Viés de Gênero em Biografias da Wikipédia em Português

Isadora A. Salles[1], **Gisele L. Pappa**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)

{isadorasalles,glpappa}@dcc.ufmg.br

Resumo. *Mulheres representam aproximadamente 15% dos editores da Wikipédia, e essa disparidade tem potencial para introduzir vieses no conteúdo dessa enciclopédia. Neste trabalho, propomos uma análise da existência de viés de gênero no conteúdo da Wikipédia comparando os textos de biografias de homens e mulheres em duas dimensões: meta-dados e linguagem. Nossos resultados mostraram que realmente existem diferenças na maneira como as mulheres são retratadas em suas biografias quando comparadas aos homens.*

Abstract. *Women represents approximately 15% of the Wikipedia contributors, and this disparity has potential to introduce biases into this encyclopedia content. In this work, we propose an analysis of gender bias existence on Wikipedia content comparing the text of biographies of men and women consdering two dimensions: meta-data and language. Our results show that there are differences in the way that women are described in biographies when compared to men.*

1. Introdução

A Web criou oportunidades únicas para a democratização da mídia, permitindo que qualquer pessoa possa ter voz. Um dos principais exemplos disso é a Wikipédia, que provê uma plataforma para compartilhamento gratuito de conhecimento entre as pessoas, em que o conteúdo é escrito por uma comunidade de voluntários. No entanto, uma pesquisa sobre os contribuidores da Wikipédia concluiu que menos de 15% dos editores são mulheres [Collier and Bear 2012]. Essa disparidade foi chamada de *gender gap*, e motivou diferentes estudos para entender esse fenômeno social. Uma vez que se há uma sub-representatividade entre os editores, é possível que haja também um viés de gênero no conteúdo disponível [Lam et al. 2011].

Em [Schmahl et al. 2020] foram analisadas biografias da Wikipédia de 2006 à 2020, e evidenciaram que o viés de gênero ainda está presente na Wikipédia. Mostrando que mesmo com tantos anos desde as primeiras pesquisas sobre o assunto, ainda há muito a ser feito para mudar as perspectivas de igualdade de gênero tanto na representatividade entre os contribuidores, quanto no conteúdo publicado.

Neste trabalho, investigamos a existência de desigualdade de gênero em biografias disponíveis na Wikipédia em português. A maioria dos estudos nessa área focam na língua inglesa, e não encontramos nenhum estudo similar para a língua portuguesa. Dadas as questões culturais de discussões de gênero no Brasil [Garcia 2020], um estudo nessa linha pode refletir a cultura do país de forma diferente que em países anglófonos. Note que a

Wikipédia em português é única para todos os países lusófonos, mas 68% dos editores são brasileiros, seguidos pelos portugueses com 10%[1].

Para realizar as análises, inicialmente preparamos uma base de dados com 25.827 biografias. A partir disso, analisamos duas dimensões: meta-dados, a partir de dados do *infobox* das biografias; e linguagem, explorando como homens e mulheres são caracterizados através de uma perspectiva léxica analisando o vocabulário usado.

2. Trabalhos relacionados

O *gender gap* [Collier and Bear 2012] motivou diversos estudos sobre a existência de viés no conteúdo da Wikipédia. Em [Wagner et al. 2015] o viés de gênero foi analisado em quatro dimensões: cobertura, estrutura, léxico e visibilidade, em seis idiomas diferentes. Evidenciando que palavras relacionadas a família, relacionamentos e gênero são mais presentes em artigos sobre mulheres. Em um trabalho subsequente [Wagner et al. 2016], os autores mostram que as mulheres presentes nas biografias da Wikipédia são mais notáveis na média do que os homens, o que sugere que mulheres devem ser mais notáveis para serem retratadas na Wikipédia. Já em [Graells-Garrido et al. 2015] analisaram a Wikipédia em inglês em relação a meta-dados, linguagem e estrutura da rede de biografias. E encontraram diferenças significantes que não podem ser atribuídas apenas aos vieses da sociedade. Por fim, em [Konieczny and Klein 2018] os autores mostraram que a proporção de biografias de mulheres em inglês e em português é bastante similar, por volta de 16%. Mostraram também que a proporção de biografias de mulheres apenas em português é bem baixa, o que significa que temos uma baixa representatividade de mulheres brasileiras notáveis na Wikipédia. Ao contrário dos trabalhos discutidos nessa seção, esse artigo foca na disparidade de gênero no conteúdo de biografias escritas em português.

3. Construção da base de dados

3.1. Extração das biografias

O *Wikipedia Dump* de primeiro de Janeiro de 2021[2] foi usado para obter os artigos da Wikipédia em Português. Os XMLs coletados foram filtrados pelo marcador do portal de biografias. Porém, percebeu-se que nem todos os artigos nesse portal tratam-se de biografias. Para contornar esse problema, filtramos os artigos também pelos marcadores de "data de nascimento". Garantindo que todos os artigos sejam de fato uma biografia. Ao final desse processo foram extraídas 25.827 biografias.

3.2. Inferência de gênero

O atributo gênero não aparece no *infobox* da grande maioria das biografias da Wikipédia. Sendo assim, é preciso realizar um processo de inferência para obter o gênero de cada uma das pessoas que compõe a base. É importante ressaltar que nesse trabalho consideramos apenas os gêneros feminino e masculino. Entretanto, o método utilizado permite que pessoas transgênero sejam classificadas de acordo com o gênero que se identificam, desde que o artigo sobre essa pessoa na Wikipédia esteja tratando-a como tal.

[1] `https://stats.wikimedia.org/#/pt.wikipedia.org/contributing/active-editors-by-country/normal|table|last-month|(activity-level)~5..99-edits|monthly`

[2] `https://dumps.wikimedia.org/ptwiki/20210101`

Inicialmente, foi construído um dicionário contendo nomes normalmente utilizados para designar pessoas do gênero feminino ou masculino. Utilizamos a base gerada por [Bamman and Smith 2014], que inferiu o gênero para 862.171 pessoas na Wikipédia em inglês, e a base de nomes do IBGE[3], que possui nomes comuns no Brasil. Nomes muito utilizados para ambos os gêneros foram descartados. Com isso, foi possível rotular 12.767 biografias, das quais 78,26% são do gênero masculino e 21,74% feminino.

Foi realizado um pré-processamento do texto das biografias a fim de retirar pontuações, acentos, *stopwords* e *tokenizar* as palavras. Assumindo que apenas com o início de uma biografia já possuímos variáveis o suficiente para distinguir o gênero da pessoa, usamos apenas as palavras dos dois primeiros parágrafos dos textos. Computamos o *term frequency* e somamos entre todos os documentos da coleção, e os termos com frequências menores que três foram eliminados. As palavras restantes foram ordenadas em ordem decrescente de informação mútua, e usadas como entrada para uma abordagem gulosa, que utiliza uma validação cruzada com cinco partições com o dado rotulado até então e um classificador *RandomForest*, para selecionar o melhor número de variáveis para treinar o modelo. O conjunto das 50 primeiras palavras apresentou melhor desempenho. E o melhor modelo na validação foi usado para prever o rótulo do restante das biografias da base. Em particular, o classificador obtido possui F1 de 0,920, com uma precisão de 0,922 e revocação de 0,918 para a classe "mulher". Já para a classe "homem", obtivemos F1 de 0,978, precisão de 0,978 e revocação de 0,979. Ao final desse processo, temos posse de uma base de biografias com 5.010 biografias de mulheres e 20.817 biografias de homens.

4. Análise de Meta-dados

Em uma caracterização inicial dos dados percebe-se que é possível que haja uma correlação entre popularidade da pessoa, ou gênero, e o número de palavras utilizadas para descrevê-la. A fim de tentar capturar esse padrão, foi gerada uma nuvem de palavras contendo o nome das pessoas que possuem as biografias com maior número de palavras na Wikipédia, conforme mostrado na Figura 1. Note que a maioria dos nomes são referentes a homens famosos, sejam jogadores de futebol, artistas, escritores, políticos, entre outros.

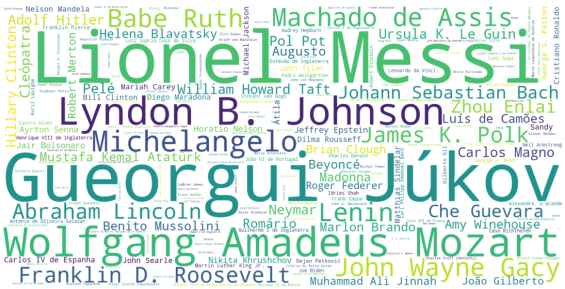


Figura 1. Nomes das pessoas com biografias mais longas.

A fim de estimar a presença e a proporção das mulheres de acordo com determinados atributos na Wikipédia, foi feita uma análise do *infobox* das biografias. O *infobox* é uma espécie de tabela que resume algumas características da pessoa sendo retratada, e inclui atributos como: *nome, data de nascimento, ocupação*, entre outros. Um *infobox* também deve se enquadrar em uma classe, sendo a mais geral denominada *Biografia*. Outros exemplos de classes são: *Futebolista, Político, Ator/Atriz*. Inicialmente, comparamos a proporção de homens e mulheres que se enquadram nas classes mais comuns (Tabela 1).

[3]https://censo2010.ibge.gov.br/nomes/#/search

Tabela 1. Número e proporção de biografias de mulheres para as classes mais comuns.

Classe	Quantidade de biografias	% Mulheres
Biografia	6422	26,38
Futebolista	3928	1,32
Político	3252	11,68
Ator/Atriz	2543	40,58
Música	1744	26,78
Ciclista	1119	10,10
Esporte	941	30,92
Cientista	772	27,97
Nobre	647	27,82
Escritor	622	21,54

Tabela 2. Porcentagem de homens e mulheres nas 10 classes mais frequentes por gênero.

Classe	% Homens
Biografia	22,71
Futebolista	18,62
Político	13,80
Ator/Atriz	7,26
Música	6,13
Ciclista	4,83
Esporte	3,12
Cientista	2,67
Treinador	2,38
Escritor	2,34

Classe	% Mulheres
Biografia	33,81
Ator/Atriz	20,60
Música	9,32
Político	7,58
Esporte	5,81
Cientista	4,31
Nobre	3,59
Escritor	2,67
Ciclista	2,26
Futebolista	1,04

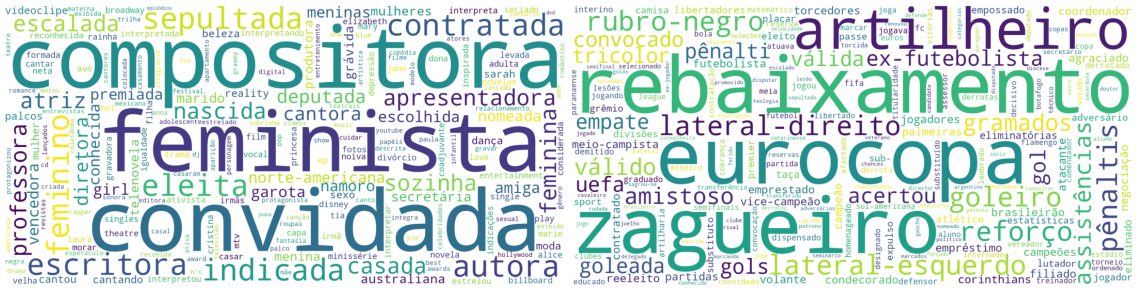
Na Tabela 1, podemos observar que as classes *Ator/Atriz*, *Música*, *Esporte*, *Cientista* e *Nobre* são as que possuem maior presença de mulheres, e a classe com maior proporção de mulheres é a *Ator/Atriz* com 40,58%. Enquanto que *Futebolista*, *Político* e *Ciclista* representam muito mais homens. A proporção relativamente alta de mulheres na categoria *Esportes* não era esperada, porém note que essa classe tem apenas 941 biografias. Enquanto que, existem outras classes também ligadas à esportes que possuem uma quantidade mais alta de biografias e uma proporção de homens também mais alta. Entretanto, isso não necessariamente configura um viés no conteúdo da Wikipédia, pode ser apenas um reflexo de um viés da sociedade. Já a Tabela 2 traz uma análise das 10 classes mais frequentes para cada gênero. Note que, a classe mais ligada aos homens é *Futebolista* e às mulheres é *Ator/Atriz* (desconsiderando a classe *Biografia*).

Para analisar a presença de diferentes atributos do *infobox* para cada gênero, utilizamos um teste chi-quadrado para escolher os atributos que apresentam uma diferença de proporções entre gêneros mais significativa. Com isso, foi possível observar que atributos relacionados a esportes, como *clubes*, *jogos* e *seleçãoNacional* são mais frequentes em biografias de homens. Nesse caso, a diferença é explicada pela maior presença de homens em classes relacionadas com esportes. O atributo *ocupação* aparece em 58,34% das biografias de mulheres, e em 31,21% de biografias de homens. Uma possível explicação para isso é que os *infoboxes* de biografias que estão em classes relacionadas com esporte não costumam conter esse atributo, porque o *template* do *infobox* já indica a ocupação da pessoa retratada. Por fim, o atributo *cônjuge* é mais frequente em mulheres, aparecendo em 26,43% das biografias de mulheres, mas não obtivemos uma explicação direta para isso, além da possível tendência a incluir esse atributo mais em biografias de mulheres.

5. Propriedades da Linguagem

Nessa seção nós exploramos a descrição de homens e mulheres na Wikipédia através de uma perspectiva léxica. Inicialmente, computamos o vocabulário comum aos dois gêneros estudados, e obtivemos um conjunto de 91.976 palavras. E então, para evidenciar as palavras mais fortemente associadas a cada gênero foi calculado o *Pointwise Mutual Information* [Church and Hanks 1990], que é uma medida de associação definida como: $\text{PMI}(g, w) = log\frac{p(g,w)}{p(g)p(w)}$, onde g é o gênero e w é uma palavra do vocabulário comum. As probabilidades foram estimadas a partir da proporção de biografias sobre homens e mulheres, e a proporção de biografias em que cada palavra aparece. Como o PMI dá mais

peso a palavras com frequências menores, foi considerado apenas palavras que aparecem em pelo menos 1% das biografias de mulheres e homens.

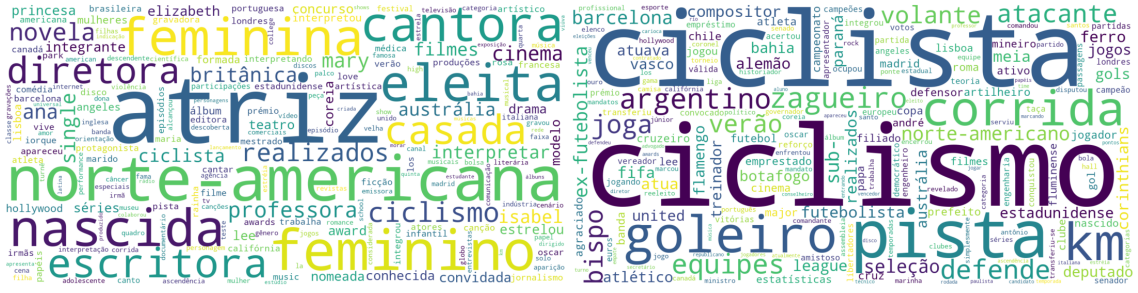


(a) Palavras mais associadas a mulheres. (b) Palavras mais associadas a homens.

Figura 2. Palavras mais associadas a cada gênero.

As palavras mais associadas com homens são relacionadas a esportes, o futebol em particular, conforme mostrado na Figura 2b. Enquanto que, para as mulheres, as palavras mais associadas são relacionadas com o meio artístico (compositora, escritora) e com gênero (feminista, feminina) (Figura 2a). A palavra *casada* é a 15° mais associada com mulheres. Isso chamou bastante nossa atenção, pois reforça a suspeita levantada anteriormente sobre o atributo *cônjuge* no *infobox* de mulheres.

Em uma segunda análise, com o propósito de determinar quais palavras são mais discriminativas para distinguir o gênero, utilizamos o tf-idf das palavras selecionadas a partir do cálculo de informação mútua como atributo de entrada para treinar um classificador *Naive Bayes*. A função de log das probabilidades dos parâmetros do modelo foi usada para comparar as relações entre as saídas do modelo. Essa função é dada por: $L(w, g) = log\dfrac{p(w|g)}{p(w)}$, onde w é a palavra e g o gênero. O valor $p(w|g)$ é a probabilidade condicional de que uma palavra apareça em um artigo sobre uma pessoa, dado que essa pessoa possui o gênero g, e foi obtida diretamente do modelo *Naive Bayes*. E $p(w)$ é a probabilidade de que a palavra apareça em qualquer artigo.



(a) Palavras mais efetivas para classificar mulheres. (b) Palavras mais efetivas para classificar homens.

Figura 3. Palavras mais discriminativas para distinguir cada gênero.

Os resultados foram similares aos obtidos até então e consistentes com [Wagner et al. 2015]. As palavras mais efetivas para classificar mulheres remetem ao ramo artístico e gênero, e a palavra *casada* aparece novamente como uma das primeiras na nossa análise (Figura 3a). Enquanto que, as palavras mais efetivas para distinguir homens são novamente relacionadas a esportes. Mas note que nesse caso temos palavras relacionadas a ciclismo e corrida entre as mais importantes (Figura 3b), diferentemente das outras análises em que tínhamos apenas o futebol em destaque.

6. Conclusão

Este trabalho propôs analisar o viés de gênero no conteúdo das biografias da Wikipédia em português. Nossos resultados indicam diferenças significativas sobre a forma como mulheres são retratadas em comparação com os homens. Em particular, percebemos que os relacionamentos amorosos de mulheres são muito mais frequentemente considerados e discutidos. Entretanto, não é possível afirmar que toda a disparidade encontrada é atribuída a um viés presente na Wikipédia ou se é apenas um reflexo de vieses da sociedade. Além disso, o foco em apenas um idioma é uma limitação desse trabalho. Como 68% dos editores da Wikipédia em português são brasileiros, temos o viés da cultura brasileira bastante presente na base de dados. Uma outra limitação foi considerar apenas o binarismo de gênero, mas acreditamos que esse seja um primeiro passo para conseguir analisar a dimensão de gênero no conteúdo da Wikipédia a partir de uma perspectiva mais inclusiva.

Referências

Bamman, D. and Smith, N. A. (2014). Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16.1:22–29.

Collier, B. and Bear, J. (2012). Conflict, criticism, or confidence: An empirical examination of the gender gap in wikipedia contributions. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW*, 12:383–392.

Garcia, L. C. (2020). Cultura do estupro: Machismo e as raízes da violência de gênero no brasil. *Diké: Revista Eletrônica de Direito, Filosofia e Política do Curso de Direito da UNIPAC Itabirito*, page 49.

Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First women, second sex: Gender bias in wikipedia. *Proceedings of the 26th ACM conference on hypertext*, pages 165–174.

Konieczny, P. and Klein, M. (2018). Gender gap through time and space: A journey through wikipedia biographies via the wikidata human gender indicator. *New Media & Society*, 20(12):4608–4633.

Lam, S. K., Uduwage, A.and Dong, Z., Sen, S.and Musicant, D. R., Terveen, L. G., and Riedl, J. (2011). Wp:clubhouse?: an exploration of wikipedia's gender imbalance. *ACM*, pages 1–10.

Schmahl, K. G., Viering, T. J., Makrodimitris, S., Jahfari, A. N., Tax, D., and Loog, M. (2020). Is wikipedia succeeding in reducing gender bias? assessing changes in gender bias in wikipedia using word embeddings. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, page 94–103.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *Ninth International AAAI Conference on Web and Social Media*.

Wagner, C., Graells-Garrido, E., Garcia, D., and Menczer, F. (2016). Women through the glass ceiling: Gender asymmetries in wikipedia. *EPJ Data Science*, 5(1).