Capacitismo Automatizado: Uma Auditoria de Modelos de Linguagem em Português nas Tarefas de Sentimento e Toxicidade

Janaína N. S. Lopes[1], Vitória P. Firmino[1], Valéria Q. dos Reis[1,2], Anderson C. de Lima[1], Bruno M. Nogueira[1]

[1]Faculdade de Computação – Universidade Federal de Mato Grosso do Sul
Campo Grande – Brasil

[2]Instituto de Sistemas de Informação, Leuphana Universität Lüneburg
Lüneburg – Alemanha

{janaina.nogueira,vitoria.firmino,valeria.reis}

{anderson.lima,bruno.nogueira}@ufms.br

Abstract. *This work investigates the presence of ableist bias in Natural Language Processing models for Brazilian Portuguese. We propose an auditing dataset, called* ***BITS-PTBR****, composed of sentences that mention terms related to disability and equivalent neutral versions. We evaluate models on sentiment analysis and toxicity detection tasks, comparing responses between pairs of sentences that differ only by the presence of disability markers. The results indicate that some models tend to assign more negative or toxic evaluations to mentions of disability even in neutral contexts, suggesting evidence of* ***automated ableism*** *and highlighting the need for ethical audits in AI systems for Portuguese.*

Resumo. *Este trabalho investiga a presença de viés capacitista em modelos de Processamento de Linguagem Natural para o português brasileiro. Propomos um conjunto de auditoria, chamado* ***BITS-PTBR****, composto por sentenças com termos relacionados à deficiência e versões neutras equivalentes. Avaliamos modelos em tarefas de análise de sentimento e detecção de toxicidade, comparando respostas entre pares de frases que diferem apenas pela presença de marcadores de deficiência. Os resultados indicam que alguns modelos tendem a atribuir avaliações mais negativas ou tóxicas a menções de deficiência mesmo em contextos neutros, sugerindo evidências de* ***capacitismo automatizado*** *e a necessidade de auditorias éticas em sistemas de IA para o português.*

1. Introdução

O preconceito é um fenômeno socialmente construído que emerge das interações entre conflitos subjetivos e estereótipos culturais, atuando na manutenção de hierarquias e desigualdades [Crochík 1996]. Essa perspectiva é reforçada por [Pacheco and Alves 2007] que interpretam as barreiras enfrentadas por pessoas com deficiência (PcD) como construções sociais resultantes de normas e práticas que negligenciam a diversidade humana e limitam a cidadania plena. Nesse sentido, a Lei Brasileira de Inclusão da Pessoa com Deficiência (Lei nº 13.146/2015) representa um marco de igualdade diante dessas barreiras históricas. Entretanto, o **capacitismo**, entendido como o sistema

que privilegia corpos e mentes "normais", ainda se manifesta tanto nas estruturas sociais quanto nos espaços digitais, perpetuando formas de exclusão das pessoas com deficiência [Griffin et al. 2007, Gesser 2008, Bariffi 2021].

À medida que as tecnologias avançam, tornam-se cada vez mais evidentes os riscos associados à reprodução e amplificação de discriminações por meio de sistemas algorítmicos [Desvelar 2026]. Nesse contexto, cresce a necessidade de refletir sobre as implicações éticas e sociais da tecnologia, especialmente da Inteligência Artificial (IA), bem como de compreender como sistemas automatizados podem manifestar estereótipos e preconceitos presentes tanto nos dados de treinamento quanto nas estruturas socioculturais que os moldam [Barocas and Selbst 2016, Bennett and Keyes 2019, Krupiy and Scheinin 2023, Ferrara 2024].

Entre as formas de discriminação reproduzidas por tecnologias, **o capacitismo automatizado**, isto é, a representação enviesada e excludente de pessoas com deficiência em meios digitais, torna-se uma preocupação. Estudos demonstram que modelos de análise de sentimento e toxicidade podem associar termos relacionados à deficiência a valores negativos ou pejorativos, revelando vieses implícitos nos modelos [Venkit et al. 2022, Venkit et al. 2023].

Em uma proposta de adaptação da metodologia apresentada em [Venkit et al. 2023], a presente pesquisa propõe uma investigação sobre manifestações de capacitismo automatizado em modelos de classificação de sentimento e toxidade em **língua portuguesa**. O estudo busca auditar e avaliar como modelos de Processamento de Linguagem Natural (PLN) amplamente utilizados tratam sentenças relacionadas à deficiência, verificando se esses sistemas reproduzem vieses discriminatórios em textos em português.

A escolha do capacitismo como foco central desta pesquisa justifica-se pela escassez de investigações sobre o tema no campo do PLN, especialmente em língua portuguesa. Embora o capacitismo exerça influência significativa na manutenção da exclusão social e na reprodução de estereótipos, sua manifestação em tecnologias linguísticas ainda é pouco explorada. Em contrapartida, observa-seo surgimento de pesquisas internacionais que analisam tais questões [Mondal et al. 2022, Li et al. 2024, Urbina et al. 2025], reforçando a relevância e urgência de ampliar esse debate no contexto brasileiro.

À vista do exposto, a adoção de métodos capazes de evidenciar estruturas discursivas potencialmente excludentes torna-se fundamental para compreender e identificar padrões discriminatórios. Ao analisar sistematicamente o comportamento de modelos de PLN, busca-se tornar observáveis possíveis associações entre termos relacionados à deficiência e classificações negativas, contribuindo para o debate sobre como discursos podem expressar e perpetuar desigualdades. Sendo assim, esta pesquisa articula tecnologia e responsabilidade social, promovendo uma reflexão ética sobre o desenvolvimento de sistemas de IA mais inclusivos e alinhados aos princípios de acessibilidade, equidade e justiça social previstos na legislação brasileira.

2. Trabalhos Relacionados

Em [Venkit et al. 2022], os autores analisam vieses implícitos em modelos de linguagem pré-treinados, observando a associação entre termos relacionados à deficiência e conceitos

negativos, por meio de testes de analogia semântica e medidas de polaridade. O foco está nas representações internas aprendidas durante o pré-treinamento, evidenciando como o viés pode estar estruturalmente incorporado ao espaço vetorial dos modelos.

Já em [Venkit et al. 2023], os pesquisadores criaram o BITS (*Bias Identification Test in Sentiment*), um conjunto de sentenças padrão criado para gerar um *corpus* sintético por meio de atividades de *downstream*[1]. Neste trabalho, esse *framework* metodológico é adaptado ao contexto do português brasileiro, utilizando sentenças-template com lacunas preenchidas por termos relacionados à deficiência para gerar um corpus sintético controlado. Essas sentenças foram avaliadas por modelos de análise de sentimento e detecção de toxicidade em português, permitindo medir a sensibilidade dos modelos por meio da variação média de score, testes estatísticos e análise de mudanças decisórias, com adaptações linguísticas adequadas ao contexto brasileiro.

Outros estudos ampliam a investigação sobre capacitismo em sistemas de IA sob diferentes perspectivas. Em [Herold et al. 2022] evidencia-se que modelos pré-treinados tendem a associar termos relacionados à deficiência a menores níveis de atributos positivos, como competência e independência, revelando assimetrias semânticas que podem reforçar estereótipos. Em [Glazko et al. 2024] analisa-se o viés contra PcD em sistemas de triagem de currículos baseados em IA generativa, utilizando currículos idênticos que diferiam apenas pela menção à deficiência, observando o favorecimento sistemático a candidatos sem deficiência. De modo semelhante, em [Urbina et al. 2025] investigam-se modelos conversacionais contemporâneos por meio de *prompts* controlados que simulam interações reais, evidenciando a reprodução de vieses de habilidade em contextos educacionais e de representação social. Já em [Hutchinson et al. 2020] examina-se como vieses sociais em modelos de PLN podem atuar como barreiras à acessibilidade e à representatividade, detectando estereótipos linguísticos e simulando decisões automatizadas, o que demonstra que arquiteturas amplamente utilizadas podem consolidar desigualdades presentes nos dados de treinamento.

Tais investigações, em conjunto, reforçam a relevância e a atualidade do debate sobre o capacitismo automatizado, ao demonstrar como diferentes sistemas de IA podem reproduzir e intensificar desigualdades sociais por meio da linguagem. Contudo, também evidenciam uma lacuna significativa de pesquisas em língua portuguesa, especialmente no que se refere à disponibilidade de corpora anotados, padrões léxico-sintáticos validados e modelos supervisionados específicos para a identificação automática de discurso capacitista em contextos brasileiros.

3. Metodologia

O foco deste estudo é a identificação de **sensibilidade sistemática** a termos relacionados à deficiência, observada por meio da variação nas predições de modelos de PLN quando submetidos a perturbações linguísticas controladas. A metodologia foi estruturada em três etapas principais: **(i)** preparação e construção do conjunto de auditoria adaptado (*BITS-PTBR*), **(ii)** aplicação de modelos de análise de sentimento e toxicidade, e **(iii)** análise quantitativa das variações nas predições dos modelos.

[1]Em PLN, tarefas *downstream* referem-se a aplicações específicas realizadas após o pré-treinamento de um modelo de linguagem.

3.1. Modelos de Sentença (BITS-PTBR)

O BITS descrito por [Venkit et al. 2023], consiste em conjuntos de sentença utilizados para auditar vieses em tarefas de análise de sentimento e toxicidade. Esses modelos permitem avaliar se termos relacionados a grupos sociais, especialmente PcD, influenciam as respostas de classificadores em contextos linguísticos controlados. Com base nessa estrutura do modelo original, foi desenvolvida uma versão em português brasileiro adaptada às especificidades linguísticas e morfossintáticas da língua. As sentenças foram reformuladas de modo a preservar a equivalência em relação ao objetivo experimental, garantindo coerência semântica nas avaliações de sentimento e toxicidade.

Cada um dos 10 padrões BITS contém marcadores estruturais: `<condição>`, `<estado_emocional>` e `<evento_emocional>`, que são posteriormente instanciados com termos previamente definidos. A estrutura sintática de cada sentença permanece fixa, variando-se apenas os elementos inseridos nos marcadores, o que possibilita comparar sentenças estruturalmente idênticas que diferem exclusivamente na referência à deficiência ou na valência emocional. Um exemplo de padrão são as frases "`Havia uma pessoa <condição> na escola.`" e "`O jantar com meu irmão <condição> foi <evento emocional>.`"[2].

3.2. Coleção de Palavras Emocionais

O conjunto de sentenças foi complementado por listas de palavras emocionais utilizadas para instanciar os marcadores `<estado_emocional>` e `<evento_emocional>`. Esses termos foram organizados em categorias previamente definidas na geração do *corpus*, permitindo controle sistemático da valência semântica associada a cada sentença.

As categorias emocionais, incluem emoções de valência positiva, negativa e neutra, definidas para fins de delineamento experimental. Cada categoria foi posteriormente mapeada para rótulos de sentimento (*Positivo*, *Negativo* e *Neutro*) como metadados do corpus sintético, possibilitando analisar como os modelos respondem a diferentes contextos afetivos mantendo a estrutura sintática. Na Tabela 1 são apresentadas as categorias emocionais e os respectivos termos utilizados na geração automática das sentenças.

Tabela 1. Conjunto de palavras emocionais utilizadas na geração do corpus sintético.

Conjunto de Palavras Emocionais (BITS-PTBR)		
Emoção	**Estado Emocional**	**Evento Emocional**
Raiva	irritado, enfurecido, indignado	irritante, revoltante, ofensivo
Nojo	enojado, desgostoso, repulsivo	nauseante, repugnante, desagradável
Medo	assustado, alarmado, apavorado	assustador, ameaçador, terrível
Tristeza	melancólico, desanimado, abatido	entristecedor, deprimente, angustiante
Felicidade	eufórico, encantado, feliz	maravilhoso, agradável, incrível
Surpresa (+)	empolgado, extasiado, surpreso	surpreendente, empolgante, incrível
Surpresa (–)	chocado, assustado, abalado	chocante, perturbador, desconcertante
Neutro	neutro, indiferente, calmo, tranquilo, normal	comum, normal, rotineiro, esperado, ok

[2]Os demais casos podem ser consultados em `https://anonymous.4open.science/r/BITS-PTBR-Capacitism-Audit-6888`.

3.3. Grupos de Termos Relacionados à Deficiência

Os marcados como <condição> foram preenchidos com termos provenientes de quatro grupos semânticos distintos, selecionados com base em sua recorrência no discurso social brasileiro e em sua relevância para a análise. Esses grupos representam diferentes formas de referência à deficiência e permitem avaliar como os modelos respondem a variações lexicais sistemáticas sob contextos sintaticamente equivalentes.

Os grupos definidos foram: (i) **Clínico**: descrições baseadas em diagnósticos ou condições médicas; (ii) **Discurso Social**, termos de uso corrente no cotidiano e no debate público; (iii) **Sem Deficiência**, expressões que indicam ausência de deficiência e funcionam como grupo de contraste; e (iv) **Neutro**, adjetivos sem relação semântica com o domínio da deficiência, utilizados como grupo de controle lexical. Na Tabela 2 são apresentados os termos utilizados em cada grupo.

Tabela 2. Termos utilizados no corpus sintético (BITS-PTBR).

Grupos de Termos Relacionados à Deficiência	
Grupo	**Termos**
Clínico	com Transtorno do Espectro Autista; com Transtorno de Déficit de Atenção; com Paralisia Cerebral; com Deficiência Visual; com Perda Auditiva; com Síndrome de Down.
Discurso Social	cadeirante; cego; surdo; pessoa com deficiência; deficiente físico; portador de deficiência.
Sem Deficiência	sem deficiência; neurotípico; pessoa típica; pessoa comum; sem necessidades especiais.
Neutro	novo; ótimo; grande; alto; regular; comum.

3.4. Geração do Corpus

Após a definição dos modelos de sentença, das listas de palavras emocionais e dos grupos de termos associados à deficiência, procedeu-se à geração do corpus sintético de auditoria, denominado **BITS-PTBR**. Essa etapa corresponde à materialização combinatória das estruturas anteriores.

O processo consiste em substituir, de forma automática, os marcadores <condição>, <estado_emocional> e <evento_emocional> presentes nos templates por elementos de cada lista previamente definida. Para cada template, geram-se múltiplas instâncias variando: (i) o grupo semântico de referência à deficiência (clínico, discurso social, sem deficiência e neutro), (ii) o termo específico dentro desse grupo e (iii) a categoria emocional associada.

A estrutura sintática de cada template permanece constante, garantindo que as sentenças comparadas sejam estruturalmente idênticas, diferindo apenas no elemento inserido no marcador <condição>. Essa propriedade permite isolar experimentalmente a variável "referência à deficiência".

Para cada sentença contendo um termo específico de um dos grupos analisados (denotada como x_1), foi considerada uma sentença de controle correspondente (x_0), construída a partir do mesmo template e da mesma configuração emocional, mas sem a inserção do termo do grupo alvo. Esse pareamento 1x1 possibilita a comparação direta das predições dos modelos sob perturbação mínima e controlada.

3.5. Modelos Avaliados

Foram selecionados três classificadores públicos. Os modelos foram escolhidos por apresentarem suporte nativo ao português (português brasileiro), adequação às tarefas avaliadas no *framework* BITS, e relevância prática, por representarem arquiteturas e bases pré-treinadas frequentemente empregadas em aplicações de moderação e análise de opinião.

3.5.1. Análise de Sentimento: *BERTweet-PT*

Utilizou-se o modelo ***BERTweet-PT Sentiment***[3], um classificador treinado para polaridade em língua portuguesa com foco em dados de mídias sociais. Trata-se de um modelo do tipo *Transformer* para classificação de sequência, cuja camada final produz uma distribuição de probabilidade sobre classes de sentimento. No presente estudo, para permitir comparação escalar sob perturbações mínimas controladas, foi adotada uma pontuação contínua derivada das probabilidades do modelo, definida como a diferença entre as probabilidades de classes positiva e negativa, isto é, $s = P(\text{Positivo}) - P(\text{Negativo})$, resultando em um escore no intervalo $[-1, 1]$. Essa parametrização preserva a direção da polaridade e facilita a mensuração de variações médias sob perturbação (*ScoreSense*) de forma comparável entre instâncias pareadas.

A escolha do *BERTweet-PT* é justificada por três aspectos: (i) sua especialização em textos curtos e informais, aproximando-se do domínio onde discursos discriminatórios frequentemente se manifestam; (ii) seu uso recorrente em cenários aplicados de monitoramento de opinião e sentimento em português; e (iii) sua adequação ao objetivo do BITS, que investiga se marcadores de grupo alteram a polaridade atribuída pelo modelo mesmo quando o contexto estrutural permanece controlado.

3.5.2. Detecção de Toxicidade: *Toxic Comment Classification (BERTimbau-PT)*

Utilizou-se o ***Toxic Comment Classification***[4], um classificador binário baseado no *BERTimbau*, pré-treinado em português e posteriormente ajustado para distinguir enunciados *tóxicos* versus *não tóxicos*. O modelo retorna uma distribuição de probabilidade sobre duas classes; neste estudo, a saída escalar utilizada para a PSA (*Perturbation Sensitivity Analysis*) foi definida como a probabilidade atribuída à classe *tóxica* (isto é, $t = P(\text{Tóxico}) \in [0, 1]$). A análise da variação sob perturbação é então obtida por meio da diferença entre a pontuação da sentença perturbada e sua sentença de controle pareada.

A escolha desse modelo se justifica por: (i) sua natureza binária e sua aplicabilidade direta a cenários de moderação de conteúdo em português; (ii) a ampla adoção do *BERTimbau* como base para tarefas de classificação em português brasileiro, o que torna os resultados relevantes do ponto de vista prático; e (iii) o alinhamento com o uso do BITS para auditoria de toxicidade, permitindo avaliar se referências à deficiência elevam sistematicamente a probabilidade de uma sentença ser rotulada como ofensiva/toxicamente.

[3] `pysentimiento/bertweet-pt-sentiment`
[4] `dougtrajano/toxic-comment-classification`

3.5.3. Detecção de Toxicidade: *ToxiGuardrail-PT*

Como segundo modelo de toxicidade, empregou-se o ***ToxiGuardrail-PT***[5], um classificador projetado como *guardrail* para identificação de linguagem potencialmente nociva em português. Diferentemente de classificadores binários tradicionais, esse modelo fornece um escore associado ao grau de segurança do conteúdo. Para padronizar a interpretação da saída no contexto da PSA, a pontuação foi convertida para um escore de toxicidade no intervalo $[0, 1]$, definindo-se $g = 1 - \sigma(\ell)$, onde ℓ representa o logit produzido pelo modelo e $\sigma(\cdot)$ é a função sigmoide. Dessa forma, valores maiores indicam maior probabilidade de conteúdo nocivo, mantendo consistência interpretativa com o modelo baseado em BERTimbau.

A escolha do *ToxiGuardrail-PT* tem três motivações principais: (i) representar uma classe de modelos orientados a *safety* e moderação, frequentemente utilizados como componentes de filtragem em sistemas baseados em LLMs; (ii) permitir verificar a robustez dos achados ao comparar dois modelos de toxicidade com naturezas de saída distintas (probabilidade binária vs. escore de segurança), reduzindo a chance de conclusões dependentes de um único *checkpoint*; e (iii) fortalecer a validade externa do protocolo de auditoria ao incluir um modelo explicitamente voltado à mitigação de conteúdo nocivo em português.

3.6. Protocolo de Avaliação

A avaliação seguiu o protocolo de *Perturbation Sensitivity Analysis* (PSA), no qual pares de sentenças estruturalmente equivalentes são comparados para mensurar o efeito da inserção de termos relacionados à deficiência. Para cada instância perturbada, contendo um termo de um grupo-alvo, foi construída uma instância de controle correspondente, oriunda do mesmo *template* e da mesma configuração emocional, diferindo apenas pela ausência do termo no marcador `<condição>`. Esse delineamento permite interpretar a inserção lexical como uma perturbação mínima e atribuir diferenças de predição à presença do termo, reduzindo a influência de variáveis de confusão.

A sensibilidade do modelo foi quantificada pela diferença entre os escores de saída das instâncias perturbada e controle, $\Delta(x_0, x_1) = f(x_1) - f(x_0)$, e resumida em nível de grupo por meio do *ScoreSense*, calculado como a média dos deltas para os pares associados a cada grupo. Complementarmente, também foram consideradas métricas de mudança decisória, como *LabelDist* e *FlipRate*, bem como um teste t uniamostral sobre os valores de Δ, com hipótese nula $H_0 : E[\Delta] = 0$, a fim de verificar se o efeito observado diferia significativamente de zero.[6]

3.6.1. Justificativa do Protocolo

A PSA aplicada ao BITS-PTBR oferece três vantagens metodológicas centrais. Primeiro, o pareamento (x_0,x_1) elimina grande parte da variabilidade contextual, permitindo atribuir

[5] `nicholasKluge/ToxiGuardrailPT`

[6] A formulação completa do protocolo, incluindo a definição formal do pareamento, das funções de escore e das métricas Δ, *ScoreSense*, *LabelDist* e *FlipRate*, pode ser consultada em: `https://anonymous.4open.science/r/BITS-PTBR-Capacitism-Audit-6888`.

variações de predição a um único elemento lexical. Segundo, a combinação de métricas contínuas (*ScoreSense*) e discretas (*LabelDist*/*FlipRate*) fornece uma caracterização mais completa do comportamento dos modelos, capturando tanto deslocamentos sutis quanto mudanças decisórias. Terceiro, a aplicação do teste estatístico por grupo aumenta a robustez inferencial, distinguindo flutuações aleatórias de efeitos sistemáticos. Em conjunto, o protocolo viabiliza uma auditoria replicável e controlada sobre como modelos em português respondem a marcadores de deficiência, contribuindo para a discussão sobre responsabilidade algorítmica e riscos de discriminação automatizada.

4. Resultados

O corpus **BITS-PTBR** contém 3.105 sentenças instanciadas distribuídas entre quatro grupos semânticos (Clínico, Discurso Social, Sem Deficiência e Neutro). Para cada instância contendo termo de grupo (x_1), foi considerada a sentença de controle correspondente (x_0), resultando em 3.105 pares avaliados sob o protocolo de *Perturbation Sensitivity Analysis* (PSA).

Os resultados são apresentados separadamente para análise de sentimento e detecção de toxicidade, seguidos de uma análise comparativa entre modelos.

Em todas as tabelas apresentadas nesta seção, a coluna **Grupo** indica o conjunto semântico ao qual o termo lexical pertence. A coluna **Mean** corresponde ao *ScoreSense*, isto é, à média do deslocamento entre a sentença contendo o termo de grupo e sua respectiva sentença de controle. A coluna **Std** apresenta o desvio padrão dos deltas observados. A estatística **t** e o respectivo **p-value** referem-se ao teste t uniamostral aplicado sob a hipótese nula. A coluna **n** indica o número de pares avaliados em cada grupo.

Nos modelos de toxicidade, são ainda reportadas as métricas **LabelDist**, que mede a divergência entre distribuições de rótulos antes e após a perturbação, e **FlipRate**, que corresponde à proporção de casos em que a decisão binária do modelo é alterada após a substituição lexical.

4.1. Análise de Sentimento — BERTweet-PT

Na Tabela 3 são apresentados o *ScoreSense* por grupo e os resultados do teste t uniamostral aplicado aos deltas para o modelo BERTweet-PT.

Tabela 3. *ScoreSense* e Teste t para Δ — BERTweet-PT

Grupo	Mean	Std	t	p-value	n
Clínico	-0.0638	0.2444	-7.43	2.81×10^{-13}	810
Discurso Social	-0.0612	0.1730	-10.07	1.54×10^{-22}	810
Sem Deficiência	-0.0021	0.1650	-0.32	0.746	675
Neutro	0.1100	0.2550	12.28	6.74×10^{-32}	810

Observa-se que os grupos **Clínico** e **Discurso Social** apresentam deslocamento médio negativo significativo na polaridade prevista, indicando que a inserção de termos relacionados à deficiência reduz sistematicamente a positividade atribuída pelo modelo. O grupo **Neutro** apresenta deslocamento positivo, enquanto o grupo **Sem Deficiência** não apresenta efeito relevante.

Os resultados indicam que os deslocamentos para os grupos Clínico e Discurso Social são estatisticamente significativos ($p-value < 0.001$), confirmando que o modelo

apresenta sensibilidade sistemática à presença desses termos. O grupo Sem Deficiência não apresenta diferença significativa, funcionando adequadamente como grupo de controle.

Esse padrão sugere que o modelo associa, em média, termos relacionados à deficiência a menor polaridade positiva, mesmo quando o contexto estrutural da sentença permanece constante.

4.2. Detecção de Toxicidade — BERTimbau (Toxic Comment Classification)

Na Tabela 4 são apresentados os resultados do protocolo de PSA para o modelo BERTimbau ajustado para detecção de toxicidade.

Tabela 4. Teste t para Δ — BERTimbau Toxic

Grupo	Mean	Std	t	p-value	n	LabelDist	FlipRate
Clínico	0.0426	0.1045	11.59	7.66×10^{-29}	810	0.1847	0.0802
Discurso Social	0.1451	0.2506	16.48	7.63×10^{-53}	810	0.4701	0.2914
Sem Deficiência	0.0678	0.2182	8.07	3.17×10^{-15}	675	0.3814	0.2000
Neutro	-0.0482	0.1145	-11.99	1.34×10^{-30}	810	0.2969	0.1173

Os resultados evidenciam aumento sistemático na probabilidade estimada de toxicidade para os grupos **Clínico**, **Discurso Social** e **Sem Deficiência**, todos com deslocamentos médios positivos e estatisticamente significativos ($p < 0.001$). O grupo **Discurso Social** apresenta o maior efeito ($\Delta = 0.1451$), superando substancialmente os demais grupos. Em contraste, o grupo **Neutro** apresenta deslocamento negativo significativo, indicando redução da toxicidade estimada quando termos neutros são introduzidos.

A análise decisória reforça esse padrão. O grupo **Discurso Social** apresenta *FlipRate* de aproximadamente 29%, indicando que quase um terço das sentenças altera sua classificação binária após a perturbação lexical. Esse valor é significativamente superior aos observados nos demais grupos, evidenciando impacto operacional relevante na decisão do modelo. Ainda que os grupos Clínico e Sem Deficiência também apresentem mudanças decisórias não desprezíveis (8% e 20%, respectivamente), o efeito do discurso social destaca-se tanto em magnitude contínua quanto em impacto categórico.

Esse comportamento sugere que termos amplamente utilizados no discurso cotidiano sobre deficiência exercem influência particularmente forte sobre o modelo, elevando de forma consistente a probabilidade estimada de conteúdo tóxico, mesmo sob controle estrutural rigoroso das sentenças.

4.3. Detecção de Toxicidade — ToxiGuardrail-PT

Na Tabela 5 são apresentados os resultados do protocolo de PSA para o modelo ToxiGuardrail-PT, incluindo os deslocamentos médios por grupo e as métricas estatísticas correspondentes.

Todos os grupos apresentam deslocamentos médios positivos e estatisticamente significativos ($p < 0.001$), indicando aumento sistemático na probabilidade estimada de toxicidade após a inserção dos termos de grupo. O maior efeito novamente é observado para o grupo **Discurso Social**, cujo deslocamento médio ($\Delta = 0.1002$) é substancialmente superior aos demais.

Observa-se que o *LabelDist* assume valor máximo (1.0) em todos os grupos, indicando que a distribuição de rótulos antes e após a perturbação nunca foi idêntica. Entretanto, a taxa efetiva de mudança decisória (*FlipRate*) permanece relativamente baixa para a maioria dos grupos, variando entre 0.2% e 10.1%. Esse comportamento sugere que o modelo ajusta sistematicamente suas probabilidades contínuas em resposta à perturbação lexical, mas raramente ultrapassa o limiar de decisão binária.

Comparativamente ao BERTimbau, o ToxiGuardrail-PT demonstra menor sensibilidade decisória, embora preserve o mesmo padrão direcional de aumento de toxicidade associado principalmente ao grupo **Discurso Social**.

Tabela 5. Teste t para Δ — ToxiGuardrail-PT

Grupo	Mean	Std	t	p-value	n	LabelDist	FlipRate
Clínico	0.0128	0.0635	5.76	1.19×10^{-8}	810	1.000	0.0025
Discurso Social	0.1002	0.2241	12.73	5.62×10^{-34}	810	1.000	0.1012
Sem Deficiência	0.0271	0.0941	7.49	2.21×10^{-13}	675	1.000	0.0207
Neutro	0.0183	0.0852	6.10	1.60×10^{-9}	810	1.000	0.0185

Todos os grupos apresentam deslocamentos positivos estatisticamente significativos ($p < 0.001$), com destaque novamente para Discurso Social.

A taxa de mudança decisória é menor que no BERTimbau, variando entre 0.2% e 10%, sugerindo menor sensibilidade decisória, embora o padrão direcional permaneça consistente.

4.4. Comparação Entre Modelos

A análise comparativa entre os modelos avaliados evidencia padrões consistentes na forma como a menção de termos relacionados à deficiência influencia as predições. Em ambos os modelos de toxicidade, o grupo **Discurso Social** apresentou os maiores deslocamentos médios de escore, indicando que expressões amplamente utilizadas no cotidiano, como aquelas pertencentes ao vocabulário social sobre deficiência, tendem a produzir maior elevação na probabilidade estimada de conteúdo tóxico. Esse padrão foi observado tanto no modelo baseado em BERTimbau quanto no ToxiGuardrail-PT, ainda que com magnitudes distintas.

Entre os modelos de toxicidade, o BERTimbau demonstrou maior sensibilidade média à perturbação lexical e maior taxa de mudança decisória (*flip rate*), sugerindo impacto mais pronunciado nas classificações finais. Já o ToxiGuardrail-PT apresentou deslocamentos médios menores e menor frequência de alteração de rótulos discretos, embora o sentido do efeito tenha permanecido consistente, especialmente para o grupo Discurso Social. Na tarefa de análise de sentimento, observou-se um deslocamento negativo sistemático na polaridade atribuída às sentenças contendo termos relacionados à deficiência, particularmente nos grupos Clínico e Discurso Social. Esse comportamento indica que, mesmo sob controle estrutural rigoroso das sentenças, a simples inserção de determinados marcadores lexicais é suficiente para reduzir a positividade prevista pelo modelo.

Em conjunto, os resultados sugerem que a menção de termos associados à deficiência, sobretudo aqueles presentes no **Discurso Social**, influencia as predições dos modelos avaliados, tanto na dimensão de polaridade quanto na de toxicidade, evidenciando sensibilidade sistemática a esse tipo de marcador lexical.

5. Discussões

Os resultados observados podem ser interpretados a partir de uma perspectiva sociológica da linguagem e da tecnologia. A consistência dos deslocamentos identificados, indica que os modelos avaliados não operam em um vazio semântico, mas refletem padrões discursivos presentes no contexto social mais amplo. Os resultados evidenciaram que a menção a termos relacionados à deficiência altera sistematicamente as predições, tanto na polaridade quanto na toxicidade, mesmo quando a estrutura sintática e o conteúdo emocional da sentença permanecem controlados. Tais questões não devem ser interpretadas apenas como variações estatísticas internas aos modelos, mas como indícios de reprodução computacional de estruturas sociais historicamente presentes.

A análise de sentimento evidenciou que os grupos **Clínico** e **Discurso Social** apresentaram escore médio negativo estatisticamente significativo, indicando que a presença de marcadores de deficiência tende a reduzir a positividade prevista pelo modelo. Por exemplo, sentenças como: **"Eu tenho um amigo cadeirante." e "Havia uma pessoa com deficiência visual na escola."** apresentaram deslocamentos negativos em relação às suas versões de controle, apesar de não conterem qualquer conteúdo avaliativo explícito. Em contextos emocionalmente positivos, como: **"O jantar com meu irmão cadeirante foi maravilhoso." e "Meu amigo cego me fez sentir feliz."** observou-se, a redução da categoria positiva prevista, sugerindo que a simples associação lexical com deficiência interfere na leitura afetiva da sentença.

Na tarefa de toxicidade, o grupo **Discurso Social** apresentou os maiores aumentos médios de probabilidade de toxicidade, além de taxas substanciais de mudança decisória. Em alguns casos, as sentenças sofreram mudança de rótulo após a inserção do termo relacionado à deficiência. Isso demonstra que sentenças estruturalmente equivalentes passaram de não tóxicas para tóxicas apenas devido à presença de um marcador lexical associado à deficiência. Entre os exemplos com maior deslocamento observado, destacam-se casos como: **"Eu sou um estudante surdo." e "A pessoa com deficiência estava em uma situação comum."** Em suas versões de controle, essas sentenças foram predominantemente classificadas como não tóxicas. Contudo, a simples inserção do marcador de deficiência elevou a probabilidade estimada de toxicidade, e em parte dos casos alterou a decisão final do modelo.

O fato de o grupo **Discurso Social** apresentar deslocamentos mais intensos que o grupo **Clínico** sugerem um paralelo com termos amplamente utilizados no cotidiano que carregam histórico cultural e usos metafóricos. Expressões como "cego" ou "surdo" são frequentemente utilizadas em contextos metafóricos ou pejorativos na linguagem social (como em expressões: **"cego de amor", "cego em tiroteio", "se faz de surdo"**), o que pode ter contribuído para associações estatísticas entre esses termos e contextos negativos.

É importante destacar que os resultados não indicam intencionalidade por parte dos modelos, mas sim a manifestação de estruturas linguísticas aprendidas a partir dos dados. No entanto, a ausência de intencionalidade não elimina os efeitos sociais potenciais, e, ao serem treinados em grandes volumes de dados extraídos de ambientes digitais, podem vir a internalizar padrões estatísticos que refletem desigualdades estruturais presentes na sociedade. A IA, nesse sentido, não cria o capacitismo, mas o codifica, sistematiza e amplifica. O que antes poderia se manifestar como preconceito no discurso social transforma-se em probabilidade numérica e limiar decisório.

6. Conclusão

Este artigo investigou a presença de padrões capacitistas em modelos de PLN para o português brasileiro por meio da adaptação do *framework* BITS ao contexto nacional, resultando na construção do corpus BITS-PTBR. A partir da aplicação da metodologia estruturada, avaliamos modelos utilizados em tarefas de análise de sentimento e detecção de toxicidade, examinando se a menção a termos relacionados à deficiência altera sistematicamente suas predições.

Os resultados indicaram que marcadores de deficiência, produzem deslocamentos estatisticamente significativos tanto na polaridade quanto na toxicidade. Observou-se redução consistente da positividade atribuída em sentenças neutras ou positivas e aumento relevante da probabilidade de toxicidade, incluindo mudanças decisórias mensuráveis em modelos de moderação. Esses achados sugerem que referências à deficiência podem funcionar como marcadores lexicais sensíveis nos modelos avaliados, evidenciando a necessidade de auditorias específicas voltadas a essa dimensão.

Dentre as limitações deste trabalho, destaca-se o uso de um corpus sintético e estruturalmente controlado. Embora esse delineamento metodológico permita isolar a variável lexical associada à deficiência, ele não captura a complexidade pragmática e discursiva de textos naturais. Além disso, o recorte adotado concentrou-se predominantemente no gênero gramatical masculino e não contemplou análises interseccionais envolvendo marcadores como identidade de gênero. Por fim, o conjunto de modelos avaliados, ainda que representativo de sistemas amplamente utilizados, não abrange a totalidade de ferramentas de PLN disponíveis, especialmente soluções ou modelos de larga escala.

Como desdobramento desta pesquisa, sugere-se a ampliação do corpus para contemplar variações de gênero, raça e outras dimensões interseccionais, permitindo investigar como múltiplos marcadores sociais podem interagir na atribuição automática de escores. Estudos futuros também podem explorar dados naturais extraídos de corpora reais e ambientes digitais, comparando-os com o corpus sintético aqui utilizado, a fim de avaliar diferenças entre viés estruturalmente isolado e viés manifestado em contextos discursivos mais complexos.

Uso de IA generativa

O uso de ferramentas de IA generativa na escrita deste trabalho restringiu-se exclusivamente ao aprimoramento linguístico do texto, incluindo reescrita, parafraseamento e lapidação da redação produzida pelos autores. Nenhuma dessas ferramentas foi empregada para sugerir, gerar ou desenvolver novo conteúdo intelectual, limitando-se a funções análogas às de corretores gramaticais, ortográficos ou dicionários.

Disponibilidade dos dados e códigos

Os dados e códigos utilizados neste trabalho estão disponíveis em `https://anonymous.4open.science/r/BITS-PTBR-Capacitism-Audit-6888/`.

Agradecimentos

O presente trabalho foi realizado com o apoio da Universidade Federal de Mato Grosso do Sul e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Bariffi, F. J. (2021). Artificial intelligence, human rights and disability. *Pensar: Revista de Ciências Jurídicas*, 26(2).

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3):671–732.

Bennett, C. L. and Keyes, O. (2019). What is the Point of Fairness? Disability, AI and The Complexity of Justice.

Crochík, J. A. L. (1996). Preconceito, individuo e sociedade. *Temas em Psicologia*, 4:47 – 70.

Desvelar, S. (2026). Danos e discriminação algorítmica: Mapeamento. Desvelar Justiça racial, IA e tecnologias digitais. Acesso em: 06/01/2026.

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1).

Gesser, A. (2008). Do patológico ao cultural na surdez: para além de um e de outro ou para uma reflexão crítica dos paradigmas. *Trabalhos em Linguística Aplicada*, 47(1):223–239.

Glazko, K., Mohammed, Y., Kosa, B., Potluri, V., and Mankoff, J. (2024). Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 687–700, New York, NY, USA. Association for Computing Machinery.

Griffin, P., Peters, M. L., and Smith, R. M. (2007). Ableism curriculum design. In Adams, M., Bell, L. A., and Griffin, P., editors, *Teaching for Diversity and Social Justice*, page 24. Routledge, 2nd edition.

Herold, B., Waller, J., and Kushalnagar, R. (2022). Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In Ebling, S., Prud'hommeaux, E., and Vaidyanathan, P., editors, *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 58–65, Dublin, Ireland. Association for Computational Linguistics.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Krupiy, T. T. and Scheinin, M. (2023). Disability Discrimination in the Digital Realm: How the ICRPD Applies to Artificial Intelligence Decision-Making Processes and Helps in Determining the State of International Human Rights Law. *Human Rights Law Review*, 23(3):ngad019.

Li, R., Kamaraj, A., Ma, J., and Ebling, S. (2024). Decoding ableism in large language models: An intersectional approach. In Dementieva, D., Ignat, O., Jin, Z., Mihalcea, R., Piatti, G., Tetreault, J., Wilson, S., and Zhao, J., editors, *Proceedings of the*

Third Workshop on NLP for Positive Impact, pages 232–249, Miami, Florida, USA. Association for Computational Linguistics.

Mondal, I., Kaur, S., Bali, K., Vashistha, A., and Swaminathan, M. (2022). "#DisabledOnIndianTwitter" : A dataset towards understanding the expression of people with disabilities on Indian Twitter. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 375–386, Online only. Association for Computational Linguistics.

Pacheco, K. M. D. B. and Alves, V. L. R. (2007). A história da deficiência, da marginalização à inclusão social: uma mudança de paradigma. *Acta Fisiátrica*, 14(4):242–248.

Urbina, J. T., Vu, P. D., and Nguyen, M. V. (2025). Disability Ethics and Education in the Age of Artificial Intelligence: Identifying Ability Bias in ChatGPT and Gemini. *Archives of Physical Medicine and Rehabilitation*, 106(1):14–19. Epub 2024-08-30.

Venkit, P. N., Srinath, M., and Wilson, S. (2022). A study of implicit bias in pretrained language models against people with disabilities. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Venkit, P. N., Srinath, M., and Wilson, S. (2023). Automated ableism: An exploration of explicit disability biases in sentiment and toxicity analysis models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 26–34, Toronto, Canada. Association for Computational Linguistics.