Um estudo sobre vieses de gênero em modelos de PLN aplicado em histórias geradas pelo GPT-3.5 e Gemini

Maria Clara Ramalho Medeiros[1], Francisco Paulo de Freitas Neto[1]

[1]INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
CAMPUS CAJAZEIRAS

mramalhomedeiros@gmail.com, f.freitas@ifpb.edu.br

Abstract. *This work addresses the importance of studying gender biases in Natural Language Processing (NLP) models, particularly in generative artificial intelligences. The research aimed to understand how these biases are reproduced in texts generated by models such as GPT and Gemini. To achieve this, the BERT NLP model was trained to infer the gender referenced in the text. The study utilized the md_gender_bias dataset to investigate these biases, highlighting the importance of analyzing the social impact of AIs, especially when used without considering these biases. Based on the analysis of the obtained results, the presence of historical bias, confirmation bias, and selection bias in these models was confirmed.*

Resumo. *Este trabalho tem como foco a análise crítica dos vieses de gênero presentes em modelos de Processamento de Linguagem Natural (PLN), dado seu impacto nas aplicações sociais da inteligência artificial, especialmente em relação às inteligências artificiais generativas. A pesquisa buscou entender como esses vieses são reproduzidos em textos gerados por modelos como GPT e Gemini. Para isso, o modelo de PLN BERT foi treinado para inferir o gênero a qual se refere o texto. O estudo utilizou a base de dados md_gender_bias para investigar esses vieses, destacando a relevância de analisar o impacto social das IAs, especialmente quando usadas sem considerar esses vieses. A partir da análise dos resultados obtidos, foi possível confirmar a presença de viés histórico, confirmação e seleção nesses modelos.*

1. Introdução

A Inteligência Artificial (IA) tem se expandido rapidamente, tornando-se uma tecnologia acessível a diversos setores, especialmente na área de Processamento de Linguagem Natural (PLN), que lida com tarefas como Análise de Sentimentos, Resolução de Correferência e Geração de Texto [Russell and Norvig 2019]. Esses modelos buscam simular a compreensão da linguagem humana, mas são vulneráveis aos vieses presentes nos dados de treinamento, uma vez que refletem as escolhas e concepções humanas.

Os vieses de gênero em modelos de PLN, como o GPT-3.5 e o Gemini, têm sido amplamente estudados, especialmente em áreas como análise de sentimentos, tradução automática e modelos gerativos de perguntas e respostas. Exemplos de impacto desses vieses incluem o algoritmo da Amazon, que em 2014 rejeitou candidatas mulheres

em processos seletivos, e falhas no Google Tradutor, que reforçaram estereótipos de gênero em traduções automáticas [Autran 2018, Stanovsky et al. 2019]. Além disso, estudos mostraram que modelos gerativos, como o GPT-4, podem reforçar estereótipos de gênero em áreas como saúde, prejudicando a precisão e a diversidade nos diagnósticos [Zack et al. 2023]. Embora as pesquisas sobre o GPT sejam extensas, o modelo Gemini ainda carece de investigação aprofundada, embora o Google tenha reconhecido os vieses presentes nele e prometido melhorias [Techrunch 2024].

Esses exemplos revelam a urgência de estudos que investiguem a presença de vieses de gênero em modelos amplamente utilizados, como o GPT-3.5 e o Gemini, especialmente quando esses modelos geram conteúdos sensíveis, como narrativas que envolvem representações de gênero.

Neste contexto, o presente trabalho visa explorar a manifestação de vieses de gênero em dois dos principais modelos de IA generativa: o GPT-3.5 e o Gemini. A pesquisa se concentrará na análise de histórias geradas com a temática de tecnologia, verificando como esses vieses se refletem nas narrativas e quais são suas implicações para a geração de conteúdo.

1.1. Problemática

O exemplo do algoritmo da Amazon, que rejeitou candidatas mulheres em 2014, ilustra de forma clara as consequências dos vieses de gênero em IA. A falha do modelo de selecionar currículos baseados em um histórico predominantemente masculino resultou em discriminação de gênero, prejudicando as chances de mulheres serem contratadas. Esse viés de gênero reflete a má representação e a estereotipação de grupos menos dominantes, como ocorre com mulheres cisgênero, mulheres trans e pessoas não binárias.

A problemática dos vieses de gênero em IA não é recente, mas ganhou destaque com o avanço de modelos de PLN baseados em Aprendizado Profundo (Deep Learning) desde a década de 1990, tornando-se um tema de pesquisa crescente a partir de 2012 [Costa-jussa 2019]. Até 2022, cerca de 200 estudos sobre o tema haviam sido publicados [Devinney et al. 2022], mas ainda há lacunas a serem exploradas, especialmente no contexto das IAs generativas.

Diante desses desafios, este estudo investiga como modelos de IA, especificamente o GPT-3.5 e o Gemini, reproduzem visões de gênero na geração de narrativas voltadas à tecnologia. Uma análise dessas histórias permite identificar padrões e compreender de que essas formações de IA refletem ou reforçam desigualdades de gênero presentes nos dados em que foram treinadas. Para isso, foi utilizado um modelo de PLN para inferir o gênero de um texto, baseando-se na arquitetura do modelo pré-treinado BERT. Esse modelo foi aplicado a textos gerados por IAs generativas, GPT-3.5 e Gemini, com histórias sobre diversas áreas da tecnologia, incluindo desenvolvimento, segurança, arquitetura de redes, engenharia de hardware, e otimização digital.

Este estudo combina uma revisão bibliográfica crítica com uma pesquisa exploratória de caráter quantitativo, utilizando o modelo BERT para análise de vieses em textos gerados pelo GPT-3.5 e Gemini. A abordagem metodológica inclui o treinamento do BERT com a base de dados MGB e a geração de narrativas temáticas em tecnologia, seguidas de análise estatística e quantitativa de viés de gênero temático e textual.

1.2. Objetivos

1.2.1. Objetivo geral

Investigar criticamente a manifestação de vieses de gênero em modelos de linguagem GPT-3.5 e Gemini, com ênfase na geração de narrativas sobre tecnologia.

1.2.2. Objetivos específicos

- Caracterizar os modelos de PLN e explicar suas principais funcionalidades.
- Avaliar as principais fontes de viés nos modelos de PLN, com ênfase nos de gênero.
- Aplicar prompts padronizados para gerar narrativas na temática de tecnologia para avaliar os vieses de gênero.
- Construir, treinar e validar um classificador baseado no modelo BERT.
- Interpretar e discutir criticamente os resultados obtidos, com foco em suas implicações sociais.

1.2.3. Trabalhos relacionados

O debate sobre vieses em modelos de linguagem de larga escala tem ganhado força nos últimos anos, principalmente com a popularização de sistemas como o chatbot do GPT. Estudos recentes têm investigado a presença de vieses sociais nesses modelos, com foco especial em vieses de gênero, ideológicos e estereotípicos, por meio de diferentes abordagens metodológicas e métricas avaliativas.

Rodrigues et al. (2023) realizaram uma análise qualitativa das respostas do GPT a perguntas com conteúdo político-ideológico, observando a presença de tendências discursivas em sua formulação linguística. Ao focarem na gramática dos adjetivos utilizados nas respostas, os autores mostram como aspectos linguísticos, mesmo sob suposta neutralidade, podem indicar posições ideológicas implícitas. Tal abordagem converge com o presente estudo ao tratar os elementos textuais — como pronomes e estrutura narrativa — como marcadores de posicionamento sociocultural [Rodrigues et al. 2023].

Assi e Caseli (2024), por sua vez, empregaram a métrica de regard para mensurar o respeito ou deferência comunicada pelo GPT-3.5-Turbo a diferentes gêneros, tanto em português quanto em inglês. Seus resultados apontam para uma leve preferência por personagens femininas. Tais achados sustentam a opção metodológica deste trabalho de realizar a análise final em inglês e reforçam a importância de considerar o idioma como variável crítica na análise de viés [Assi and Caseli 2024].

O presente estudo diferencia-se dos trabalhos citados ao integrar duas frentes analíticas complementares: (i) uma análise temática de gênero inferida por meio de modelo BERT supervisionado e (ii) uma análise textual de pronomes e estrutura linguística das narrativas. A proposta de utilizar histórias geradas pelas IAS a partir de prompts neutros como corpus também representa uma contribuição metodológica, ao permitir avaliar os vieses gerados espontaneamente pelo modelo em um cenário controlado, mas com alta variabilidade discursiva.

2. Referencial teórico

2.1. PLN

Modelos de PLN lidam com tudo que está relacionado com linguagem. O termo PLN em si é um termo guarda-chuva que abriga todas as possibilidades relacionadas à capacidade da máquina reconhecer, interpretar e gerar linguagem [Cristina Lopes Perna 2010]. Para isso, é preciso antes compreender a linguagem em toda sua complexidade em semântica, sintaxe, regras gramaticais e variações.

O campo específico responsável por isso dentro da PLN é Compreensão de Linguagem Natural (CLN), que é mais conhecido pelo termo em inglês *NLU*, que se refere a *Natural Language Understanding* [Caseli and Nunes 2024]. Esse campo é essencial para que o modelo consiga compreender as linguagens em sua totalidade, num processo que para o ser humano é muito fluído, mas que deve ser convertido para o que é compreensível para máquinas.

Outro campo paralelo ao CLN é o de Geração de Linguagem Natural (GLN), que esse é o responsável por atribuir sua capacidade de gerar respostas [Torfi et al. 2020].

As duas andam lado a lado para compôr um modelo completo capaz de compreender a linguagem como também de criar por conta própria artefatos linguísticos de forma que mais se assemelhe a um ser humano. Para atingir tal propósito, o PLN utiliza a Aprendizagem Profunda (Deep Learning), uma subárea da Aprendizagem de Máquina (Machine Learning), o campo das IAs, para suprir as necessidades e dificuldades provindas da sua complexidade. A Aprendizagem Profunda se baseia no uso de redes neurais que simulam a função dos neurônios no sistema nervoso central humano [LeCun et al. 2015].

2.2. BERT

O BERT (Bidirectional Encoder Representations from Transformers), proposto por Devlin et al. (2019), utiliza uma arquitetura bidirecional baseada em transformers para pré-treinar representações contextuais profundas. Diferentemente de modelos como word2vec e GloVe, que atribuem representações estáticas a palavras, o BERT analisa o contexto completo (palavras anteriores e posteriores), capturando nuances semânticas em múltiplas camadas da rede neural. Sua estratégia de pré-treinamento inclui o Masked Language Modeling (MLM), que oculta tokens aleatoriamente, obrigando o modelo a inferir termos ausentes com base no contexto ([Devlin et al. 2019].

Essa abordagem bidirecional viabiliza transfer learning eficiente: o modelo, pré-treinado em grandes corpora não rotulados, é adaptado via fine-tuning para tarefas específicas como classificação de texto, detecção de viés e question answering [Qasim et al. 2022]. Sua eficácia consolidou-o como referência em PLN, inspirando variantes como XLNet e DistilBERT, que otimizam eficiência computacional e generalização para domínios diversos [Koroteev 2021].

2.3. GPT e Gemini

Ambos os modelos são generativos, capazes de gerar e validar novos dados, com aplicações em textos, imagens, vídeos e áudios. Neste estudo, serão avaliadas suas capacidades no Processamento de Linguagem Natural (PLN), especificamente em compreensão

e geração de linguagem. Para isso, é essencial compreender o que os torna populares, inclusive fora do campo tecnológico, alcançando um público leigo.

O GPT (*Generative Pre-trained Transformers*) [OpenAI 2022] é uma família de modelos desenvolvida pela *OpenAI* desde 2018. Sua arquitetura de transformadores, associada a métodos autodecodificadores e unidimensionais, potencializa seu desempenho [Radford et al. 2018]. O *ChatGPT*, um desses produtos, ganhou popularidade por gerar textos coerentes e interpretar conteúdo, realizando tarefas como síntese, resumo e criação de ideias. Isso facilita atividades que antes demandariam tempo e criatividade humana.

O Gemini [Pichai and Hassabis 2023], anteriormente conhecido como "Bard", é um modelo do Google, também baseado em transformadores e autodecodificadores, funcionando de maneira similar ao GPT. Em dezembro de 2023, o Gemini Ultra prometeu ser 4% superior ao GPT-4 em interpretação e geração de respostas, superando em 93% outros modelos de PLN em benchmarks. O Gemini Pro, embora mais rápido que o GPT-4 em traduções, oferece menor qualidade. Em tarefas de perguntas e respostas, o Gemini Pro se destaca em relação ao GPT-3.5. Contudo, sua geração de conteúdo pode ser tendenciosa aos interesses do Google [II 2023].

2.4. Conhecendo a base de dados

A base de dados MGB é um produto da pesquisa "Multi-Dimensional Gender Bias Classification" [Nemani et al. 2024]. Ela possui diversas configurações, variando entre dados obtidos por inferência de anotadores e dados rotulados por modelos de PLN, que identificam o gênero temático do texto. Para esta pesquisa, foi escolhida a configuração *opensubtitles_inferred*, baseada em legendas de filmes e TV [Lison and Tiedemann 2016], contendo apenas legendas com o nome ou identidade de personagens. Esta configuração inclui 442 mil tuplas (351 mil para treino, 49 mil para teste e 42 mil para validação). Os rótulos de gênero foram atribuídos com base em termos de parentesco, distribuição probabilística de nomes e dinâmica da conversa entre personagens, utilizando um classificador treinado em datasets anteriores [Dinan et al. 2020].

Os anotadores, todos trabalhadores da Amazon Mechanical Turk nos Estados Unidos e falantes de inglês, tiveram seu gênero reportado da seguinte forma: 67,38% se identificaram como homem, 18,34% como mulher, 0,21% como não-binário e 14,07% preferiram não dizer.

A predominância de anotadores do gênero masculino pode introduzir viés nos dados, afetando a imparcialidade do classificador gerado e contribuindo para padrões indesejados. Além disso, o uso de nomes como indicadores de gênero apresenta desafios, uma vez que nomes não refletem necessariamente a identidade de uma pessoa. A abordagem binária, embora comum, limita a representação de identidades de gênero diversas, levantando questões sobre seu impacto no desempenho do modelo.

Uma característica relevante dessa configuração é a inclusão da categoria "gender-neutral", que permite rotulações ternárias, contrastando com a abordagem binária. Embora essa categoria possa ser vista como prejudicial [Dev et al. 2021], ela é a melhor escolha para esta pesquisa, pois outras bases de dados não consideram opções além da binariedade.

A configuração inclui rótulos binários (0 para feminino, 1 para masculino) e o

rótulo 2 para o gênero neutro. Além disso, a estrutura contém o percentual atribuído à rotulação e a frase analisada.

2.5. Entendendo vieses e quais suas fontes

No contexto do PLN, fontes de viés em modelos incluem o Viés de Dado e o Viés de Rotularização [Nemani et al. 2024]. O Viés de Dado refere-se ao viés nos dados utilizados para treinar o modelo, podendo ser originado por viés do desenvolvedor, padrões distorcidos, representações limitadas de grupos minoritários e dados desatualizados. Esses vieses são influenciados por vieses cognitivos, como: Viés de Confirmação (informações que sustentam crenças preexistentes, como a ideia de que mulheres não são aptas para cargos de liderança); Viés Histórico (decorrente de crenças culturais sistemáticas que distorcem os dados e não refletem a realidade); e Viés de Seleção (quando a amostra de dados não representa adequadamente o grupo-alvo, podendo ocorrer o Viés de Amostragem, com dados não aleatorizados, ou Viés de Convergência, quando a amostra não é coletada corretamente) [Mehrabi et al. 2022].

Um exemplo notável de viés de dados é o caso da *Amazon*, em que o modelo de recrutamento reproduziu vieses históricos e de seleção, já que o treinamento se baseou em dados predominantemente masculinos. Isso reflete o Viés Histórico, pois homens têm mais oportunidades na área de tecnologia, e o Viés de Seleção, já que a amostra não representava adequadamente o público-alvo, resultando em discriminação implícita.

O Viés de Rotularização ocorre no processo de anotação dos dados, realizado por anotadores que inserem seus próprios vieses nos dados de treinamento e validação. Esse viés é exacerbado quando os dados rotulados já são enviesados, como no caso das bases de dados utilizadas nesta pesquisa, que passaram por esse processo de rotulação.

Independentemente da fonte, todo o desenvolvimento de um modelo está intrinsecamente ligado aos vieses que o permeiam: na escolha, origem e pré-processamento dos dados, e na finalidade do modelo. Como os dados refletem representações humanas, eles carregam os vieses das vivências e culturas, o que influencia os resultados. Para entender como os vieses de gênero surgem e impactam os modelos, é necessário compreender como o gênero é interpretado, uma vez que as distorções de gênero nos dados geram resultados enviesados.

3. Metodologia

Esta pesquisa foi conduzida em cinco etapas principais, conforme ilustrado na Figura 1:

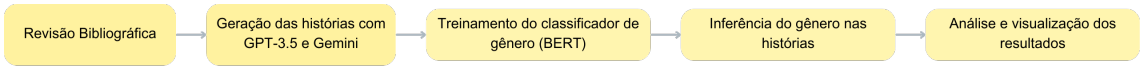

Figure 1. Fluxograma das etapas metodológicas da pesquisa

A primeira etapa consistiu na realização de uma **Revisão Bibliográfica**, onde foi realizada uma revisão teórica sobre Processamento de Linguagem Natural (PLN) e as fontes de vieses que podem influenciar os modelos de linguagem.

Na segunda etapa, intitulada **Geração das Histórias com Modelos GPT-3.5 e Gemini**, foram elaborados 12 *prompts* (instruções) com temática voltada para áreas tecnológicas, como ”desenvolvimento de front end” e ”engenharia de dados”. Os *prompts*

estão disponíveis no repositório do projeto [1]. Inicialmente, foi realizada uma geração exploratória em português com um número reduzido de amostras, permitindo avaliar se os modelos produziam narrativas coerentes e se os *prompts* precisavam de ajustes. Essa fase não foi utilizada na análise quantitativa. As análises finais foram conduzidas exclusivamente com histórias geradas em inglês, visto que os modelos analisados têm como base de treinamento majoritariamente esse idioma. Embora os *prompts* tenham sido formulados de forma neutra, reconhece-se a possibilidade de viés já nesta etapa, devido à associação inconsciente de determinadas áreas técnicas a estereótipos de gênero, o que motivou uma análise crítica posterior sobre esse aspecto.

A terceira etapa, **Treinamento do Classificador de Gênero**, consistiu no desenvolvimento de um classificador de gênero textual utilizando o modelo `bert-base-uncased`, treinado sobre a base de dados MGB. O treinamento teve duração de 64 horas e alcançou uma acurácia de 89% segundo o F1 Score. Os hiperparâmetros utilizados foram definidos com base na documentação da HuggingFace[2]: taxa de aprendizado de 2e-5, tamanho de lote (batch size) de 16, 2 épocas de treinamento e decaimento de peso de 0.01.

Na quarta etapa, denominada **Inferência do Gênero nas Histórias**, cada história gerada foi segmentada em frases menores. Essa fragmentação visou compatibilizar a granularidade da entrada com os dados utilizados para treinar o modelo, uma vez que a base MGB era composta por frases curtas. Para cada fragmento, foi inferido um rótulo de gênero. Em seguida, foi identificado o rótulo predominante em cada história e o percentual de ocorrência dos demais rótulos, permitindo também a atribuição múltipla.

Por fim, a quinta etapa, **Análise e Visualização dos Resultados**, envolveu o uso de bibliotecas em Python para visualização gráfica dos resultados. A análise incluiu: (i) contagem dos pronomes utilizados; (ii) identificação do gênero classificado como dominante; e (iii) análise temática e linguística para investigação de viés, agregando inferências quantitativas e observações qualitativas.

4. Resultado e Discussões

Os resultados obtidos consistem em dois datasets a serem analisados, além dos dois datasets das histórias geradas pelo GPT e Gemini.

A estrutura desses dados é composta por um primeiro argumento, que indica a porcentagem atribuída a cada rótulo, e um segundo argumento, que representa o rótulo dominante identificado pelo algoritmo. No caso do MGB, os rótulos são 0 (gênero feminino), 1 (gênero masculino) e 2 (gênero neutro), além do gênero dominante.

Foram gerados gráficos de *heatmap* (intensidade) baseados no MGB (ver Figuras 2 e 3) que analisam a distribuição de histórias classificadas de acordo o percentual de cada rótulo, onde os labels classificados entre os percentis 76% e 100% possuem as menores frequências (a não ser o label_2 no Gemini) comparadas ao restante.

Essa análise se torna especialmente relevante, pois, durante a observação inicial de uma amostra dos dados, foi identificada uma inconsistência nos pronomes e na atribuição

[1] `https://encurtador.com.br/ggMGd`
[2] `https://encurtador.com.br/ueqPC`

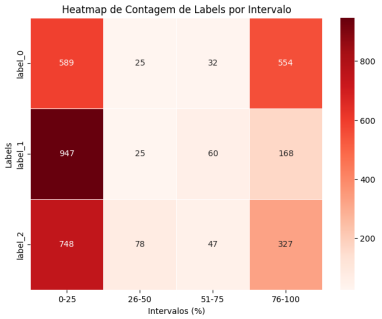

Figure 2. Heatmap de rotulação dos labels (BERT no MGB – histórias do GPT)

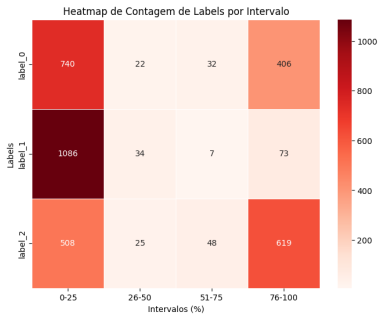

Figure 3. Heatmap de rotulação dos labels (BERT no MGB – histórias do Gemini)

de gênero utilizados nas histórias geradas. Em um dos textos gerados pelo GPT, uma pessoa desenvolvedora front-end é inicialmente descrita no masculino, mas a pessoa recebe pronomes e concordâncias femininas posteriormente, como ”desafiada” e ”uma profissional admirada”. Já no outro texto gerado pelo Gemini, uma pessoa desenvolvedora back-end é introduzida no masculino, mas depois é tratada no feminino. Essa variação sugere um possível viés na associação de gênero a diferentes papéis na área de tecnologia.

No entanto, ao expandir a análise para um volume maior de dados, torna-se inviável identificar com precisão os fatores que levaram a essa variação nos percentis rotulados pelos modelos – principalmente considerando que a base de dados original está em inglês. Apesar disso, ao observar uma amostragem menor, essa falha na geração das histórias se torna perceptível: muitos textos começam utilizando um pronome masculino, mas terminam com um pronome feminino, que, por sua vez, concorda com o nome da personagem.

Esse comportamento sugere uma possível limitação dos modelos na manutenção da coerência textual ao longo da narrativa, o que pode impactar diretamente na forma como os rótulos são atribuídos e interpretados. Porém, como as histórias da amostra são em português, pela base de dados de estudo ser em uma língua não generificada, é incerto como essa inconsistência pode ocorrer no texto.

Na análise de gênero, a contagem de histórias que utilizam o pronome neutro they/their se destacou como um ponto de atenção. Foi necessário um cuidado maior na identificação e rotulação dessas histórias, pois esses pronomes poderiam estar sendo empregados simplesmente no plural, sem necessariamente indicar neutralidade de gênero. No entanto, observou-se que, em diversos casos, seu uso estava intencionalmente associado à representação de personagens de gênero neutro, reforçando a importância de um tratamento criterioso na categorização.

Isso comprova que os modelos têm noção de identidade de gênero, mesmo que de forma limitada, mas que é capaz de desenvolver uma história tentando manter o gênero do personagem de forma neutra.

No GPT, foram produzidas 64 histórias utilizando os pronomes masculinos ”he/him”, com um total de 159 ocorrências desses pronomes ao longo dos textos. Já para os pronomes femininos ”she/her”, o número foi significativamente maior: 627 histórias,

contabilizando 4.733 repetições desses pronomes. No Gemini, os números foram ligeiramente diferentes: 99 histórias com pronomes masculinos, totalizando 295 ocorrências, e 469 histórias com pronomes femininos, somando 3.501 ocorrências. Essa distribuição é semelhante aos resultados gerados pelos modelos. De acordo com o BERT, no Gemini, 445 histórias foram rotuladas com gênero feminino e 85 com gênero masculino. No GPT, os números foram 592 histórias com gênero feminino e 234 com gênero masculino, como pode ser visto nas Figuras 4 e 5.

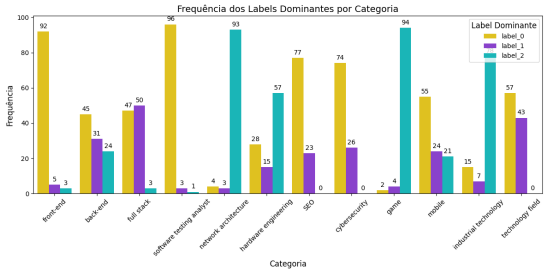

Figure 4. Barras do modelo MGB sobre histórias do GPT

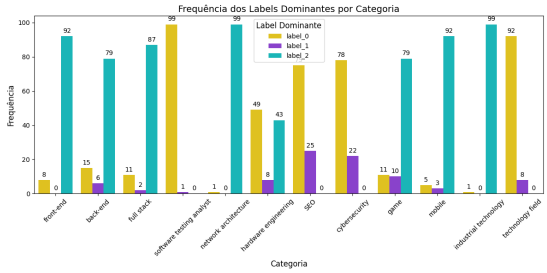

Figure 5. Barras do modelo MGB sobre histórias do Gemini

No caso do Gemini, a distribuição foi diferente, com um número maior de histórias rotuladas com gênero neutro. Independentemente do modelo analisado, o que se destaca é que as histórias com protagonistas masculinos não tiveram um volume expressivo, reforçando um padrão observado ao longo dos resultados.

Essa discrepância na quantidade de histórias geradas com pronomes femininos surpreende, especialmente por ir na contramão do que se observa na realidade. No contexto da área de tecnologia, há um esforço contínuo para promover maior diversidade e inclusão feminina, dado que as mulheres ainda enfrentam barreiras significativas no setor. No entanto, a diferença observada nos modelos não parece refletir um equilíbrio natural, mas sim um direcionamento intencional para amplificar a presença feminina.

Isso pode ser interpretado como uma tentativa de mitigar vieses por meio da super-representação, preenchendo a base com histórias desse tipo a ponto de torná-las dominantes. Contudo, esse esforço acaba tendo um efeito colateral: em vez de soar como um reflexo genuíno da diversidade, a abordagem se torna artificial, como se houvesse uma necessidade constante de reafirmar a ausência de viés.

Algo interessante que pode ser analisado é quando nem mesmo pronomes são usados, mas são utilizadas outras questões que influenciam nisso. Por exemplo, quando uma história gerada pelo próprio GPT foi apresentada a ele novamente, o modelo automaticamente assumiu que a personagem principal era ”claramente identificada como mulher”, mesmo que tivesse sido escrito num cenário neutro e utilizado os pronomes neutros da língua inglesa ”they/them”. Ou seja, a única indicação utilizada para categorizar foi o nome ”Dr. Aurora Chang”, o que evidencia como os modelos de IA associam nomes a gêneros de forma implícita.

Em outra situação semelhante, a IA afirmou que ”o gênero de Dr. Chen é feminino, pois o pronome usado para se referir a ela é ’their’”. Esses comportamentos refletem a influência dos dados de treinamento e a forma como a IA internaliza padrões sociais, reforçando a ideia de que nomes, por si só, são vistos como marcadores de gênero mesmo

que não tenha nenhum pronome explícito ou sejam utilizados pronomes neutros.

Ao comparar as contagens de pronomes com os rótulos atribuídos pelos modelos, percebe-se que essa discrepância pode estar diretamente relacionada. Quando os pronomes não estavam fortemente presentes no texto, o modelo pode ter inferido o gênero com base na associação entre nome e identidade de gênero. Isso só reforça o ponto de que, mesmo quando o objetivo era manter o texto o mais neutro possível, o modelo ainda tendia a atribuir rótulos de acordo com seus próprios vieses, reforçando estereótipos sobre a relação entre nomes e identidades de gênero, o que consequentemente contribui para exclusão de identidades de gênero não binárias ao invalidar sua identidade.

5. Considerações finais

Neste trabalho é investigado como modelos de PLN, mais especificamente as IAs generativas, como GPT-3.5 e Gemini, são capazes de reproduzir vieses. Com esse objetivo, esta pesquisa foi dividida em dois módulos principais: uma revisão bibliográfica no tema e uma pesquisa quantitativa, onde foram analisados 12 resultados obtidos a partir das aplicações dos modelos desenvolvidos, além das próprias histórias geradas pelo GPT-3.5 e Gemini.

Conclui-se que os resultados sugerem a presença de vieses nos modelos analisados, possivelmente influenciados por fatores como viés de confirmação, viés histórico e viés de seleção. O viés de confirmação pode estar refletido na maior atribuição de personagens femininas em certas áreas, possivelmente como uma tentativa dos modelos de corrigir desigualdades pré-existentes, mas sem necessariamente representar a realidade de forma balanceada. O viés histórico se manifesta na associação automática entre nomes e gênero, levando os modelos a atribuírem rótulos mesmo quando os textos foram escritos de forma neutra. Já o viés de seleção pode ter influenciado os resultados devido à composição dos dados de treinamento, fazendo com que os modelos reforcem padrões distorcidos e imponham classificações de gênero mesmo em cenários não generificados.

Este estudo foi conduzido com bases de dados em inglês, o que limita a análise dos vieses específicos da língua portuguesa, especialmente no que diz respeito à generificação inerente ao idioma. Um estudo futuro poderia investigar como esses vieses se manifestam em histórias geradas em português pelo GPT-3.5 e Gemini. No entanto, para que essa análise fosse viável, seria necessário que as bases de dados Stereoset e MGB estivessem disponíveis em português, exigindo um processo de tradução e adaptação desses conjuntos de dados para garantir uma avaliação precisa dos modelos no contexto da língua portuguesa. Foi considerado ser feita a tradução via IA, mas o que poderia garantir outra camada de vieses pelo mesmo produtor.

Além disso, esta pesquisa apresenta algumas limitações, como as mudanças nas políticas dos modelos de IA ao longo do tempo, o que comprova sua natureza dinâmica e a possibilidade de atualizações constantes; o tamanho da amostra de histórias geradas, que se restringiu a 1200 textos; e fatores contextuais na análise de viés, como as diferenças culturais no treinamento dos modelos. Além disso, a análise de vieses foi limitada ao uso dessas duas ferramentas, que já carregam vieses inerentes. Essas não são as condições ideais para estudar um tema sensível e complexo, o que sugere que o tempo e os recursos disponíveis para a pesquisa foram limitados.

Portanto, esse trabalho é essencial para compreender como os modelos de IA,

especialmente os de PLN, influenciam a construção social da linguagem e reforçam desigualdades de gênero. A análise desses vieses não apenas contribui para o aprimoramento técnico das inteligências artificiais, mas também permite um debate mais amplo sobre ética, responsabilidade e transparência na Computação. Ao propor uma abordagem crítica sobre IA e viés, este estudo se alinha à interseção entre Computação e Sociedade, ressaltando a necessidade de regulamentações e práticas mais inclusivas no desenvolvimento de tecnologias. A pesquisa abre caminhos para investigações futuras que aprofundem a relação entre aprendizado de máquina e justiça social, explorando como políticas públicas e diretrizes podem mitigar esses efeitos negativos.

Embora esta pesquisa mencione IAs generativas como o GPT-3.5 e o Gemini, amplamente utilizadas atualmente por diferentes perfis de usuários, o foco não se restringe a essas ferramentas. Pelo contrário, busca-se justamente investigar como o viés se manifesta de forma mais ampla em modelos generativos, utilizando diferentes abordagens que permitam reflexões que transcendam versões e ferramentas específicas.

References

Assi, F. and Caseli, H. (2024). Biases in gpt-3.5 turbo model: a case study regarding gender and language. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 294–305, Porto Alegre, RS, Brasil. SBC.

Autran, F. (2018). Ia da amazon usada em análise de currículos discriminava mulheres.

Caseli, H. and Nunes, M. (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português – 3ª Edição*. BPLN, São Carlos.

Costa-jussa, M. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1.

Cristina Lopes Perna, H. (2010). *Linguagens especializadas em corpora : modos de dizer e interfaces de pesquisa*. Edipucrs.

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J., and Chang, K.-W. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devinney, H., Björklund, J., and Björklund, H. (2022). Theories of "gender" in nlp bias research.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., and Williams, A. (2020). Multi-dimensional gender bias classification.

II, S. M. W. (2023). Comparative analysis: Google gemini pro vs. openai gpt-3.5.

Koroteev, M. V. (2021). BERT: A review of applications in natural language processing and understanding. *CoRR*, abs/2103.11943.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2022). A survey on bias and fairness in machine learning.

Nemani, P., Joel, Y., Vijay, P., and Liza, F. (2024). Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6.

OpenAI (2022). Chatgpt: Language models are few-shot learners.

Pichai, S. and Hassabis, D. (2023). Apresentando o gemini: nosso maior e mais hábil modelo de ia.

Qasim, R., Bangyal, W. H., Alqarni, M. A., and Almazroi, A. A. (2022). A fine-tuned bert-based transfer learning approach for text classification. *Journal of Healthcare Engineering*, 2022:1–17.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rodrigues, G., Albuquerque, D., and Chagas, J. (2023). Análise de vieses ideológicos em produções textuais do assistente de bate-papo chatgpt. In *Anais do IV Workshop sobre as Implicações da Computação na Sociedade*, pages 148–155, Porto Alegre, RS, Brasil. SBC.

Russell, S. and Norvig, P. (2019). *Artificial Intelligence: A Modern Approach*. Pearson, Harlow, England, 3rd edition.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Techrunch (2024). Google still hasn't fixed gemini's biased image generator.

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. *CoRR*, abs/2003.01200.

Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., and Alsentzer, E. (2023). Coding inequity: Assessing gpt-4's potential for perpetuating racial and gender biases in healthcare. *medRxiv*.