Profissões e estereótipos: avaliando o viés de gênero em versões do BERTimbau

Gabriel Kuster de Azevedo[1], **Brenda Salenave Santana**[1]

[1]Centro de Desenvolvimento Tecnológico (CDTec)
Universidade Federal de Pelotas (UFPel) – Pelotas, RS – Brasil

{gkuster,brenda}@inf.ufpel.edu.br

Abstract. *This study investigates the presence of gender bias in the contextual embeddings of the BERTimbau model (BASE and LARGE versions), focusing on professions stereotypically associated with men and women. The results from the WEAT metric and analogy tests indicate a strong gender bias in both models. The findings highlight the risk that these models reinforce stereotypes in real-world applications, highlighting the need for mitigation methods.*

Resumo. *Este estudo investiga a presença de viés de gênero nos embeddings contextuais do modelo BERTimbau (versões BASE e LARGE), com foco em profissões estereotipicamente associadas a homens e mulheres. Os resultados, obtidos através da métrica WEAT e testes de analogia, indicam um forte viés de gênero em ambos os modelos. Os achados apontam para o risco de que esses modelos reforcem estereótipos em aplicações reais, destacando a necessidade de métodos de mitigação.*

1. Introdução

Atualmente o termo Inteligência Artificial vem se popularizado cada vez mais, trazendo novas tecnologias e diferentes aplicações para a comunidade em geral. Entre as áreas que ganharam destaque nesse contexto está o Processamento de Linguagem Natural (PLN) e os grandes modelos de linguagem (do inglês *Large Language Models* – LLMs), com avanços recentes creditados à arquitetura Transformers [Vaswani et al. 2023]. Consequentemente, modelos focados no português, como o BERTimbau [Souza et al. 2020] — treinado em corpus brasileiro a partir do BERT [Devlin et al. 2019] —, e modelos multilíngues como o XLM-RoBERTa [Conneau et al. 2020], ganham cada vez mais espaço em diversas aplicações.

Entretanto, segundo [Yucong Duan et al. 2024], esses modelos podem herdar e amplificar preconceitos humanos presentes no mundo real. Para investigar um potencial impacto do uso desses modelos, este trabalho investiga a presença de vieses de gênero nos modelos BERTimbau (versões LARGE e BASE). Para isso, geramos embeddings contextuais em um conjunto de profissões com estereótipos de gênero e quantificamos o viés utilizando o *Word Embedding Association Test* (WEAT) [Caliskan et al. 2017], que se baseia na similaridade de cosseno para medir a associação entre grupos de palavras.

2. Trabalhos Relacionados

Para embeddings estáticos em português, alguns trabalhos investigam o viés de gênero com abordagens distintas. O estudo de [Santana et al. 2018], por exemplo, evidencia a

severidade do viés ao demonstrar que, antes de qualquer correção, a analogia "homem está para fotógrafo assim como mulher está para stripper" era considerada válida pelos modelos. Para mitigar esse problema, os autores aplicam o algoritmo de mitigação de vieses (*debias*) [Bolukbasi et al. 2016], que busca identificar e remover o subespaço de gênero dos vetores. No entanto, o trabalho destaca que, embora eficaz, a aplicação do *debias* pode levar a uma perda de acurácia em tarefas de PLN subsequentes, revelando um importante *trade-off* entre justiça e desempenho do modelo.

No campo de embeddings estáticos para o português, o trabalho de [Taso et al. 2023] já demonstrava vieses de gênero no GloVe utilizando as métricas WEAT e WEFAT. A análise de viés se estendeu aos embeddings contextuais, predominantemente baseados no modelo BERT. Estudos como [Jentzsch and Turan 2022, Puttick et al. 2024] aplicam a métrica WEAT, enquanto [May et al. 2019] propõe a SEAT, que avalia sentenças em vez de palavras isoladas [Yarrabelly et al. 2024]. Um achado crucial nesses trabalhos é que o viés de gênero pode se intensificar com o aumento do tamanho do modelo, como demonstrado por [Jentzsch and Turan 2022].

Apesar desses avanços, a literatura carece de uma análise similar focada no português brasileiro e, especificamente, em embeddings contextuais (como o modelo BERTimbau). Não foram encontrados estudos que apliquem essas métricas para comparar sistematicamente o viés de gênero entre suas diferentes versões. Esta lacuna motivou o desenvolvimento do presente trabalho.

3. Metodologia

A metodologia escolhida neste trabalho é similar a de [Bolukbasi et al. 2016, Santana et al. 2018], onde seleciona-se uma lista de profissões e a partir dessas são analisadas as analogias sugeridas pelo modelo. É importante destacar que, diferentemente de estudos referidos na Seção 3, neste trabalho empregamos embeddings contextuais, capazes de capturar variações semânticas de acordo com o contexto em que a palavra aparece. Essa característica permite uma análise mais refinada das associações entre gênero e profissão, já que não consideramos apenas uma representação fixa para cada termo, mas sim diferentes ocorrências e usos em sentenças reais. Para investigar o viés de gênero, o modelo BERTimbau foi utilizado para realizar testes de analogia semântica. A análise consistiu em avaliar o vetor resultante de operações como `[(<profissão> + ela) - ele]`, um método clássico para verificar se o modelo associa determinadas profissões a um gênero específico, abordagem também conhecida como analogias extremas. A seleção das profissões seguiu dois critérios principais: (i) Priorizaram-se ocupações comuns no contexto brasileiro, com base na Classificação Brasileira de Ocupações (CBO)[1]; (ii) Foram incluídas profissões já analisadas no estudo de [Santana et al. 2018] com embeddings estáticos, a fim de permitir uma análise comparativa entre os resultados dos diferentes tipos de modelos. Utilizamos a métrica WEAT para avaliar a incidência e a intensidade dos viéses de genêro nos modelos BERTimbau, buscando observar também se este modelo segue o padrão observado em [Jentzsch and Turan 2022].

[1]Listagem de profissões: http://www.mtecbo.gov.br/cbosite/pages/downloads.jsf

4. Resultados

Para avaliar a presença de estereótipos de gênero, foram realizados testes de analogia com as profissões selecionadas. A Tabela 1 ilustra alguns dos resultados observados.

Tabela 1. Analogias extremas nos Modelos Large e Base

	Masculino			**Feminino**		
Modelo	Profissão	Similaridade	Analogia	Profissão	Similaridade	Analogia
LARGE	arquiteto	0.9486	arquiteto	arquiteta	0.7461	arquitetura
	blogueiro	0.8020	blog	blogueira	0.7012	blog
	cantor	0.9433	cantor	cantora	0.9377	cantora
	garçom	0.6840	restaurante	garçonete	0.6665	ana
	historiador	0.9442	historiador	historiadora	0.7177	historia
	pintor	0.9474	pintor	pintora	0.7138	pintura
BASE	arquiteto	0.9190	arquiteto	arquiteta	0.5951	ana
	blogueiro	0.5939	blog	blogueira	0.4943	blog
	cantor	0.9270	cantor	cantora	0.9377	cantora
	garçom	0.6840	gerente	garçonete	0.4393	obrigada
	historiador	0.9193	historiador	historiadora	0.6300	ela
	pintor	0.9271	pintor	pintora	0.6368	ela

A análise das analogias revelou uma disparidade de desempenho entre os gêneros. O modelo BASE acertou 19 analogias de profissões masculinas, mas apenas 7 femininas. O modelo LARGE apresentou padrão similar, com 20 acertos masculinos e 11 femininos, sugerindo viés consistente.

Para quantificar formalmente essa tendência, aplicamos a métrica WEAT. Os resultados confirmaram a presença de um forte viés de gênero associado a profissões em ambos os modelos. O tamanho do efeito (*Effect Size*) foi acentuado tanto para o modelo BASE (d = 1.358) quanto para o LARGE (d = 1.595), com alta significância estatística (p = 0.0010 para ambos). Esses valores indicam uma forte associação implícita entre profissões e estereótipos de gênero nas representações do modelo.

Adicionalmente, foi observado um artefato metodológico com palavras desconhecidas pelo vocabulário do modelo. Nesses casos, como ilustrado na Tabela 1, a operação de analogia frequentemente resultava em um vetor próximo a `'ela'`. Isso ocorre porque, na ausência do vetor da profissão, o resultado da operação é dominado pela direção do próprio vetor de gênero (`ela - ele`).

5. Considerações

Mesmo sendo um trabalho em desenvolvimento, este estudo aponta que os modelos BERTimbau, tanto na versão BASE quanto na LARGE, apresentam um forte viés de gênero associado a profissões. Essa conclusão é sustentada pelos resultados dos testes de analogia e, principalmente, pelos valores obtidos na métrica WEAT, que indicam uma forte associação implícita entre palavras de profissão e de gênero

Esses resultados implicam que, quando aplicado em contextos reais, o modelo pode direcionar suas respostas de forma enviesada, reforçando estereótipos de gênero em tarefas como sistemas de recomendação, assistentes virtuais e ferramentas educacionais.

Isso ressalta a necessidade crítica de avaliações e da aplicação de métodos de mitigação de viés antes de empregar tais tecnologias em sistemas que impactam diretamente as pessoas. Adicionalmente, observou-se uma representação vocabular mais limitada para profissões femininas em ambos os modelos, especialmente na versão BASE. Em trabalhos futuros, propomos a ampliação do conjunto de modelos avaliados e, adicionalmente, a aplicação e avaliação de algoritmos de debias reportados na literatura.

Referências

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Caliskan, A., Bryson, J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Jentzsch, S. and Turan, C. (2022). Gender bias in bert - measuring and analysing biases through sentiment rating in a realistic downstream classification task.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders.

Puttick, A., Rankwiler, L., Ikae, C., and Kurpicz-Briki, M. (2024). The bias detection framework: Bias detection in word embeddings and language models for european languages.

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

Souza, F., Nogueira, R., and Lotufo, R. (2020). *BERTimbau: Pretrained BERT Models for Brazilian Portuguese*.

Taso, F., Reis, V., and Martinez, F. (2023). Sexismo no brasil: análise de um word embedding por meio de testes baseados em associação implícita.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Yarrabelly, N., Damodaran, V., and Su, F.-G. (2024). Mitigating gender bias in contextual word embeddings.

Yucong Duan, Fuliang Tang, Kunguang Wu, Zhendong Guo, Shuaishuai Huang, Yingtian Mei, Yuxing Wang, Zeyu Yang, and Shiming Gong (2024). "The Large Language Model (LLM) Bias Evaluation (Age Bias)- DIKWP Research Group International Standard Evaluation.