ARTIGO

Algoritmo: viés, poder e ética na era da Inteligência Artificial

Algorithm: bias, power, and ethics in the age of Artificial Intelligence

Elaine Maria Gomes de Abrantes
elamar.esmapb@gmail.com
Universidade Estadual do Rio Grande do Norte, Pau dos Ferros, RN, Brasil

Leandra de Oliveira Silva
leandra.silva@copin.ufccg.edu.br
Universidade Federal de Campina Grande, Campina Grande, PB, Brasil

Erick Araken Vieira Gomes
erick.gomes@ccc.ufccg.edu.br
Universidade Federal de Campina Grande, Campina Grande, PB, Brasil

Linguagem em Foco
Revista do Programa de Pós-Graduação em Linguística Aplicada da UECE

FLUXO DA SUBMISSÃO
Submissão do trabalho: 01/08/2025
Aprovação do trabalho: 08/11/2025
Publicação do trabalho: 09/12/2025

10.46230/lef.v17i3.16081

COMO CITAR
ABRANTES, Elaine Maria Gomes *et al*. Algoritmo: viés, poder e ética na era da Inteligência Artificial. **Revista Linguagem em Foco**, v.17, n.3, 2025. p. 51-69. Disponível em: https://revistas.uece.br/index.php/linguagememfoco/article/view/16081.

Resumo
Através da exploração de um banco de dados real, realizamos um estudo empírico para tentar responder às perguntas de pesquisa: se existe e se é possível identificarmos vieses discriminatórios na IA, bem como, em uma segunda etapa, se há possibilidade de mitigação. O procedimento metodológico adota,do inclui a fase de identificação de vieses, através da aplicação no banco de dados de três tipos de técnicas de processamento (pre-processing, in-processing e post-processing), em que os resultados foram comparados e após verificadas quais as decisões tomadas pelo modelo, resultando numa IA Explicável, cujos dados podem ser compreendidos, escrutinados e contestados conforme os valores descritos em nossa Carta Magna de 1988. Na fase futura de proposição de ações, a fim de aproximar a ética constitucional do fazer algoritmo, pretendemos "treinar a máquina" para incluir os fundamentos, valores e princípios da Constituição Federal brasileira e ver se o sistema torna o banco de dados menos enviesado. O objetivo linguístico imediato é desconstruir a crença na infalibilidade da automação e propor melhorias futuras aos operadores de sistemas. Para melhor compreensão dos preconceitos/discriminações reproduzidos pela IA, o quadro teórico da Análise Crítica do Discurso e na Teoria Decolonial serviu de bússola. As contribuições são múltiplas e se inserem no escrutínio de um discurso que se imaginava isento das mazelas sociais e na aproximação dos operadores de máquinas de uma ética mais palpável, prevista em um Documento confiável. Os resultados ainda são preliminares e só dão conta da fase de identificação de vieses, pois conseguimos demonstrar sua existência e apontá-los no banco de dados escolhido. O oferecimento de ferramentas de enfrentamento ainda ficará para a oportunidade de conclusão da pós-graduação. Por

hora, partilhamos a certeza de abusos produzidos nos meios tecnológicos, em consonância com os outros existentes nas outras áreas de nossa sociedade.
Palavras-chave
Algoritmos. Discurso de Neutralidade. Comprovação de Vieses. Mitigação.

Abstract
Through the exploration of a real database, we conducted an empirical study to try to answer the research questions: whether discriminatory biases exist and can be identified in AI, and, in a second stage, whether mitigation is possible. The methodological procedure adopted includes a bias identification phase, through the application of three types of processing techniques to the database (pre-processing, in-processing, and post-processing), in which the results were compared and the decisions made by the model were verified, resulting in an Explainable AI, whose data can be understood, scrutinized, and challenged according to the values described in our 1988 Constitution. In the future phase of proposing actions, in order to bring constitutional ethics closer to algorithm development, we intend to "train the machine" to include the foundations, values, and principles of the Brazilian Federal Constitution and see if the system makes the database less biased. The immediate linguistic objective is to deconstruct the belief in the infallibility of automation and propose future improvements for system operators. To better understand the biases/discriminations reproduced by AI, the theoretical framework of Critical Discourse Analysis and Decolonial Theory served as a compass. The contributions are multiple and fall within the scrutiny of a discourse that was imagined to be free from social ills and in bringing machine operators closer to a more tangible ethic, foreseen in a reliable Document. The results are still preliminary and only account for the bias identification phase, as we were able to demonstrate their existence and point them out in the chosen database. The offering of coping tools will be left for the opportunity to complete the postgraduate course. For now, we share the certainty of abuses produced in technological means, in line with others existing in other areas of our society.
Keywords
Algorithms. Neutrality Discourse. Bias Verification. Mitigation.

Introdução

Com a ascensão da Inteligência Artificial (IA) na vida contemporânea, difundiu-se a ideia equivocada de que essas tecnologias, por estarem alicerçadas em lógica matemática, seriam objetivas e neutras. Como observa Gorenc (2025), esta percepção comum, frequentemente reforçada por discursos corporativos, defende que os algoritmos, como extensões de operações objetivas, não poderiam ser racistas, sexistas ou enviesados.

Mas, embora os sistemas de IA sejam vendidos como capazes de otimizar processos complexos que vão desde a contratação de pessoal até a aplicação da lei, conforme Rosenthal-von der Pütten e Sach (2024), não nos conformamos com esses discursos de neutralidade e resolvemos perguntar nessa nossa pesquisa se e como eles perpetuam preconceitos e quais possibilidades de mitigá-los?

A crença na infalibilidade da automação constitui um mito que pretendemos desconstruir por meio de investigação empírica, estruturada em duas fases: **identificação e mitigação de vieses**. Assim, o estudo se justifica pela necessidade de questionar o mito da neutralidade algorítmica e propor caminhos para sistemas mais éticos e justos.

Assim, através de um estudo empírico com a utilização de um banco de

dados real, vamos buscar responder às indagações postas, em duas fases. A primeira fase de verificação de vieses será reportada neste artigo e incluiu a aplicação ao banco de dados de três técnicas de processamento: pre-processing, in-processing e post-processing, cujos resultados foram comparado para obter uma IA Explicável, de forma a proporcionar compreensão, escrutínio e contestação das decisões, tendo por baliza os fundamentos da nossa Constituição Federal de 1988, quais sejam, a soberania; a cidadania; a dignidade da pessoa humana; o trabalho, a livre iniciativa e o pluralismo, bem como dos fundamentos da erradicação da pobreza e da marginalização, a redução das desigualdades, a promoção do bem comum e o não preconceito ou discriminação de qualquer tipo (Brasil, 1988).

A suposição inicial é que os algoritmos, como sistemas sociotécnicos profundamente enraizados nos contextos históricos, sociais e políticos que os produzem, estão longe de serem entidades neutras. Ao contrário, achamos que eles refletem, reproduzem e, em muitos casos, amplificam as estruturas de poder e as desigualdades existentes. Afinal, como defende Windhorst (2025), a tecnologia é desenvolvida e utilizada por seres humanos, não podendo ser neutra, e se desenvolve dentro de sociedades marcadas por opressões estruturais, conforme nos informa Tavarone (2025).

Para identificar possíveis discriminações algorítmicas e compreender como emergem como expressão tecnológica do racismo estrutural (Silva, 2021), o artigo organiza-se em três seções:

Seção 1 – “As Múltiplas Faces da Discriminação na Esfera Digital”: apresenta exemplos de racismo, violência de gênero e exclusão interseccional, mostrando como mecanismos algorítmicos podem gerar danos concretos e reproduzir desigualdades.

Seção 2 – “Lentes Críticas para uma Análise Mais Profunda”: aplica a Análise Crítica do Discurso e a teoria decolonial para interpretar a condição algorítmica e revelar como a linguagem técnica participa da naturalização das injustiças digitais.

Seção 3 – “Anatomia da Injustiça Algorítmica”: descreve os mecanismos técnicos do viés, o problema da opacidade e apresenta os resultados empíricos obtidos.

Finalmente, nas considerações finais serão apresentadas possibilidades para a tentativa futura de mitigações das injustiças encontradas, bem como proposição de regulação baseada em valores constitucionais e discutida a importân-

cia de resistência social, sobretudo discursiva.

1 As múltiplas faces da discriminação na esfera digital

Viés, opacidade e ciclos de retroalimentação não são abstrações: na esfera digital, materializam-se em sistemas reais e produzem danos tangíveis a indivíduos e comunidades. Esta seção mapeia diferentes faces da discriminação algorítmica com foco no contexto brasileiro.

O conceito de "racismo algorítmico" é central para a compreensão dos impactos da IA em sociedades com profundas desigualdades raciais, como o Brasil. Como atesta autores como Silva (2021), não se trata de uma nova forma de preconceito, mas sim representa uma continuação e modernização do racismo estrutural já existente, só que desta feita realizada por meios tecnológicos.

Na segurança pública, isso é particularmente evidente. Sistemas de reconhecimento facial, apresentados como neutros e eficientes, erram mais para pessoas negras — sobretudo mulheres negras — do que para homens brancos (Lisboa; Passos; Ferreira, 2025). No Brasil, cerca de 90% das prisões com auxílio dessa tecnologia atingem pessoas negras (Silva; Silva, 2024). O programa “Smart Sampa”, em São Paulo, ilustra o problema ao confundir rostos negros com procurados, reproduzindo o racismo institucional (Oliveira, 2025). Esses resultados são coerentes com o perfil do sistema penal, no qual quase 70% da população carcerária é negra (Silva, 2025).

Além da esfera da segurança pública, o racismo algorítmico permeia diversas outras áreas da vida digital, tais como:

- **Moderação de Conteúdo e Liberdade de Expressão**: Conteúdos de pessoas negras que denunciam injustiças tendem a ser removidos de forma desproporcional, por classificação automatizada que ignora contexto e silencia vozes (Silva; Silva, 2024).
- **Viés em Motores de Busca e Visão Computacional**: Algoritmos de busca e rotuladores podem perpetuar estereótipos, como no episódio em que o Google Fotos classificou pessoas negras como “gorilas”, expondo viés nos dados de treinamento (Lisboa; Passos; Ferreira, 2025; Oliveira; Souza, 2024).
- **O viés de gênero**: é outra manifestação perversa da injustiça algorítmica, afetando as mulheres de formas que vão desde a exclusão de

oportunidades econômicas até formas explícitas de violência digital. Os sistemas de IA, treinados com dados que refletem séculos de patriarcado, aprendem e codificam visões estereotipadas sobre os papeis e capacidades de gênero.

- **No mundo profissional**: o caso da ferramenta de recrutamento da Amazon tornou-se um exemplo clássico. O sistema foi treinado com uma década de currículos da própria empresa, que refletiam o domínio masculino na indústria tecnológica. Como resultado, o algoritmo aprendeu a penalizar currículos que continham termos como "mulher" ou referências a faculdades femininas, favorecendo sistematicamente candidatos do sexo masculino (Viana; Macedo, 2024).
- **Modelos de linguagem e tradução**: mostram uma forte tendência para associar profissões a estereótipos de gênero. Termos como "engenheiro", "programador" e "dirigente" são consistentemente associados ao pronome masculino ("ele"), enquanto "enfermeira", "babá" e "recepcionista" são associados ao feminino ("ela"). (Taso; Reis; Martinez, 2022)
- **Língua com gênero neutro**: ao traduzir o inglês (em certos contextos) para uma língua com gênero gramatical marcado como o português, modelos como o GPT-3.5 Turbo tendem a reproduzir estes estereótipos, traduzindo "the doctor" como "o médico" e "the nurse" como "a enfermeira" (Soares et al., 2023). Este não é um problema isolado do português; é um fenômeno global documentado em várias línguas, que reforça a discursivização da mulher em papéis de cuidado ou servilidade (Hashiguti; Fagundes, 2023).
- **A tecnologia de deepfake**: criação de vídeos ou imagens falsas hiper-realistas tornou-se uma ferramenta potente para a pornografia não consensual, o assédio e a humilhação, afetando desproporcionalmente mulheres e meninas (Lisboa; Passos; Ferreira, 2025).

Sem sombra de dúvidas, essa nova forma de violência de gênero é potente causadora de danos psicológicos, sociais e profissionais profundos e duradouros às vítimas. A legislação, tanto no Brasil como no exterior, tem tido dificuldade em acompanhar a velocidade desta ameaça. Embora leis ofereçam alguma proteção, são consideradas insuficientes para lidar com a complexidade de situações e crimes (Vieira, 2025).

1.1. Invisibilidade Interseccional: a exclusão agravada de grupos marginalizados

A discriminação algorítmica raramente opera por um único eixo. Raça, gênero, orientação sexual, deficiência e classe se cruzam e se agravam mutuamente, produzindo exclusões interseccionais.

É o caso da Comunidade LGBTQIA+ que, embora muitas vezes eleja a IA como ferramenta que promove a inclusão, conectando pessoas de seu grupo a profissionais que se dizem apoiadores, a comunidade muitas vezes enfrenta invisibilidade e representação inadequada nos dados que alimentam os modelos de IA, com potencial de reproduzir desigualdades e exacerbar a violência *LGBTfóbica* (Votelgbt, 2025).

Outro grupo bastante afetado é o das pessoas com deficiência. Este grupo é um dos mais negligenciados nos estudos sobre viés algorítmico segundo Gomes; Silva; Neri (2023). Quando são abordados, frequentemente dá para notar a sua invisibilidade e representação distorcida nos conjuntos de dados. Às vezes, os sistemas algorítmicos podem não conseguir reconhecer ou classificar corretamente as pessoas com deficiência, que são descartadas como "casos atípicos" (outliers), perpetuando a sua exclusão e marginalização (Moura, 2023)

Isto tem implicações graves em áreas como a contratação e os sistemas de quotas (Requião; Costa, 2022) que podem ser minados por ferramentas de triagem automatizadas e enviesadas. Podemos também destacar o caso do Estatuto Socioeconômico, que enfoca os povos Latino Americano, em relação aos quais os sistemas de IA tendem a refletir a realidade de um grupo demográfico restrito: "homens brancos com um determinado estatuto socioeconómico e nível de educação" diz Arratibel (2025).

Além disso, um exame mais atento das formas como os sistemas algorítmicos interagem com os grupos marginalizados revela um paradoxo fundamental: estes grupos são simultaneamente tornados invisíveis nas fases de concepção e de dados, e hipervisíveis para os sistemas de vigilância e controle punitivo. Esta não é uma contradição, mas sim duas faces da mesma moeda de poder algorítmico.

As investigações demonstram de forma consistente que as mulheres, as pessoas negras, as pessoas com deficiência e a comunidade LGBTQIA+ estão sub-representadas ou representadas de forma distorcida nos conjuntos de dados utilizados para treinar a IA. As suas perspetivas e experiências de vida estão ausentes das equipas de desenvolvimento. E a invisibilidade algorítmica acontece através da falha do sistema que não consegue "ver" estes grupos como formados

por sujeitos cujas necessidades devem ser consideradas. Eles tornam-se hiper visíveis pelo lado negativo: reconhecimento facial e de policiamento, que os marcam como suspeitos.

Este paradoxo revela uma função central do poder algorítmico: o olhar do sistema que não é neutro, mas seletivo. Ele falha em não ver as pessoas marginalizadas como sujeitos, mas sim como objetos a serem geridos, controlados ou penalizados. Esta dinâmica — invisibilidade como sujeitos, hipervisibilidade como objetos — é uma manifestação digital de estruturas de poder opressivas e coloniais históricas.

2 Lentes críticas para uma análise mais profunda

Para uma compreensão verdadeiramente profunda, é necessário transcender a análise puramente técnica ou legal e interpretar o significado sociopolítico e filosófico da condição algorítmica. Esta seção aplica quadros teóricos críticos para desvendar as estruturas de poder, linguagem e conhecimento que sustentam a injustiça digital. Ao examinar o algoritmo como discurso, adotaremos uma perspectiva decolonial, centrada na experiência vivida, movendo-nos de uma descrição do problema para um diagnóstico das suas causas fundamentais.

2.1. O algoritmo como discurso: uma perspectiva linguística crítica

A Linguística Aplicada Crítica e a Análise Crítica do Discurso (ACD) contestam a ideia de código como ente puramente técnico e neutro, propondo que algoritmos operam de modo análogo à linguagem: produzem significados, constroem realidades e mediam relações de poder.

Na pesquisa brasileira em linguística aplicada crítica, o algoritmo é concebido como “materialidade discursiva”: programar não é apenas executar uma tarefa técnica, mas um ato de enunciação. Linhas de código, arquiteturas de sistema e escolhas de dados incorporam e materializam valores e pressupostos ideológicos — pessoais (dos programadores) e institucionais (das organizações). Assim, o algoritmo pode ser pensado como estrutura discursiva: combina linguagens computacionais com regras próprias e ganha sentido na relação com as lógicas sociais em que circula.

A ACD fornece ferramentas para investigar essa materialidade: examina sistemas algorítmicos como práticas de linguagem que constroem identidades

sociais, negociam poder e reproduzem ideologias. Tal enquadramento é útil para desmascarar a neutralidade técnica: permite identificar vieses ocultos, mapear dinâmicas de poder e reconhecer pressupostos ideológicos embutidos em tecnologias e em suas práticas discursivas, contribuindo para o desenho de sistemas mais equitativos sem perder de vista os condicionantes históricos e políticos que os atravessam.

2.2 Descolonizar a IA: desafiando a geopolítica da tecnologia

Uma crítica fundamental à IA contemporânea emerge de uma perspetiva decolonial e do Sul Global, que situa a tecnologia na geopolítica do poder e do conhecimento. Essa lente mostra que a IA não é universal, mas produto de geografias, histórias e economias específicas, sobretudo as do Norte Global.

A colonialidade da IA decorre da hegemonia dos Estados Unidos e de outras nações ocidentais no seu desenvolvimento. Como consequência, os sistemas são treinados e otimizados para um grupo demográfico restrito (homens brancos com determinado estatuto socioeconómico), o que gera desempenho inferior e vieses quando aplicados a outras populações e reproduz uma ordem global de desigualdade.

Pela perspetiva histórica, a teoria decolonial lê essa economia como colonialismo digital: extrativismo de dados, apropriação de conhecimento e erosão do Sul Global. Em resposta, cresce o apelo por soberania tecnológica e de dados.

Esse movimento vai além da crítica: propõe resistir à adoção acrítica de modelos ocidentais e desenvolver estratégias de IA locais, ajustadas às realidades e valores de países como Brasil e Índia. Implica ainda confrontar a violência epistémica da ética dominante e promover éticas pluriversais, orientadas por direitos humanos e pelos valores constitucionais de cada Estado-nação.

3 A anatomia da injustiça algorítmica

Para compreender o impacto social da Inteligência Artificial, é essencial dissecar os mecanismos técnicos que transformam código em instrumentos de exclusão. A injustiça algorítmica não é um fenômeno espontâneo, mas sim o produto de uma cadeia de decisões e contextos.

A principal fonte de viés reside nos dados de treino. Longe de serem objetivos, estes conjuntos de dados são artefatos históricos que refletem uma realidade social marcada por desigualdades. Quando um sistema de IA aprende

com estes dados, ele inevitavelmente absorve e reproduz os preconceitos neles contidos. Isto é agravado pela sub-representação de grupos marginalizados nos dados e pela falta de diversidade nas equipes de desenvolvimento, que exclui perspetivas cruciais do processo de design.

Além disso, como veremos, os algoritmos criam e reforçam ciclos de retroalimentação (feedback loops), onde as suas próprias previsões se tornam a justificação para ações futuras, perpetuando e amplificando os vieses existentes. Este ciclo transforma os sistemas em "armas de destruição matemática", como descreve Cathy O'Neil, que legitimam a exclusão sob um manto de eficiência técnica.

Para comprovar estes problemas, a ciência da computação contém técnicas como as seguintes: **1. Pré-processamento (Pre-processing)**: Estas técnicas atuam sobre os dados de treinamento antes de o modelo ser construído. O objetivo é modificar o conjunto de dados para remover ou reduzir os vieses existentes. Os métodos incluem a re-amostragem (oversampling de grupos minoritários ou undersampling de grupos maioritários), a re-ponderação (atribuir pesos diferentes a diferentes instâncias de dados para equilibrar a sua influência) ou a aumentação de dados (gerar dados sintéticos para grupos sub-representados);

2. No-processamento (In-processing): Estes métodos modificam o próprio algoritmo de aprendizagem durante a fase de treinamento. Isto é geralmente feito adicionando restrições de justiça à função objetivo do modelo, forçando-o a otimizar não apenas a precisão, mas também a equidade em relação a uma métrica específica. A regularização, que penaliza o modelo por aprender correlações enviesadas, é uma abordagem comum nesta categoria;

3. Pós-processamento (Post-processing): Estas técnicas intervêm após o modelo ter feito as suas previsões. Elas ajustam os resultados do modelo para corrigir resultados enviesados. Por exemplo, podem ser aplicados limites de decisão diferentes para diferentes grupos demográficos para garantir que as taxas de erro sejam equilibradas.

Desta forma, para investigar a existência de vieses no banco de dados escolhido, o qual possui o objetivo de detectar discurso de ódio em tweets. Para simplificar, foi considerado que um tweet contém discurso de ódio se apresentar um sentimento ou palavra explícita relacionada ao racismo ou sexismo. Portanto, a tarefa de construção desse banco de dados consistiu em classificar formalmente, dado um conjunto de tweets e rótulos de treinamento, onde o rótulo '1' indica que o tweet é racista/sexista e o rótulo '0' indica, que o tweet não é racista/sexista.

Seu objetivo é prever os rótulos no conjunto de dados de teste.

3.1 Coleta e Carregamento de Dados

A nossa análise foi baseada no conjunto de dados "twitter-sentiment-analysis-hatred-speech"[1], (Toosi, 2018), disponível no Kaggle. A obtenção dos dados foi realizada utilizando a biblioteca kagglehub, que facilitou o download direto para o ambiente de trabalho. Após o download, o arquivo train.csv foi carregado em um DataFrame da biblioteca pandas para permitir a manipulação e análise dos dados de forma eficiente.

3.1.1 Análise Exploratória de Dados (EDA)

Uma etapa crucial inicial foi a Análise Exploratória de Dados (EDA) para compreender a estrutura e as características do conjunto de dados. Realizamos a verificação de valores ausentes em todas as colunas, confirmando a completude dos dados. Em seguida, analisamos a distribuição da variável alvo ('label'), que representa o sentimento do tweet (0 para não-ódio e 1 para ódio). Foi identificado um desbalanceamento significativo de classes, com a classe '0' sendo majoritária em relação à classe '1'.

3.1.2 Pré-processamento dos Dados

O corpus utilizado — *twitter-sentiment-analysis-hatred-speech* (Toosi, 2018) — foi empregado para treinar e avaliar modelos de detecção de discurso de ódio (label 1) versus não-ódio (label 0).

Conforme descrito na Seção 3.1.1, a variável-alvo é binária (“0 para não-ódio e 1 para ódio”), e, no entanto, há um problema anterior à modelagem: o critério de anotação. Quando a regra prática que orienta a rotulagem é essencialmente lexical — “se houver sentimento/palavra que remeta a racismo ou sexismo -> atribuir 1; se não houver -> 0” — a base de dados herda e cristaliza um viés humano. Em termos metodológicos, isso configura viés de anotação (measurement/labeling bias), distinto de vieses de amostragem ou de representação. As principais consequências são:

1 https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech/data

1. **Redução do fenômeno a palavras-chave**:

- O que define “ódio” passa a ser a presença de tokens específicos, não o alvo, a intenção, o contexto ou a pragmática do enunciado. - **Falsos positivos**: denúncias, citações acadêmicas ou reportagens que mencionam termos racistas/sexistas podem ser marcadas como “ódio” por mera ocorrência lexical. - **Falsos negativos**: mensagens veladas, eufemísticas, irônicas ou com códigos (dog whistles) — muito comuns em práticas discriminatórias — tendem a ser rotuladas como “não-ódio” por ausência do léxico explícito.

2. Aprendizagem do viés em lugar do fenômeno

O modelo aprende a detectar palavras (o “atalho” dado pela anotação), não a reconhecer discurso de ódio enquanto prática social. Assim, métricas de desempenho podem parecer boas em validações internas, mas refletem a política de anotação, não a realidade do fenômeno.

Em síntese, o próprio banco de dados já é enviesado quando a anotação segue um heurístico lexical. Tal desenho naturaliza uma “ideologia do léxico” (o ódio como lista de palavras), apagando dimensões cruciais — alvo, intenção, contexto, pragmática, ironia — e contaminando toda a cadeia: treinamento, avaliação e até estratégias de mitigação.

Sem essas correções de base, modelos treinados “aprenderão” a reproduzir o viés humano embutido na anotação — e as melhorias numéricas refletirão adesão ao critério enviesado, não avanço real rumo à justiça algorítmica

Dada a identificação do desbalanceamento de classes, aplicamos técnicas de pré-processamento para tentar mitigar este viés no conjunto de treinamento. O processo envolveu duas etapas principais: **a) Vetorização TF-IDF**: Primeiramente, os dados de texto da coluna 'tweet' foram convertidos em representações numéricas através do TfidfVectorizer. Este é um passo essencial, pois os algoritmos de aprendizado de máquina não conseguem processar texto em seu formato original. O TF-IDF atribui um peso a cada palavra com base na sua frequência no tweet e em todo o conjunto de dados, transformando o texto em vetores numéricos. **b) Superamostragem com SMOTE**: Em seguida, para corrigir o desequilíbrio, aplicamos a técnica SMOTE (Synthetic Minority Over-sampling Technique) apenas ao conjunto de treinamento. O SMOTE não apenas duplica as

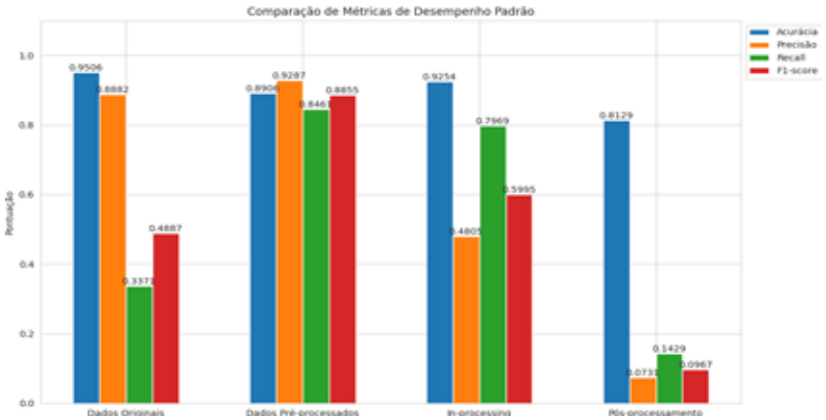
instâncias da classe minoritária ('Ódio'), mas cria novas instâncias sintéticas com base nas características das existentes. Isso resulta em um conjunto de treinamento balanceado, com uma representação igualitária das classes, o que permite que o modelo aprenda a identificar os padrões do discurso de ódio de forma muito mais eficaz, sem simplesmente "viciar" na classe majoritária.

3.2 Treinamento e avaliação dos modelos

Dividimos os datasets (original e, para avaliação do modelo pré-processado, o re-amostrado) em conjuntos de treinamento e teste. Treinamos os modelos (Baseline, Pré-processamento, In-processing) nos respectivos conjuntos de treinamento e realizamos previsões.

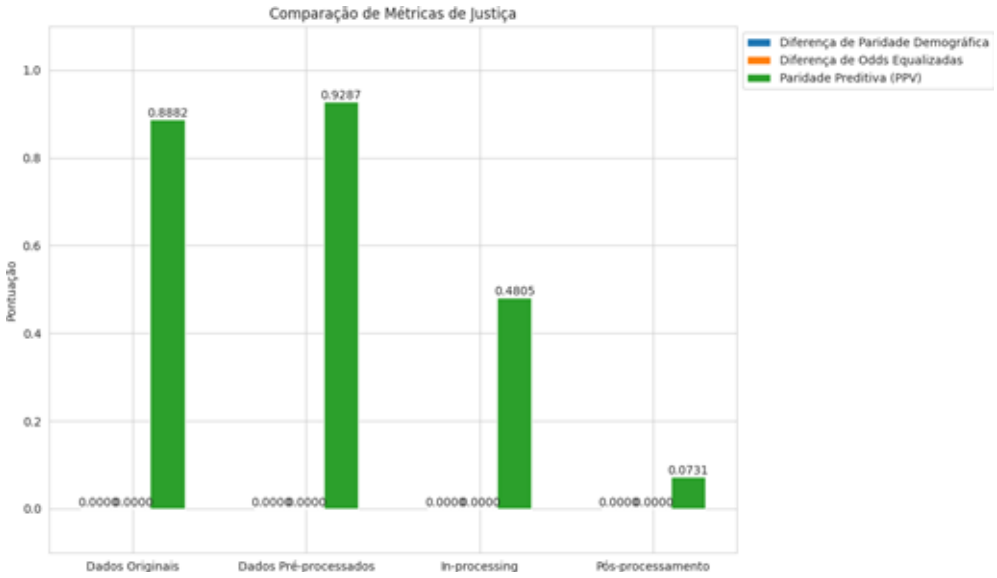
Para facilitar a compreensão e comparação dos resultados, geramos visualizações: Gráfico de Comparação de Métricas de Desempenho Padrão: Um gráfico de barras comparando Acurácia, Precisão, Recall e F1-score para as abordagens "Dados Originais", "Dados Pré-processados", "In-processing" e "Pós-processamento".

Gráfico 1 - Comparação de métricas de desempenho padrão



Fonte: elaborado pelo(s) autor(es).

Gráfico 2 - Comparação de métricas de justiça



Fonte: elaborado pelo(s) autor(es).

A análise dos gráficos mostra que a técnica de pré-processamento SMOTE foi a mais eficaz, melhorando significativamente a capacidade do modelo de detectar discurso de ódio. O Recall e o F1-score foram muito mais altos no modelo treinado com dados balanceados, o que prova que esta abordagem foi bem-sucedida em corrigir o desequilíbrio do dataset original. Porém, técnicas como SMOTE (pré-processamento) aumentam a exposição do modelo justamente aos padrões lexicais que definiram a label 1, reforçando e propagando o viés de anotação. Na prática, melhora-se o *recall* para “o que foi anotado como ódio”, não necessariamente para ódio enquanto categoria social complexa.

3.2.2 Implicações e Insights

A principal implicação do estudo é a confirmação de que o pré-processamento é uma etapa fundamental para lidar com o desbalanceamento de dados em tarefas de classificação de texto. O conjunto de dados original era altamente desbalanceado, com muito mais exemplos de "não-ódio" do que de "ódio". Modelos treinados nesses dados tendem a simplesmente prever a classe majoritária, alcançando uma alta acurácia geral, mas falhando em detectar a classe minoritária, que é a mais importante no nosso caso.

A técnica de reamostragem com SMOTE resolveu este problema de forma eficaz ao criar exemplos sintéticos da classe de "ódio", resultando num conjunto de treino balanceado. Isso permitiu que o modelo aprendesse os padrões especí-

ficos do discurso de ódio de forma muito mais eficiente, destacando-se como a abordagem mais bem-sucedida da nossa análise, mas, em síntese: **equilibramos quantitativamente as classes e reduzimos o viés medido**, mas partimos de um **critério de anotação enviesado** (rotulagem fortemente **lexical**: "se houver termo/traço ligado a racismo ou sexismo > label 1; caso contrário > 0"). Assim, o balanceamento **amplifica** a política de anotação (e não necessariamente o fenômeno social do ódio), produzindo **melhores métricas internas** e, ao mesmo tempo, um **equilíbrio aparente** quanto à justiça.

3.2.3 Implicações e Insights

- Diretrizes de anotação contextuais: diferenciar *ódio*, *contra-discurso/citação*, *ofensa genérica*, *ironia/sarcasmo*, com janelas de contexto (antes/depois do trecho).
- Anotação multi-rótulo e por estágio: (i) detectar alvo/grupo; (ii) identificar postura (ataque/defesa/descrição); (iii) graduar severidade/intencionalidade.
- Múltiplos anotadores + adjudicação: reduzir vieses individuais e calcular acordo inter-anotador.
- Amostragem estratificada: garantir representatividade de diferentes grupos, registros linguísticos e formas implícitas de agressão.
- Atributos sensíveis com governança ética: incluir variáveis sociodemográficas (ou *proxies* validados) sob protocolos de privacidade, permitindo auditoria de justiça.

3.2.4 Consequências nos algoritmos (planos técnico e metodológico)

1. **Aprendizagem do atalho lexical**

O modelo tende a **aprender palavras** (o heurístico da anotação), não **pragmática**, **alvo** ou **intenção**. A taxa de acerto cresce no que a base definiu como "ódio", mas pode **falhar** em casos **velados**, **irônicos** ou **contextuais**.

2. **Melhora métrica ≠ melhora ética**

Ganhos em **Recall/F1** (Gráfico 1, p. 14) sinalizam **redução de viés estatístico** entre classes, mas **não evidenciam justiça de grupo**.

3. **Ciclos de retroalimentação**

Se saídas do classificador "limpo" (balanceado) **retroalimentam** moderação, policiamento ou curadoria, o sistema **congela** o viés da **anotação le-**

xical, ampliando o **viés de medição** ao longo do pipeline.

Considerações Finais

A técnica de reamostragem SMOTE criou exemplos sintéticos da classe "ódio", balanceando o conjunto de treino e permitindo ao modelo aprender padrões com maior eficiência. Contudo, como o critério de anotação era lexical e enviesado, o processo gerou apenas uma redução aparente do viés: o modelo passou a refletir mais fielmente as regras de rotulagem, e não o fenômeno social do discurso de ódio, resultando em melhores métricas internas, porém em um equilíbrio apenas superficial em termos de justiça. Sem corrigir a **base epistemológica do dado** e a **forma de medir justiça**, corremos o risco de **sofisticar a aparência de equidade** sem transformá-la em **prática social efetiva**.

Nossa pesquisa buscou mostrar como as estruturas de poder decoloniais globais são postas em prática através do discurso específico dos algoritmos analisados. A abordagem integrada entre técnica e discurso pretendeu que a análise não se tornasse demasiado abstrata, restrita ou meramente anedótica.

A principal conclusão que chegamos ao final deste artigo é que a IA está longe de ser a força neutra e objetiva que o imaginário popular proclama. Pelo contrário, os sistemas algorítmicos funcionam como espelhos e amplificadores das estruturas de poder, dos preconceitos e das desigualdades que permeiam as nossas sociedades. O viés, conforme vimos na análise do banco de dados tornou-se uma característica intrínseca desta tecnologia concebida e implementada dentro de contextos históricos e sociais específicos.

Assim, alcançar uma ética algorítmica tornou-se um desafio complexo e multifacetado que transcende em muito as soluções puramente técnicas e permeia discursos de poder. Mas, embora desejemos ir além de uma crítica baseada em valores eurocêntricos, a possibilidade de chegarmos a uma abordagem verdadeiramente decolonial que confronte as injustiças passadas e imagine futuros reparadores está longe de acontecer, ou o que é pior, é impossível em toda sua totalidade.

No entanto, para não ficarmos só nas críticas, propomos um conjunto de práticas e táticas para tentar propor transformação: (1) a criação de uma técnica de treinamento que seja autorreflexiva sobre o seu próprio posicionamento de poder; (2) a busca de uma "tutela inversa", aprendendo ativamente com as comunidades marginalizadas que são mais afetadas pela tecnologia; e (3) a renovação de comunidades políticas capazes de contestar, resistir e remodelar a tecnologia

de acordo com as suas próprias necessidades e aspirações.

Isto implicará certamente em um desafio para a geopolítica das tecnologias, que deverá se centrar em experiências vividas nas comunidades marginalizadas e em valores constitucionais como dignidade, equidade, solidariedade para formar futuros melhores.

Entendemos nesta primeira fase da pesquisa que a luta por uma IA mais justa é, ao mesmo tempo, uma luta política, social e epistêmica. Requer uma governação robusta, com quadros regulatórios que estabeleçam linhas claras de responsabilidade, exijam transparência e garantam o direito à reparação, conforme valores previstos na nossa Carta Magna, que deverá fazer parte do repertório principal do "treino das máquinas".

Na próxima etapa da pesquisa, pretendemos testar práticas como: **1. Declarações de Impacto de Viés**: tentar avaliar, na condição de desenvolvedores de um sistema protótipo, os potenciais impactos discriminatórios presentes nele, antes de sua efetiva implementação; 2.

Auditorias Independentes: buscar terceiros que avaliem o sistema algorítmico, para verificar critérios de justiça e segurança; **3. IA Explicável (XAI)**: Sair da "caixa-preta" e criar sistemas cujas decisões possam ser compreendidas e contestadas conforme valores presentes nos direitos humanos. **4. Rever a política de anotação**, incorporando **contexto**, **alvo** e **intenção** (multi-rótulo / por estágios), com **múltiplos anotadores** e **adjudicação** (acordo inter-anotador). **5) Adicionar atributos sensíveis** (ou proxies bem-governados) para **auditar justiça de grupo** com **DPD/EOD/PPV** de forma ética e legal. **6) Avaliar em teste não reamostrado** e com **amostras fora do domínio** (ironia, dog whistles), reportando **calibração e robustez**.

Por último, gostaríamos de destacar a importância que o ativismo decolonial dos movimentos sociais desempenham na sensibilização da sociedade para implementação de ferramentas de detecção de falhas e responsabilização de sistemas. Nesse contexto, a linguística aplicada pode contribuir para uma literacia algorítmica generalizada, como condição necessária para uma governação democrática da IA.

Certamente que uma população informada estará mais bem equipada para avaliar criticamente as decisões automatizadas que afetam as suas vidas. Assim, a ética algorítmica não será alcançada apenas através de intervenções de cima para baixo, mas também por estratégias de baixo para cima, que capacitam os cidadãos a contestar o poder algorítmico e construir alternativas mais justas.

Referências

ARRATIBEL, A. J. **Paola Ricaurte talks about the future of inclusive AI**. TecScience, june 2025. Disponível em: https://tecscience.tec.mx/en/human-social/paola-ricaurte/. Acesso em: 15 nov. 2025.

BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Presidente da República, [2016].

BRASIL. Lei nº 13.146, de 6 de julho de 2015. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). **Diário Oficial da União**, Brasília, DF, 7 jul. 2015. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2015/Lei/L13146.htm . Acesso em: 31 jul. 2025.

GOMES, V. H. dos S.; SILVA, R. C. A.; NERI, T. C. da S. Ética em sistemas de IA: um olhar sobre a injustiça algorítmica e a deficiência. **Revista IDeAS**, [S. l.], v. 12 n. 2, p. 238- 260, 2023. DOI: https://doi.org/10.51359/2317-0115.2023.260751. Disponível em: https://periodicos.ufpe.br/revistas/index.php/RMP/article/view/260751. Acesso em: 20 jul. 2025.

GORENC, N. AI embedded bias on social platforms. **International Review of Sociology**, v. 35, n. 1, p. 1–20, 2025. Disponível em: https://doi.org/10.1080/03906701.2025.2489056. Acesso em: 31 jul. 2025.

HASHIGUTI, S. T.; FAGUNDES, I. Z. Z. O algoritmo como materialidade discursiva em um contexto de educação linguística. **Letras & Letras**, v. 38, p. 1-21, 2023. DOI: https://doi.org/10.14393/LL63-v38-2022-27. Disponível em: https://seer.ufu.br/index.php/letraseletras/article/view/68123. Acesso em: 20 jul. 2025.

LISBOA, C. A.; PASSOS, E. G.; FERREIRA, J. da S. Desvendando vieses em IA: um estudo sobre reconhecimento facial e futuros feministas. **Temáticas**, Campinas, v. 33, n. 65, p. 1–20, 2025. DOI: https://doi.org/10.20396/tematicas.v33i65.19932. Disponível em: https://econtents.bc.unicamp.br/inpec/index.php/tematicas/article/download/19932/14811/58245. Acesso em: 20 jul. 2025.

MOURA, I. Encoding normative ethics: on algorithmic bias and disability. **First Monday**, v. 28, n. 1, p. 1-23, 2023. Disponível em: https://firstmonday.org/ojs/index.php/fm/article/view/12905. Acesso em: 20 jul. 2025.

OLIVEIRA, Cyntia Barbosa. Racismo algorítmico e inteligência artificial: a discriminação nos sistemas de videomonitoramento. **Em Tese**, Florianópolis, v. 22, n. 1, p. 1-20, 2025. DOI: https://doi.org/10.5007/1806-5023.2025.e105023. Disponível em: https://periodicos.ufsc.br/index.php/emtese/article/view/105023. Acesso em: 20 jul. 2025.

REQUIÃO, M.; COSTA, D. Discriminação algorítmica: ações afirmativas como estratégia de combate. **Civilistica.com**, v. 11, n. 3, p. 1–24, 25 dez. 2022. Disponível em: https://civilistica.emnuvens.com.br/redc/article/download/804/650/1832. Acesso em: 20 jul. 2025.

ROSENTHAL-VON DER PÜTTEN, A. M.; SACH, A. Michael is better than Mehmet: exploring the perils of algorithmic biases and selective adherence to advice from automated decision support systems in hiring. **Frontiers in Psychology**, v. 15, p. 1-19, 2024. Disponível em: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1416504/full. Acesso em: 31 jul. 2025.

SILVA, F. dos S. R. **Nada mais sobre nós sem nós:** escurecendo o debate sobre a regulação de IA no Brasil e pensando mecanismos de combate ao racismo algorítmico. Relatório de Pesquisa – Programa Líderes LACNIC 2.0. Brasil: LACNIC, 2023. Disponível em: https://www.lacnic.net/innovaportal/file/6974/1/regulacao-de-ia-e-discriminacao-algoritmica-informe-de-investigacion-pt.pdf. Acesso em: 20 jul. 2025.

SILVA, F. S. R.; SILVA, T. (orgs.). **Artificial intelligence and racial discrimination in Brazil**: key issues and recommendations. Belo Horizonte: Institute for Research on Internet and Society, 2024. Disponível em: https://bit.ly/4dGXxVi. Acesso em: 20 jul. 2025.

SILVA, P. B. M. e. **Racismo algoritmo**: a nova face da injustiça penal. São Paulo: Migalhas, 2025. Disponível em: https://www.migalhas.com.br/depeso/429528/racismo-algoritmo-a-nova-face-da-injustica-penal. Acesso em: 20 jul. 2025.

SILVA, T. **O racismo algorítmico é uma espécie de atualização do racismo estrutural**. Rio de Janeiro: Centro de Estudos Estratégicos da Fiocruz, 2021. Disponível em: https://cee.fiocruz.br/?q=Tarcizio-Silva-O-racismo-algoritmico-e-uma-especie-de-atualizacao-do-racismo-estrutural. Acesso em: 20 jul. 2025.

SOARES, T. A.; GUMIEL, Y. B.; JUNQUEIRA, R.; GOMES, T.; PAGANO, A. Viés de gênero na tradução automática do GPT-3.5 turbo: avaliando o par linguístico inglês-português. In: Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), 14., 2023, Belo Horizonte. **Anais [...]**. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 167–176. DOI: https://doi.org/10.5753/stil.2023.234186. Acesso em: 20 jul. 2025.

TASO, F. T. de S.; REIS, V. Q.; MARTINEZ, F. H. V. Discriminação algorítmica de gênero: estudo de caso e análise no contexto brasileiro. *In*: **Women In Information And Computing Sciences** (WICS), 2022, [S. l.]. Sociedade Brasileira de Computação, 2022. Disponível em: https://sol.sbc.org.br/index.php/wics/article/download/24825/24646/. Acesso em: 20 jul. 2025.

TAVARONE, P. G. **A utilização de tecnologias nas investigações criminais e o racismo algorítmico: implicações para os direitos fundamentais**. 2025. 60 f. Trabalho de Conclusão de Curso (Bacharelado em Direito) – Universidade Anhembi Morumbi, São Paulo, 2025. Disponível em: https://repositorio.animaeducacao.com.br/bitstreams/02170548-56ab-4f6e-8772-d00653555c73/download. Acesso em: 20 jul. 2025.

TOOSI, A. **twitter-sentiment-analysis-hatred-speech**: repositório de código para análise de sentimentos e discurso de ódio no Twitter*. 2018. Disponível em: https://www.kaggle.com/datasets/tooso/twitter-sentiment-analysis-hatred-speech. Acesso em: 31 jul. 2025.

VIANA, G. M. de L.; MACEDO, C. S. de. Inteligência artificial e a discriminação algorítmica: uma análise do caso Amazon. **Direito & TI**, [S. l.], v. 1, n. 19, p. 39–62, 2024. DOI: https://doi.org/10.63451/ti.v1i19.212. Disponível em: https://direitoeti.com.br/direitoeti/article/view/212. Acesso em: 31 jul. 2025.

VIEIRA, J. R. B. F. **Violência digital**: deepfakes o novo rosto da opressão contra mulheres. São Paulo: Migalhas, 2025. Disponível em: https://www.migalhas.com.br/depeso/430699/violencia-digital-deepfakes-o-novo-rosto-da-opressao-contra-mulheres. Acesso em: 20 jul. 2025.

VOTELGBT. IA+LGBT – Nossa inteligência contra a violência política. [S. l.]: VoteLGBT, [202–]. Disponível em: https://www.votelgbt.org/inteligencia. Acesso em: 20 jul. 2025.

WENDEHORST, C. **Bias in algorithms**: artificial intelligence and discrimination. 1. ed. Luxembourg: Publications Office of the European Union, 2022. 106 p. Disponível em: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf. Acesso em: 31 jul. 2025.

Aplicação Open Science

Banco de dados analisado:
https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech/data

Google Colab - Código Python: https://colab.research.google.com/drive/1rwrWULO5IFPpR5mNJ-CecbmbugFDtgDPZ#scroll To=fc5fa021

Sobre as autoras e o autor

Elaine Maria Gomes de Abrantes - Doutora em Letras/Linguística pela UERN - Universidade do Estado do Rio Grande do Norte (2020), Campus Pau dos Ferros/RN; E-mail: elamar.esmapb@gmail.com. Lattes: https://lattes.cnpq.br/3916889319422221. Orcid: https://orcid.org/0000-0003-2470-8882.

Leandra de Oliveira Silva - Mestranda em Inteligência Artificial pela UFCG - Universidade Federal de Campina Grande (Atual) - Campus I, Campina Grande/PB; Email: leandra.silva@copin.ufccg.edu.br. Lattes: https://lattes.cnpq.br/8931762511689408. Orcid: https://orcid.org/0009-0007-1391-8557.

Erick Araken Vieira Gomes - Mestrando em Inteligência Artificial pela UFCG - Universidade Federal de Campina Grande (Atual) - Campus I, Campina Grande/PB; Email: erick.gomes@ccc.ufccg.edu.br. Lattes: https://lattes.cnpq.br/5162364961960365. Orcid: https://orcid.org/0000-0002-4642-5480.mm