Uma análise sobre viés de gênero na Wikipédia

Isadora Alves de Salles[1]
Orientadora: Gisele Lobo Pappa[1]

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

{isadorasalles, glpappa}@dcc.ufmg.br

Resumo. *A Wikipédia é uma enciclopédia "criada pela comunidade", onde qualquer pessoa pode contribuir para o conteúdo, visando documentar o mundo de um ponto de vista neutro. Entretanto, a diversidade limitada da comunidade de voluntários da Wikipédia tem potencial para introduzir vieses, como viés de gênero, no conteúdo dessa enciclopédia. Nesse trabalho decidimos explorar uma parte desse problema, buscando verificar a existência de viés de gênero especificamente nas biografias da Wikipédia em português. E descobrimos que a maneira como as mulheres são retratadas nas biografias se difere bastante dos homens.*

1. Introdução

Atualmente, a internet tem criado oportunidades para a democratização da mídia, onde qualquer pessoa pode ter voz. Um dos principais exemplos disso é a Wikipédia, que provê uma plataforma para compartilhamento gratuito de conhecimento entre as pessoas. Diferentemente de outras enciclopédias tradicionais, onde pessoas experientes nas áreas específicas escrevem, editam e validam o conteúdo, na Wikipédia isso é feito por uma comunidade de voluntários. Porém, é sabido que essa comunidade consiste predominantemente de homens brancos. E essa falta de representatividade entre os editores da Wikipédia tem potencial para introduzir vieses no conteúdo.

Em 2012, uma pesquisa sobre os contribuidores da Wikipédia concluiu que menos de 15% dos editores são mulheres [1]. Essa disparidade foi chamada de *gender gap*, e motivou diferentes estudos para entender esse fenômeno social e o que pode ser feito para aumentar a quantidade de mulheres que contribuem para a Wikipédia. Além disso, o *gender gap* motivou pesquisas sobre o conteúdo dessa enciclopédia, pois, se há uma sub-representatividade entre os editores, é possível que haja também um viés de gênero no conteúdo disponível.

Um artigo publicado em novembro de 2020 analisou as mudanças no viés de gênero da Wikipédia usando *word embeddings*, tendo em vista que houve um aumento de 18% na fração de biografias de mulheres (em inglês). Eles evidenciaram que o viés de gênero decresceu para as áreas de Ciência e Família, enquanto que aumentou para Artes [2]. Esse artigo mostra que, mesmo com tantos anos desde as primeiras pesquisas sobre viés de gênero na Wikipédia, nós ainda temos muito a fazer para mudar as perspectivas de igualdade de gênero tanto na representatividade entre os contribuidores, quanto no conteúdo publicado.

Devido a influência da Wikipédia é importante avaliar e corrigir esses vieses. Assim, nesse trabalho decidimos explorar uma parte desse problema a fim de investigar a

existência de desigualdade de gênero especificamente em biografias disponíveis na Wikipédia em português, utilizando conhecimentos de processamento de linguagem natural. Essa escolha se dá pelo fato de que existem muitas possibilidades a serem exploradas, e como a maioria dos estudos na área focam na língua inglesa acreditamos que ao final desse trabalho será possível propor novas contribuições.

Tendo em vistas essas motivações, a abordagem que seguimos para a construção da análise final foi avaliar os artigos de biografias da Wikipédia em duas dimensões: meta-dado, comparando como homens e mulheres proporcionalmente possuem alguns atributos em seu *infobox* (tabela usada para apresentar um resumo da vida da pessoa contendo um conjunto de atributos); e linguagem, explorando como homens e mulheres são caracterizados através de uma perspectiva léxica analisando o vocabulário usado.

2. Referencial Teórico

Existem várias pesquisas focadas em reconhecer a importância de entender o viés de gênero na Wikipédia, trazendo análises exploratórias sobre os dados e utilizando técnicas de processamento de linguagem natural para expor a presença desse viés de diversas formas. Uma pesquisa de 2015 analisou desigualdade de gênero na Wikipédia, em quatro dimensões sobre seis idiomas diferentes [3]. As dimensões analisadas foram: viés de cobertura, viés estrutural, viés léxico e viés de visibilidade. A dimensão que foi chamada de viés léxico nesse artigo está bastante correlacionada com o estudo sobre a linguagem que fizemos nesse trabalho, sendo assim é importante contextualizar como essa análise foi feita.

Claudia, David, Mohsen e Markus [3] computaram o viés léxico seguindo a seguinte abordagem: foi computado o tf-idf (*term frequency - inverse document frequency*) de cada radical de palavras obtidos a partir do algoritmo *Snowball Stemmer*, que é um pequeno processamento de strings para reduzir palavras ao seu radical, e isso foi usado como *features* para treinar um classificador *Naive Bayes*. O classificador determina quais palavras são mais efetivas para diferenciar o gênero da pessoa que o artigo descreve. A função de log das probabilidades dos parâmetros do modelo é usada para comparar as relações entre as predições.

A partir dessa predição é possível responder algumas perguntas, como por exemplo: "Palavras referentes a relacionamentos (por exemplo, esposa ou viúva) são mais frequentes nas biografias de mulheres?". Os mesmos autores desse primeiro artigo deram sequência com os trabalhos em 2016 publicando um outro artigo sobre o tema [4].

Também em 2015, foi publicado um outro artigo que foca na análise de viés de gênero na Wikipédia, apenas para a língua inglesa, através de três dimensões: meta-dado, linguagem e estrutura da rede [5]. Esse estudo se relaciona bastante com nosso trabalho, no qual também analisamos as duas primeiras dimensões. No artigo, para a análise das propriedades dos meta-dados foi utilizado o *infobox* das biografias de forma a quantificar a presença, a proporção e a distribuição dos atributos para cada gênero. Já a dimensão de propriedades de linguagem pode ser comparada com o viés léxico, proposto no primeiro artigo citado nessa seção [3], onde a ideia é avaliar quais são as palavras e conceitos utilizados nas biografias de mulheres e homens, e como essas biografias se diferem em termos de uso da língua. Eduardo, Mounia e Filippo [5] utilizaram o *Pointwise Mutual Information* (PMI)[6] sobre o vocabulário de palavras comum aos dois gêneros, a fim de

explorar quais palavras são mais fortemente associadas a cada gênero.

Além disso, existem alguns trabalhos mais recentes na área que visam entender se o viés de gênero na Wikipédia mudou no decorrer dos anos [2, 7], ou que visam entender porque as mulheres contribuem menos para essa enciclopédia [8, 9]. Um estudo de 2018 sobre o *gender gap* através do tempo [7], e em diferentes línguas, mostrou que a proporção de biografias de mulheres em inglês e em português é bastante similar, por volta de 16%. E também mostrou que a proporção de biografias de mulheres apenas em português é bem baixa, o que significa que temos uma baixa representatividade sobre mulheres brasileiras notáveis na Wikipédia.

No nosso trabalho, nós buscamos quantificar viés de gênero na caracterização de homens e mulheres. Para isso, nossa abordagem consiste de fazer uma análise tanto de atributos quanto da linguagem. Os passos necessários para tal incluem uma coleta e extração das biografias, inferência do gênero das pessoas descritas e caracterização e análise dos dados. Sendo assim, e levando em consideração os trabalhos relacionados, as seguintes áreas da computação foram necessárias para a conclusão desse trabalho: ciência dos dados, aprendizado de máquina, processamento de linguagem natural e recuperação de informação.

3. Base de dados

Para estudar a presença de viés nas biografias da Wikipédia foi necessário montar a base de dados. Para tal, os dados foram coletados a partir da plataforma *Wikimedia*. Esse projeto tem o objetivo de tornar os artigos da Wikipédia fáceis de acessar e baixar. Mensalmente, os arquivos são atualizados, subindo para a plataforma a versão mais recente de todos os artigos da Wikipédia. A base utilizada para esse trabalho foi montada através do *Wikipedia Dump* de primeiro de Janeiro de 2021 [10].

3.1. Extração das biografias

A coleta através do *Wikimedia* contempla todos os tipos de artigos da Wikipédia, sendo assim, foi necessário selecionar aqueles que se tratam de uma biografia. Todas as biografias da Wikipédia encontram-se no portal biografias, de maneira que o arquivo XML da página possui uma "tag" com o nome do portal. Dessa forma, foi possível fazer uma busca nos arquivos pela "tag" referente a biografias utilizando uma expressão regular. Porém, como a Wikipédia é uma enciclopédia em que qualquer pessoa pode adicionar um conteúdo novo, aluns erros podem ocorrer. Por exemplo, algumas páginas não referentes a uma biografia podem estar com o marcador de biografia, como o artigo "Lista de gols de Lionel Messi pela Seleção Argentina de Futebol" [11].

Além disso, algumas biografias são muito pequenas e nem mesmo seguem o padrão definido pela Wikipédia para esse tipo de artigo. Sendo assim, como tentativa de contornar esses problemas buscamos também pelos marcadores de data de nascimento nos arquivos do portal biografias, assim temos mais certeza de que filtramos apenas por biografias e também excluímos textos muito pequenos que não contém nem mesmo a data de nascimento da pessoa descrita. Ao final desse processo foram extraídas 25.827 biografias.

Após selecionar quais artigos se tratam de biografias foi necessário fazer uma filtragem do texto e do *infobox*. Para extrair apenas o texto, eliminando os marcadores e

comentários do XML, foi utilizado o *WikiExtractor* [12] que é uma biblioteca disponível em PyPI contendo um *script* para extrair e limpar textos originados do *Wkipedia Database Dumps*. Já a extração dos *infobox* foi feita através do uso de expressões regulares para encontrar a tabela dentro do artigo XML. A limpeza foi feita por uma sequência de pré-processamentos como retirar pontuações e acentos, e criar um dicionário contendo as chaves e valores presentes no *infobox*.

3.2. Inferência de gênero

Para seguir com as análises é preciso saber o gênero de cada uma das pessoas que compõe a base de biografias. É importante ressaltar que nesse trabalho vamos tratar apenas dois gêneros: feminino e masculino. O atributo gênero ou sexo não aparece no *infobox* da grande maioria das biografias da Wikipédia. Sendo assim, foi necessário criar um método para inferir o gênero de cada pessoa. Essa tarefa demandou bastante tempo e pesquisa.

A primeira abordagem seguida foi basear-se no número de palavras que remetem à algum gênero presentes no texto, por exemplo: ele, ela, dele, dela, como proposto por Bamman e Smith [13]. Para isso, buscamos uma maneira de extrair os pronomes usados no texto. Porém, foi difícil estabelecer um método assertivo, visto que, processamento de linguagem natural com textos em português ainda é um desafio pois existem poucas referências na área.

Sendo assim, optamos por seguir um caminho mais simples mas que trouxe bons resultados. O método utilizado baseia-se no que foi proposto por Renato Miranda em sua dissertação de mestrado [14], utilizando as duas etapas a seguir em sequência:

1. Dicionário de nomes: conjunto de nomes normalmente utilizados para designar pessoas do sexo feminino ou masculino.
2. Modelo de classificação supervisionada: o classificador foi treinado a partir das biografias previamente rotuladas com o dicionário de nomes.

3.2.1. Dicionário de nomes

Na primeira etapa, foram utilizados dicionários de nomes para rotular parte da base de biografias. Inicialmente foi utilizado uma base de dados com os nomes e os gêneros inferidos para 862.171 pessoas na Wikipédia em inglês obtida por Bamman e Smith [13]. Após isso, foi utilizada a base de nomes do IBGE [15] para rotular o gênero de brasileiros e portugueses. Para ampliar um pouco mais o dicionário corrente, foi utilizado também o pacote Python gender-guesser [16] para prever o gênero de americanos apenas a partir do primeiro nome. A previsão gerada por essa biblioteca foi desconsiderada quando a resposta do detector foi “mostly_male” ou “mostly_female”. De forma que, nomes que são muito usados para ambos os gêneros só serão inferidos na segunda etapa (por exemplo, Taylor). Ao final desse processo conseguimos rotular 12.767 biografias, das quais 78,26% são do gênero masculino e 21,74% feminino.

3.2.2. Classificador

Na segunda etapa, utilizamos o pedaço da base rotulado até então para formar o conjunto de treinamento para um modelo de classificação. Foi utilizado apenas os dois primeiros

parágrafos do texto das biografias para fazer a previsão do gênero. Visto que, apenas com o início do texto de uma biografia já possuímos variáveis o suficiente para distinguir o gênero da pessoa, principalmente na língua portuguesa, em que grande parte das palavras possuem gênero (por exemplo adjetivos: bonito e bonita). Sendo assim, primeiramente, foi feito um pré-processamento do texto a fim de retirar pontuações, acentos, *stopwords* e *tokenizar* os dois primeiros parágrafos das biografias, obtendo no final uma lista com as palavras usadas.

Depois de pré-processar o texto, foi feita uma seleção de variáveis para utilizar no treinamento do modelo. Essa seleção constituiu-se de 3 passos:

1. **Term frequency:** foi calculado o *term frequency* (frequência do termo em cada documento) para cada palavra de cada texto, e esse valor foi somado entre os documentos. Resultando em um dicionário em que as chaves são todas as palavras que aparecem nos textos. Dessa forma, foi possível eliminar as palavras com menor *term frequency*. E então foi criado um *dataset* em que as palavras restantes os nomes das colunas, e a matriz é preenchida com 1 se a palavra existe no documento da linha i, e 0 caso contrário.
2. **Mutual information:** a partir do *dataset* criado no passo anterior, foi feito o cálculo de dependência mútua entre duas palavras. Isso será utilizado para selecionar o conjunto de palavras que terá melhor desempenho para a classificação.
3. **Abordagem gulosa:** as palavras foram ordenadas em ordem decrescente de *mutual information* e partir disso, foi seguida uma abordagem gulosa, em que o objetivo é encontrar o melhor número de variáveis (palavras) para treinar o modelo. O melhor número encontrado foi 50 variáveis.

A abordagem gulosa foi feita a partir do treinamento do classificador *RandomForest* para várias quantidades de *features*, ou seja, vários *datasets* de treino de diferentes dimensões, até que o melhor ponto foi encontrado. Assim, o classificador final consiste de um *RandomForest* com número de estimadores igual a 18, e utilizando um *dataset* com as 50 primeiras palavras com maior *mutual information*. Na imagem a seguir temos a matriz de confusão gerada pelo modelo ao classificar o conjunto da base rotulada separado para teste:

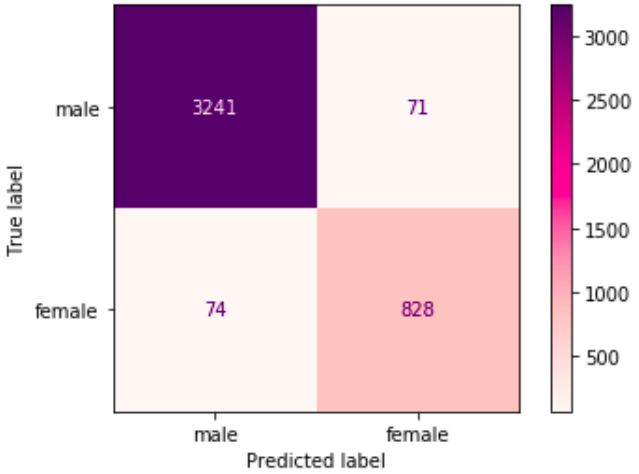

Figura 1. Matriz de confusão do modelo

Um dos grandes desafios da construção desse modelo foi trabalhar com uma base de dados bastante desbalanceada. Dessa forma, não foi possível se basear apenas na acurácia para validar se o classificador obtido é bom. Nesse caso é melhor avaliar métricas como F1, precisão e revocação. Na Figura 1 podemos observar que temos precisão = 0,922, revocação = 0,918 e f1 = 0,92, para a classe "female". Após a criação e validação da eficácia desse classificador, ele foi utilizado para prever o rótulo do restante das biografias da base.

4. Análise de Meta-dados

Antes de realizar uma análise mais aprofundada a cerca das propriedades da linguagem, foi feita uma caracterização inicial dos dados e um estudo de atributos para que fosse possível compreender melhor os dados que estamos trabalhando. Após o processo de inferência de gênero, temos posse de uma base de biografias rotulada e bastante desbalanceada. A base possui 5.010 biografias de mulheres e 20.817 biografias de homens.

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 a seguinte nuvem de palavras contendo o nome das pessoas que possuem as maiores biografias na Wikipédia:

Figura 2. Nuvem de palavras com os nomes das pessoas com maior biografia

Vale ressaltar que para essa análise foram excluídas palavras irrelevantes (*stopwords*), como artigos. Note que na Figura 2 a maioria dos nomes são referentes a homens famosos, seja jogador de futebol, artista, escritor, político entre outros. Apenas com esse exemplo inicial é possível perceber indícios de viés na base.

Tendo em vista a distribuição da base, foi feita uma análise do *infobox* das biografias, a fim de estimar a presença e a proporção das mulheres de acordo com determinados atributos na Wikipédia. Comparamos a proporção de homens e mulheres que se enquadram em determinadas classes contidas nos *infoboxes*. Essas classes podem ser por exemplo *Futebolista* para descrever pessoas ligadas ao futebol. A classe mais geral é chamada apenas de *Biografia*. Na Tabela 1 temos as 10 classes mais comuns, e a proporção de mulheres para cada classe.

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

Podemos observar que as classes *Ator/Atriz*, *Música*, *Esporte*, *Cientista* e *Nobre* são as que possuem maior presença de mulheres. 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 temos apenas 941 biografias nessa classe. Enquanto que, existem outras classes menos abrangentes também ligadas à esportes (como *Futebolista* e *Ciclista*) e que possuem uma quantidade mais alta de biografias e uma proporção de homens bastante alta.

Quando olhamos para esses dados de outra forma, analisando as 10 classes mais frequentes em cada gênero, é possível observar essas análises de forma mais clara. Esses dados podem ser vistos na Tabela 2, em que temos que a classe mais ligada aos homens é *Futebolista* e às mulheres é *Ator/Atriz* (desconsiderando a classe *Biografia*).

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

Tabela 2. Porcentagem de homens e mulheres para as 10 classes mais frequentes em cada gênero

Tendo que existem diferentes classes de *infobox*, existem também vários atributos que podem ser incluídos nas biografias. Assim, tendo feito essa exploração inicial sobre a distribuição das biografias entre as classes, decidimos explorar um pouco mais o restante dos atributos presentes no *infobox*. Esses atributos podem ser, por exemplo: *nome, data de nascimento, país natal, ocupação*, entre outros.

A fim de analisar a presença dos atributos do *infobox* para cada gênero, foram escolhidos os atributos que apresentam uma diferença de proporções entre gêneros mais significativa. Se fossem escolhidos os atributos que aparecem em mais biografias, por exemplo, teríamos nome e data de nascimento com proporções altas e muito semelhantes entre os gêneros. Porém esse não é nosso interesse, com essa análise a intenção é verificar as diferenças entre os gêneros. Sendo assim, para cada atributo foi computado o número de biografias que o contém, e então foi feita uma comparação das proporções relativas entre gêneros através do uso de um teste chi-quadrado. Esse teste é usado para determinar se há uma diferença estatisticamente significativa entre as frequências esperadas e frequências observadas em uma ou mais categoria de um conjunto de dados.

A partir disso, ordenamos os atributos pela maior diferença e selecionamos os 12 primeiros. Na Tabela 3 esses atributos podem ser vistos juntamente com a respectiva porcentagem de homens e mulheres que os possui.

Atributo	% Homens	% Mulheres
clubes	21,97	1,20
ano	20,40	1,24
posição	22,28	2,25
jogos	17,21	0,80
atualClube	16,88	0,01
paísNatal	18,13	1,90
anoSeleção	14,53	0,74
seleçãoNacional	14,85	0,01
partidasSeleção	13,99	0,64
ocupação	31,21	58,34
cônjuge	14,63	26,43
altura	25,63	14,19

Tabela 3. Porcentagem de Homens e Mulheres que possuem um atributo específico no seu *infobox*

É possível fazer algumas observações a partir da Tabela 3:

- Atributos *clubes*, *ano*, *posição*, *jogos*, *atualClube*, *anoSeleção*, *seleçãoNacional* e *partidasSeleção* são mais frequentes em homens. Note que todos esses atributos são relacionados com esportes, provavelmente muitos deles relacionados especificamente com futebol por terem a palavra "seleção". Nesse caso, a diferença entre os gêneros pode ser explicada pela maior presença de homens em classes relacionadas com esportes (por exemplo, *Futebolista* na Tabela 2).
- O atributo *altura* também pode estar correlacionado com esportes como vôlei e basquete, e é mais frequente em biografias de homens.
- O atributo *ocupação* é mais frequente em mulheres. 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 sua ocupação. Assim, por exemplo a classe *Futebolista* (que é composta por maioria masculina) não contém o atributo *ocupação*.

O atributo *cônjuge* é mais frequente em mulheres, mas não obtivemos uma explicação direta para isso, além da possível tendência a incluir esse atributo mais em biografias de mulheres. Isso levantou suspeitas de viés sobre a forma como as mulheres são apresentadas nas biografias.

5. Propriedades da Linguagem

Nessa seção nós exploramos a descrição de homens e mulheres na Wikipédia através de uma perspectiva léxica. Primeiramente, para analisar as diferenças entre o vocabulário usado para descrever pessoas dos dois gêneros, foi necessário computar o vocabulário comum entre os gêneros. Obtivemos os seguintes tamanhos para cada vocabulário: mulheres = 153.520, homens = 216.533 e interseção = 91.976. A partir disso, foi calculada uma medida de associação entre as palavras no vocabulário comum e cada um dos dois gêneros, a fim de explorar quais palavras são mais fortemente associadas a cada gênero. A medida utilizada foi o *Pointwise Mutual Information*, que é definido como:

$$\text{PMI}(g, w) = log\frac{p(g, w)}{p(g)p(w)}$$

onde g é o gênero (feminino ou masculino) e w é a palavra. As probabilidades foram estimadas a partir da proporção de biografias sobre homens e mulheres, e a correspondente proporção de palavras, que pode ser obtida computando o *document frequency*, que mede em quantos documentos da coleção a palavra apareceu pelo menos uma vez. O PMI dá mais peso a palavras com frequências menores, dessa forma, foi considerado apenas palavras que aparecem em pelo menos 1% das biografias de mulheres e homens.

Os resultados da associatividade das palavras estão apresentados nas seguintes nuvens de palavras, no qual quanto maior a palavra maior seu nível de associação ao gênero maior a palavra.

Figura 3. Nuvem de palavras com as palavras mais associadas a mulheres

Figura 4. Nuvem de palavras com as palavras mais associadas a homens

A seguir temos o top-15 das palavras mais associadas com cada gênero, e a frequência relativa em parênteses:

- Mulheres: compositora (7,88%), convidada (9,28%), feminista (4,53%), sepultada (4,41%), indicada (6,49%), autora (7,46%), escritora (11,69%), feminino (12,31%), eleita (10,24%), contratada (4,45%), apresentadora (5,87%), escalada (3,91%), feminina (10,50%), nascida (28,04%), casada (11,32%).
- Homens: rebaixamento (1,50%), eurocopa (1,26%), zagueiro (3,95%), artilheiro (2,73%), lateral-esquerdo (1,35%), rubo-negro (1,33%), lateral-direito (1,37%), reforço (2,67%), goleiro (3,41%), pênaltis (1,69%), gramados (1,54%), válido (2,26%), amistoso (3,01%), assistências (1,43%), ex-futebolista (5,31%).

As palavras mais associadas com homens são relacionadas a esportes, o futebol em particular, isso pode ser visto tanto na Figura 4 quanto no top-15 de palavras associadas a homens. Enquanto que, para as mulheres, as palavras mais associadas são relacionadas com o meio artístico (compositora, escritora, apresentadora) e com gênero (feminista, feminino, feminina). Há também presença da palavra *casada* no top-15 de palavras mais associadas com mulheres, isso chamou bastante atenção pois reforça a suspeita levantada anteriormente sobre o atributo *cônjuge* no *infobox* de mulheres (Tabela 3). Além disso, toda essa análise é consistente com os resultados obtidos por Eduardo, Mounia e Filippo [5], onde foram analisadas biografias da Wikipédia em inglês.

Propomos também uma outra análise léxica, com o propósito de determinar quais palavras são mais efetivas para distinguir o gênero da pessoa ao qual um texto se refere. Para tal, nos baseamos no calculo do viés léxico proposto por Claudia, David, Mohsen e Markus [3]. Nossa abordagem constitui-se, inicialmente, dos seguintes passos:

1. **Term frequency:** foi calculado o *term frequency* (frequência do termo em cada documento) para cada palavra de cada biografia, depois esse valor foi somado entre os documentos. Resultando em um dicionário em que as chaves são todas as

palavras que aparecem nos textos. Selecionamos apenas as palavras que aparecem no vocabulário de palavras comuns aos dois gêneros, e também excluímos palavras que possuem o *term frequency* muito baixo.

2. **Mutual information:** a partir das palavras selecionadas com a etapa anterior, foi feito o cálculo de dependência mútua entre duas palavras, de forma análoga à apresentada na seção 3.2.2.
3. **Term frequency - inverse document frequency:** foi computado o tf-idf para cada uma das palavras presentes nos textos.
4. **Dataset:** a partir do *mutual information* selecionamos as primeiras 1500 palavras para realizar a classificação e montamos um *dataset* preenchendo cada linha com tf-idf das palavras (representadas nas colunas) para a aquele documento.

Por fim, o *dataset* criado é usado para treinar um classificador *Naive Bayes*, que irá determinar quais as melhores palavras para diferenciar os gêneros. Usamos a função de log das probabilidades dos parâmetros do modelo para comparar as relações entre as variáveis. Essa função é dada por:

$$L(w, g) = log\frac{p(w|g)}{p(w)}$$

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 seguir temos as nuvens de palavras criadas a partir dessa análise:

Figura 5. Nuvem de palavras com as palavras mais efetivas para classificar mulheres

Figura 6. Nuvem de palavras com as palavras mais efetivas para classificar homens

Os resultados para essa análise foram bastante similares aos obtidos até então. As palavras que são mais efetivas para classificar mulheres são palavras que remetem ao ramo artístico (atriz, cantora, escritora, novela, cinema) e gênero (feminino, feminina). A palavra *casada* aparece novamente como uma das primeiras na nossa análise, como pode ser visto na Figura 5. Enquanto que, as palavras mais efetivas para distinguir homens são novamente relacionadas com esportes. Uma diferença significativa nessa análise, é que temos palavras relacionadas a ciclismo e corrida entre as mais importantes (Figura 6), diferentemente das outras análises em que temos apenas o futebol em destaque.

6. Conclusão e Trabalhos futuros

Nesse trabalho nós estudamos viés de gênero nas biografias da Wikipédia, fazendo uma análise de meta-dados e da linguagem utilizada. Aplicamos uma série de métodos para tentar quantificar e detectar viés de gênero no conteúdo da Wikipédia em português, de maneira que foi possível também comparar os resultados obtidos com pesquisas feitas na Wikipédia em inglês. A princípio os resultados não são suficientes para afirmar a existência do viés, mas indicam uma diferença significativa sobre a forma como mulheres são representadas na Wikipédia, reforçando a hipótese de que existe viés de gênero no conteúdo da mesma.

O principal desafio para a realização desse trabalho foi a criação de um bom método de inferência de gênero. Esse passo é muito importante para todas as análises, e quanto melhor a inferência mais significativos são os resultados obtidos. Dessa forma, como trabalho futuro buscaremos melhorar o método criado. Um outro desafio, foi o uso de textos apenas em português visto que existem poucas referências ao uso de métodos de processamento de linguagem natural para textos em português. Além disso, isso também representa uma limitação do nosso trabalho por inserir vieses culturais na base de dados.

Em todas as análises realizadas percebemos uma presença muito forte de palavras ligadas ao futebol em textos sobre homens. Isso mostra que, enquanto mulheres

possuem palavras e atributos relacionados a gênero e relacionamentos em sua biografia, homens possuem mais palavras relacionadas à sua profissão. Sendo também um indicativo forte de viés sobre a maneira como as pessoas estão sendo retratadas. Porém, essas observações levaram também à questionamentos sobre a falta de representatividade de pessoas notáveis na Wikipédia. Como grande parte das biografias de homens analisadas tratam-se de "futebolistas", e sabendo que a base é bastante desbalanceada (apenas 19,4% das biografias são sobre mulheres), é possível que haja viés de cobertura nessa enciclopédia. Assim, em trabalhos futuros pretendemos analisar quais personalidades notáveis brasileiras e estrangeiras são cobertas pela Wikipédia [3].

Além disso, a fim de avaliar melhor nossa hipótese de existência de viés de gênero na Wikipédia e buscar obter um resultado mais concreto que a análise inicial apresentada, nos trabalhos futuros temos a intenção de analisar os dados a partir de mais duas dimensões: imagens e estrutura da rede. É padrão nas biografias da Wikipédia a presença de uma imagem da pessoa. Pretendemos utilizar alguns recursos de processamento de imagem e computação visual para tentar investigar a presença de sexismo nessas imagens.

E finalmente, para a análise da estrutura da rede de biografias pretendemos construir um grafo de biografias [17] a partir dos links entre os artigos. Sobre isso, a ideia será estimar o *PageRank* (medida de centralidade de um nó baseado na conectividade da rede). Assim, podemos computar o viés estrutural comparando a importância dada pelo *PageRank* no grafo de biografias com os modelos nulos, ou seja, grafos que por construção não são enviesados mas que mantém certas propriedades de uma rede de biografias. Em contextos similares, o *PageRank* foi usado para prover uma aproximação de importância histórica [17, 18] e para estudar o viés gerado pelo *gender gap* [18].

Referências

[1] Collier, B. e J. Bear: *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, 2012.

[2] Schmahl, K. G., T. J. Viering, S. Makrodimitris, A. N. Jahfari, D. Tax e M. Loog: *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, página 94–103, 2020.

[3] Wagner, C., D. Garcia, M. Jadidi e M. Strohmaier: *It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia*. Ninth International AAAI Conference on Web and Social Media, 2015.

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

[5] Graells-Garrido, E., M. Lalmas e F. Menczer: *First Women, Second Sex: Gender Bias in Wikipedia*. Proceedings of the 26th ACM conference on hypertext, páginas 165–174, 2015.

[6] Church, K. W. e P. Hanks: *Word association norms, mutual information, and lexicography*. Computational linguistics, 16.1:22–29, 1990.

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

[8] Hinnosaar, M.: *Gender Inequality in New Media: Evidence from Wikipedia*. Journal of Economic Behavior & Organization, 163:262–276, 2019.

[9] Shaw, A. e E. Hargittai: *The pipeline of online participation inequalities: the case of Wikipedia editing*. J Commun, 68(1):143–168, 2018.

[10] *ptwiki dump progress on 20210101*. `https://dumps.wikimedia.org/ptwiki/20210101`. Data de acesso: 02/01/2021.

[11] *Lista de gols de Lionel Messi pela Seleção Argentina de Futebol*. `https://pt.wikipedia.org/wiki/Lista_de_gols_de_Lionel_Messi_pela_Sele%C3%A7%C3%A3o_Argentina_de_Futebol`.

[12] *wikiextractor PyPI*. `https://pypi.org/project/wikiextractor/`. Data de acesso: 05/01/2021.

[13] Bamman, D. e N. A. Smith: *Unsupervised Discovery of Biographical Structure from Text*. Transactions of the Association for Computational Linguistics, 2:363–376, 2014.

[14] Filho, Renato Miranda: *Um arcabouço para pesquisas de opinião em redes sociais*. Tese de Mestrado, Universidade Federal de Minas Gerais, 2014.

[15] *Nomes no Brasil*. `https://censo2010.ibge.gov.br/nomes/#/search`. Data de acesso: 15/01/2021.

[16] *gender-guesser PyPI*. `https://pypi.org/project/gender-guesser/`. Data de acesso: 26/01/2021.

[17] Aragón, P., A. Kaltenbrunner, D. Laniado e Y. Volkovich: *Biographical social networks on Wikipedia: a cross-cultural study of links that made history*. Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration. ACM, página 19, 2012.

[18] Skiena, S. S. e C. B. Ward: *Who's Bigger?: Where Historical Figures Really Rank*. Cambridge Univ. Press, 2014.