Identificando Padrões de Sexismo na Música Brasileira através do Processamento de Linguagem Natural

Vitória Pereira Firmino[1], **Janaina Nogueira de S. Lopes**[1], **Valéria Q. Reis**[12]

[1]Faculdade de Computação – Universidade Federal de Mato Grosso do Sul do Sul (UFMS)
Campo Grande – Brasil

[2]Instituto de Sistemas de Informação, Leuphana Universität Lüneburg
Lüneburg – Alemanha

{vitoria.firmino,janaina.nogueira,valeria.reis}@ufms.br

Abstract. *This paper presents an analysis of a corpus composed of* $138,368$ *Brazilian lyrics in order to identify gender biases. Natural Language Processing methods were used to define the set of adjectives most often used to characterize men and women in songs. The results show that the female gender is often described using words that refer to physical appearance, while the opposite gender is constantly described based on their good personality. Our results corroborate other studies in the literature and shed light on the discussion about the perpetuation of sexism in our society and the need for intervention to provide equal opportunities for women.*

Resumo. *Este trabalho apresenta uma análise de um corpus composto por* 138.368 *músicas brasileiras a fim de identificar vieses de gênero. Para isso, foram utilizados métodos de Processamento de Linguagem Natural para definir o conjunto de adjetivos mais utilizados para caracterizar homens e mulheres nas canções. Os resultados mostram que o gênero feminino é frequentemente descrito utilizando predicativos que remetem à aparência física, enquanto o gênero oposto é constantemente descrito a partir de sua boa personalidade. Nossos resultados corroboram outros estudos da literatura e jogam luz na discussão sobre a perpetuação do sexismo em nossa sociedade e na necessidade de intervenção para proporcionar equidade de oportunidade para as mulheres.*

1. Introdução

O Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial (IA) que se dedica ao desenvolvimento de algoritmos capazes de compreender, interpretar e gerar linguagem humana de forma natural [Caseli and Nunes 2024]. Desde sua criação até os avanços recentes impulsionados pelo Aprendizado de Máquina e pela Linguística Computacional, o PLN tem desempenhado um papel crucial na automatização de tarefas linguísticas complexas.

Ao discorrer sobre PLN, [Caseli and Nunes 2024] ressaltam que ao se processar dados textuais para interação com computadores, não se está simplesmente organizando um conjunto de caracteres ou ordenando uma representação ortográfica formal. No processamento de texto se faz presente também um contato com recursos linguísticos que representam a experiência humana. Quando essa experiência carrega vieses sociais, as

decisões tomadas com base nelas também são suscetíveis a tendências, muitas vezes discriminatórias.

Com os avanços da aplicação de IA, tem sido cada vez mais evidente a ocorrência de casos em que algoritmos levam a discriminação. Consequentemente, projetos com o intuito de apoiar o desenvolvimento de sistemas digitais justos começam a surgir. O projeto "Desvelar – justiça racial na Inteligência Artificial e TICs", por exemplo, mantém uma página em que artigos e notícias relacionadas a casos de discriminação algorítmica são dispostos [Desvelar 2024]. Na página há uma linha do tempo, na qual pessoas falantes de Português podem conhecer detalhes de acontecimentos que explicitaram o tratamento diferenciado dado por sistemas de recomendação a grupos vulneráveis. A título de ilustração dentro do contexto racial, pode-se citar os casos em que: a busca por "garotas negras" no Google resultava em conteúdo pornográfico; pessoas negras eram rotuladas como gorilas pelo aplicativo de fotos; e a análise de reincidência criminal prejudicava réus negros.

De acordo com [Barocas and Selbst 2016], a eficácia do uso de dados por sistemas computacionais está inerentemente relacionada à qualidade dos dados analisados. Se esses dados tiverem comportamento tendencioso, os modelos treinados a partir deles aprenderão com o mau exemplo que estas decisões estabeleceram. Assim, observa-se que uma decisão algorítmica é tão boa quanto os dados que a alimentam. Neste sentido, o viés pode se manifestar quando, por exemplo, análises são feitas e modelos são treinados em conjuntos de dados que refletem desigualdades existentes na sociedade. Isso pode afetar grupos específicos de maneira desigual, especialmente quando se trata de características como gênero, raça, idade, entre outras.

Utilizando modelos de representação textual, chamados de *Word Embeddings* (WE), [Caliskan et al. 2022] constatam diversas situações de preconceitos humanos na língua inglesa. Dentre algumas: 'mulheres estão mais associadas à família e homens, à carreira'; 'termos femininos estavam mais associados às artes e termos masculinos, às ciências'. Resultados semelhantes são obtidos em análises de um modelo para a língua Portuguesa [Taso et al. 2023a, Taso et al. 2023b]. Novamente, os autores concluem que se a IA aprende o suficiente sobre as propriedades da linguagem para ser capaz de a compreender e produzir, também adquire associações culturais que podem ser ofensivas, questionáveis ou prejudiciais.

A identificação de vieses através de métricas aplicadas a WE são amplamente empregadas, porém questionáveis devido à dificuldade de se separar o impacto do estereótipo social do que é simplesmente parte da similaridade entre palavras [Zhang et al. 2020]. Uma alternativa a esse tipo de método é realizar a análise com foco no padrão linguístico dos textos originais, assim como realizado em [Freitas and Martins 2023]. As autoras analisam um corpus composto por obras da literatura brasileira em domínio público, majoritariamente dos séculos XIX e XX. A análise dos dados foi feita utilizando métodos de PLN e os resultados revelam objetivamente uma construção estereotipada dos gêneros, com o feminino sendo caracterizado principalmente pela forte associação com o corpo, especificamente a beleza. O masculino, por sua vez, é constantemente apresentado em características de papel social e caráter, com muito menos destaque para a aparência.

Além da literatura, as expressões culturais em forma de música têm um enorme

peso no retrato contemporâneo e são muito úteis para estudar questões que se referem às relações de gênero, sobretudo na linguagem [Feijó and Macedo 2013]. Neste sentido, [Kong 1995] expressa que a música possui uma instigante estrutura: tanto como meio quanto como resultado da experiência, ela serve para produzir e reproduzir sistemas sociais. A autora reforça a ideia de que a música também é um meio através do qual as pessoas transmitem suas experiências do cotidiano e do extraordinário, trazendo à tona uma construção social, sentimentos e outros aspectos.

Em seu trabalho, [Duprat 2008] evidencia a representação da mulher na música brasileira, por muito sendo representada de maneira sexualizada e mais associada a adjetivos que remetem à beleza. O autor afirma que "...o poder do Homem tem se manifestado na Música pela produção. O da Mulher manifestou-se pela inspiração e pela sedução. É como que a consagração de um mundo em que o feminino não se oporia ao masculino, mas que o seduz".

Em consonância com estudos anteriores que identificaram vieses de gênero em diferentes contextos sociais, é hipotetizado que as mulheres são frequentemente retratadas de maneira estereotipada também em músicas, mais vinculadas a situações que atribuam às mulheres características relacionadas à sua aparência física e adjetivos gerais. Dessa maneira, este trabalho tem como objetivo demonstrar a existência de vieses relacionados às mulheres em músicas brasileiras de diferentes gêneros, a partir da análise de um corpus com aproximadamente 130 mil títulos.

Como resultado deste trabalho espera-se quantificar e qualificar os padrões de sexismo encontrados nas letras de canções brasileiras, contribuindo para a conscientização sobre a criação e persistência de estereótipos prejudiciais e preconceitos nas representações das mulheres na música nacional. A cultura popular, em especial a música, tem um impacto significativo na formação de construções sociais, percepção individual e na reprodução de sistemas sociais.

No Brasil, a quebra do sexismo é especialmente relevante, pois, de acordo com Instituto Brasileiro de Geografia e Estatística, as mulheres representam $51,5\%$ da população. Apesar disso, em 2022, elas ocupavam apenas 39% dos cargos gerenciais, os quais eram frequentemente centrados em ocupações relacionadas a cuidados e beleza. Essas questões reforçam a importância da discussão da desigualdade de gênero e da expressão e representatividade da mulher na sociedade brasileira. Além disso, é preciso jogar luz na qualidade e composição de dados sabidamente utilizados no treinamento de grandes modelos de linguagem, os quais estão sendo utilizados sem regulamentação na tomada de decisões que impactam direta e indiretamente a vida dos cidadãos. Nesse aspecto, é inquestionável o papel que o PLN tem dentro das Ciências Aplicadas, mais concretamente na área social, onde a análise de grandes quantidades de dados podem revelar inequalidades de tratamento e oportunidades.

2. Fundamentação Teórica

Nesta seção são apresentados alguns termos do PLN utilizados na implementação do trabalho.

Um **corpus** é um conjunto de dados linguísticos [Caseli and Nunes 2024]. Trata-se de uma coleção de textos que pode ser processada por computadores, sendo o material

que compõe um corpus coletado com propósito, por exemplo, de investigar ou explorar aspectos da linguagem.

A **tokenização** de palavras é parte do pré-processamento da análise linguística em PLN. Nela ocorre o processo de dividir um texto em palavras ou unidades menores denominadas *tokens*. A tokenização facilita a posterior análise do texto, permitindo que algoritmos e modelos processem informações utilizando o significado de cada palavra individualmente.

Part-of-Speech Tagging ou somente **POS-Tagging** (em tradução livre, marcação de classe gramatical) é uma técnica que atribui etiquetas, chamadas de *tags*, a cada palavra em um texto, indicando sua classe gramatical (como pronome, substantivo, verbo, adjetivo, etc).

No contexto de PLN, as **expressões regulares** são notações algébricas frequentemente utilizadas para encontrar padrões específicos em textos [Jurafsky and Martin 2023]. A **análise de dependência**, por sua vez, consiste em relações direcionadas entre palavras. Nela são analisadas e classificadas as relações de dependência entre as palavras, tipos de relações gramaticais e funções gramaticais desempenhadas. Salienta-se que na abordagem da análise de dependência "uma palavra é vista como subordinada a outra ou regida por ela, de acordo com relações sintáticas tais como sujeito-verbo; sujeito-objeto; verbo-objeto; coordenação; subordinação etc." [Caseli and Nunes 2024].

Uma **árvore de dependência** é uma representação gráfica da estrutura gramatical de uma frase, destacando as relações sintáticas entre as palavras. Cada nó na árvore representa uma palavra, e as setas unidirecionais indicam as dependências gramaticais entre elas. A Figura 1 apresenta a árvore de dependência da frase 'Ela é bonita'. A partir dela, observam-se as tags *PRON*, *AUX*, *ADJ* e as relação entre elas, onde *nsubj* informa que 'bonita' se refere ao sujeito 'ela' e *cop* se refere ao verbo 'ser', o qual indica uma relação entre o sujeito (ela) e o predicativo (bonita).

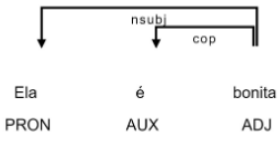

Figura 1. Exemplo de árvore de dependência.

3. Metodologia

O estudo fez uso de um corpus de música de acesso livre e bibliotecas de análise de linguagem para o Português. O código para fins de validação e replicação está disponibilizado online[1].

3.1. Aquisição do Corpus

O conjunto de dados utilizado neste estudo foi obtido a partir de um trabalho para classificação de gêneros musicais, no qual os autores [Lima et al. 2020] coletaram uma

[1]Código disponibilizado em: https://github.com/vitoriafirmino/busca-padroes-sexismo-com-pln

extensa variedade de letras de músicas escritas em Português. O conjunto de dados resultante consistiu em um total de 138.368 canções distribuídas em 14 estilos musicais representativos da música brasileira, contendo ainda os títulos e nomes dos artistas. Os estilos musicais incluem: Gospel, Sertanejo, MPB, Forró, Pagode, Rock, Samba, Pop, Axé, Funk-carioca, Infantil, Velha-guarda, Bossa-nova e Jovem-guarda.

Além disso, uma filtragem inicial do DataFrame foi conduzida para excluir letras do gênero gospel, visando focar a análise nos dados mais pertinentes aos objetivos do estudo. Especificamente, para analisar vieses de gênero, onde o gênero gospel não se aplica diretamente a um ser feminino ou masculino, optou-se por sua exclusão.

Até onde vai nosso conhecimento, este é o maior corpus de música disponível na Língua Portuguesa e suas características nos permitem diferentes análises sob o ponto de vista de gênero, temática, artistas e tempo.

3.2. Pre-processamento e Padrões de Busca

O Spacy foi a biblioteca de PLN escolhida para o pré-processamento das canções, devido principalmente aos seus reconhecidos métodos para segmentação e análise sintática. Foi utilizado o modelo pré-treinado de grande porte para processamento de linguagem natural em português.

No pré-processamento, cada letra de música foi dividida em frases para posterior verificação do teor de cada sentença. Buscavam-se frases que qualificassem homens e mulheres. Para isso, utilizaram-se padrões de busca nas sentenças.

Os padrões de busca foram elaborados por meio de um processo de tentativa e erro. Inicialmente, foram selecionadas amostras representativas de frases contendo adjetivos femininos e masculinos nas letras de música. Essas frases foram compiladas em um dicionário para servir como base de análise.

As frases do dicionário foram submetidas a uma análise detalhada, levando em consideração suas estruturas morfológicas e sintáticas. Esse processo envolveu a identificação das classes gramaticais das palavras e as relações sintáticas entre elas. Com base na análise inicial das frases do dicionário, foram formuladas hipóteses sobre os padrões de busca que poderiam capturar de forma precisa os adjetivos femininos e masculinos. Essas hipóteses foram testadas no próprio conjunto para avaliar sua eficácia.

Foram empregadas duas abordagens para a identificação de frases qualitativas: Expressões Regulares e Árvore de Dependência. Enquanto as Expressões Regulares permitem uma busca mais precisa com base em padrões específicos, a Árvore de Dependência possibilita a captura de relações sintáticas mais complexas entre as palavras. Essa combinação de técnicas visa aumentar a eficácia do processo de identificação dos adjetivos de interesse.

3.2.1. Expressões Regulares

Os padrões de busca foram elaborados com base em combinações de pronomes ou sujeitos pré definidos, verbos e adjetivos,. Cada padrão foi definido como uma lista de dicionários, onde cada dicionário especifica os *tokens* que devem ser encontrados na ordem especificada.

Portanto, o padrão a seguir é um exemplo de busca realizado, este identifica frases que iniciam com um sujeito (como ”ele”, ”homem”, ”menino”, etc.), seguido por um verbo auxiliar e, em seguida, um adjetivo. Por exemplo: ”Ele é ligeiro”. Essa abordagem é útil para encontrar padrões específicos de estrutura gramatical em um texto.

```
sujeito_aux_adj = [
    {"LOWER":
        {"IN": ["homem", "homens", "ele", "eles", "menino", "meninos","garoto",
        "garotos", "senhor", "senhores", "moços", "moços", "rapaz", "rapazes", "joão",
        "cavalheiro", "cavalheiros", "rei", "reis", "esposo", "esposos", "marido",
        "maridos", "namorado", "namorados"]}},
    {'POS': 'AUX'}, {'POS': 'ADJ'}]
```

3.2.2. Árvore de Dependência

Para cada segmento de música, a análise sintática das palavras foi realizada para detectar padrões específicos de relacionamento entre elas. Foram definidos os padrões de busca em forma de dicionários, representando os tipos de relacionamentos entre os elementos da frase.

Por exemplo, o início do padrão a seguir procura por um nó que representa um adjetivo, o qual é considerado como a raiz da árvore de dependência da sequência que se está buscando. A segunda parte do padrão especifica que o adjetivo encontrado deve estar relacionado a um sujeito por meio de uma relação de dependência chamada “NSUBJ”. Isso significa que o adjetivo descreve ou modifica o sujeito da sentença. Por fim, o padrão procura por um verbo auxiliar que esteja conectado à raiz da árvore de dependência (o adjetivo) por meio de uma relação de dependência. Portanto, esse padrão de busca identifica frases como: ”Ele é bom.”, ”Ele está sendo bom.”, ”Ele é o bom.”, ”Ele é muito bom.”, ”Ele é realmente muito bom.”.

```
sujeitoauxadjetivo = [
    {"RIGHT_ID": "adjetivo", "RIGHT_ATTRS": {"POS": "ADJ"}},
    {"LEFT_ID": "adjetivo", "REL_OP": ">", "RIGHT_ID": "sujeito", "RIGHT_ATTRS":
        {"DEP": "nsubj", "LOWER": {"IN": ["homem", "homens", "ele", "eles", "menino",
        "meninos", "garoto", "garotos", "senhor", "senhores", "moços", "moços", "rapaz",
        "rapazes", "joão", "cavalheiro", "cavalheiros", "rei", "reis", "esposo",
        "esposos", "marido", "maridos", "namorado", "namorados"]}}},
    {"LEFT_ID": "adjetivo", "REL_OP": ">", "RIGHT_ID": "auxiliar", "RIGHT_ATTRS":
        {"POS":"AUX"}}]
```

Extensões ao padrão de busca estão sendo feitas para contemplarem descrições que utilizem verbos no particípio. Tais estruturas também têm papel predicativo e podem indicar subordinação quando utilizadas na voz passiva.

3.3. Análise Quantitativa dos Adjetivos

Um código Python foi desenvolvido para contar e ranquear a frequência dos adjetivos. Os resultados foram armazenados em uma estrutura para análises posteriores.

4. Resultados

Por meio da análise por expressão regular, foram encontradas 8.691 sentenças relacionadas às mulheres e 5.705 associadas aos homens, resultando em 1104 adjetivos para o gênero feminino e 1048 para o masculino. A diferença observada foi de $52,34\%$ a mais de sentenças atribuídas à mulheres (2.986) em comparação com o gênero masculino.

Expressão Regular				
Feminino		Masculino		
Adjetivo	Ocorrência	Adjetivo	Ocorrência	Posição Feminino
bonita	956	feliz	284	7°
linda	634	bom	206	4°
louca	255	forte	123	61°
boa	225	capaz	111	34°
feia	191	bonito	92	1°
gostosa	150	triste	89	24°
feliz	147	louco	84	3°
top	127	diferente	83	14°
nova	116	casado	82	16°
brasileira	88	fiel	73	74°
Total de Sentenças:	8691	Total de Sentenças:		5705
Total de Adjetivos:	1104	Total de Adjetivos:		1048

Tabela 1. Ranking dos dez adjetivos mais frequentes encontrados por Expressão Regular.

A análise da árvore de dependência identificou 8.634 sentenças atribuídas às mulheres e 7.477 atribuídas aos homens. Os adjetivos para esses gêneros foram contabilizados em 1.123 e 1233, respectivamente. A diferença de sentenças observada foi de $15,47\%$ (1.157), ou seja, há mais fragmentos de músicas associadas ao gênero feminino.

A partir dos resultados de ambos os padrões de busca, nota-se a dominância da temática relacionada à mulher no contexto das músicas, sendo as mulheres mais adjetivadas do que os homens. Para determinar o grau dessa adjetivação, foi analisada o número de ocorrências para cada adjetivo. A Tabela 1 apresenta os resultados obtidos a partir de expressões regulares. A primeira e a terceira coluna apresentam os 10 adjetivos mais frequentemente encontrados nas músicas para os gêneros feminino e masculino, respectivamente. A segunda e a quarta coluna expressam o número de ocorrências para cada adjetivo. A quinta coluna apresenta para cada um dos top-10 adjetivos utilizados para a descrição de homens a respectiva posição do mesmo adjetivo flexionado para o gênero feminino.

É notável a existência de uma tendência de uso de adjetivos associados a características físicas e estereotipadas ao descrever mulheres e adjetivos mais relacionados a características de personalidade ao descrever homens. Observa-se que os adjetivos mais frequentes associados ao feminino são principalmente relacionados à beleza e ao aspecto físico, como ”linda”, ”bonita”e ”gostosa”, enquanto os adjetivos associados ao masculino estão mais relacionados ao caráter e aos aspectos emocionais, como ”feliz”, ”bom”, ”forte”e ”capaz”. Os adjetivos mais masculinos não são tão frequentes nas descrições femininas. Por exemplo, o adjetivo ”forte”foi utilizado 123 vezes. Para mulheres, esse adjetivo está ranqueado na posição 61^{a}, sendo utilizado apenas 23 vezes. ”Capaz”aparece apenas na posição 34^{a}. Há uma exceção para a qualidade ‘boa’, a qual aparece como a 4^{a} mais utilizada para descrever o gênero feminino. No entanto, na língua portuguesa coloquial, esse adjetivo também é empregado para descrever mulheres atraentes.

Árvore de Dependência				
Feminino		Masculino		
Adjetivo	Ocorrência	Adjetivo	Ocorrência	Posição Feminino
bonita	1039	feliz	541	7°
linda	638	bom	265	4°
louca	271	capaz	190	23°
boa	213	forte	143	85°
feia	188	triste	105	31°
gostosa	155	melhor	104	75°
feliz	121	diferente	101	14°
nova	115	louco	99	3°
bela	106	bonito	93	1°
top	104	sozinho	86	12°
Total de Sentenças:	8634	Total de Sentenças:		7477
Total de Adjetivos:	1123	Total de Adjetivos:		1233

Tabela 2. Ranking dos dez adjetivos mais frequentes encontrados por Árvore de Dependência.

As tendências observadas através das expressões regulares são corroboradas pela análise com árvore de dependência, cujos resultados são apresentados na Figura 2. Nota-se também uma predominância de adjetivos relacionados à aparência física no feminino, como ”linda”e ”bonita”, enquanto no masculino há uma variedade maior de adjetivos, que incluem não apenas aspectos físicos, mas também qualidades pessoais e emocionais, como ”melhor”e ”capaz”.

Há de se salientar que as ocorrências dos principais adjetivos femininos são substancialmente maiores do que as dos adjetivos masculinos. O reforço de um discurso é sabidamente uma prática capaz de moldar julgamentos. Por exemplo, o simples ato de ver um estímulo múltiplas vezes, aumenta a probabilidade de gostarmos desse estímulo.

Também é importante destacar que o adjetivo ”louco”foi reproduzido 99 vezes, enquanto ”louca”foi reproduzido 271 vezes, o que representa quase três vezes mais o uso no feminino do que no masculino. Isso revela disparidade no uso de certos adjetivos que descrevem comportamentos ou características emocionais, refletindo estereótipos de gênero que perpetuam desigualdades e preconceitos.

Por fim, é importante destacar ainda a presença do adjetivo ”nova”, que foi reproduzido 116 vezes, contribuindo para uma representação etarista das mulheres. Essa observação evidencia a percepção estereotipada das mulheres na música brasileira.

5. Trabalhos Correlatos

O trabalho de [Freitas and Martins 2023] caracteriza personagens dos gêneros masculino e feminino em textos literários. As autoras exploram um corpus de obras da literatura brasileira com o uso de padrões léxico-sintáticos para busca e classificações semânticas, observando predicação e organização quantitativa das ocorrências. Elas realizaram a distribuição dos predicadores por gênero e eixo (aparência, caráter, emoção e papel social). O trabalho atual se assemelha a este, pois segue o mesmo método

de busca para a caracterização do viés. Nossos resultados corroboram aqueles obtidos por [Freitas and Martins 2023], mesmo utilizando um corpus com diferenças relevantes de domínio, linguagem e tempo.

Nossos resultados relacionando mulheres à aparência física também vão ao encontro do exposto por [Taso et al. 2023a, Taso et al. 2023b] na análise de WE treinados a partir de dados obtidos principalmente em portais de notícias do Brasil. Os autores mostram que existe um sexismo de profissões no mercado brasileiro que se reflete no modelo estudado. Mulheres são frequentemente empregadas em ocupações que envolvem estética ou cuidados. Buscas preliminares nos padrões de sentenças que encontramos evidenciam que também há um sexismo de profissões nas letras de músicas. A confirmação desse resultado tem contribuição não somente para estudos sociais, mas também para a validação da métrica de identificação de viés em WE utilizada em [Taso et al. 2023a, Taso et al. 2023b].

Por fim, [Salles and Pappa 2021] analisaram o viés de gênero nas biografias da Wikipédia em Português, comparando a representação de homens e mulheres em duas dimensões: meta-dados e linguagem. A pesquisa revelou diferenças significativas na forma como as mulheres são retratadas em comparação com os homens. Apenas 16% das biografias da Wikipedia são referentes a mulheres. Além disso, as mulheres descritas nas páginas são mais notáveis na média do que os homens, o que sugere que mulheres devem ter feitos vultosos para serem retratadas na enciclopédia. Na análise das palavras mais associadas a cada gênero, observa-se que os homens são relacionadas a esportes. Para as mulheres, as palavras mais associadas são relacionadas com o meio artístico. Profissões tradicionalmente dominadas por mulheres, tal como enfermagem, têm pouca expressão na Wikipedia. As autoras destacam que os relacionamentos amorosos das mulheres são mais frequentemente discutidos. Nossas conclusões não estão diretamente relacionadas aos resultados obtidos por [Salles and Pappa 2021], mas apontam para a mesma direção no tratamento subalterno dado a mulheres na sociedade e perpetuado nos meios digitais.

6. Limitações do Trabalho

Nosso trabalho apresenta limitações, principalmente relacionadas à biblioteca de PLN, utilizada. As características e os algoritmos específicos da biblioteca podem influenciar diretamente a qualidade e a profundidade das informações extraídas das letras das músicas. A segmentação das músicas, bem como a análise sintática das frases, são processos fundamentais para a compreensão do conteúdo textual. No entanto, essas etapas podem ser desafiadoras, especialmente quando lidamos com letras de músicas, que muitas vezes apresentam estruturas linguísticas não convencionais e criativas. Como resultado, a precisão da extração de informações pode ser comprometida, impactando a interpretação dos dados. Além disso, nossa análise se restringe aos gêneros femininos e masculinos.

Ao reconhecer essas limitações, buscamos fornecer uma análise cuidadosa e transparente, destacando as áreas onde os resultados podem ser interpretados com cautela e apontando direções para futuras investigações que possam abordar essas questões de maneira mais abrangente e aprofundada.

7. Conclusões

Este trabalho tinha como objetivo identificar vieses de gênero oriundos do cenário musical brasileiro. Após a execução de experimentos baseados em buscas de padrões utilizando POS-tagging e árvore de dependência, os resultados confirmaram a existência de uma representação estereotipada das mulheres na música brasileira, onde elas são frequentemente descritas com base em sua aparência física, enquanto os homens são retratados de forma mais diversificada, muitas vezes focando na descrição de uma boa índole. Essa discrepância reflete e reforça normas de gênero e expectativas sociais, destacando a importância de análises críticas sobre representações de gênero na cultura popular.

A discussão sobre viés histórico é crucial, especialmente considerando o contexto demográfico do Brasil, onde as mulheres constituem uma parcela significativa da população e enfrentam desafios persistentes em termos de participação no mercado de trabalho, cargos de liderança e diferenças salariais. Destacar essas disparidades é fundamental para promover uma sociedade mais igualitária e para ampliar a representatividade e expressão das mulheres na sociedade.

Para trabalhos futuros apontam-se naturalmente análises de vieses por gênero musical e períodos de tempo. Vislumbra-se também uma anotação manual de parte das sentenças descritivas presentes no corpus. A rotulagem permitiria uma compreensão mais detalhada das nuances e sutilezas das emoções expressas nas letras e tornaria viável o treinamento de um modelo para classificação automática das sentenças referentes à existência de vieses. Além disso, sugere-se uma comparação dos resultados obtidos por meio das técnicas de PLN com os referentes a métricas de identificação de vieses utilizando WordEmbeddings.

Por fim, é importante destacar o papel que a Computação tem na análise social de diferentes domínios. Através do PLN um grande volume de dados são processados e analisados, trazendo luz a injustiças muitas vezes cometidas de forma velada. É preciso urgência no reconhecimento de vieses para que intervenções de representação de dados e algoritmicas possam ser propostas e validadas a fim de evitar que discrepâncias de tratamento de diferentes origens sejam perpetuadas por modelos que amparam os sistemas de apoio à decisão.

Agradecimentos

O presente trabalho foi realizado com apoio da Universidade Federal de Mato Grosso do Sul e da Leuphana Universität Lüneburg.

Referências

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3):671–732.

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., and Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22. ACM.

Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.

Desvelar, S. (2024). Danos e discriminação algorítmica: Mapeamento. Desvelar Justiça racial, IA e tecnologias digitais. Acesso em: 22/03/2024.

Duprat, R. (2008). Fruição, sedução e produção: o papel da mulher na música. *Música em Perspectiva*, 1(1).

Feijó, M. and Macedo, R. M. S. d. (2013). Gênero, cultura e rede social: a construção social da desigualdade de gênero por meio da linguagem. *Nova Perspectiva Sistêmica*, 21(44):21–34.

Freitas, C. and Martins, F. (2023). Bela, recatada e do lar: o que a mineração de textos literários nos diz sobre a caracterização de personagens femininas e masculinas. *Fórum Linguístico*, 20:9118–9138.

Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing*, volume 3.

Kong, L. (1995). *Popular Music in Geographical Analyses*. Progress in Human Geography, 15th edition.

Lima, R. A., de Sousa, R. C. C., Barbosa, S. D. J., and Lopes, H. C. V. (2020). Brazilian lyrics-based music genre classification using a BLSTM network. *CoRR*, abs/2003.05377.

Salles, I. and Pappa, G. (2021). Viés de gênero em biografias da wikipédia em português. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 211–216, Porto Alegre, RS, Brasil. SBC.

Taso, F., Reis, V., and Martinez, F. (2023a). Discriminação algorítmica de gênero: Estudo de caso e análise no contexto brasileiro. In *Anais do IV Workshop sobre as Implicações da Computação na Sociedade*, pages 13–25, Porto Alegre, RS, Brasil. SBC.

Taso, F., Reis, V., and Martinez, F. (2023b). Sexismo no brasil: análise de um word embedding por meio de testes baseados em associação implícita. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 53–62, Porto Alegre, RS, Brasil. SBC.

Zhang, H., Sneyd, A., and Stevenson, M. (2020). Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 759–769, Suzhou, China. Association for Computational Linguistics.