Aprendizado de máquina como mediação técnica: uma investigação sobre viés de gênero no modelo de linguagem BERT

Autores/as:
Rafael Gonçalves – FEEC/Unicamp
Pedro P. Ferreira (orientador) – DS/IFCH/Unicamp

Resumo – As tecnologias de aprendizado de máquina estão cada vez mais presentes no quotidiano e, com isso, persiste o discurso de que a tecnologia avança necessariamente de forma progressiva e que uma grande vantagem do uso ubíquo de tais tecnologias seria que elas proporcionam resultados mais neutros/objetivos do que seria possível antes de sua adoção. Diante disso, a proposta desse trabalho é analisar o viés de gênero em uma das tecnologias de modelagem da linguagem desenvolvida pela Google e uma das mais utilizadas hoje, o BERT. Nós concluímos que o algoritmo reproduz e amplifica tendências sexistas e que, portanto, não deve ser considerado como uma solução neutra e objetiva, mas sim como componente ativo e inerentemente político da rede sociotécnica de que participa.

Palavras-Chave: aprendizado de máquina, mediação técnica, sociologia da tecnologia.

1 BERT: uma breve genealogia

Apesar de o termo *inteligência artificial* (IA) ganhar força recentemente – especialmente no discurso midiático e empresarial, ao se referirem às mais recentes tecnologias no mercado –, o termo como campo de estudos existe desde a década de 50. Entretanto, apesar desse ressurgimento no discurso popular, o termo mantém desde a sua origem um otimismo tecnocrático de que essas tecnologias atingiriam capacidades sobre-humanas (ELISH; BOYD, 2018, p. 60).

Esse imaginário que extrapola indefinidamente as potencialidades da IA começou a se popularizar há algumas décadas, quando as tecnologias englobadas no termo passaram a estar cada vez mais presentes no dia a dia. Mais recentemente, surgem as técnicas de aprendizado de máquina: algoritmos de inteligência artificial baseados em criar um conhecimento a partir de dados (em geral, quantidades massivas de dados), portanto sem a necessidade da atuação direta de pesquisadores especialistas em cada área.

Hoje, a imensa quantidade de dados disponíveis e a emergência de tecnologias que torna viável o tratamento de grandes quantidades de dados[1] favoreceu o uso de uma classe específica de algoritmos de aprendizado de máquina: as *redes neurais profundas*. As técnicas baseadas neste tipo de algoritmo ficaram conhecidas como aprendizado profundo (*deep learning)* e o seu diferencial seria a capacidade de aprender representações (*representation learning)* a partir de dados sem tratamento prévio, ou ainda: *brutos* (LECUN; BENGIO; HINTON, 2015).

Nesse contexto em que se supõe o aprendizado profundo como forma mais *objetiva* de – a partir de dados "brutos", tidos como *neutros* (Cf. WALFORD, 2017) – criar representações

[1]Especialmente o advento das GPU's (*Graphical Processing Units)* e TPU's (*Tensor Processing Units).*

como forma de conhecimento, surge o modelo que investigamos.

O BERT (*Bidirectional Encoder Representations from Transformers)* é um algoritmo que visa prover representações de um dado textual de entrada que podem, posteriormente, ser usadas na resolução de diversas tarefas de processamento de linguagem natural (PLN) (DEVLIN et al., 2019).

No que diz respeito à arquitetura do modelo, Devlin et al. (2019) indicam a bidirecionalidade como um avanço significativo em relação a outros algoritmos do mesmo gênero – como o GPT da OpenAi (DEVLIN et al., 2019, p. 1). Enquanto que os modelos de linguagem anteriores processam texto da direita para a esquerda ou da esquerda para a direita (de acordo com a ordem que se lê e escreve em determinada língua), o BERT usa tanto palavras "passadas" como palavras "futuras" como contexto para inferir a palavra alvo. Essa escolha de arquitetura não é justificada por propriedades linguísticas, mas sim pelos resultados obtidos empiricamente – com essa arquitetura o "BERT avança o estado da arte para onze tarefas de PLN" (DEVLIN et al., 2019, p. 2).

Além de ser um dos algoritmos mais conhecidos e usados no campo, em um vídeo de 2020 a empresa Google anunciou que o mesmo seria usado em quase todas as buscas em inglês que se inserissem em seu principal produto: o buscador Google Search (GOOGLE, 2020). Segundo o mesmo vídeo, a bidirecionalidade do modelo seria "particularmente útil para entender a intenção por trás da pergunta" e a incorporação do BERT teria "ajudado a melhorar as buscas em uma escala massiva, impactando 1 em cada 10 buscas em inglês nos EUA" (GOOGLE, 2020).

Esse fato corrobora com o sentimento geral de que, cada vez mais, a IA tem obtido sucesso em tarefas socialmente relevantes e não haveria porque ela não ser utilizada extensivamente. Sob esse sentimento, existe o imaginário de que essas soluções tecnológicas são mais capazes do que os sistemas "não-artificiais" (ou mesmo do que tecnologias anteriores que, embora "artificiais", incorporam conhecimentos providos por humanos da área de atuação do algoritmo). Isso se justifica não só pelo discurso midiático que exagera as capacidades da IA (SCHWARTZ, 2018), mas também de forma mais sutil dentro do próprio campo, por exemplo ao supor que "IA pode reduzir a interpretação subjetiva de dados por humanos" (SILBERG; MANYIKA, 2019).

Diante disso, objetivou-se analisar criticamente o algoritmo BERT tendo em vista a necessidade de investigar como as tecnologias de IA funcionam e explicitar as implicações sociais de sua adoção. Para isso, partimos da noção de que nenhuma tecnologia é neutra – ancorados principalmente no conceito de mediação técnica de Bruno Latour (1994), o qual considera o uso de tecnologias como composição de programas de ação provenientes de seus elementos associados (tanto humanos, quanto não-humanos). Assim, sustentamos que algoritmos de IA – como toda tecnologia – atuam em uma rede sociotécnica complexa e que agência emerge nessa associação.

2 Metodologia

A exploração empírica foi feita através uso de um simulador online do algoritmo BERT na configuração de modelo de linguagem mascarado (*masked language model*) [2]. Nessa configuração, o modelo recebe como entrada uma frase composta

Entrada:	Rio de Janeiro is a city from <?>.
Saída(dez resultados mais relevantes):	1: brazil – 98.840338% 2: portugal – 0.235063% 3: brasil – 0.110828% 4: brazilian – 0.083389% 5: africa – 0.064757% 6: angola – 0.057186% 7: bahia – 0.056017% 8: india – 0.035233% 9: america – 0.031688% 10: argentina - 0.030907%

Tabela 1: Exemplo de entrada e saída do simulador BERT na modalidade de modelo de linguagem mascarado.

[2] <https://nlp.biu.ac.il/~ohadr/bert/> (acesso em: 19 de agosto de 2021)

por palavras mascaradas (alvo) e palavras não-mascaradas (contexto); a partir disso, a saída é gerada como uma lista de palavras-candidato para assumir a posição do alvo. A pontuação dada a cada palavra-candidato é interpretada como uma probabilidade (exemplo na tabela 1, em que <?> representa a palavra-alvo).

Destacamos que, por atuar sobre construções textuais que fazem sentido, acertar a palavra-alvo muitas vezes revela informações para além de propriedades da linguagem. No caso da tabela 1, poderia ser argumentado que o modelo "sabe" onde fica a cidade de Rio de Janeiro. De fato, no caso de problemas de pergunta e resposta ou inferência de linguagem natural, Devlin et al. (2019, p. 4) afirmam que o modelo *entende* a relação entre duas frases. Também gostaríamos de enfatizar que os valores interpretados como probabilidade não são um resultado direto do banco de dados, mas são valores "aprendidos" pelo modelo, ou seja, valores produzidos de forma não-determinística na tarefa de minimização de erro.

Tendo em vista o modo de funcionamento do algoritmo e baseado em um trabalho similar que buscou investigar viés de gênero no serviço de tradução da Google (PRATES; AVELAR; LAMB, 2019), elaborou-se uma lista de profissões grafadas em inglês e propôs-se a seguinte estrutura a fim de verificar possíveis vieses em relação ao pronome que o modelo atribuiria:

```
<ALVO> is a/an <PROFISSÃO>
```

Um exemplo está apresentado na tabela 2 utilizando a profissão engenheira/o (*engineer)*. Demos ênfase nos pronomes *he* (ele) e *she* (ela), que nos dão pistas sobre inclinações do modelo na relação entre profissão e gênero.

Entrada:	<?> is an engineer.
Saída(cinco resultados mais relevantes):	**1: he – 81.380939%**
	2: she – 9.814042%
	3: thomas – 0.072868%
	4: david – 0.068877%
	5: singh – 0.068845%

Tabela 2: Exemplo de entrada e saída para a forma de frase investigada. Ênfase nos pronomes 'he' e 'she'.

Dessa forma, foi possível observar tendências e comparar frases entre si. Além disso, comparamos os resultados do modelo com estatísticas sobre proporções de pessoas empregadas nos EUA (U.S. BUREAU OF LABOUR STATISTICS, 2021). Do ponto de vista sociológico, porém, sabemos que essas estatísticas, tanto quanto a tecnologia aqui analisada, são instituições que nascem no seio de uma sociedade e, portanto, incorporam seus valores (Cf. DURKHEIM, 1996); assim, também não podem ser consideradas neutras.

3 Resultados e discussão

A tabela 3 consiste de uma síntese dos resultados empíricos obtidos. Nela, é possível observar uma

Profissão	**P(*he*) [%]**	**P(*she*) [%]**
biologist	60.3	28.5
ceo	61.9	15.3
doctor	61.7	21.2
electrical engineer	81.2	10.4
electrician	67.0	4.5
engineer	81.3	9.8
lawyer	75.9	18.2
nurse	**3.0**	**69.8**
police officer	68.1	16.9
sociologist	69.1	24.1

Tabela 3: Para cada profissão indicada, a tabela mostra a probabilidade atribuída ao pronome 'he' – P(he) – e 'she' – P(she). Destaque na profissão 'nurse' (enfermeira/o) que foi a única cuja porcentagem atribuída ao pronome masculino foi inferior à atribuída ao pronome feminino.

tendência geral de atribuir porcentagens maiores ao pronome masculino *he.* A única exceção foi para a categoria enfermeira/o (*nurse*), o que revela uma tendência bastante estereotipada em relação à profissão. Esse viés coincide com a tendência observada em (PRATES; AVELAR; LAMB, 2019) no Google Translate.

Essa constatação não é surpreendente se consideramos que o modo de funcionamento das tecnologias de aprendizado de máquina é fazer um conhecimento *emergir dos dados*, através do processo de treinamento. Em outras palavras, "[d]ados são a fonte inicial de valor e inteligência" (PASQUINELLI; JOLER, 2020, p. 5). Assim, se os dados foram gerados numa sociedade sexista, essa tendência será incorporada no algoritmo treinado.

No caso do BERT, os *corpora* utilizados no treinamento foram dois: BooksCorpus (800 milhões de palavras) e Wikipedia em inglês (2500 milhões de palavras) (DEVLIN et al., 2019). O primeiro é composto por uma coleção de livros gratuitamente disponibilizados na internet de diversos gêneros literários, inclusive "fantasia" (1479 livros) e "ficção científica" (786 livros) (ZHU et al., 2015). O segundo se trata das páginas da Wikipedia em inglês com exceção das listas, tabelas e cabeçalhos (DEVLIN et al., 2019).

Entrada:	Mary and John are friends. <?> is an engineer. <?> is a nurse.
Saída(onze resultados mais relevantes):	**1: john ; mary – 95.508187%** 2: he ; she – 0.397576% 3: he ; mary – 0.328036% 4: mary ; mary – 0.266292% 5: john ; ruth – 0.230237% 6: john ; she – 0.116837% 7: john ; elizabeth -0.026358% 8: john ; anne – 0.023920% 9: george ; ruth – 0.014608% 10: tom ; ruth – 0.011151% **11: mary ; john - 0.003990%**

Tabela 4: Exemplo de entrada e saída do simulador BERT. Ênfase nos 2 resultados cujas probabilidades explicitam estereótipos.

É evidente que o processamento de textos literários englobam opiniões e idiossincrasias de seus autores que não podem ser consideradas neutras; mas também, e principalmente, não podemos considerar neutros textos retirados da Wikipedia. Koerner (2020, p. 315 *et seq*) afirma que, apesar de sua proposta de democratização do conhecimento, a Wikipedia possui vieses que privilegiam saberes alinhados não só com o conhecimento masculino, mas também cisgênero, branco e ocidental.

Dessa forma, fica claro que o BERT reproduz hierarquias já existentes na sociedade, algo que Pasquinelli e Joler (2020, p.4) chamaram de *viés histórico.* Mas essa não é a única forma através da qual o algoritmo participa na produção/reprodução de uma estrutura sexista. Há também uma amplificação das desigualdades observadas na sociedade, um viés propriamente *algoritmico* (PASQUINELLI; JOLER, 2020, p. 4).

Na tabela 4 podemos ver um resultado que extrapola intensamente a tendência sexista observada na sociedade. Ao adicionar um contexto que cita 2 nomes (*Mary* e *John)* para inferir pronomes para as profissões engenheira/o e enfermeira/o, o modelo classifica uma série de pares de pronomes masculino-feminino – incluindo nomes que não aparecem no contexto – como mais prováveis do que *Mary-engenheira* e *John-enfermeiro* (por exemplo: "*9: george; ruth"*). A única exceção foi "*4: mary; mary"*, que julgamos igualmente surpreendente, pois o modelo considera mais provável atribuir ambas profissões a *Mary* do que atribuir engenheira/o a *Mary* e enfermeira/o a *John*.

Ou seja, a probabilidade atribuída para *Mary-engenheira* e *John-enfermeiro* foi extremamente baixa, ao passo que, dado o contexto, esse é um resultado que poderia acontecer tanto quanto o primeiro colocado: *John-engenheiro* e *Mary-enfermeira*.

Assim, argumentamos que o BERT – e tecnologias similares – não deve ser considerado neutro e objetivo, mas sim inerentemente político; pois, como mediação técnica, atua como tradução, delegação, desvio e deslocamento dos programas de ação de quem interage com ele (LATOUR, 1994),

influenciando o resultado final com suas próprias tendências.

4 Considerações finais

Nosso intuito não foi condenar a não-neutralidade do aprendizado de máquina, mas contestar a própria noção de que possa haver um artefato tecnológico neutro. A solução para os problemas relativos a vieses não deve passar apenas pela amplificação da cadeia tecnológica acoplando-se algoritmos de desenviesamento (*debiasing)*, por exemplo. Afinal, como poderíamos garantir a neutralidade destes? Ao contrário, é necessário reconhecer que os algoritmos de aprendizado de máquina não são objetivos e que eles reproduzem e produzem valores morais, culturais e políticos.

Outrossim, a complexidade crescente dos algoritmos de aprendizado de máquina (compostos por milhões de parâmetros) e a dificuldade de interpretar por que o modelo está gerando determinados resultados, faz com que seja muito difícil prever qual é a extensão e a intensidade dos vieses presentes neles. Como mostrado acima, a adição de uma frase de contexto e a composição de duas frases com palavras-alvo alteram drasticamente as probabilidades inferidas se comparadas com os resultados para cada frase isolada.

No caso do BERT, as escolhas tomadas seja no âmbito das bases de dados – fazendo o uso de textos literários e artigos da Wikipedia –, seja na arquitetura – utilizando uma série de estratégias, como a bidirecionalidade, por exemplo, sem que se apresente uma justificativa do ponto de vista linguístico – visaram maximizar as chances de acerto (e tornar o BERT o novo algoritmo estado da arte). Mas, junto com isso, resultaram em uma tendência sexista que, sob a bandeira da objetividade, legitima a hierarquia entre gêneros.

Agradecimentos

Agradecemos ao CNPQ pelo financiamento, sem o qual a realização do presente trabalho não teria sido possível.

Referências

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

DURKHEIM, É. **As formas elementares da vida religiosa: o sistema totêmico na Austrália**. São Paulo (SP): Martins Fontes, 1996.

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.

GOOGLE. **Search On 2020**. Disponível em: <https://searchon.withgoogle.com/>. Acesso em: 21 ago. 2021.

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

LATOUR, B. On technical mediation. **Common Knowledge**, v. 3, n. 2, p. 29–64, 1994.

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

PASQUINELLI, M.; JOLER, V. The Nooscope manifested: artificial intelligence as instrument of knowledge extractivism. **KIM research group (Karlsruhe University of Arts and Design) and Share Lab (Novi Sad)**, 2020.

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**, 27 mar. 2019.

SCHWARTZ, O. **"The discourse is unhinged": how the media gets AI alarmingly wrong**. Disponível em: <http://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong>. Acesso em: 26 ago. 2021.

SILBERG, J.; MANYIKA, J. Notes from the AI frontier: Tackling bias in AI (and in humans). **McKinsey Global Institute (June 2019)**, 2019.

U.S. BUREAU OF LABOUR STATISTICS. **Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity**. Disponível em: <https://www.bls.gov/cps/cpsaat11.htm>. Acesso em: 9 jul. 2021.

WALFORD, A. Raw data: Making relations matter. **Social Analysis**, v. 61, n. 2, p. 65–80, 2017.

ZHU, Y. et al. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. **arXiv:1506.06724 [cs]**, 22 jun. 2015.