Discriminação Algorítmica de Gênero: Estudo de Caso e Análise no Contexto Brasileiro

Fernanda Tiemi de S. Taso[1], **Valéria Q. Reis**[1,2], **Fábio H. V. Martinez**[1]

[1]Faculdade de Computação – Universidade Federal de Mato Grosso do Sul (UFMS)

[2]Institute of Information Systems – Leuphana University Lüneburg

{tiemi.taso, valeria.reis, fabio.martinez}@ufms.br

Abstract. *This paper aims to identify discriminatory trends in Natural Language Processing models that represent words through vectors called Word Embeddings. Pre-defined metrics for identifying bias were adapted and exposed the existence of gender stereotypes in traditional occupations and their correlation with the women's proportion in the national labor market. Moreover, stereotyped analogies between feminine and masculine pronouns were found. Results reveal sexism similar to other studies and allow us to discuss the impact of the use of language models in our society. Finally, the work paves the way for the use of metrics to identify other types of discrimination in the Brazilian context.*

Resumo. *Este trabalho visa identificar tendências discriminatórias em modelos de Processamento de Linguagem Natural que representam palavras por meio de vetores chamados Word Embeddings (WE), buscando vieses de gênero no domínio de profissões encontradas em WE em português. Para isso, métricas pré-definidas para identificação de viés foram adaptadas e revelaram a existência de estereótipos de gênero em ocupações tradicionais e sua correlação com a proporção de mulheres no mercado de trabalho nacional. Também verificaram-se analogias preconceituosas entre pronomes femininos e masculinos. Os resultados evidenciam sexismos semelhantes aos de outros estudos e permitem discutir sobre o impacto do uso de modelos de linguagem em nossa sociedade. Por fim, o trabalho abre caminho para o uso das métricas para identificação de outros tipos de discriminação no contexto brasileiro.*

1. Introdução

Muitas atividades cotidianas estão intrinsecamente conectadas a tecnologias que empregam inteligência artificial (IA) para resolver tarefas específicas. Exemplos são algoritmos de recomendação de conteúdo, reconhecimento facial e atendimento ao cliente. A análise do impacto da tecnologia na sociedade se faz importante a fim de levantar questões éticas e problemas causados pela sua interação. Efeitos negativos têm se proliferado com a adoção dessas ferramentas nos serviços de segurança pública nos EUA, Europa e recentemente no Brasil [Falcão 2021], atingindo grupos específicos da sociedade. Notícias sobre o uso de reconhecimento facial em câmeras de segurança nas ruas de cidades mostraram que prisões de pessoas negras inocentes foram feitas com o auxílio da tecnologia [Werneck 2019]. A repercussão nas redes sociais dos efeitos do algoritmo de

recorte de imagens da plataforma Twitter, que favorecia rostos brancos a negros, fez com que o mesmo fosse alterado [Yee et al. 2021]. Muitos outros casos de discriminação algorítmica, inclusive em serviços assistenciais, são encontrados na literatura [Silva 2022].

Críticas aos monopólios de tecnologia e seus algoritmos, usados por bilhões de pessoas diariamente, ajudam a ter dimensão do impacto que podem causar. Ferramentas de pesquisa em conjunto com algoritmos de ranqueamento de sites e imagens, como o Google Search e Google Images, respectivamente, foram acusadas de racismo algorítmico após inúmeros casos de recomendação de sites pornôs quando se pesquisava por "meninas negras". O caso foi classificado pela empresa como lapso não intencional e posteriormente o erro foi corrigido. A resposta evasiva do Google a essas acusações ilustra um comportamento relutante em reconhecer que seus algoritmos possam precisar de mudanças para não discriminarem grupos específicos [Noble and Damorim 2022].

A questão da neutralidade algorítmica surge como possível solução para diminuição do viés humano em decisões nas quais o pensamento das pessoas sofre influência indesejada. Apesar disso, modelos computacionais não são isentos de viés, pois são construídos e elaborados a partir do raciocínio e decisões humanas, seja ele individual, coletivo por parte de uma equipe, ou hierárquico, no qual descende de uma autoridade maior.

Processamento de Linguagem Natural (PLN) lida com a compreensão de grandes quantidades de dados textuais e de fala por meio de sistemas computacionais. Nessa área, [Suresh and Guttag 2021] descreveram diferentes tipos de vieses que podem surgir na coleta de dados, no desenvolvimento do modelo ou na sua implantação. O viés histórico surge quando os sistemas produzem resultados prejudiciais e discriminatórios, apesar das medições e amostras nos dados terem sido feitas corretamente, refletindo os dados do mundo real. Modelos de *Word Embeddings* (WE) já foram estudados e são portadores de vieses históricos [Bolukbasi et al. 2016, Caliskan et al. 2017, Sogancioglu et al. 2022]. WE é uma técnica de PLN muito utilizada na qual palavras de um *corpus* são transformadas em vetores, preservando seus significados e contextos.

[Bolukbasi et al. 2016] foram pioneiros na proposta e análise de uma métrica para a quantificação de vieses em WE. Usando modelos treinados em textos de notícias, os autores mostram a farta ocorrência de estereótipos no contexto de profissões, criando assim a relação "*O homem está para o programador de computador, assim como a mulher está para a dona de casa*". [Caliskan et al. 2022] trabalharam nesse mesmo tema ao analisarem o viés de gênero em termos de frequência, sintaxe e semântica em dados obtidos da Web. Para os autores, o vocabulário usado nos modelos de WE analisados apresentam forte inclinação para termos masculinos: verbos estão mais associados com gênero masculino e adjetivos com feminino; grupos de palavras como engenharia e esportes agrupam-se perto do gênero masculino, enquanto aparência e termos relativos à cozinha estão mais associados ao gênero feminino.

Por fim, [Sogancioglu et al. 2022] estudaram um *corpus* de medicina e evidenciaram a presença de viés histórico em dois modelos de WE clínicos, principalmente em doenças sexualmente transmissíveis e mentais. Esses três trabalhos foram realizados em textos de língua inglesa e levantam questionamentos se os fenômenos também podem ser encontrados no vocabulário da língua portuguesa. Infelizmente, a pesquisa da discriminação algorítmica de gênero com WE em português é ainda pouco expressiva,

apesar de um conjunto com 31 modelos de WE ter sido disponibilizado ainda no ano de 2017 [Hartmann et al. 2017]. Até onde temos conhecimento, apenas [Santana et al. 2018] usaram preliminarmente parte desse conjunto.

O objetivo deste trabalho é analisar vieses de gênero em um dos modelos de WE criados por [Hartmann et al. 2017], usando as métricas descritas e disponibilizadas por [Bolukbasi et al. 2016]. Para isso, estereótipos de profissões são casos de estudo. Como resultado, este trabalho mostra a existência e caracterização de vieses em modelos da língua portuguesa. Ademais, também incentiva a minimização do sexismo e outras formas de discriminação presentes em modelos de WE.

Este artigo está estruturado como segue. A Seção 2 descreve os procedimentos, métricas e dados que foram usados. Os resultados dos experimentos são expostos na Seção 3. A Seção 4 evidencia os impactos da discriminação de gênero na sociedade, principalmente com o surgimento e adoção de grandes modelos de linguagem. A Seção 5 apresenta discussões sobre as limitações das métricas usadas e sobre o processo de análise de viés de gênero na área de PLN. A Seção 6 conclui e elenca oportunidades de pesquisa.

2. Procedimentos metodológicos

Para prosseguir com a análise de vieses de gênero, o estudo fez uso de modelos e métricas existentes, que são amplamente utilizadas para análise em outros idiomas. Para isso, foi preciso escolher um modelo de WE em português com um desempenho significativo, além da necessidade de um grande número de *tokens* para ser possível analisar palavras nos mais variados escopos. Métricas foram remodeladas para satisfazer regras gramaticais necessárias para a língua portuguesa.

2.1. Escolha do modelo

[Hartmann et al. 2017] treinaram 31 modelos de WE em um grande *corpus* de língua portuguesa, composto por 1,2 bilhões de *tokens* obtidos a partir de 17 provedores digitais de conteúdo, brasileiros e portugueses. Exemplos de fontes do *corpus* são Wikipedia, GoogleNews, Folhinha e G1. Entre os 5 diferentes métodos de criação de WE utilizados, os autores concluíram que o GloVe teve o melhor resultado para tarefas envolvendo analogias de sintaxe e semântica, as quais são importantes para o bom desempenho nas análises de similaridade e analogias.

Os modelos criados por [Hartmann et al. 2017] são um dos poucos WE disponíveis para a língua portuguesa, sendo amplamente usados em diversos contextos e tipos de aplicações [Grave et al. 2018, Garcia and Berton 2021, Silva et al. 2020]. Por esse motivo, o GloVe foi escolhido para as análises deste trabalho. O modelo adotado utiliza uma representação de 300 dimensões, sendo este o tamanho mais utilizado para análise por apresentar um custo-desempenho mais eficiente que dimensões maiores.

2.2. Escolha das métricas

[Bolukbasi et al. 2016] analisam o sexismo encontrado em um modelo Word2vec treinado em artigos do Google News. Para isso, são utilizadas duas métricas: a primeira, dada por similaridade entre palavras, e a segunda, por pares de analogia. A análise foi feita usando listas de profissões e exemplos de analogias. O artigo indica um potencial viés de gênero no modelo. Profissões como “dona de casa”, “enfermeira” e “recepcionista” foram

as palavras mais relacionadas com o pronome “ela”, enquanto “maestro”, “capitão” e “filósofo” foram mais relacionadas com pronome “ele”. Algumas analogias revelam mais dos preconceitos, tal como, *enfermeira-médico* (leia-se “ela está para enfermeira assim como ele está para médico”), *decoradora-arquiteto* ou mesmo *vocalista-guitarrista*.

As duas métricas propostas por [Bolukbasi et al. 2016] foram consideradas essenciais para entender os princípios básicos de cálculos com WE e suas aplicações na análise do sexismo algorítmico. A relevância acadêmica, a consolidação das métricas em trabalhos posteriores e a disponibilização do código-fonte pelos autores tornaram oportuna a replicação de sua análise e suas métricas para modelos ainda pouco estudados, como modelos de WE em português.

2.2.1. Similaridade

A similaridade por cosseno é tipicamente usada em trabalhos que buscam semelhança entre pares. Nela, o cosseno do ângulo entre os vetores é o produto interno entre dois vetores de palavras, dividido pela multiplicação de seus módulos. Com a utilização de um modelo normalizado, a divisão se torna redundante e a fórmula resultante é $\cos(\vec{u}, \vec{v}) = \vec{u} \cdot \vec{v}$, onde $\vec{u}$ e $\vec{v}$ representam dois vetores de palavras. A aplicação da similaridade por cosseno à frase “a palavra *ator* é mais similar a *filme* do que a *avião*” resultaria na seguinte inequação:

$$\overrightarrow{\text{ator}} \cdot \overrightarrow{\text{filme}} > \overrightarrow{\text{ator}} \cdot \overrightarrow{\text{avião}}\,.$$

Dada a particularidade de marcação de gênero da língua portuguesa, neste trabalho a similaridade é adaptada para ser obtida para cada flexão de uma profissão e seu respectivo gênero. Prezando pela comparação dos resultados obtidos por [Bolukbasi et al. 2016], a direção de gênero será dada pelos pronomes pessoais *ela* e *ele*. Assim, considerando $\overrightarrow{p_f}$, $\overrightarrow{p_m}$ como pares de vetores de profissões com concordância nominal feminina e masculina (ex: advogada, advogado), e sendo $\overrightarrow{\text{ela}}$ e $\overrightarrow{\text{ele}}$ seus vetores de pronomes correspondentes, a similaridade entre a profissão com marcador feminino (advogada) e o pronome (ela), s_f, e a similaridade entre a profissão com marcador masculino (advogado) e o pronome (ele), s_m, são dadas por:

$$(s_f, s_m) = (\cos(\overrightarrow{p_f}, \overrightarrow{\text{ela}}), \cos(\overrightarrow{p_m}, \overrightarrow{\text{ele}})) = (\overrightarrow{p_f} \cdot \overrightarrow{\text{ela}}, \overrightarrow{p_m} \cdot \overrightarrow{\text{ele}})\,. \tag{1}$$

A métrica de similaridade foi utilizada para 114 pares de profissões de diferentes grupos e famílias ocupacionais existentes na Classificação Brasileira de Ocupações (CBO). O conjunto foi obtido abrangendo as famílias ocupacionais existentes dentro do vocabulário do modelo. Depois, a filtragem foi direcionada para as profissões onde a flexão de ambos os gêneros, feminino e masculino, estivessem presentes no vocabulário para a criação dos pares.[1]

[1] A listagem de profissões pode ser acessada em: https://github.com/nandayot/gender-bias-portuguese-replication-bolukbasi

Usando a Equação 1, foi construída uma lista de profissões mais similares aos vetores $\overrightarrow{\text{ela}}$ e $\overrightarrow{\text{ele}}$ e, com uniformização, a diferença entre similaridades é dada por:

$$\text{diferença}_{\text{pares}} = s_f - s_m \,. \tag{2}$$

Ou seja, profissões mais similares ao pronome "ela" apresentam os maiores resultados positivos e profissões mais similares ao pronome "ele", os maiores resultados negativos.

2.2.2. Analogias

O cálculo das melhores analogias verifica quais são os "melhores"pares de palavras paralelos ao valor da direção do gênero. Isto é, o menor valor do cosseno das diferenças entre os pares de gênero e os pares das palavras a partir de um limite pré-definido. Isto é feito para achar a distância mínima entre os pares e o valor da direção, mantendo também essa mesma distância entre as palavras correspondentes para preservar a coerência semântica. Considere (a, b) sendo (ela, ele), $(a - b)$ a direção do gênero e $(x - y)$ a diferença entre duas palavras quaisquer do conjunto, a fórmula utilizada é:

$$S_{(a,b)}(x, y) = \begin{cases} \cos(\vec{a} - \vec{b}, \vec{x} - \vec{y}) \,, & \text{se } ||\vec{x} - \vec{y}|| \leq \delta \\ 0 \,, & \text{caso contrário} \,. \end{cases} \tag{3}$$

Dado uma direção e o tamanho do subconjunto, o resultado para cada analogia é a primeira maior pontuação positiva de $S_{(a,b)}$. O cálculo exclui analogias em que uma palavra é usada mais de uma vez. Com isso, conseguimos calcular as n-melhores analogias em um subconjunto de palavras de diferentes tamanhos.

3. Resultados

Os resultados foram divididos em duas categorias. Primeiro foi conduzida a análise de similaridades entre os pronomes "ela" e "ele" com os seus respectivos pares de profissões. Em seguida, estes resultados foram organizados em grupos de profissões para identificação de padrões e correlações. Por fim, foi analisado as 500 melhores analogias do modelo agrupado por cinco conjuntos de tamanhos diferentes e discussões foram levantadas a respeito das diferenças e semelhanças entre os grupos.

3.1. Similaridades

A Tabela 1 apresenta as 18 ocupações mais próximas à representação do pronome "ela" e a Tabela 2, os 18 grupos mais próximos à representação do pronome "ele".

Melhores similaridades entre profissões e “ela”			
Profissão	Grupo/Profissão	Profissão	Grupo/Profissão
1 Enfermeira	Enfermeira	10 Taquígrafa	Apoio administrativo
2 Bailarina	Espetáculo e artes	11 Psicóloga	Ciências Sociais
3 Babá	Serviços domésticos	12 Estatística	Matemáticos e estatísticos
4 Fonoaudióloga	Métodos pedagógicos	13 Nutricionista	Nutricionista
5 Juíza	Advogados e juristas	14 Recepcionista	Balconistas
6 Maquiadora	Tratamento de beleza	15 Vidraceira	Construção Civil
7 Química	Físicos e químicos	16 Fisioterapeuta	Fisioterapeuta
8 Dermatologista	Dermatologista	17 Repórter	Jornalistas
9 Atriz	Técnicos em audiovisual	18 Farmacêutica	Farmacêutica

Tabela 1. Maiores similaridades de cada grupo/profissão baseadas nas diferenças entre as similaridades de ocupações com pronomes “ela-ele”.

Melhores similaridades entre profissões e “ele”			
Profissão	Grupo/Profissão	Profissão	Grupo/Profissão
1 Soldado	Forças armadas, policiais militares	10 Geógrafo	Ciências Sociais
2 Carpinteiro	Indústria da madeira	11 Corretor	Negócios
3 Marinheiro	Operadores de máquinas	12 Senador	Senador
4 Piloto	Piloto	13 Dirigente	Diretores e gerentes
5 Bombeiro	Bombeiro	14 Prefeito	Prefeito
6 Encanador	Construção Civil	15 Vereador	Vereador
7 Serralheiro	Usinagem de metais	16 Programador	Programador
8 Garimpeiro	Agropecuária, florestais, caça e pesca	17 Limpador	Serviços domésticos
9 Procurador	Advogados e juristas	18 Sonoplasta	Técnicos em audiovisual

Tabela 2. Maiores similaridades de cada grupo/profissão baseadas nas diferenças entre as similaridades de ocupações com pronomes “ela-ele”.

Nota-se que a palavra “Enfermeira” teve a melhor similaridade com o pronome “ela”, seguido de “Bailarina” e “Babá” em seus respectivos grupos. Neste trabalho, “Dona de casa” não fez parte do conjunto de ocupações analisadas. No artigo de [Bolukbasi et al. 2016], as profissões “Dona de casa”, “Enfermeira”, “Recepcionista” e “Babá” aparecem em 1º, 2º, 3º e 7º respectivamente. Para as similaridades com o pronome “ele”, profissões semelhantes aos resultados originais aparecem na Tabela 2. As profissões “Piloto” e “Dirigente” são similares, respectivamente, às profissões “Piloto de caça” e “Chefe”, classificadas na 11º e 12º posição no artigo original. Também foi possível verificar que para a profissão de tecnologia da informação como “Programador” e “Programadora”, a primeira aparece em 16º mais similar ao pronome “ele”, enquanto a segunda não aparece na tabela ficando na 31º posição.

Pode-se também identificar quais são os grupos ocupacionais (e não uma profissão isolada) mais similares aos termos femininos e masculinos. Para isso, obtêm-se

as médias das diferenças de similaridade dentro de cada grupo ocupacional. A Figura 1 apresenta um resumo dessa similaridade por grupos. No eixo horizontal dessa figura, há informações sobre a concentração de mulheres em cada grupo de acordo com dados do Instituto Brasileiro de Geografia e Estatísticas (IBGE) e de outros institutos de pesquisa[2]. Os dois primeiros grupos mais similares aos termos femininos (1.a) apresentam alta proporção de mulheres no mercado de trabalho nacional. O terceiro grupo apresenta uma proporção um pouco acima de 50%. Os três grupos mais similares aos termos masculinos (1.b) são grupos com alta proporção de homens em todos os casos. É possível notar ainda que para os grupos em que a diferença de similaridade entre os gêneros é pequena (1.c), a proporção das mulheres nessas áreas de atuação também é equilibrada (40-60%), com uma leve inclinação para o lado feminino nos valores de similaridade. Tais relações podem demonstrar a absorção de estereótipos do mundo real pelos modelos.

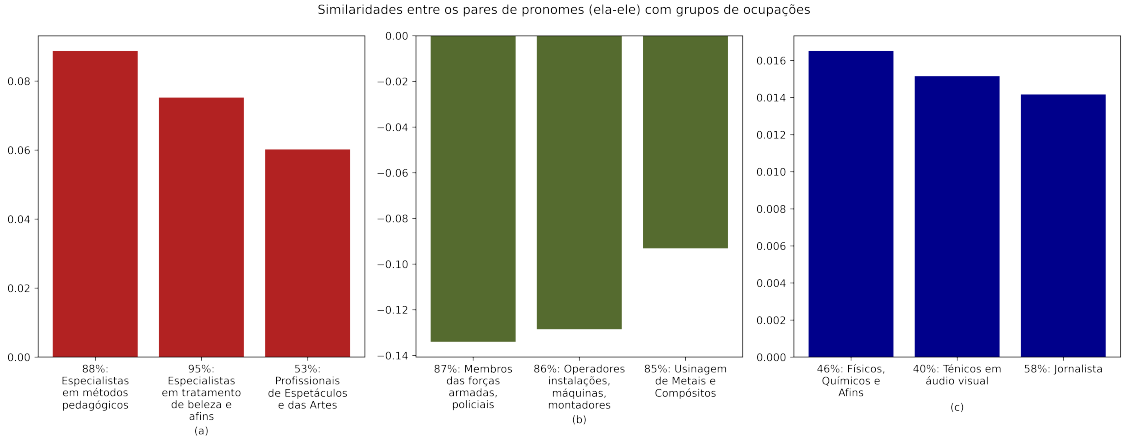

Figura 1. (a) Os três grupos ocupacionais mais similares aos termos femininos, (b) masculinos e (c) grupos sem similaridade significativa entre os termos. Em (a) e (c), a proporção de mulheres no mercado de trabalho está descrita à frente de cada grupo e em (b) a proporção de homens. O eixo y representa a média das diferenças de similaridades para cada grupo.

3.2. Analogias

As melhores analogias consideram apenas a direção do gênero em um subconjunto de palavras. Neste experimento, foram analisadas as 500 melhores analogias encontradas em conjuntos de 10, 20, 30, 40 e 50 mil palavras. A Tabela 3 apresenta as 10 melhores analogias que se mostraram constantes em todos os conjuntos de palavras e que captam a semântica dos pares corretamente. A analogia *mulher-homem* aparece como segunda melhor na listagem. De fato, alguns artigos utilizam este par de gênero como referência na análise de similaridade. A Tabela 4 mostra, para cada conjunto de palavras (10K a 50K), analogias que não capturaram corretamente o par morfológico e/ou semântico e que demonstram supostas características de estereótipo de gênero[3]. Analogias corretas aparecem em conjuntos de diferentes tamanhos. Por outro lado, analogias incorretas têm a similaridade reduzida conforme o conjunto de palavras aumenta, corroborando o que [Hartmann et al. 2017] diz a respeito do aumento de desempenho à medida

[2] https://github.com/nandayot/gender-bias-portuguese-replication-bolukbasi

[3] A classificação de estereótipo de gênero é subjetiva e aqui está apresentada na perspectiva dos autores.

que se aumenta o volume de dados no *corpus*. É importante também destacar a presença de xingamentos de conotação sexual dentro dos conjuntos de melhores analogias e que estão presentes nos 3 maiores conjuntos, o que pode sinalizar a presença de palavras de xingamentos perto do subespaço do vetor "ela".

10 melhores analogias ("ela-ele")	
1 ela-ele	2 mulher-homem
3 sozinha-sozinho	4 escolhida-escolhido
5 amiga-amigo	6 conhecida-conhecido
7 internada-internado	8 garota-garoto
9 nascida-nascido	10 convidada-convidado

Tabela 3. Melhores analogias usando a métrica $S_{(a,b)}(x, y)$ com a direção do gênero considerando pares $(\overrightarrow{\text{ela}}, \overrightarrow{\text{ele}})$. As melhores analogias permanecem constantes em todos os conjuntos analisados.

Analogias incorretas ("ela-ele")				
10K	20K	30K	40K	50K
117: grávida-doente	82: chateada-irritado	253: marido-companheiro	250: compositora-instrumentista	262: compositora-instrumentista
119: marido-companheiro	202: marido-companheiro	299: vadia-porra	270: marido-companheiro	282: marido-companheiro
223: feminina-juvenil	236: vadia-porra	312: mamãe-vovô	325: vadia-porra	344: vadia-porra
261: voleibol-futebol	248: mamãe-vovô	356: blusa-boné	340: mamãe-vovô	412: blusa-boné
262: canção-rock	307: grávida-doente	386: grávida-doente	387: blusa-boné	445: grávida-doente
275: jornalista-historiador	380: estreante-veterano	459: psicóloga-psiquiatra	419: grávida-doente	
288: seio-membros	397: feminina-juvenil	471: pianista-saxofonista	494: psicóloga-psiquiatra	
322: dançar-tocar	468: voleibol-futebol	480: estreante-veterano		
330: gravidez-aborto	469: canção-rock	499: feminina-juvenil		
379: cantando-tocando	491: jornalista-historiador			
413: vocalista-guitarrista				

Tabela 4. Analogias que produziram pares incorretos em todos grupos analisados, com estereótipos de gênero. Colunas 10K a 50K mostram a quantidade de palavras usadas no cálculo das analogias. Os números em cada analogia denotam a posição da analogia entre as 500 melhores.

Entender a origem de cada analogia incorreta foge do escopo deste trabalho por envolver questões sociais e linguísticas. No entanto, analisar as relações entre palavras no WE pode trazer luz ao entendimento. Por exemplo, a analogia "grávida-doente" está presente em todos os conjuntos e provoca questionamentos, já que expressa gravidez

como doença, assim como a analogia “gravidez-aborto” no primeiro conjunto de palavras. Usando a similaridade por cosseno foi possível extrair as 20 palavras mais similares à “gravidez” (Tabela 5). Note que “gravidez” é permeada por palavras que caracterizam problemas de saúde que podem ser contraídos durante a gestação. Isto pode indicar o fato da palavra “doente” estar tanto próxima com “ela-gravidez” assim como com “ele”.

20 palavras mais similares à “gravidez”					
1	gestação	2	parto	3	aborto
4	grávida	5	feto	6	complicações
7	amamentação	8	grávidas	9	infecção
10	doença	11	útero	12	bebê
13	microcefalia	14	maternidade	15	mães
16	zika	17	gestante	18	sexual
19	interrupção	20	abortos		

Tabela 5. As 20 palavras mais similares à palavra ”gravidez”.

Os resultados obtidos neste experimento mostraram em detalhes a presença de diversas analogias com viés de gênero e um *ranking* de similaridade de profissões que não foram mostradas por [Santana et al. 2018]. Em comparação com o trabalho original de [Bolukbasi et al. 2016], foi possível identificar a presença de duas analogias idênticas: *vocalista-guitarrista*, que aparece na 413º posição e *voleibol-futebol* presente nos conjuntos 10K e 20K nas posições 261º e 468º, respectivamente.

4. Impactos da discriminação de gênero na sociedade

A Convenção sobre a eliminação de todas as formas de discriminação contra as mulheres, aprovada e adotada pela Resolução nº 34/180 da Assembleia Geral das Nações Unidas em 18 de dezembro de 1979 e ratificada pelo Brasil em 1º de fevereiro de 1984, define os diversos tipos de discriminação sofridos pelas mulheres e estabelece estratégias para sua mitigação pelos Estados-partes dessa Convenção. Apesar disso, é possível verificar que os avanços no cumprimento dessas medidas são ainda modestos e tímidos, quando existem. No Brasil, os últimos anos têm apresentado dados preocupantes de aumento de casos de feminicídio, violência de gênero, discriminação e preconceito, segundo o Monitor da Violência[4]. Por exemplo, o país bateu o recorde de assassinatos de mulheres em 2022, com uma mulher morta a cada 6 horas.

[Garg et al. 2018] mostraram como a dinâmica temporal de WE ajuda a quantificar as mudanças nos estereótipos e atitudes em relação às mulheres e minorias étnicas nos séculos 20 e 21 nos Estados Unidos, com a integração de WE treinadas em 100 anos de textos a partir do censo dos EUA. Os autores mostraram que as mudanças refletidas no modelo são muito similares às mudanças demográficas e de ocupação do país ao longo do tempo. WE captaram as mudanças sociais tais como o movimento das mulheres na década de 1960 e também mostraram como adjetivos e ocupações específicas tornaram-se mais associados a certas populações ao longo do tempo.

Dados do IBGE mostram que as mulheres trabalham, em média, três horas por semana a mais do que os homens, combinando trabalhos remunerados, afazeres domésticos

[4] http://bit.ly/3JrAJKJ

e cuidados de pessoas. Mesmo assim, e ainda contando com um nível educacional mais alto, elas ganham em média 76,5% do rendimento dos homens, ou seja, as mulheres estudam mais, trabalham mais e ganham menos que os homens[5].

Este trabalho mostra tendências discriminatórias em modelos de WE, apresentando vieses de gênero no domínio de profissões encontradas em português, revelando a existência de estereótipos de gênero em ocupações tradicionais e a sua correlação com a proporção de mulheres no mercado de trabalho nacional (Tabelas 1 e 2). Adicionalmente, também foram verificadas analogias entre pronomes femininos e masculinos e a existência de preconceitos (Tabelas 3 e 4) e analogias associadas a doenças quando a palavra "gravidez" é usada (Tabela 5). Esses resultados evidenciam sexismos semelhantes aos descritos nos estudos que são referências para este e levantam discussões sobre o impacto do uso de modelos de linguagem (LM) em nossa sociedade. Os estudos de analogias em conjuntos de diferentes tamanhos mostrados aqui e os resultados obtidos por [Hartmann et al. 2017] evidenciam a relação entre um bom desempenho de modelos de WE com o tamanho dos dados textuais para o treinamento. Isto também se aplica para os LM, usados para predição de *tokens* através de seu contexto, como BERT e GPT-3. Esses modelos pré-treinados estão sendo amplamente usados não apenas por pesquisadores da área, mas ampliaram-se para uso em massa pela população.

A criação e a atualização de grandes LM através da extração de dados da Internet sem a devida atenção para seu conteúdo abre questionamentos sobre os limites de tamanho que LM podem ou devem ter. [Bender et al. 2021] abre horizontes sobre os danos ambientais, sociais, econômicos e culturais que a criação e atualização de LM, utilizando enormes conjuntos de treinamento, podem causar para grupos marginalizados. Os autores discutem sobre a escolha e a forma que dados textuais retirados na Internet são inseridos no modelo, visto que, a geração de dados na Web não é diversa e global, mas sim hegemônica e limitada (principalmente gerada a partir de textos de jovens estadunidenses), tendo como base um conjunto não-diversificado para treinamentos de modelos.

Mesmo que o desempenho de modelos de WE e LM seja proporcional ao aumento de dados textuais, [Bender et al. 2021] afirmam que a distribuição dos custos e benefícios dos resultados desta tecnologia não se revertem para as mesmas pessoas. Isto é, a geração de textos sexistas, e outros tipos de discriminação e ideologias violentas, e a implantação em massa de LM e modelos de WE em sistemas de classificação, podem ocasionar danos alocativos, já que a representação desses vieses impactam nas suas decisões. Assim, os autores discutem soluções contra-hegemônicas que precisam acontecer para que sistemas de linguagem quebrem a retroalimentação de vieses gerados, espalhados pela Web e reinseridos em novos treinamentos, com o princípio metodológico de anotação dos conjuntos de treinamento e curadoria criteriosa da extração de dados da Internet como forma de prevenção de possíveis danos que modelos de LM e WE enviesados podem causar.

5. Ameaças à validação

As métricas de [Bolukbasi et al. 2016], apesar de serem muito utilizadas em trabalhos da literatura, também já foram criticadas. [Zhang et al. 2020] discutem a confiabilidade e robustez de análises que utilizam pares de gênero. O trabalho argumenta que essas análises não necessariamente refletem o viés social e que o tipo de analogia proposto por

[5] http://bit.ly/40jZBLz

[Bolukbasi et al. 2016] poderia apenas apontar uma alta similaridade por cosseno entre os pares. Os autores propõem testes para verificar se as métricas de identificação de vieses são robustas. Entre os testes estão: 1) estabilidade entre pares: a métrica se demonstra confiável caso o valor do viés não mude significativamente entre os pares de gênero (como [“ela”, “ele”] e [“mulher”, “homem”]); 2) estabilidade morfológica: o viés social entre palavras morfológica e semanticamente similares (como plurais de palavras) devem se manter constante; 3) correspondência linguística: explora se as métricas conseguem predizer os termos de gênero de palavras corretamente (exemplo: leão–leoa).

Outras discussões são feitas a respeito das análises de discriminação algorítmica dentro da área de PLN. [Blodgett et al. 2020] apresentam três recomendações que devem ser consideradas na análise de discriminação em PLN. Primeiro, deve-se analisar os estudos de discriminação fora da área, envolvendo estudos sociolinguísticos e antropológicos, que exploram as relações hierárquicas com a linguagem. Segundo, ser claro sobre como o viés pode ser prejudicial e para quem. Por fim, a terceira recomendação é analisar o uso prático desses sistemas e entender os impactos que podem causar para comunidades afetadas, as relações de poder que existem entre a tecnologia e a sociedade, além de não somente focar em suas falhas tecnológicas. Entendemos que parte dessas recomendações foram cumpridas na Seção 4.

Existem poucos estudos sobre novas métricas e alternativas para contrapor as críticas apresentadas. Dessa forma, as métricas usadas nesta análise podem ser consideradas uma das principais ao se analisar discriminação de gênero, o qual ainda é um conceito em construção dentro da área de PLN, como [Blodgett et al. 2020] constatou. De todo modo, as recomendações elencadas por [Zhang et al. 2020] e [Blodgett et al. 2020] nos experimentos deste trabalho fazem parte dos planos para trabalhos futuros.

6. Conclusões

Este artigo revisou estudos sobre a discriminação algorítmica em aplicações de inteligência artificial, dando ênfase à inexistência da neutralidade tecnológica nas decisões automatizadas. Constatou-se que o sexismo algorítmico é extensamente estudado na área de Processamento de Linguagem Natural (PLN) e que há muitos estereótipos e sexismo em modelos de Word Embeddings (WE). Apesar disso, pesquisas com a língua portuguesa são escassas e este trabalho pretende jogar luz sobre o tema dentro do cenário brasileiro.

Por esse motivo, este trabalho também contribuiu com a análise de um modelo de WE em português utilizando métricas descritas em [Bolukbasi et al. 2016] que foram adaptadas para a língua portuguesa[6]. Assim como ocorrido com modelos em inglês, estereótipos de gênero também foram identificados em modelos de WE em português. Esses modelos conseguem, inclusive, captar estereótipos do mercado de trabalho brasileiro, confirmando a presença de vieses históricos nos *corpora* usados nos modelos. Estas confirmações evidenciam a possibilidade de replicação de métricas de outro idioma para o português, trazendo discussões sobre a presença de discriminação de gênero em dados textuais usados como *corpora* de treinamento de modelos de PLN e, consequentemente, impactando nas aplicações desses modelos.

[6] O código de replicação e os resultados completos estão disponíveis publicamente em: https://github.com/nandayot/gender-bias-portuguese-replication-bolukbasi

Possíveis soluções para a mitigação da estereotipação em WE devem ser pensadas para trabalhos futuros. Também propomos analisar outros tipos de discriminação e traçar a interseccionalidade entre eles.

Alguns estudiosos alegam que métricas de identificação de viés podem ser manipuladas para produzir um efeito desejado. Além disso, análises de viés com utilização de pares de gênero podem apresentar discordância com pares diferentes sintaticamente, mas semântica e morfologicamente iguais. Entretanto, as métricas utilizadas neste trabalho seguem como referências nas análises de viés de gênero atualmente dentro do contexto de Word Embeddings.

Ainda assim, outras análises podem ser realizadas e outras abordagens podem ser construídas dentro da área para além das métricas de estudo. Foi visto que a interdisciplinaridade deve ser utilizada para abrir o escopo sobre como o sexismo algorítmico em PLN pode ser entendido por meio de estudos da Sociolinguística e Ciências Sociais. Além disso, o estudo sobre como aplicações de PLN impactam as comunidades que as utilizam deve ser essencial para os objetivos de pesquisa.

Referências

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proc. of ACM-FAccT*, page 610–623.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proc. of NIPS*, pages 4349–4357.

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., and Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proc. of AAAI/ACM AEIS*, pages 156–170.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Falcão, C. (2021). Lentes racistas: Rui Costa está transformando a Bahia em um laboratório de vigilância com reconhecimento facial. https://interc.pt/3nKVrw9. [Online: acesso em 20-09-2021].

Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Appl Soft Comput*, 101:107057.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, 115(16):E3635–E3644.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Noble, S. and Damorim, F. (2022). *Algoritmos da Opressão: Como os mecanismos de busca reforçam o racismo*. Editora Rua do Sabão.

Santana, B. S., Woloszyn, V., and Wives, L. K. (2018). Is there gender bias and stereotype in portuguese word embeddings? *arXiv preprint arXiv:1810.04528*.

Silva, R. M., Santos, R. L., Almeida, T. A., and Pardo, T. A. (2020). Towards automatically filtering fake news in portuguese. *Expert Syst Appl*, 146:113199.

Silva, T. (2022). Linha do tempo do racismo algorítmico. http://bit.ly/3yFFrzw. [Online: acesso em 04-05-2022].

Sogancioglu, G., Mijsters, F., van Uden, A., and Peperzak, J. (2022). Gender bias in (non)-contextual clinical word embeddings for stereotypical medical categories. *arXiv preprint arXiv:2208.01341*.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proc. of EAAMO*, volume 17, pages 1–9.

Werneck, A. (2019). Reconhecimento facial falha em segundo dia, e mulher inocente é confundida com criminosa já presa. http://bit.ly/3mSoNKy. [Online: acesso em 10-11-2021].

Yee, K., Tantipongpipat, U., and Mishra, S. (2021). Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proc. of HCI*, 5:1–24.

Zhang, H., Sneyd, A., and Stevenson, M. (2020). Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. *arXiv preprint arXiv:2010.02847*.