XXIX Congresso de Iniciação Científica UNICAMP 2021

Detecção de Viés de Gênero via Processamento de Linguagem Natural

Palavras-Chave: Vieses de Gênero, Processamento de Linguagem Natural, Aprendizado de Máquina

Marcela Medicina Ferreira, IMECC/Unicamp
Profa. Dra. Esther Luna Colombini (orientadora), IC/Unicamp
Profa. Dra. Sandra Avila (orientadora), IC/Unicamp

Resumo: A Inteligência Artificial (IA) tem proporcionado contribuições em diversas áreas da ciência e na sociedade nos últimos anos. No entanto, esses algoritmos são moldados com base em dados que refletem o mundo e crenças pela óptica apenas de quem os desenvolve, que são um pequeno grupo de pessoas e que, em geral, culturalmente já possuem privilégios levando em conta que é uma área ainda dominada por homens. Essa falta de diversidade em IA resulta em algoritmos enviesados reforçando estereótipos de gênero, racismo, religiões e outras formas de discriminação em suas aplicações. Por isso, esse projeto está investigando vieses de gênero em textos em português com protocolo de classificação contextualizada fracamente supervisionada denominado ConWea. Neste processo, a partir de palavras-sementes, classificamos textos sem viés retirados de biografias da Wikipédia e textos enviesados retirados de redes sociais com uma boa acurácia média.

1 Introdução

Cada vez mais a inteligência artificial tem sido usada na tomada de decisões importantes, sejam elas em empresas privadas ou mesmo em sistemas empregados pelos Governos ao redor do mundo. Com isso em mente, devemos pensar no uso responsável e ético das tecnologias que estamos desenvolvendo, minimizando os impactos sobre os grupos sub-representados, que são aqueles que mais sofrem os efeitos desses algoritmos.

Como dito pela matemática e autora americana Cathy O'Neil em seu TED Talk [1], "Algoritmos são opiniões embarcadas em código", ou seja, os algoritmos muitas vezes, acabam refletindo a sociedade do ponto de vista, comportamentos e crenças de quem os desenvolve.

Entre os casos que demonstram esses vieses algorítmico podemos citar, os anúncios de vagas de emprego no Google, que quando mostrados a mulheres tendiam a mostrar com menos frequência vagas bem remuneradas, como descoberto por pesquisadores da Universidade de Carnegie Mellon [2]. Situações como essa acabam reforçando problemas estruturais, como por exemplo, diferença salarial entre homens e mulheres, que só no Brasil (2019), na média, as mulheres receberam 77,7% da remuneração dos homens, de acordo com a Agência Senado [3].

Preocupadas com esses vieses algorítmicos, em especial os de gênero, adaptamos a abirdagem ConWea [4] para classificar textos em português, que se baseia em palavras-sementes (que podem ser entendidas também como palavras-chave) que representam bem as classes a serem previstas. Dessa maneira, pudemos classificar, dado um texto se ele continha algum tipo de estereótipo de gênero ou não. Para isso, utilizamos como base de dados biografias de mulheres e homens cientistas retirados da Wikipédia, e também textos de redes sociais como Twitter, por exemplo. Este último foi cedido pelo projeto MINA-BR.

Atualmente, em português, não existem muitos trabalhos sobre identificação de vieses de gênero utilizando técnicas de Processamento de Linguagem Natural, e alguns dos que existem, também estão na língua inglesa, e focam em analisar viés de gênero a partir da diminuição do viés nos *word embeddings*, isto é, nos vetores com componentes numéricos que representam palavras individuais obtidas por técnicas de modelagem de linguagem [5]. Com isso, pretendemos contribuir com a criação de um modelo para detecção automática de viés de gênero em textos em português.

2 Metodologia

Dado que o contexto no qual as palavras estão inseridas é fundamental para determinar o sentido da frase e que em português existem muitas palavras polissêmicas, técnicas baseadas em classificações por frequência de palavras no texto poderiam não ser tão benéficas. Assim, neste trabalho empregamos o ConWea [4], uma abordagem para classificação contextualizada fracamente supervisionada. Para nossos experimentos, coletamos textos da Wikipédia em português e utilizamos textos em português de usuários do Twitter, como melhor descrito nas próximas subseções.

2.1 ConWea

A abordagem ConWea, proposta para classificação de textos, tem como principal objetivo contextualizar textos de acordo com palavras-sementes (*seed words*) fornecidas pelo usuário. São fornecidas m classes $C = C_1, C_2, \ldots, C_m$, suas respectivas *seed words*, que podem ser entendidas como um pequeno conjunto de palavras representativo em relação ao tópico de cada classe, e n documentos $D = D_1, D_2, \ldots, D_n$. Com isso, deseja-se atribuir $C_j \mapsto D_i$ para cada $D_i \in D$ e $C_j \in C$, como mostrado na Figura 1.

As técnicas utilizadas no texto artigo [4] foram executadas com textos em inglês, mais especificamente, notícias das bases de dados 20Newsgroup (20News) e New York Times (NYT). Além da abordagem ConWea, também foram utilizados o algoritmo de clustering k-means para agrupar as ocorrências de uma mesma palavra com k significados identificadas nos documentos, o modelo BERT pré-treinado para os *embeddings* e *Hierarchical Attention Networks* (HAN) para a classificação dos documentos.

Para que os experimentos pudessem ser executados em textos em português, utilizamos uma versão do BERT multilínguas, na qual, o modelo foi treinado para as 104 top línguas com maior acervo na Wikipédia e disponibilizado pelo Hugging Face[1]. Para o pré-processamento, usamos recursos de algumas bibliotecas em Python para remover sinais de pontuação '@' dos usuários, entre outros sinais gráficos.

[1] `https://huggingface.co/bert-base-multilingual-cased`

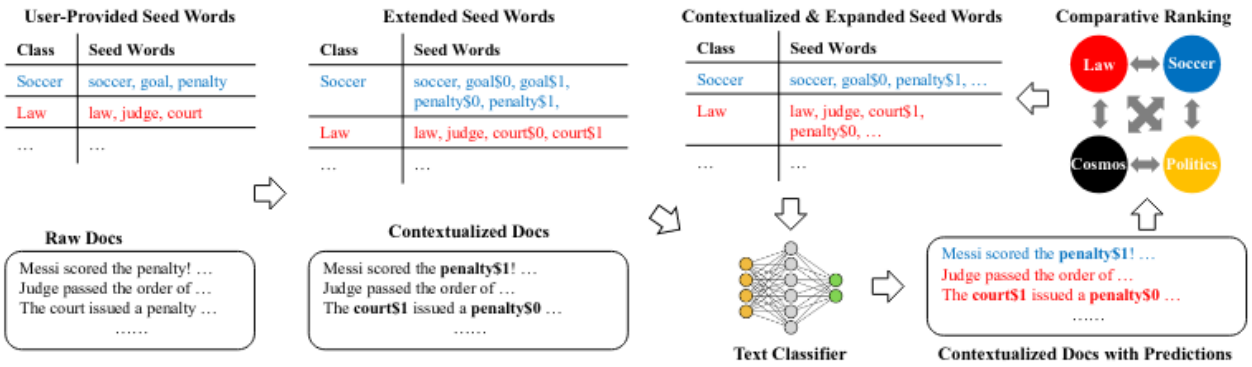

Figure 1: Esquema de funcionamento do ConWea. Figura reproduzida Shang *et al.* [4].

2.2 Base de Dados

Nossos dados foram divididos em duas classes: viés e sem viés. Para a classe com viés, utilizamos os dados do projeto MINA-BR[2], que foi criado para realização de pesquisas sobre discursos de ódio contra mulheres com textos de redes sociais.

Para a classe sem viés, coletamos textos biográficos de homens e mulheres cientistas da Wikipédia. Essa decisão foi tomada após realizarmos diversos testes sobre os dados para analisar se era possível inferir estereótipos de gênero destes textos. Foram escolhidas *seed-words* a partir das palavras mais comuns na língua inglesa para descrever homens e mulheres — como por exemplo *beauty*, *husband* e *gossip* para mulheres e *brilliant*, *arrogant* e *logical* para homens. Mesmo assim, não foi possível classificar sobre os vieses, já que nesse caso específico não havia uma quantidade significativa dessas palavras nos textos ou até mesmo, algumas delas não estavam presentes por se tratar de um texto mais impessoal, que não costuma descrever aspectos físicos e psicológicos.

O fluxo da extração dos dados funciona como mostrado na Figura 2. Através do Wikidata Query Service[3], escrevemos uma consulta em SPARQL (vários exemplos estão disponíveis no website da Wikidata[4]). Após realizar a consulta, podemos gerar pelo próprio serviço de *queries* um script em python, e com ele obtemos uma lista das páginas encontradas. Em nosso caso, especificamente, essa lista corresponde a lista de nomes de cientistas.

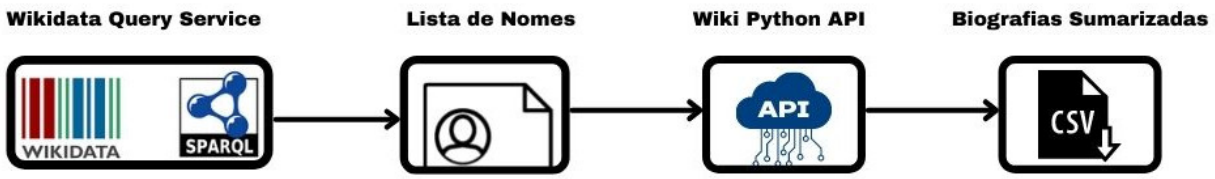

Figure 2: Passo a passo da coleta dos dados para a categoria "sem viés".

Em seguida, através da biblioteca em python 'wikipedia', fizemos o scrapping das biografias. Essa biblioteca possui uma função de sumarização dos textos (summary) que dado o nome da página (neste caso, o nome de uma pessoa), já retorna como output o texto sumarizado. Isso

[2] `https://mina-br.netlify.app`
[3] `https://query.wikidata.org`
[4] `https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples#Cats`

foi muito importante para nós porque alguns textos da Wikipédia são longos, o que seria muito discrepante nas postagens das redes sociais e até mesmo a outros textos de cientistas menos famosos, por exemplo, e com esse recurso pudemos delimitar o número máximo de sentenças no texto. A seguir apresentamos duas amostras do nosso conjunto de dados, nas categorias viés e sem viés, respectivamente.

> viés; @NOME @NOME @NOME Entender a cabeça de mulher é simples, elas são hipergamicas e emocionais, mas a cabeça da feminazi aí já é loucura pura porque elas não possuem o mínimo de racionalidade e lógica.

> sem viés; "Zaida Muxí Martínez (Buenos Aires, 1964) é arquiteta e urbanista graduada pela Faculdade de Arquitetura, Design e Urbanismo da Universidade de Buenos Aires, doutora pela Escola Técnica Superior de Arquitetura de Sevilha e professora na Escola Técnica Superior de Arquitetura de Barcelona. Coordena junto a Josep Maria Montaner o Mestrado Laboratório da Habitação no século XXI da Universidade Politécnica da Catalunha. É conhecida por sua atuação nos temas de espaço e gênero."

Na categoria viés, vimos que o usuário afirma que as mulheres podem ser descritas de maneira "simples", são hipergâmicas e emocionais e as 'feminazis' (termo pejorativo para mulheres feministas) não tem lógica e racionalidade, todos esses termos reforçam estereótipos de comportamento e desmerecem as lutas e os papéis das mulheres na sociedade que foram conquistados pelo feminismo. Já a classe sem viés, descreve uma arquiteta de forma direta, valorizando seu trabalho e sua formação, de maneira impessoal.

2.3 Validação da Técnica

Como descrito anteriormente, não obtivemos sucesso ao tentar classificar os textos da Wikipédia utilizando palavras comuns em frases que reforçam estereótipos de gênero. Por isso, contamos a frequência de palavras nesses textos e vimos que de fato não havia uma discrepância grande por gênero entre as palavras utilizadas nas biografias. Apesar disso, obtivemos que, em média, os textos masculinos são mais longos.

Diante disso, para validar a técnica que utilizamos, repetimos os experimentos com pronomes pessoais masculinos (como *his*, *he*, *him*) e femininos (como *her*, *she*, *hers*) como *seed-words* com objetivo de analisar se o modelo era capaz de diferenciar se o autor do texto estava descrevendo a biografia de uma mulher ou de um homem cientista. Com isso, alcançamos resultados positivos e este processo permitiu a classificação correta entre textos de homens e mulheres, permitindo assim a validação da técnica.

3 Resultados e Discussão

Baseado na abordagem ConWea, repetimos nossos experimentos com variações na quantidade de seed-words:

- 1: "abortista", "feminazi"‘, "feminista", "antifeminista", "estupro", "mal amada" e "cientista", "inteligente", "pesquisa", "ciência", "trabalho", "carreira"
- 2: "abortista", "feminazi", "antifeminista", "estupro", "mal mada" e "inteligente", "pesquisa", "ciência", "trabalho", "carreira".
- 3: "feminazi", "feminista", "antifeminista", "mal amada" e "inteligente", "pesquisa", "trabalho", "carreira".

Mesmo com a variação na quantidade de *seed-words* obtivemos acurácia média superior a 90%. Agora, pretendemos continuar coletando dados e verificar o desempenho do modelo que obtivemos em cenários que cuja o tema principal textos seja o mesmo, como por exemplo "tweets" de usuários antifeministas e textos de cientistas feministas. Pretendemos também adaptar o protocolo para o modelo do BERT pré-treinado no BrWaC (Brazilian Web as Corpus), o BERTimbau — Portuguese BERT [6, 7].

Agradecimentos

Agradecemos ao PIBIC/CNPq/Unicamp pelo suporte financeiro para a realização do trabalho.

Referências

[1] C. Felschen, "Algorithms are opinions embedded in code." `https://wpmu.mah.se/nmict201group5/2020/03/01/algorithm-discrimination-inequality-usa-credit-score/`, Mar. 2020.

[2] S. Gibbs, "Women less likely to be shown ads for high-paid jobs on google, study shows." `https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study`, July 2015.

[3] A. Senado, "Combate à diferença salarial entre homens e mulheres está na pauta do plenário." `https://www12.senado.leg.br/noticias/materias/2021/03/12/combate-a-diferenca-salarial-entre-homens-e-mulheres-esta-na-pauta-do-plenario`, Mar. 2021.

[4] D. Mekala and J. Shang, "Contextualized weak supervision for text classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 323–333, Association for Computational Linguistics, July 2020.

[5] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[6] F. Souza, R. Nogueira, and R. Lotufo, "Portuguese named entity recognition using bert-crf," *arXiv preprint arXiv:1909.10649*, 2019.

[7] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: pretrained BERT models for Brazilian Portuguese," in *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.