GRANDES MODELOS DE LINGUAGEM: UMA ANÁLISE SOBRE SEUS WORD EMBEDDINGS

GABRIEL KUSTER DE AZEVEDO[1]; BRENDA SALENAVE SANTANA[2]

[1] *Universidade Federal de Pelotas – gkuster@inf.ufpel.edu.br*
[2] *Universidade Federal de Pelotas – brenda@inf.ufpel.edu.br*

1. INTRODUÇÃO

Processamento de linguagem natural (do inglês *Natural Language Processing* - NLP) é uma subárea da Inteligencia Artificial (do ingles *Artificial Intelligence* - AI) que habilita maquinas a pensarem e entender com a linguagem humana, essa área introduziu os Grande modelos de linguagem (do inglês *Large Language Models* - LLMs) que são modelos treinados em grandes quantidades de dados para serem capazes de entender e gerar linguagem natural. Recentemente os LLMs vem ganhando um grande espaço nas nossas vidas, em diferentes tarefas, seja para pesquisas, estudos em diferentes áreas, tradução e também para realizar tarefas simples do dia-a-dia (FARIAS et al., 2024). No Brasil esse termo vem sendo utilizado notavelmente nos últimos anos pelo lançamento e popularização dos chatbots, dessa forma surgiram outros avanços no âmbito de LLMs, como, por exemplo, para o português temos o modelo BERTimbau (SOUZA et al., 2020), que é um modelo pré treinado em corpus brasileiros em cima do modelo BERT (DEVLIN et al., 2019) além desses, ainda existem modelos multilíngue que é o caso do XLM-RoBERTa (CONNEAU et al., 2020). Ainda neste contexto, podemos notar grandes avanços em pesquisas e desenvolvimentos nestes modelos de linguagem, porém temos que nos atentar com vieses nesses modelos, segundo DUAN et al. (2024) os LLMs podem herdar e amplificar preconceitos humanos presentes no mundo real.

Diversas pesquisas vêm sendo feitas para mitigar esses vieses presentes dentro desses modelos, como SANTANA et al. (2018) e TASO et al. (2023). Estes trabalhos focaram em uma análise de embeddings estáticos porém com algumas particularidades, por exemplo o trabalho de SANTANA et al. (2018) utiliza a abordagem de analogias extremas, além disso ainda é aplicado um algoritmo de *debias* BOLUKBASI et al., 2016, esse algoritmo busca identificar e remover o subespaço de gênero dos vetores. Do outro lado, o trabalho de TASO et al. (2023) faz o uso da métrica *Word Embedding Evaluation Test* (WEAT) e *Word Embedding Factual Association Test* (WEFAT) (CALISKAN et al., 2017) para analisar os embeddings escolhidos em seu estudo. Sobre embeddings contextuais analisados em modelos brasileiros, os trabalhos ainda são escassos e bastante limitados no sentido de que não existem tantas variedades de modelos para a nossa língua como existem para outras. Baseado nesta motivação nós propomos um estudo nos embeddings contextuais do modelo BERTimbau, utilizando a métrica WEAT em que fizemos uma análise buscando viés de gênero baseado em profissões estereotipadas.

2. METODOLOGIA

A metodologia adotada para este trabalho foi inspirada em pesquisas anteriores que investigaram vieses em representações vetoriais de palavras, como nos trabalhos de BOLUKBASI et al. (2016) e SANTANA et al. (2018). De maneira similar,

realizamos a análise de analogias extremas, explorando combinações que evidenciam possíveis distorções de gênero. Um exemplo deste tipo de operação vetorial é dado por `(<profissão> + ela) - ele`, no qual buscamos verificar como os modelos de linguagem posicionam semanticamente profissões em relação a marcadores de gênero.

Diferentemente de estudos anteriores que utilizaram embeddings estáticos, 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. A seleção das profissões seguiu dois critérios principais. Primeiramente, priorizaram-se ocupações comuns no contexto brasileiro, com base na Classificação Brasileira de Ocupações (CBO)[1] e em estudos como o de ÜNAL et al. (2018). Adicionalmente, 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. Conforme mencionado anteriormente, utilizou-se a expressão matemática apresentada nesta seção. Para viabilizar o cálculo, foram obtidos os embeddings das palavras ele e ela, bem como das profissões consideradas. Em seguida, as relações foram extraídas por meio da similaridade do cosseno, a qual mensura a proximidade entre dois vetores em um espaço multidimensional (STECK et al., 2024).

3. RESULTADOS E DISCUSSÃO

Nos testes que adotamos utilizamos o modelo BERTimbau BASE para gerar os embeddings contextuais utilizados. O algoritmo que utilizamos serviu para verificar a presença de preconceitos de gênero e também dentro dos resultados obtidos analisar se havia algum tipo de estereotipo ligado as profissões adotadas. Na Tabela 1 trouxemos alguns resultados que obtivemos utilizando o algoritmo de analogias extremas.

Tabela 1: Analogias Extremas

Masculino			**Feminino**		
Profissão	**Similaridade**	**Analogia**	**Profissão**	**Similaridade**	**Analogia**
arquiteto	0.9190	arquiteto	arquiteta	0.5951	ana
blogueiro	0.5939	blog	blogueira	0.4943	blog
cantor	0.9270	cantor	cantora	0.5550	ela
escritor	0.9180	escritor	escritora	0.9075	escritora
garçom	0.6840	gerente	garçonete	0.4393	obrigada
professor	0.9095	professor	professora	0.9057	professora

Fonte: Elaborado pelo autor (2025).

No trabalho de SANTANA et al. (2018), que utilizou embeddings estáticos

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

e um corpus relativamente menor em comparação ao empregado em nossos experimentos, foi observado um resultado interessante para a profissão garçonete: a maior similaridade encontrada foi com a profissão stripper. Nos resultados obtidos neste estudo, esse fenômeno não se repetiu. Tal diferença pode ser explicada por diversos fatores, sendo o principal deles a utilização de um dataset significativamente maior. Ademais, a origem das informações exerce papel crucial, uma vez que a escolha adequada do corpus pode mitigar a inclusão de preconceitos presentes na linguagem humana.

4. CONCLUSÕES

Mesmo sendo um trabalho em desenvolvimento, com esses resultados que obtivemos foi possível observar como esse modelo se comporta analisando ele nesse contexto. A partir dos resultados obtidos, concluímos que, o modelo apresenta um forte indício de viés de gênero no contexto de profissões, principalmente nas 48 que foram executadas nos testes, podemos assegurar essa afirmação visto que foram extraídas 7 relações corretas para profissões femininas e 19 relações masculinas. Além disso, foi verificado que o modelo gerava e nessa análise notamos que as palavras que o modelo realmente não encontrava relação corretamente, não existia no modelo, ou seja, os embeddings tendem a achar mais profissões masculinas do que femininas por que o modelo tem mais dados masculinos. 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 de métodos de mitigação de viés e de uma avaliação crítica antes de empregar tais modelos em soluções que impactam diretamente pessoas.

Como trabalhos futuros, pretende-se realizar uma análise comparativa entre modelos de diferentes tamanhos, de modo semelhante ao proposto por JENTZSCH; TURAN (2022), em que foi utilizada uma gama de modelos com variações de porte, aplicando-se a métrica WEAT para fins de comparação. Ademais, considera-se a aplicação de relações extremas em modelos diversos, tanto em relação ao tamanho quanto ao tipo, a fim de obter resultados mais robustos e permitir uma avaliação comparativa mais precisa, possibilitando, assim, a classificação dos modelos que apresentam melhor desempenho.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BOLUKBASI, T.; CHANG, K.-W.; ZOU, J.; SALIGRAMA, V.; KALAI, A. (2016). **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings**. arXiv: 1607.06520 [cs.CL].

CALISKAN, A.; BRYSON, J. J.; NARAYANAN, A. **Semantics derived automatically from language corpora contain human-like biases**. **Science** 356.6334, vol. 356, n. 6334, p. pp. 183–186, 2017.

CONNEAU, A.; KHANDELWAL, K.; GOYAL, N.; CHAUDHARY, V.; WENZEK, G.; GUZMÁN, F.; GRAVE, E.; OTT, M.; ZETTLEMOYER, L.; STOYANOV, V. **Unsupervised Cross-lingual Representation Learning at Scale**. Em: **Proceedings of the**

58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. Pp. 8440–8451.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. Em: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. Pp. 4171–4186.

DUAN, Y.; TANG, F.; WU, K.; GUO, Z.; HUANG, S.; MEI, Y.; WANG, Y.; YANG, Z.; GONG, S. (2024). **"The Large Language Model (LLM) Bias Evaluation (Age Bias)--DIKWP Research Group International Standard Evaluation**.

FARIAS, I.; ALBUQUERQUE, D.; RODRIGUES, G.; FILHO, E.; XAVIER, K.; SILVA, J. **Investigando o Uso de Ferramentas baseadas em Grandes Modelos de Linguagem no Contexto Acadêmico**. Em: **Anais do XXXII Workshop sobre Educação em Computação**. Brasília/DF: SBC, 2024. Pp. 489–500.

JENTZSCH, S.; TURAN, C. **Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task**. Em: **Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)**. Seattle, Washington: Association for Computational Linguistics, 2022. Pp. 184–199.

SANTANA, B. S.; WOLOSZYN, V.; WIVES, L. K. (2018). **Is there Gender bias and stereotype in Portuguese Word Embeddings?** arXiv: 1810.04528 [cs.CL].

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **BERTimbau: Pretrained BERT Models for Brazilian Portuguese**. Em: **Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I**. Rio Grande, Brazil: Springer-Verlag, 2020. Pp. 403–417.

STECK, H.; EKANADHAM, C.; KALLUS, N. **Is Cosine-Similarity of Embeddings Really About Similarity?** Em: **Companion Proceedings of the ACM Web Conference 2024**. Singapore, Singapore: Association for Computing Machinery, 2024. Pp. 887–890.

TASO, F.; REIS, V.; MARTINEZ, F. **Sexismo no Brasil: análise de um Word Embedding por meio de testes baseados em associação implícita**. Em: **Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Belo Horizonte/MG: SBC, 2023. Pp. 53–62.

ÜNAL, F.; TARHAN, S.; KÖKSAL, E. Ç. **Gender and Perception of Profession. Journal of education and training studies** 6.n3a, vol. 6, n. n3a, p. pp. 35–44, 2018.