et al., 2013), contornando parcialmente a chamada "maldição da dimensionalidade", o crescimento exponencial do uso de recursos com o aumento das dimensões do dado de entrada (PASQUINELLI; JOLER, 2021, p. 1273). Dessa forma, o artigo é pioneiro em propor uma técnica capaz de representar milhões de palavras no vocabulário, enquanto tentativas anteriores usariam vocabulários de no máximo centenas de milhares de palavras (MIKOLOV et al., 2013).

Apesar de seu sucesso e amplitude de aplicações, a técnica de codificação vetorial ainda é um funcionamento particular de alguns usos das tecnologias de IA. Mas seu entendimento nos ajuda a explicar um outro funcionamento similar e que é comum a todo algoritmo de redes neurais profundas: a representação do conhecimento "aprendido" na forma de parâmetros numéricos (vetores de pesos) (LECUN; BENGIO; HINTON, 2015).

Para melhor caracterizar essa operação, é necessário separar o funcionamento das redes neurais profundas em dois momentos: inferência e treinamento. A inferência é quando o modelo é utilizado para produzir um resultado novo a partir de uma informação inédita; mas, para que isso seja possível, o modelo precisa ser antes *treinado* em um processo iterativo de cálculo de parâmetros numéricos. No treinamento, exemplos advindos de uma base de dados e uma função de otimização (função-objetivo) são mobilizados para calcular um erro entre o resultado produzido e o resultado visado, esse erro é realimentado no algoritmo que melhora sucessivamente seus resultados. No fim desse processo, considera-se que o modelo "aprendeu" a realizar uma tarefa a partir dos exemplos da base de dados (pois atingiu-se uma situação em que o erro é consistentemente baixo). Assim, comprime-se o conhecimento presente nos exemplos da base de dados para os vetores de peso[4].

Dessa forma, o conhecimento de um modelo de aprendizado de máquina é armazenado como um conjunto de parâmetros numéricos e o seu funcionamento pode ser aproximado ao de uma função matemática configurável (pois seu resultado depende dos parâmetros calculados no treinamento) (GONÇALVES, 2022). Por outro lado, o cálculo desses parâmetros é feito de modo a minimizar o erro na tarefa descrita pela função-objetivo de modo que os parâmetros que determinam o funcionamento do modelo são o resultado de uma operação de otimização.

OTIMIZAÇÃO E CONCORRÊNCIA

A partir do esquema do funcionamento da operação de compressão de informação exposto acima, entendemos que a principal ação executada por um algoritmo de aprendizado de máquina é a otimização, isto é, a maximização ou a minimização de uma função-objetivo. Associaremos essa operação com a ideia de *eficiência*: a máxima realização de uma tarefa (ou a mínima produção de erros) a partir de uma quantidade definida de recursos.

Embora seja muito difícil precisar as relações de causalidade, é possível constatar que existe uma ressonância em múltiplos níveis em torno dessa noção de eficiência nos artigos analisados. O funcionamento de um

[4] Para uma referência atual e relevante sobre as redes neurais profundas, conferir (GOODFELLOW; BENGIO; COURVILLE, 2016).

algoritmo de IA se realiza através de uma otimização que maximizará a taxa de acertos e minimizará o erro em uma determinada tarefa. Os próprios artigos constroem sua argumentação sobre métricas de acurácia, sendo que os principais argumentos presentes nos artigos analisados são que o modelo proposto teria o "melhor desempenho" (MIKOLOV et al., 2013, p. 4), "superaria [*outperform*]" um modelo concorrente (MNIH et al., 2015, p. 530-1; SUTSKEVER; VINYALS; LE, 2014, p. 2) ou "avançaria o estado-da-arte" na resolução de tarefas determinadas (DEVLIN et al., 2019, p. 2; SZEGEDY et al., 2014, p. 2). Dessa forma, estabelece-se também um campo concorrencial em que os diversos modelos de aprendizado de máquina disputariam entre si em busca daquele que produz a maior taxa de acertos (a concorrência como meio para atingir a eficiência). Por fim, é uma característica das sociedades capitalistas o incentivo a situações de concorrência e a comportamentos que levem ao máximo desempenho.

Dardot e Laval (2016) propõem uma caracterização das sociedades neoliberais – entendendo neoliberalismo não apenas como uma ideologia ou como um conjunto de políticas, mas como um modo de funcionamento da sociedade. Para os autores, ele se define pela implantação – inclusive por meio de políticas estatais, mas também pelo uso de ferramentas de microgestão e uma produção cultural que incentiva o autoaperfeiçoamento – de uma ordem generalizada de concorrência e do privilégio da empresa como agente social de referência em todas as esferas da vida. O próprio indivíduo passa a ser visto como sendo uma "empresa de si" em situação de concorrência. Isso se reflete no campo subjetivo como um incentivo desenfreado pela produção e pelo prazer, que os autores denominarão "dispositivos de desempenho e gozo":

> Até então, essa exigência própria do regime de acumulação do capital não havia desdobrado todos os seus efeitos. Isso aconteceu quando o comprometimento subjetivo foi tal que a procura desse "além de si mesmo" tornou-se a condição de funcionamento tanto dos sujeitos como das empresas. Daí o interesse da identificação do sujeito como empresa de si mesmo e capital humano: a extração de um "mais-de-gozar", tirado de si mesmo, do prazer de viver, do simples fato de viver, é que faz funcionar o novo sujeito e o novo sistema de concorrência. Em última análise, subjetivação "contábil" e subjetivação "financeira" definem uma *subjetivação pelo excesso de si em si* ou, ainda, pela *superação indefinida de si*. (DARDOT; LAVAL, 2016, p. 356–7)

Assim, é possível notar um movimento de afirmação da noção de eficiência que se realimenta. A construção de tecnologias num contexto social determinado, promove a incorporação naquelas dos valores deste. Mas também, o uso de tecnologias na sociedade naturaliza e concretiza valores associados com o modo específico de funcionamento técnico. Particularmente, a racionalidade neoliberal incentiva a valorização de técnicas que maximizem a taxa de acertos, enquanto que tecnologias de otimização naturalizam e intervêm no mundo a partir da lógica neoliberal.

Por mais que pareça óbvio visar maximizar a taxa de acertos, esse privilégio da noção de eficiência não ocorre isento de problemas. Sua constante reafirmação é feita em substituição a outras valorações possíveis. É por isso que a incorporação dessa lógica nos sistemas de IA amplifica as opressões históricas. Privilegiar a maior taxa de acerto é, por exemplo, privilegiar a tradução de termos neutros em relação ao gênero para o masculino (PRATES; AVELAR; LAMB, 2020), ou relacionar pronomes com profissões estereotipadas em relação ao gênero (GONÇALVES; FERREIRA, 2021). Ou seja, a ênfase na eficiência é uma decisão sociotécnica, embutida nas técnicas de aprendizado de máquina e que tem como consequências éticas e políticas a reprodução de formas históricas de discriminação e a redução da diversidade cultural.

CONCLUSÃO

A emergência de novas tecnologias traz consigo novos problemas. A expansão do uso de tecnologias ditas "inteligentes" e especialmente das técnicas de aprendizado de máquina para o processamento automático de dados, colocou luz sobre a questão do *viés*. Pasquinelli e Joler (2021) propõe a divisão deste em três níveis: o *viés histórico*, opressões estruturais presentes na sociedade antes da intervenção tecnológica, o *viés de base de dados*, desbalanços causados na coleta e no tratamento de dados e o *viés algorítmico*, amplificação de tendências dos

vieses anteriores causada pelo modelo de aprendizado de máquina.

Neste trabalho, focamos no terceiro tipo de viés, argumentando que, como consequência do processo de compressão de informação, *existe um viés de eficiência nas técnicas de aprendizado de máquina*. Embora tenhamos mostrado que isso é principalmente uma forma de viés algorítmico, mostramos também como a tendência por resultados eficientes está relacionada com a naturalização ou a incorporação de um viés histórico que busca a maximização do desempenho – ideia presente, sobretudo, no discurso neoliberal. Assim, propomos que existe uma ressonância entre os funcionamentos do aprendizado de máquina e do neoliberalismo que reafirmam a eficiência como valor a ser reproduzido a despeito das consequências disso. Por fim, argumentamos que esse viés pode ser entendido como uma das causas da discriminação social e da diminuição da diversidade cultural observadas nos resultados de sistemas de IA.

AGRADECIMENTOS

Agradecemos ao PIBIC/CNPq pelo financiamento da nossa pesquisa de iniciação científica, sem o qual o presente trabalho não poderia ter sido realizado.

BIBLIOGRAFIA

ANGWIN, J. et al. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. **ProPublica**, 23 maio 2016.

CARRERA, F. Racismo e sexismo em bancos de imagens digitais: análise de resultados de busca e atribuição de relevância na dimensão financeira/profissional. In: SILVA, T. (Ed.). . **Comunidades, algoritmos e ativismos digitais: Olhares afrodiaspóricos**. 1. ed. São Paulo: LiteraRUA, 2020. p. 139–153.

CRAWFORD, K.; PAGLEN, T. Excavating AI: The politics of images in machine learning training sets. **The AI Now Institute, NYU**, 19 set. 2019.

DARDOT, P.; LAVAL, C. **A nova razão do mundo**. Tradução: Mariana Echalar. 1. ed. São Paulo: Boitempo, 2016.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv:1810.04805 [cs]**, 24 maio 2019.

ELISH, M. C.; BOYD, DANAH. Situating methods in the magic of Big Data and AI. **Communication Monographs**, v. 85, n. 1, p. 57–80, 2 jan. 2018.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] MIT Press, 2016. Disponível em: <www.deeplearningbook.org>. Acesso em: 22 jul. 2022.

GONÇALVES, R. Automatismo ontem e hoje: reflexões sobre os limites da inteligência artificial a partir de Simondon. **Ideias**, v. 13, e022008, p. 1–22, 21 jun. 2022.

GONÇALVES, R.; FERREIRA, P. P. Aprendizado de máquina como mediação técnica: uma investigação sobre viés de gênero no modelo de linguagem BERT. **XXIX Congresso de Iniciação Científica da UNICAMP**, 2021. Disponível em: <https://proceedings.science/pibic-2021/papers/aprendizado-de-maquina-como-mediacao-tecnica--uma-investigacao-sobre-vies-de-genero-no-modelo-de-linguagem-bert>. Acesso em: 7 jul. 2022.

KOERNER, J. Wikipedia Has a Bias Problem. In: REAGLE, J.; KOERNER, J. (Eds.). **Wikipedia@ 20: Stories of an Incomplete Revolution.** Cambridge, Massachusetts: The MIT Press, 2020. p. 311–321.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, maio 2015.

MIKOLOV, T. et al. Efficient Estimation of Word Representations in Vector Space. **arXiv:1301.3781 [cs]**, 6 set. 2013.

MNIH, V. et al. Human-level control through deep reinforcement learning. **Nature**, v. 518, n. 7540, p. 529–533, 26 fev. 2015.

PASQUINELLI, M.; JOLER, V. The Nooscope manifested: AI as instrument of knowledge extractivism. **AI & SOCIETY**, v. 36, n. 4, p. 1263–1280, 1 dez. 2021.

PRATES, M. O. R.; AVELAR, P. H.; LAMB, L. C. Assessing gender bias in machine translation: a case study with Google Translate. **Neural Computing and Applications**, v. 32, n. 10, p. 6363–6381, maio 2020.

SANTOS, L. G. DOS. A informação após a virada cibernética. In: **Revolução tecnológica, internet e socialismo**. Socialismo em discussão. 1a. ed ed. São Paulo, SP, Brasil: Editora Fundação Perseu Abramo, 2003. p. 9–34.

SILVA, T. Racismo algorítmico em plataformas digitais: microagressões e discriminação em código. In: SILVA, T. (Ed.). . **Comunidades, algoritmos e ativismos digitais: Olhares afrodiaspóricos**. 1. ed. São Paulo: LiteraRUA, 2020. p. 121–137.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to Sequence Learning with Neural Networks. **arXiv:1409.3215 [cs]**, 14 dez. 2014.

SZEGEDY, C. et al. Going Deeper with Convolutions. **arXiv:1409.4842 [cs]**, 16 set. 2014.

ZUBOFF, S. Big other: capitalismo de vigilância e perspectivas para uma civilização de informação. In: BRUNO, F. et al. (Eds.). . **Tecnopolíticas de Vigilância: perspectivas da margem**. 1. ed. São Paulo: Boitempo, 2018. p. 17–68.