Sexismo no Brasil: análise de um *Word Embedding* por meio de testes baseados em associação implícita

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

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

[2]Institute of Information Systems, Leuphana University, Lüneburg, NS, Germany

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

Abstract. *This work reports experiments based on the Psychology Implicit Association Test to identify and quantify biases in a Word Embeding (WE) of the Portuguese language. For this, we use a GloVe model trained on an Internet corpus collection. The results show that several common sense and gender stereotypes can be found in WE. Within the context of professions, we note a historical sexism, since the identified bias often reflects the statistics of gender performance in occupation groups in Brazil. The results show discrimination similar to those of international studies and allow discussing the impact of the use of language models in our society.*

Resumo. *Este artigo relata experimentos baseados no Teste de Associação Implícita da Psicologia para identificar e quantificar vieses em um Word Embeding (WE) de língua portuguesa. Para isso, usamos um modelo GloVe treinado em uma coleção de corpus da Internet. Os resultados mostram que diversos estereótipos de senso comum e de gênero podem ser encontrados no WE. Dentro do contexto de profissões, notamos um sexismo histórico, pois frequentemente o viés identificado reflete as estatísticas da atuação dos gêneros em grupos de ocupação do Brasil. Os resultados evidenciam discriminações semelhantes aos de estudos internacionais e permitem discutir sobre o impacto do uso de modelos de linguagem em nossa sociedade.*

1. Introdução

A discriminação de gênero, também chamada sexismo ou viés de gênero, é extensamente analisada na área de Processamento de Linguagem Natural (PLN). [Sun et al. 2019] classificaram os tipos de discriminação de gênero no PLN em quatro categorias, sendo: (a) difamação ou uso cultural ou histórico de termos depreciativos; (b) estereotipagem, que intensifica estereótipos sociais já existentes; (c) reconhecimento, que se refere à desproporção ou imprecisão de certo algoritmo em alguma tarefa de reconhecimento e (d) sub-representação, que define a baixa representação de certos grupos. Todas as quatro categorias são encontradas em modelos de *Word Embeddings* (WE). Exemplos de reconhecimento podem ser encontrados em máquinas de tradução [Tatman 2017, Prates et al. 2020].

[Prates et al. 2020] mostraram que o Google Tradutor exibe uma tendência a traduzir frases como "X é engenheiro", onde "X" é um pronome neutro de idiomas que

não possuem flexão de gênero como o Húngaro, para o pronome masculino no inglês. Traduções desse tipo tornam-se mais frequentes quando nas sentenças são utilizadas profissões com sub-representação de mulheres na sociedade. São os casos das áreas de Ciência, Tecnologia, Engenharia e Matemática (do inglês *STEM*). Similarmente, essa desproporcionalidade de representação pode ser vista em modelos de linguagem onde a probabilidade condicional para profissões consideradas femininas ou masculinas são maiores para seus respectivos pronomes, amplificando padrões sexistas.

Ademais, [Suresh and Guttag 2021] demonstraram que modelos de WE podem apresentar viés histórico, o qual 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. Em [Caliskan et al. 2017], os autores mostraram que modelos de WE conseguem captar relações implícitas de gênero, assim como ocorre nas respostas de participantes do Teste de Associação Implícita (IAT, do inglês) da Psicologia [Greenwald et al. 1998].

O teste IAT segue um paradigma de tempo de reação no qual os participantes são encorajados a classificar palavras rapidamente e o tempo de resposta observado quantifica a saída do teste. A reprodução do IAT com WE e PLN usa os mesmos atributos e palavras-alvo do trabalho original. No entanto, em vez de usar o tempo de resposta para a associação de palavras, usam a similaridade entre os vetores que as representam.

Além dos trabalhos que avaliam sexismo na língua inglesa usando PLN e modelos de WE, diversos trabalhos relacionados têm sido propostos em outras línguas tais como a chinesa [Chen et al. 2022, Li et al. 2022, Jiang et al. 2023, Qin et al. 2023], espanhola [Torres Berrú et al. 2023], alemã [Wagner and Zarrieß 2022], filipina [Gamboa and Justina Estuar 2023], línguas africanas [Wairagala et al. 2022] e línguas indígenas [Hansal et al. 2022].

Carecem estudos sobre sexismo na língua portuguesa, mesmo com a disponibilização de diversos modelos de WE por [Hartmann et al. 2017] ainda em 2017. Assim, inspirado no trabalho de [Caliskan et al. 2017] e dando continuidade ao trabalho iniciado em [Taso et al. 2023], este artigo tem como objetivo verificar a existência de vieses de gênero por estereotipagem e sub-representação, utilizando metodologias similares às que foram empregadas em [Caliskan et al. 2017, Greenwald et al. 1998], mas utilizando um dos modelos de WE criados por [Hartmann et al. 2017]. Deve-se ainda analisar a relação dos vieses sexistas encontrados em profissões tradicionais com a proporção de mulheres no mercado de trabalho nacional.

Como resultado, este trabalho valida uma metodologia de identificação de vieses para WE da língua portuguesa, apresentando associações estereotipadas e, no caso da área profissional, sua relação com dados do mundo real. Adicionalmente, também abre caminho para o uso da metodologia no diagnóstico de outros tipos de discriminação.

2. Testes de associação implícita

O Teste de Associação Implícita (IAT) é um instrumento da Psicologia utilizado para quantificar o posicionamento de pessoas de maneira indireta, tal como o nome sugere. O seu uso é indicado em pesquisas onde os participantes não devem ou não querem expressar suas opiniões, mas as evidenciam ao associarem em tempos muito distintos pares de conceitos que consideram similares ou opostos [Greenwald et al. 1998].

Os criadores do IAT usaram o teste para comprovar alguns vieses humanos. Durante os experimentos, os participantes deviam associar dois conceitos, tais como flores e insetos, a um atributo, tal como agradável. Notaram-se tempos de resposta menores quando os envolvidos tiveram que classificar, por exemplo, um tipo de flor como agradável e um tipo de inseto como desagradável do que quando foram solicitados a fazer a classificação com os objetos trocados (flores como desagradáveis e insetos como agradáveis). O fato de um emparelhamento ser mais rápido indicaria que as partes envolvidas estão relacionadas no cognitivo dos indivíduos. Essa premissa motivou o uso do IAT para identificar e quantificar preconceitos étnicos e de estereótipo [Kiefer and Sekaquaptewa 2007, Nosek BA 2002].

2.1. WEAT

Word Embedding Association Test (WEAT) é um método variante do IAT, proposto para o diagnóstico de vieses em WE [Caliskan et al. 2017]. O WEAT assume que a similaridade por cosseno, métrica frequentemente utilizada para medir a semelhança semântica entre palavras representadas no espaço vetorial, é análoga ao tempo de reação do IAT, isto é, quanto menor o tempo de decisão, maior a proximidade semântica.

Em cada teste WEAT há dois conjuntos de palavras-alvo e dois conjuntos de atributos. Verifica-se então se o primeiro conjunto de palavras-alvo está mais associado ao primeiro conjunto de atributos e se o segundo conjunto de palavras-alvo está mais associado ao segundo conjunto de atributos. A hipótese nula é que não existe diferença de similaridade entre os conjuntos e seus respectivos atributos. O valor-p é utilizado para testar esta hipótese por meio do teste de permutação, e verificar a possibilidade de rejeitar a hipótese nula, ou seja, quanto menor o valor-p, maior a chance de rejeição. O valor-p de 10^{-2} foi mantido para rejeitar a hipótese assim como sugerido por [Caliskan et al. 2017].

Mais formalmente, considere o conjunto $W = X \cup Y$, onde X e Y são conjuntos-alvo de uma associação. Considere A e B seus respectivos conjuntos de atributos. A diferença entre as médias (μ) de similaridade entre uma palavra-alvo w, onde $w \in W$, e os conjuntos de atributos A e B é dada pela Equação 1:

$$s(w, A, B) = \mu_{a \in A} \{\cos(\vec{w}, \vec{a})\} - \mu_{b \in B} \{\cos(\vec{w}, \vec{b})\}. \tag{1}$$

Para exemplificar a Equação 1, considere $X = \{\text{rosa, azaléia, orquídea}\}$ o conjunto de palavras de flores, $Y = \{\text{formiga, pulga, mosca}\}$ o conjunto de insetos, $A = \{\text{paz, paraíso, arco-íris}\}$ o conjunto de palavras que denotam o conceito de agradável e $B = \{\text{fedor, veneno, agonia}\}$ o conjunto de palavras que denotam o conceito desagradável. A equação para $w = \text{rosa}$ pode ser lida da seguinte forma: $s(w, A, B)$ representa a média do cosseno entre $\vec{\text{rosa}}$ e todos os conceitos agradáveis de A menos a média do cosseno entre rosa e todos os conceitos desagradáveis de B. A mesma lógica pode ser utilizada para $w \in Y$.

O *effect size* d, medida que determina o tamanho da significância entre os conjuntos-alvos e seus respectivos atributos, é dado pela Equação 2, onde σ denota o desvio padrão:

$$d = \frac{\mu_{x \in X} \{s(x, A, B)\} - \mu_{y \in Y} \{s(y, A, B)\}}{\sigma_{w \in X \cup Y} \{s(w, A, B)\}}. \tag{2}$$

A medida da associação diferencial de dois conjuntos de palavras-alvo com os atributos é dada pela Equação 3:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)\,. \tag{3}$$

Por fim, o valor-p, medida estatística para determinar a probabilidade da significância do valor d para os conjuntos e seus atributos é dado pela Equação 4. Nela, os valores $\{(X_i, Y_i)\}_i$ representam a união dos conjuntos X e Y distribuídos aleatoriamente. A equação retorna o valor-p aproximado utilizando dez mil iterações com função de distribuição normal:

$$p = \mathrm{Pr}_i[s(X_i, Y_i, A, B) > s(X, Y, A, B)]\,. \tag{4}$$

No estudo de [Caliskan et al. 2017] são usados os mesmos conjuntos de palavras adotados em [Greenwald et al. 1998] para a análise de um modelo GloVe treinado em textos encontrados na rede mundial de computadores. Naquele trabalho, todos os vieses linguísticos descritos no artigo original são também identificados, incluindo preconceitos raciais e de gênero. Por esse motivo, os proponentes do WEAT argumentaram que as associações nos vetores de representação dos WE não poderiam existir por casualidade e que seriam reflexo da perspectiva cultural da população.

A relevância acadêmica do trabalho de [Caliskan et al. 2017], a boa documentação dos experimentos realizados e a consolidação das métricas em trabalhos posteriores tornaram oportuna a validação da metodologia em modelos ainda pouco estudados, como de WE em português.

2.2. WEFAT

Word Embedding Factual Association Test (WEFAT) é um teste de associação implícita também proposto por [Caliskan et al. 2017]. Ele busca extrair informações empíricas sobre o mundo dentro de modelos de WE. Para isso, consideram-se, assim como no WEAT, um conjunto de palavras-alvo W e dois conjuntos de atributos A e B como definidos pela Equação 2.

Observe que o WEAT é usado para verificar se existem diferenças entre conjuntos de palavras-alvo em termos de sua similaridade relativa com conjuntos de atributo. Com o WEFAT, uma propriedade factual que pode ser valorada é associada a cada palavra-alvo e se deseja testar se os vetores correspondentes às palavras-alvo incorporam o conhecimento dessa propriedade, isto é, se é possível extrair ou prever a propriedade dado o vetor. Assim, o valor do WEFAT é utilizado para verificar a correlação com as informações do mundo real que foram informadas.

[Caliskan et al. 2017] utilizaram o WEFAT para mostrar que a representação de profissões em WE de língua inglesa embutem conhecimento sobre a composição da força de trabalho em ocupações nos Estados Unidos, ou seja, os estereótipos de profissão encontrados no WEAT apresentavam alta correlação com a proporção de mulheres atuantes naquela atividade. Neste trabalho, usaremos o teste para o mesmo objetivo dentro do contexto brasileiro e da língua portuguesa.

3. Procedimentos metodológicos

A metodologia seguida neste trabalho é consolidada na literatura no contexto americano e da língua inglesa. Para realizar experimentos no cenário brasileiro, foi preciso escolher um modelo de WE em português que já tivesse sido validado pela comunidade de PLN e contivesse um número suficiente de *tokens* para validar palavras nos mais variados escopos. Além disso, foram necessárias adaptações nas métricas adotadas para contemplar a marcação de gênero nas palavras.

3.1. Escolha do modelo

O modelo GloVe com 300 dimensões, avaliado por [Hartmann et al. 2017], demonstrou ter um bom desempenho em tarefas de analogias de sintaxe e semântica. Ele é formado pela junção de diferentes corpus, em português brasileiro e europeu, tendo mais de 1,2 bilhão de tokens e sendo amplamente utilizado para diversos tipos de aplicações [Grave et al. 2018, Fortuna et al. 2019, Garcia and Berton 2021, Silva et al. 2020]. Diante de seu bom resultado e uso no trabalho de [Caliskan et al. 2017], o modelo GloVe foi escolhido para ser usado na análise de discriminação de gênero em português.

3.2. Associações

Seis associações são analisadas: *Flores vs Insetos*, *Instrumentos vs Armas*, *Carreira vs Família*, *Matemática vs Artes*, *Ciência vs Artes*, e *Atuações Femininas vs Atuações Masculinas*. As duas primeiras fazem parte do grupo de associações universalmente aceitas como agradáveis ou desagradáveis. Elas servem principalmente para validar o WEAT em assuntos neutros, sobre os quais não há nenhuma questão social a ser discutida. As demais associações contêm supostos vieses de gênero e empregam os seguintes grupos de atributos:

- **Termos Femininos**: feminino, mulher, menina, irmã, ela, dela, delas, filha;
- **Termos Masculinos**: masculino, homem, menino, irmão, ele, dele, deles, filho.

Todas as associações, exceto as relacionadas a carreiras e áreas de atuação profissional, foram traduzidas de [Caliskan et al. 2017] para a língua portuguesa.

Para que o WEAT verificasse a existência de viés em profissões, foram construídos dois conjuntos de áreas de atuação. *Atuações Femininas* e *Atuações Masculinas* referem-se, respectivamente, às áreas de atuação com maior e menor proporção de mulheres no mercado de trabalho brasileiro de acordo com pesquisas de órgãos oficiais do Brasil[1]. Esses conjuntos apresentam a seguinte formação[2]:

- **Atuações Femininas**: culinária, artes, educação, psicologia, pedagogia, enfermagem, assistência, estética, limpeza, farmácia, jornalismo, biblioteconomia, gastronomia, comunicação, literatura, sociologia, antropologia, nutrição, fisioterapia, música;
- **Atuações Masculinas**: atletismo, pesca, mecânica, comércio, indústria, agropecuária, ciência, economia, engenharia, física, medicina, diretoria, construção, administração, biologia, polícia, gerência, aviação, computação, direito.

[1] `https://bit.ly/3XjhZnw`, `https://bit.ly/46dBzp5`

[2] Os dados e códigos utilizados nos experimentos estão disponíveis em `https://github.com/nandayot/WEAT-WEFAT`.

3.3. Grupos ocupacionais

Para os experimentos com o WEFAT, foi criado um conjunto com 104 profissões cadastradas na Classificação Brasileira de Ocupações (CBO) agrupadas em 35 categorias ocupacionais de acordo com a similaridade de atuação. Paralelamente, obteve-se a proporção de mulheres atuantes em cada grupo, de acordo com dados do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2018 e outras instituições. Um exemplo de grupo ocupacional é o "Especialistas em métodos pedagógicos" , o qual apresenta 88% de mão de obra feminina. Nele constam profissões tais como pedagoga, psicopedagoga, fonoaudióloga e educadora.

4. Resultados

Os resultados foram divididos em duas categorias, onde se discute primeiramente o teste WEAT e, em seguida, o teste WEFAT.

4.1. WEAT

Na Tabela 1 são apresentados os resultados dos testes WEAT. Os valores obtidos para *d* apontam para vieses em todos os grupos de associação, indicando que os conjuntos de palavras-alvo possuem significativas diferenças de similaridade de acordo com os diferentes grupos de atributos[3]. Portanto, para todos os resultados, o 1º conjunto de palavras-alvo está mais associado ao 1º conjunto de atributos assim como o 2º está mais associado ao 2º conjunto de atributos.

Tabela 1. Resultados do teste WEAT. "A vs D" representam Agradável vs Desagradável e "TM vs TF" Termos Masculinos vs Termos Femininos.

Palavras-alvo	Atributos	Resultado	
		d	p
Flores vs Insetos	A vs D	$0{,}87$	10^{-3}
Instrumentos vs Armas	A vs D	$0{,}91$	10^{-4}
Carreira vs Família	TM vs TF	$1{,}62$	10^{-4}
Matemática vs Artes	TM vs TF	$1{,}38$	10^{-3}
Ciência vs Artes	TM vs TF	$0{,}86$	10^{-2}
Atuações Masculinas vs Atuações Femininas	TM vs TF	$0{,}93$	10^{-3}

Foi possível identificar vieses universais nas associações *Flores vs Insetos* e *Instrumentos vs Armas*. Dessa maneira, o WEAT está validado para relações assumidamente fortes, onde não há necessidade de discussão sobre posicionamentos, e abre espaço para a análise de de associações que evidenciam discriminação de gênero.

Os grupos *Carreira vs Família* e *Matemática vs Artes* foram os que obtiveram os maiores *effect size* (*d*), com valores maiores do que os obtidos para os grupos de vieses universais. Esse resultado demonstra a existência de profundos estereótipos de gênero no campo de ocupações, com uma grande força de associações entre os conceitos e seus atributos. Assim, palavras tais como *executivo* e *carreira* estariam mais associadas a termos masculinos e palavras tais como *casa* e *filhos*, a termos femininos.

[3]Segundo [Caliskan et al. 2017], valores de *d* maiores que 0,8 indicam grande diferença de associação.

Ainda no contexto de profissões, notou-se uma clara associação das *Atuações Masculinas* com termos masculinos e das *Atuações Femininas* com termos femininos. Todos os vieses de gênero identificados colaboram na perpetuação de discriminação, principalmente contra as mulheres que, na sociedade atual, ainda são associadas a papeis relacionados à família, artes ou a profissões estereotipadas que envolvem, em sua maioria, cuidados e educação.

Todos os grupos obtiveram valores-p suficientes para refutar a hipótese de que não há diferença de associações entre os conjuntos-alvos. Dessa maneira, conclui-se que no WE utilizado identificam-se relações inquestionáveis, assim como associações carregadas de estereótipos de gênero. Os resultados vão ao encontro dos relatados em [Caliskan et al. 2017].

4.2. WEFAT

O teste WEFAT foi alterado para satisfazer as particularidades linguísticas da língua portuguesa, na concordância de gênero das palavras. Considere $\overrightarrow{p_f}$ e $\overrightarrow{p_m}$ palavras de profissões com flexão de gênero feminino e masculino, respectivamente (ex: advogada-advogado) e A e B os conjuntos de palavras que denotam os respectivos gêneros (Termos Femininos e Termos Masculinos, respectivamente). A nova fórmula para o cálculo da diferença entre as médias de similaridade entre palavras-alvo e atributos é dada pela Equação 5:

$$s(\overrightarrow{p_f}, \overrightarrow{p_m}, A, B) = \frac{\mu_{a \in A}\{\cos(\overrightarrow{p_f}, \vec{a})\} - \mu_{b \in B}\{\cos(\overrightarrow{p_m}, \vec{b})\}}{\sigma_{x \in A \cup B, w \in F \cup M}\{\cos(\vec{w}, \vec{x})\}}. \tag{5}$$

A Figura 1 ilustra a correlação entre a força de associação média de cada grupo ocupacional com a proporção de mulheres atuantes nele. O coeficiente de Pearson obtido foi de $0,86$.

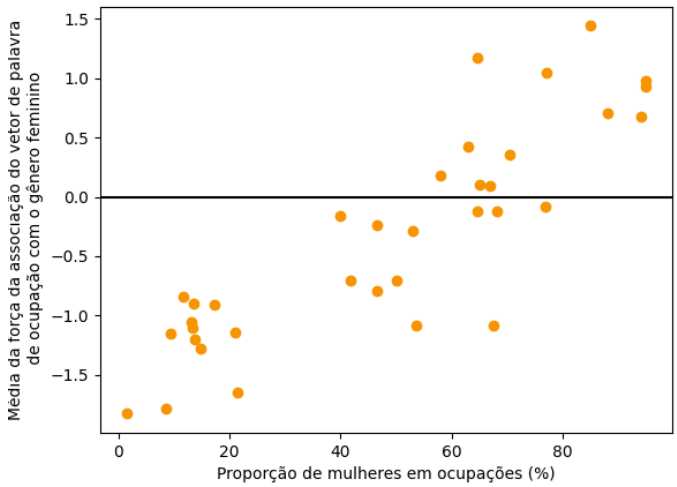

Figura 1. Relação entre a proporção de mulheres em ocupações no mercado de trabalho e a média de valores do teste WEFAT para ocupações com termos femininos e masculinos. Coeficiente de correlação de Pearson $p = 0{,}86$.

Observa-se que grande parte dos valores WEFAT estão abaixo do eixo x, indicando baixa força de associação com os termos femininos. De fato, muitas profissões, apesar de possuírem forte atuação de mulheres, possuem *embeddings* fracamente associados com termos femininos, o que pode indicar baixa co-ocorrência dessas palavras com estes termos dentro do modelo. A análise das frequências de palavras em *embeddings* feita por [Caliskan et al. 2022] corrobora essa hipótese. Os autores quantificaram o viés de gênero em grupos de 100, 1000, 10000 e 100000 mil palavras mais frequentes no modelo de língua inglesa e puderam verificar que 77% das mil palavras mais frequentes estão mais associadas a termos masculinos do que femininos.

Nota-se que a replicação do WEFAT consegue captar características reais do mercado de trabalho feminino no Brasil com um nível significativamente positivo de correlação.

5. Conclusões

Considerando que a Inteligência Artificial compreende a linguagem humana a partir de textos do mundo real, espera-se que os modelos de aprendizado gerados apresentem preconceitos encontrados nas sociedades onde os corpus se originaram. Tratando dessa hipótese, [Caliskan et al. 2017] mostraram que vieses universais, raciais e de gênero existem em WE da língua inglesa treinados em corpus obtidos na Internet. Para isso, os autores validam uma metodologia que utiliza duas novas métricas, WEAT e WEFAT, baseadas no Teste de Associação Implícita da Psicologia.

Neste trabalho, estendemos a proposta de [Caliskan et al. 2017] para o contexto brasileiro. Foi utilizado um modelo de Aprendizado de Máquina puramente estatístico treinado em diversos corpora com textos de páginas da Internet. As análises incluíram associações neutras, tidas como universais, assim como associações de gênero, dentro do contexto de profissões.

Os resultados mostraram a existência dos mesmos vieses humanos identificados em [Caliskan et al. 2017]. Além disso, é possível identificar vieses históricos no ramo de profissões do Brasil. Assim, concluímos que o uso de similaridade por cosseno é uma boa aproximação para a associação implícita de conceitos também na língua portuguesa. Nosso trabalho é a continuação do estudo iniciado em [Taso et al. 2023], sendo ambos pioneiros na detecção de vieses de dados no contexto brasileiro.

Críticas sobre o uso de associações e pares de gênero existem e devem ser levadas em consideração nas análises e discussões, mas reconhecemos que outras alternativas que contrapõem as métricas utilizadas ainda não são unanimidade dentro da área [Ethayarajh et al. 2019, Gonen and Goldberg 2019, Zhang et al. 2020]. Possíveis soluções para a mitigação de esteriótipos em WE devem ser pensadas para os próximos trabalhos. Também propomos analisar outros tipos de modelo, assim como tratar diferentes tipos de discriminação e traçar a interseccionalidade entre eles.

A contribuição deste trabalho extrapola os limites da Computação. Entendemos que a interdisciplinaridade deve ser utilizada para abrir o escopo sobre como o sexismo em PLN pode ser entendido por meio de estudos da Sociolinguística e Ciências Sociais [Blodgett et al. 2020]. Além disso, o estudo sobre como aplicações de PLN impactam as comunidades que as utilizam deve ser essencial para os objetivos de pesquisa.

Agradecimentos

O presente trabalho foi realizado com apoio da Universidade Federal de Mato Grosso do Sul e da Universidade Leuphana de Lüneburg.

Referências

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*.

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 AIES*, 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.

Chen, X., Li, M., Yan, R., Gao, X., and Zhang, X. (2022). Unsupervised mitigating gender bias by character components: A case study of Chinese word embedding. In *Proc. of GeBNLP*, pages 121–128. ACL.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. *arXiv preprint arXiv:1908.06361*.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019). A hierarchically-labeled portuguese hate speech dataset. In *Proc. of AWL*, pages 94–104.

Gamboa, L. C. and Justina Estuar, M. R. (2023). Evaluating gender bias in pre-trained filipino fasttext embeddings. In *Procc. of ITIKD*, pages 1–7.

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.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

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

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *J Pers Soc Psychol*, 74(6):1464–80.

Hansal, O., Le, N. T., and Sadat, F. (2022). Indigenous language revitalization and the dilemma of gender bias. In *Proc. of GeBNLP*, pages 244–254. ACL.

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*.

Jiang, T., Li, Y., Fu, S., and Chen, Y. (2023). Creating a Chinese gender lexicon for detecting gendered wording in job advertisements. *Inform Process Manag*, 60(5):103424.

Kiefer, A. K. and Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. *J Exp Soc Psychol*, 43(5):825–832.

Li, J., Zhu, S., Liu, Y., and Liu, P. (2022). Analysis of gender bias in social perception and judgement using Chinese word embeddings. In *Proc. of GeBNLP*, pages 8–16. ACL.

Nosek BA, Banaji MR, G. A. (2002). Math = male, me = female, therefore math not = me. *J Pers Soc Psychol*, 83(1):44–59.

Prates, M. O., Avelar, P. H., and Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google translate. *Neural Comput Appl*, 32(10):6363–6381.

Qin, C., Zhang, X., Zhou, C., and Liu, Y. (2023). An interactive method for measuring gender bias and evaluating bias in Chinese word embeddings. In Imane, H., editor, *Proc. of CVAA*, volume 12613, page 126130U.

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.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

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.

Taso, F. T., Reis, V. Q., and Martinez, F. V. (2023). Discriminação algorítmica de gênero: Estudo de caso e análise no contexto brasileiro. In *Anais do WICS*, pages 13–25. SBC.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proc. of EthNLP*, pages 53–59. ACL.

Torres Berrú, Y., Batista, V., and Zhingre, L. (2023). A data mining approach to detecting bias and favoritism in public procurement. *Intell Autom Soft Co*, 36(3):3501–3516.

Wagner, J. and Zarrieß, S. (2022). Do gender neutral affixes naturally reduce gender bias in static word embeddings? In *Proc. of KONVENS*, pages 88–97.

Wairagala, E. P., Mukiibi, J., Tusubira, J. F., Babirye, C., Nakatumba-Nabende, J., Katumba, A., and Ssenkungu, I. (2022). Gender bias evaluation in Luganda-English machine translation. In *Proc. of AMTA*, pages 274–286. AMTA.

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*.