

A Caracterização de Mulheres na Música Brasileira: Identificando Padrões com Processamento de Linguagem Natural

Vitória Pereira Firmino¹, Janaina Nogueira de S. Lopes¹, Valéria Q. Reis¹²

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

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

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

Resumo. *Este trabalho investiga a presença de vieses de gênero nas letras de músicas brasileiras desde meados do último século. As análises, obtidas por meio de uso de técnicas de PLN, revelam que as mulheres são frequentemente descritas de maneira estereotipada e baseada principalmente em suas aparências físicas. Quando descritas com base no caráter e emoção, a ênfase está em adjetivos que remetem à instabilidade emocional. Além disso, em alguns estilos musicais, as ações relacionadas às mulheres estão frequentemente ligadas ao ato sexual. Os resultados obtidos na análise de profissões vão ao encontro de dados sociais disponibilizados por institutos de pesquisa do Brasil. As conclusões deste trabalho são pioneiras na área musical, mas corroboram estudos de identificação de sexismo realizados anteriormente em outros domínios, evidenciando a necessidade de uma reflexão crítica sobre a perpetuação da discriminação de gênero na sociedade.*

1. Introdução

O preconceito de gênero, também conhecido como sexismo, refere-se ao julgamento, consciente ou não, de um indivíduo baseado unicamente em seu gênero [Moss-Racusin et al. 2012, Sun et al. 2019]. Essa forma de discriminação é amplamente documentada, inclusive no Brasil, onde indicadores de monitoramento do [IBGE 2024], disponibilizados em 2024, revelam desigualdades expressivas entre homens e mulheres. As mulheres representam 51,5% da população, porém ocupam 39,3% dos cargos gerenciais. Além disso, a representatividade feminina está concentrada nos grupos ocupacionais que demandam atenção às necessidades e ao bem-estar, tais como Educação, Saúde e Serviços Sociais. A pesquisa aponta ainda que as mulheres recebem 78,8% do salário dos homens para o mesmo cargo. Esses dados demonstram as dificuldades enfrentadas pelas mulheres em sua inserção em ambientes profissionais e educacionais diversos, evidenciando a persistência de barreiras de gênero na sociedade brasileira.

A mídia e a cultura desempenham um papel fundamental na luta contra a desigualdade de gênero. No entanto, alguns veículos de comunicação e manifestações culturais carregam vieses históricos de limitação da representatividade feminina, enfatizando as características físicas e emocionais das mulheres em detrimento de suas qualificações profissionais ou caráter. Em seu estudo, [Salles and Pappa 2021] analisam o viés de gênero nas

biografias da Wikipédia em Português, onde apenas 16% das biografias da Wikipédia são referentes a mulheres. Os preconceitos se manifestam também através da objetificação feminina, a qual exalta a aparência.

Em [Freitas and Martins 2023], os autores analisam um corpus de obras da literatura em língua portuguesa, disponibilizado em domínio público. A análise dos dados foi feita utilizando métodos de Processamento de Linguagem Natural (PLN) e os resultados revelam uma construção estereotipada dos gêneros. O feminino é caracterizado principalmente por uma forte associação com o corpo, especialmente com a beleza. Em contraste, o masculino é apresentado com ênfase nas características de papel social e caráter.

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 que a música também é um meio através do qual as pessoas transmitem suas experiências, trazendo à tona uma construção social, sentimentos e outros aspectos.

Diversos autores, por meio de análises manuais, já destacaram a presença de preconceitos, sexismo e violência de gênero em músicas brasileiras [Sanchez 2009, Assis 2014, Schlösser and Fantin 2022]. No entanto, abordagens automatizadas capazes de processar e avaliar grandes volumes de dados ainda são raras, especialmente no contexto da língua portuguesa, sendo a maioria dos estudos concentrada em músicas em língua inglesa.

O trabalho de [Firmino et al. 2024] demonstra, por meio do uso de técnicas de PLN, que existe uma representação estereotipada das mulheres na música brasileira, onde elas são mais adjetivadas e frequentemente descritas com base em sua aparência física. Em contrapartida, os homens são retratados com ênfase em uma boa índole. Os resultados obtidos instigaram as seguintes perguntas que são o cerne de nossa pesquisa: i) *Como é a distribuição dos adjetivos utilizados para caracterizar as mulheres?*; ii) *Os adjetivos empregados para caracterizar mulheres são utilizados de forma positiva ou negativa?*; iii) *Quais são as ocupações mais citadas para cada gênero?*; iv) *Os estereótipos de gênero mudam de acordo com o estilo musical?*; e v) *Qual proporção das músicas analisadas apresentam sexismo?*.

Em consonância com os estudos descritos, este trabalho levanta a hipótese de que estereótipos sexistas contra as mulheres existem no cenário musical brasileiro e variam de acordo com os gêneros musicais. O objetivo central é demonstrar a presença desses vieses utilizando técnicas de PLN que automatizam tarefas linguísticas complexas e com grande quantidade de dados. Os objetivos específicos consistem em responder as perguntas anteriormente descritas. Para isso, foi construído um corpus com 146.612 músicas escritas em português, divididas em 19 gêneros estilos musicais. Uma filtragem do corpus obteve 30.304 sentenças que descreviam a caracterização feminina e a masculina. Com base nesses conjuntos, foram conduzidas diversas análises, incluindo cálculo de frequência, categorização e análise de sentimentos.

Os resultados obtidos quantificam e qualificam padrões de sexismo encontrados

nas letras de canções brasileiras com atenção ao gênero das composições, indo além do senso comum e permitindo uma análise sistemática e abrangente da linguagem utilizada nas músicas. Foram identificadas tendências negativas nos adjetivos utilizados na descrição de mulheres. Também ficou evidente a limitação da caracterização feminina no campo profissional. Uma última contribuição deste trabalho para a promoção da igualdade de gênero consiste na disponibilização de um modelo de análise de sentimentos capaz de identificar linguagem sexista.

O restante deste artigo está organizado em seções conforme descrito a seguir. Na Seção 2, são abordados os principais conceitos utilizados na implementação deste estudo. Na Seção 3 são discutidos os trabalhos relacionados, oferecendo um panorama do estado da arte e sua comparação com este estudo. Os recursos empregados e a metodologia adotada em nossos experimentos são apresentados nas Seções 4 e 5, respectivamente. Os resultados obtidos são detalhados na Seção 6. A Seção 7 traz uma análise sobre o preconceito de gênero observado na sociedade e os impactos do uso de dados advindos dessa entidade no aprendizado de máquina. Perspectivas de trabalhos futuros são apresentados na Seção 8. Finalmente, as conclusões são discutidas na Seção 9.

2. Fundamentação teórica

Nesta seção são apresentados conceitos técnicos utilizados na implementação do trabalho.

2.1. Processamento de Linguagem Natural

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]. Para isso, são utilizadas coleções de textos, coletados com um propósito comum e que podem ser processadas por computadores. A uma coleção de dados linguísticos, denomina-se **corpus**.

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 relações 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 a predicação (bonita).

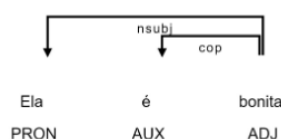


Figura 1. Exemplo de árvore de dependência.

Técnicas de PLN são naturalmente empregadas em análises automatizadas de grandes volumes de dados a fim de identificar padrões de linguagem, fenômenos literários, sociais e culturais. Esse tipo de análise literária que se concentra em observar tendências amplas e padrões estruturais em vastas quantidades de textos ao invés de analisar minuciosamente uma única fonte de informação é chamada **leitura distante**[Moretti 2008].

2.2. Aprendizado de Máquina

Aprendizado de Máquina é uma área da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos que permitem aos computadores identificar padrões em dados e tomar decisões baseadas nesses padrões, sem serem explicitamente programados para essa tarefa. Os algoritmos aprendem a partir de exemplos e experiências anteriores, ajustando-se automaticamente para melhorar seu desempenho ao longo do tempo[Bishop 2006].

Conforme exposto por [Caseli and Nunes 2024], diversas fontes são utilizadas como base de conhecimento para “ensinar” as máquinas. Com o avanço na capacidade de memória e processamento dos computadores, tornou-se viável o tratamento de volumes cada vez maiores de dados por meio de estruturas complexas. Um **modelo**, nesse contexto, é definido como uma simplificação de um fenômeno intrincado — no caso do PLN, a própria linguagem — para que seja representado por ferramentas computacionais. Embora busque capturar nuances do fenômeno real, seu objetivo não é substituir a realidade, mas sim oferecer uma representação que contribua para o entendimento e a solução de tarefas, respeitando as especificidades da linguagem.

Referente à **análise de sentimento** (AS), [Caseli and Nunes 2024] apontam que a área estuda aspectos como as opiniões, sentimentos e emoções em relação a entidades como produtos, serviços, tópicos e etc. Busca-se, então, encontrar soluções computacionais para extrair e analisar as opiniões das pessoas sobre uma entidade e seus diferentes aspectos. Como se trata de uma tarefa de classificação de texto e essas opiniões podem ser categorizadas de acordo com suas polaridades, como neutro, positivo ou negativo, várias abordagens e modelos podem ser utilizados.

Como elencado pelos autores do livro[Caseli and Nunes 2024], no campo dos modelos de linguagem, o **BERT** (Bidirectional Encoder Representations for Transformers) se destaca como um dos mais relevantes. Um ponto importante sobre o BERT é que ele segue um vocabulário específico da língua inglesa. Embora o processo de tokenização permita identificar algumas palavras de outros idiomas, todo o aprendizado do modelo é fundamentado nas regras e características linguísticas e culturais do inglês. Não apenas o vocabulário é relevante em um idioma, mas também as suas regras sintáticas, os significados contextuais e até aspectos culturais da sociedade onde a língua é falada. Em busca de fornecer bases sólidas para futuras pesquisas de NLP, um modelo de linguagem adaptado ao português, foi criado: o BERTimbau[Souza et al. 2020]. Esse modelo, cujo nome faz alusão ao instrumento musical brasileiro berimbau, foi desenvolvido a partir da arquitetura do BERT, mas com foco na língua portuguesa¹.

Entre outros termos empregados no texto para descrever o treinamento de modelos de classificação, destacam-se:

- **Dataset:** para [Caseli and Nunes 2024] pode ser descrito como um conjunto de dados estruturados e anotados, que serve para auxiliar o aprendizado de máquina e a avaliação de modelos de PLN. Diferentemente de um corpus, um dataset é preparado com exemplos cuidadosamente anotados, o que permite que modelos aprendam de maneira eficiente e sejam avaliados com base na comparação entre o desempenho das máquinas e a compreensão humana.
- **Fine-tuning:** é o processo de ajuste fino de um modelo previamente treinado para adaptá-lo a novas tarefas específicas, utilizando uma quantidade menor de dados. Esse método otimiza parâmetros do modelo, permitindo a adaptação eficiente sem a necessidade de treinamento completo desde o início[Howard and Ruder 2018, Aghajanyan et al. 2020]
- **Validação cruzada:** A técnica avalia o desempenho de um modelo de forma robusta, dividindo os dados em k subconjuntos (*folds*). O modelo é treinado em $k - 1$ subconjuntos e testado no restante, repetindo esse processo para cada *fold*. Isso garante que todos os dados sejam usados para treino e teste, proporcionando uma avaliação mais confiável e generalizada.²
- **Validação cruzada estratificada:** É uma variação da validação cruzada *K-Fold* que garante que a proporção de amostras de cada classe seja preservada em cada *fold*, mantendo a distribuição original das classes no conjunto de dados durante a divisão de treino e teste³.

3. Trabalhos correlatos

A análise de corpus com o auxílio de PLN é crucial para detectar e mitigar vieses em dados textuais. O trabalho de [Freitas and Martins 2023] combina esse método com a leitura distante para caracterizar 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

¹<https://github.com/neuralmind-ai/portuguese-bert>

²https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.KFold.html

³https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.StratifiedKFold.html

predicação e organização quantitativa das ocorrências. Elas realizaram a distribuição dos predicadores por gênero e categorias, denominados eixos, (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.

Nossas conclusões relacionando mulheres à aparência física também vão ao encontro do exposto por [Taso et al. 2023a] na análise de WE treinados a partir de dados obtidos principalmente em portais de notícias do Brasil. Em [Taso et al. 2023b], 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].

[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 figuras femininas. Além disso, as mulheres descritas nas páginas são mais notáveis na média do que os homens, o que sugere que pessoas do gênero feminino 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.

Por fim, o trabalho de [Benatti et al. 2024] investiga o preconceito de gênero em decisões judiciais brasileiras, com enfoque em casos de violência. Utilizando técnicas de PLN, foi desenvolvido um sistema automatizado para detectar preconceitos de gênero em decisões emitidas em português. Assim como em nosso trabalho, a metodologia envolveu a coleta e anotação de dados e treino de modelo de aprendizado. Os resultados indicaram que a abordagem automatizada ampliou a capacidade de análise, permitindo a identificação de sexismo de forma mais eficiente do que os métodos tradicionais. A análise em larga escala forneceu perspectivas sobre como esses preconceitos afetam o julgamento de casos, especialmente no tratamento de vítimas de violência. O estudo reforça a importância da combinação entre técnicas automatizadas e julgamento humano para garantir que decisões jurídicas sejam mais justas e menos influenciadas por estereótipos.

3.1. Sexismo nas músicas

Em [Schlösser and Fantin 2022], por meio de entrevistas, os autores identificam que as letras musicais brasileiras frequentemente apresentam conteúdos sexuais explícitos que

reforçam a objetificação da mulher e perpetuam estereótipos de gênero. A maioria das mulheres entrevistadas expressou descontentamento em relação à forma como a figura feminina é retratada nessas canções, descrevendo-as como expressões de uma cultura patriarcal e machista. Relataram sentir-se desrespeitadas e invadidas ao ouvir músicas que sexualizam e desvalorizam a mulher, especialmente em gêneros como funk e forró, que foram associados a uma erotização explícita do corpo feminino. Elas identificaram que esses conteúdos sexuais reforçam a submissão feminina e legitimam comportamentos masculinos abusivos e violentos, contribuindo para a normalização de situações de violência de gênero nas relações interpessoais. Além disso, o estudo aponta que, embora algumas mulheres não considerem essas letras ofensivas, a maioria das entrevistadas associa esses conteúdos à perpetuação de preconceitos de gênero, sublinhando a necessidade de continuar lutando por igualdade de direitos e respeito às mulheres na sociedade.

Sob uma perspectiva psicológica e social, o estudo de [Assis 2014] revela como as letras de músicas brasileiras influenciam e reforçam dinâmicas emocionais e comportamentais associadas à violência de gênero. O autor destaca que as representações sociais de amor nessas canções frequentemente mesclam afetos positivos com agressividade, perpetuando a ideia de que o conflito e a violência são componentes naturais das relações amorosas. Assis argumenta que essa ambivalência emocional nas letras musicais, que mistura “amor e ódio” ou “carinho e violência”, colabora para a normalização da violência de gênero. Além disso, o estudo mostra que essas representações ajudam a solidificar uma visão de papéis de gênero desiguais, especialmente ao reforçar a submissão feminina e o domínio masculino. A romantização da violência nas músicas, segundo o autor, legitima comportamentos agressivos e controladores, contribuindo para a manutenção de uma cultura patriarcal que naturaliza a violência contra a mulher como parte do relacionamento amoroso.

Em [Firmino et al. 2024], por meio do uso de técnicas de Processamento de PLN, foi realizada uma análise de um corpus de músicas brasileiras para identificar vieses de gênero. O estudo revelou que as mulheres são frequentemente descritas com adjetivos relacionados à aparência física, como “bonita”, “linda” e “gostosa”, enquanto os homens são retratados principalmente com adjetivos que destacam traços de caráter e habilidades, como “feliz”, “forte” e “capaz”. Esses resultados evidenciam uma construção estereotipada de gênero nas letras de músicas, onde as mulheres são objetificadas e valorizadas pela aparência, e os homens são representados por suas qualidades pessoais e sociais.

Em [Betti et al. 2023], também a partir de técnicas de PLN, realizou-se a análise de 377.808 letras de músicas em inglês, com foco na expressão do sexismo e na medição de preconceitos de gênero ao longo de cinco décadas (1960–2010). Os resultados demonstraram que as letras de artistas masculinos solos se tornaram progressivamente mais sexistas ao longo dos anos, enquanto esse comportamento foi menos perceptível para outras categorias de artistas. Além disso, os vieses linguísticos revelaram que as letras de artistas masculinos continham mais preconceitos de gênero do que as de artistas femininas, especialmente no que se refere aos estereótipos de homens como fortes e focados no sucesso, e mulheres como mais ligadas à família.

No trabalho de [Huang 2022] foram analisadas mais de 237 mil músicas em inglês de diferentes gêneros entre as décadas de 1.970 e 2.010, utilizando representações de palavras como base. Foram aplicadas técnicas de embeddings dinâmicos de palavras,

treinados com BERT, a fim de identificar mudanças nos sentidos e nas conotações desse termo ao longo do tempo. A análise destacou a diversificação dos significados de *bitch*, com o termo sendo usado tanto de forma positiva quanto negativa. Artistas de diferentes gêneros passaram a usar o insulto para diferentes fins, refletindo tanto a aceitação como a contestação de normas patriarcais. Entre as descobertas, foi identificado um viés claro no uso do termo por artistas homens, que frequentemente usavam *bitch* de forma pejorativa para descrever mulheres, enquanto artistas mulheres o reapropriaram para significar empoderamento feminino.

Não foi possível encontrar no contexto brasileiro trabalhos similares a [Betti et al. 2023, Huang 2022].

4. Recursos de pesquisa

Todas as análises descritas neste trabalho são feitas considerando um grande corpus de músicas e as sentenças nele contidas, as quais qualificam homens e mulheres. Esta seção apresenta os métodos utilizados na criação deste corpus e na identificação de suas sentenças predadoras.

4.1. Corpus e coleta de dados

Para o desenvolvimento da pesquisa, utilizamos quatro corpora de músicas brasileiras previamente disponibilizados na Internet: *Song Lyrics from 79 Musical Genres*⁴, *Bossa Nova Lyrics*⁵, *Funk Carioca Dataset*⁶ e o *Corpus Vagalume* [Lima et al. 2020]. Esses corpora foram aglutinados com o corpus *MusicSpotsBr*, desenvolvido exclusivamente para este estudo. Chamaremos o corpus resultante das agregações de *Corpus União*.

Durante o processo de unificação de *datasets*, as colunas dos diferentes corpora foram padronizadas e músicas duplicadas foram removidas. Canções do gênero Gospel também foram desconsideradas, pois apresentavam um padrão onde o pronome “Ele” e adjetivos masculinos eram majoritariamente utilizados para referenciar a Deus. Ao final, o *Corpus União* resultante continha um total de 146.612 canções, contendo as colunas: *Nome da música*, *Artista*, *Gênero da Música* e *Letra da Música*.

4.1.1. O corpus MusicSpotsBr

A fim de enriquecer o *Corpus União* com músicas mais recentes e também apresentar informações a respeito da popularidade e data de lançamento das músicas, foi criado o *MusicSpotsBr*, um corpus das músicas brasileiras mais tocadas e baixadas desde o ano de 1.947. Os dados foram extraídos do Spotify⁷, o maior serviço de *streaming* de música no Brasil, o qual implementa um modelo de negócios com opções gratuitas e pagas. Foram selecionadas 55 das *playlists* mais populares da plataforma, garantindo uma amostra diversificada e representativa. No total, o corpus apresenta 1.627 letras distribuídas em 19 gêneros musicais distintos, que podem ser vistos na Figura 2.

⁴<https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres>

⁵<https://www.kaggle.com/datasets/mcarujo/bossa-nova-lyrics/data>

⁶<https://www.kaggle.com/datasets/sleshes/funk-carioca-dataset>

⁷<https://open.spotify.com/>

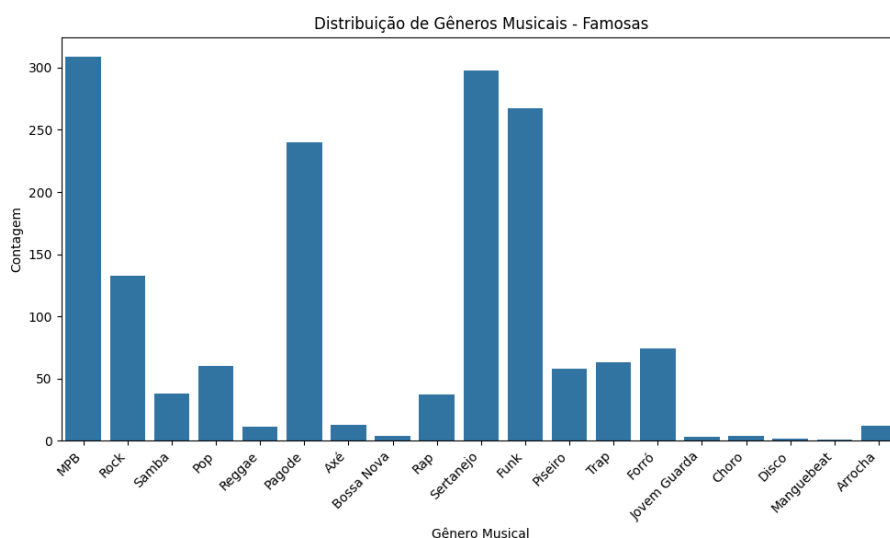


Figura 2. Distribuição das músicas por gênero musical

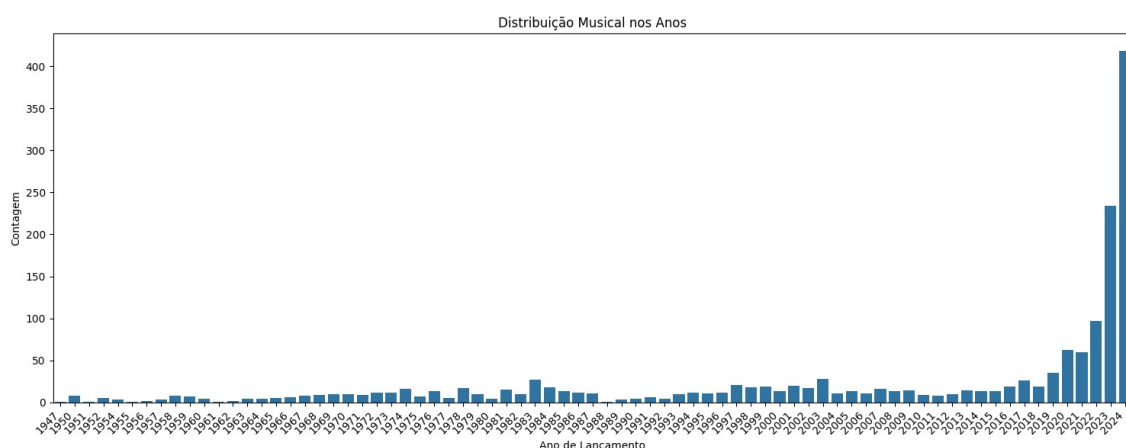


Figura 3. Distribuição das músicas por ano de lançamento

A extração das letras das músicas foi realizada por meio da integração entre a API do Spotify e a API do Vagalume, um dos principais repositórios de letras de músicas em português. Essa integração permitiu a coleta automatizada das canções, assegurando a consistência dos dados. Músicas em idiomas diferentes do português foram excluídas, concentrando a coletânea exclusivamente na produção musical na língua portuguesa. A Figura 3 apresenta a distribuição das músicas quanto ao ano de lançamento. O corpus foi estruturado em seis colunas: *Nome da Música*, *Artista*, *Ano de Lançamento*, *Gênero da Música*, *Playlist* e *Letra da Música*.

4.2. Pré-processamento e padrões de busca

O Spacy⁸ 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.

⁸<https://spacy.io/>

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 empírico. 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. Foram obtidas 15.223 sentenças com predições femininas e 15.081 sentenças com predições masculinas.

4.2.1. Expressões regulares

O *Matcher* é uma ferramenta de correspondência baseada em padrões de *tokens*. Ela utiliza padrões definidos por listas de dicionários, onde cada dicionário representa um token com atributos como POS, LEMMA, LOWER entre outros. O *Matcher* verifica se a sequência de tokens no texto atende a todas essas características na ordem especificada.

No total, 104 padrões de busca foram elaborados utilizando uma combinação de pronomes, sujeitos, verbos, advérbios, adjetivos, determinantes e substantivos, para identificar diferentes estruturas gramaticais dentro das frases. Cada padrão é definido como uma lista de dicionários, onde cada dicionário especifica os tokens que devem ser encontrados na ordem determinada para atender a uma estrutura gramatical específica. Essa organização permite analisar com precisão variações de construções linguísticas, possibilitando a detecção de padrões e suas variações complexas.

Além disso, os padrões de busca incluem estruturas com múltiplos verbos auxiliares e advérbios intercalados e permitem a análise de construções com nomes próprios e aposto descritivo. Esse detalhamento garante uma cobertura mais ampla e a possibilidade de identificar não apenas adjetivos isolados, mas também o papel de advérbios de intensidade e os efeitos de construções que envolvem nomes próprios e sujeitos múltiplos. Alguns exemplos de padrões de busca são:

Sujeito/Nome Próprio/Você/Nome Próprio/Você + Auxiliar + Adjetivo/Substantivo

Sujeito/Nome Próprio/Você + Auxiliar + muito/pouco + Adjetivo/Substantivo

Sujeito/Nome Próprio/Você + Auxiliar + Auxiliar + Adjetivo/Substantivo

Os padrões listados retornariam sentenças do tipo:

```
Você é/está bonita.  
Você é puta/vaca.  
Você é/está muito/pouco bonita.  
Você tem sido desobediente.  
Você está sendo educada.
```

Um dos pontos fortes da implementação é a capacidade de extração específica de adjetivos, substantivos e verbos no particípio. Para alcançar essa análise, utilizam-se padrões de busca baseados em características morfológicas dos tokens, como gênero, número, forma nominal do verbo e voz do verbo. Por exemplo, a busca por uma sequência do tipo

```
Sujeito/Nome Próprio/Você + Auxiliar(Ser/Estar) + verbo no particípio
```

retornaria frases do tipo

```
Ela foi absolvida.
```

4.2.2. Árvore de dependência

O *DependencyMatcher* é uma ferramenta que utiliza a árvore de dependência sintática para buscar padrões com base em como as palavras se relacionam dentro de uma frase. Ao contrário do *Matcher*, ele permite capturar informações complexas sobre as conexões hierárquicas entre termos.

O *DependencyMatcher* usa padrões definidos por listas de dicionários que representam nós da árvore de dependência. Cada nó é identificado por atributos como `RIGHT_ID` (nó atual) e `LEFT_ID` (nó pai), além de características como `DEP` (tipo de dependência sintática), `POS` e `LEMMA`. O *DependencyMatcher* avalia como esses nós estão conectados para verificar se correspondem ao padrão.

A análise sintática dos segmentos de música é realizada utilizando árvores de dependência para detectar padrões específicos de relacionamento entre os elementos da frase. Esses padrões são definidos por dicionários que representam a estrutura de dependência entre adjetivos, sujeitos e verbos. A ideia é capturar como as palavras estão relacionadas dentro da hierarquia sintática, identificando como certos termos qualificam, descrevem ou interagem com outros componentes gramaticais.

Por exemplo, um dos padrões implementados busca um nó que representa um adjetivo como raiz da árvore de dependência. Esse adjetivo deve estar conectado a um sujeito através de uma relação chamada “`NSUBJ`”, indicando que ele modifica ou descreve esse sujeito. Em seguida, o padrão procura por um verbo auxiliar associado ao adjetivo, de modo a capturar construções como “Ele é bom” e variações mais complexas como “Ele está sendo muito bom” ou “Ela sempre foi muito determinada”.

```
sujeitoauxadjetivo = [{"RIGHT_ID": "adjetivo", "RIGHT_ATTRS": {"POS": "ADJ"}},  
  {"LEFT_ID": "adjetivo", "REL_OP": ">", "RIGHT_ID": "sujeito",  
    "RIGHT_ATTRS": {"DEP": "nsubj",  
      "LOWER": {"IN": sujeitos_femininos}}},  
  {"LEFT_ID": "adjetivo", "REL_OP": ">", "RIGHT_ID": "auxiliar",  
    "RIGHT_ATTRS": {"POS": "AUX"}}]
```

Extensões aos padrões de busca foram implementadas para incluir descrições que utilizem verbos no particípio. Essas estruturas desempenham um papel predicador e podem indicar subordinação, especialmente quando usadas na voz passiva, como em “Ela é amada”. De forma semelhante, substantivos também podem ter função predicadora, como em “Ela é amor”.

5. Métodos de pesquisa

Este estudo adota técnicas de PLN para analisar padrões linguísticos sexistas em letras de músicas brasileiras, com ênfase na sexualização de mulheres. Os códigos utilizados para a análise estão disponíveis em um repositório no GitHub⁹.

Após a seleção do corpus, a metodologia seguiu um fluxo de etapas estruturadas para atender aos objetivos propostos. Inicialmente, foram identificados padrões recorrentes nas letras, os quais serviram de base para análises mais detalhadas.

O fluxo metodológico seguido, bem como as perguntas de pesquisa abordadas em cada etapa, estão detalhados no diagrama ilustrado na Figura 4.

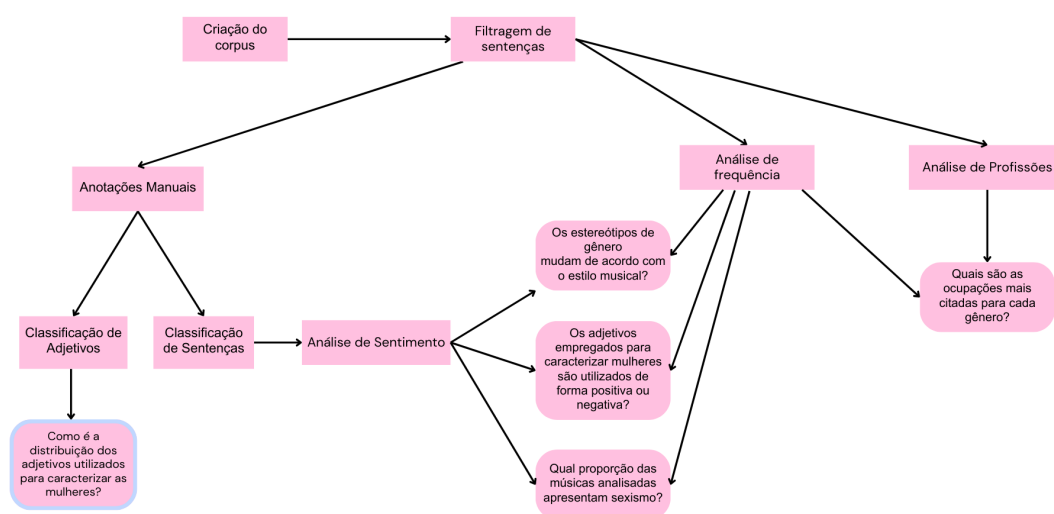


Figura 4. Diagrama do fluxo metodológico e questões de pesquisa abordadas por tópico.

5.1. Anotações manuais

Anotações foram realizadas de forma manual, a fim de garantir maior confiabilidade às análises. Parte das sentenças obtidas por meio dos padrões de busca foram rotuladas, com o objetivo de treinar um modelo de classificação de sentimentos. Além disso, adjetivos presentes nessas sentenças foram manualmente anotados para categorizar as qualidades atribuídas aos gêneros feminino e masculino.

⁹<https://github.com/firminovitoria/identificando-padroes>

Definir o que é sexismo é uma tarefa subjetiva, mas a compreensão de como comportamentos, atitudes e linguagem contribuem para desvalorizar, marginalizar ou reforçar estereótipos negativos baseados em gênero é fundamental. Muitas vezes, o sexismo se manifesta de maneira sutil, em expressões que podem parecer inofensivas ou ser disfarçadas, mas que carregam um viés de desvalorização e reforço de desigualdades de gênero. Exemplos de comentários como “Você é tão gostosa”, “Mulher bunduda”, “Mulher maluca” e “Você é puta” ilustram a objetificação e o estigma relacionados a aspectos físicos ou características emocionais atribuídas a uma pessoa. Esses e outros fatores foram considerados neste trabalho para definir o que caracteriza o sexismo e suas diversas manifestações.

5.1.1. Classificação das sentenças

Foram anotadas manualmente 1999 sentenças referentes ao gênero masculino e 1999 sentenças referentes ao gênero feminino em três categorias: Positivo, Negativo e Neutro. As considerações para cada categoria foram as seguintes:

- **Positivo:** Sentenças que expressam opiniões ou informações favoráveis sobre o gênero. Por exemplo: “*Você é muito inteligente.*”.
- **Negativo:** Sentenças que contêm comentários desfavoráveis, degradantes ou pejorativos sobre o gênero. Por exemplo: “*Ela é feia.*” ou “*Ela é fogosa.*”
- **Neutro:** Sentenças que não possuem uma conotação claramente positiva ou negativa, ou que apresentam informações objetivas sem juízo de valor. Por exemplo: “*A moça está presente.*”.

No total, foram encontradas 1.550 sentenças negativas, 1.498 sentenças positivas e 950 sentenças neutras. Frases desconexas foram desconsideradas.

5.1.2. Categorização de adjetivos

Para a análise considerando o interesse na caracterização humana, assim como descrito em [Freitas and Martins 2023], foi feita a categorização manual de 1.000 predicções para cada gênero em seis categorias: *Aparência*, *Caráter*, *Emoção*, *Papel Social*, *Sexual* e *Neutro*. Na contabilização, foram desconsideradas palavras desconhecidas, no plural, bem como abreviações e variações ortográficas. A categoria *Sexual* foi criada como uma nova classe de análise, a fim de capturar aspectos relacionados à sexualidade, que não se encaixavam nas demais categorias.

Além disso, foi criada a categoria *Neutro*, que agrupou termos com múltiplos significados ou que não se encaixavam claramente nas outras categorias. Esses termos neutros foram indicados ao longo do processo de categorização devido à sua ambiguidade, mas foram omitidos nas análises para evitar distorções nas comparações entre as outras categorias.

5.2. Análise de sentimentos

A fim de classificar as emoções transmitidas pelas sentenças predadoras obtidas a partir do processamento das letras músicas, foi feito o *fine-tuning* de um modelo de aprendizado de máquina capaz de rotular sentenças como positivas, negativas e neutras.

No treinamento do modelo, foi utilizada a anotação descrita na Subseção 5.1.1. Um processo de balanceamento das classes foi realizado utilizando o algoritmo *RandomOverSampler* da biblioteca *imblearn*. Isso assegurou que o modelo fosse treinado com uma representação equitativa das classes.

Para a tokenização e o modelo de classificação de sentimentos, utilizamos o *BER-Timbau*, um modelo pré-treinado de *BERT* para a língua portuguesa. Mais especificamente, foi utilizada a versão *bert-base-portuguese-cased*, que é adequada para tarefas de classificação de sentenças em português com diferenciação entre maiúsculas e minúsculas.

Para garantir a robustez dos resultados e avaliar o desempenho do modelo em diferentes subdivisões dos dados, utilizamos validação cruzada estratificada com 30 *folds*.

Ao final da avaliação, foram computadas quatro métricas principais: i) Acurácia, que mede a proporção de previsões corretas; ii) Precisão, que avalia a proporção de verdadeiros positivos entre todas as predições positivas; iii) Revocação, que mede a capacidade do modelo de identificar corretamente todas as instâncias positivas; e iv) *F1-Score*, que consiste na harmônica entre a precisão e a revocação, proporcionando uma visão equilibrada do desempenho geral.

Essas métricas foram calculadas utilizando funções do *sklearn*, com a média ponderada pelas classes, garantindo uma avaliação justa, mesmo em cenários de classes desbalanceadas.

Ao final do processo de validação cruzada, as métricas de desempenho foram calculadas como médias para todas as iterações. O desempenho final do modelo apresentou uma acurácia média de 97,16%, uma precisão média de 97,27%, um *recall* médio de 97,16% e um F1-score médio de 97,15%. Essas métricas demonstram a eficácia do modelo em classificar corretamente as sentenças em suas respectivas categorias de sentimento, evidenciando sua capacidade preditiva.

5.3. Análise de profissões

Nesta análise, investigamos a relação entre representação profissional e gênero. Elaboramos uma lista de profissões com base na Classificação Brasileira de Ocupações Domiciliar (CBO-Domiciliar)¹⁰, complementada por outras ocupações populares no Brasil. Nosso objetivo foi compreender como as profissões são atribuídas a homens e mulheres nas letras musicais, buscando padrões de gênero nas associações ocupacionais por meio de expressões regulares e árvores de dependência.

Para estruturar a análise, as profissões foram organizadas em três categorias: flexionadas no feminino, flexionadas no masculino e comuns de dois gêneros (onde o gênero era inconclusivo sem suporte gramatical ou semântico adicional). No último caso, a distinção de gênero ocorre pela anteposição de determinantes (pronomes, artigos) masculinos ou femininos como, por exemplo, *o presidente* ou *a presidente*.

Nas buscas para profissões flexionadas no masculino, usamos termos como “doutor” e “professor”, enquanto para o feminino, foram utilizados termos como “doutora” e “professora”. As profissões comuns de dois gêneros foram incluídas apenas quando

¹⁰<https://concla.ibge.gov.br/estrutura/ocupacao-estrutura>

o gênero do sujeito pôde ser identificado claramente, seja como masculino ou feminino, com base na morfologia do sujeito envolvido.

Nos casos em que a identificação do gênero sujeito não era possível, como em “você é piloto”, realizamos buscas onde a profissão era investigada necessariamente junto com a flexão para o masculino ou feminino. Para garantir maior precisão nos resultados, sentenças duplicadas dentro da mesma música foram removidas.

6. Resultados

Os resultados são apresentados como respostas a cada questão de pesquisa deste estudo.

6.1. *Como é a distribuição dos adjetivos utilizados para caracterizar as mulheres?*

No ranking dos 10 predicadores mais utilizados para cada gênero, apresentando na Tabela 1 com adjetivo e ocorrência, observa-se que as mulheres estão fortemente associadas a adjetivos relacionados à aparência e características físicas. Termos como “bonita”, “linda” e “gostosa” são frequentes, sugerindo um foco predominante na estética feminina e a priorização da aparência como principal atributo.

Feminino		Masculino	
Adjetivo	Ocorrência	Adjetivo	Ocorrência
bonita	852	feliz	335
linda	794	bom	299
louca	263	pobre	189
novinha	207	forte	158
gostosa	183	capaz	141
boa	159	triste	124
feia	153	louco	114
feliz	145	amor	114
dona	123	diferente	114
mãe	110	casado	95
Total:	2989	Total:	1683

Tabela 1. Ranking dos dez adjetivos mais frequentes encontrados nas sentenças predadoras.

Para os homens, por outro lado, os adjetivos mais comuns estão associados a características emocionais e de caráter. Termos como “feliz”, “bom”, “capaz” e “forte” destacam traços internos e habilidades, apontando para uma representação que valoriza a personalidade e a competência masculina. Essa tendência reforça a ideia de que, enquanto as descrições femininas focam em atributos externos, as masculinas enfatizam qualidades subjetivas e capacidades. É importante destacar ainda a frequência que os predicadores mais comuns acontecem: 2.989 no caso feminino e 1.683 no caso masculino.

Na Figura 5 são apresentadas, para cada gênero, as frequências das categorias analisadas anteriormente (*Aparência, Caráter, Emoção, Papel Social e Sexual*). Essa distribuição permite visualizar a predominância de cada categoria na caracterização de homens e mulheres. Na Tabela 2, por sua vez, são detalhados os 10 termos mais frequentes em cada uma dessas categorias, evidenciando um agrupamento de características que reforça as diferenças nas representações de gênero.

Nos resultados para o masculino, destaca-se uma predominância significativa da categoria “Caráter”, que representa 36,3% das ocorrências em relação às outras categorias analisadas. Essa categoria inclui descrições que ressaltam traços de personalidade ou características gerais associadas aos homens, como “bom” e “capaz”. A segunda categoria mais frequente é “Emoção”, com 25,1%. Em seguida, aparece “Papel Social” 24,2%, que aborda funções sociais atribuídas aos homens, como “pai” e “casado”, e a categoria “Aparência” 12,6%, que destaca aspectos físicos como “forte” e “bonito”. A categoria menos frequente é “Sexual”, com apenas 1,7%. Os termos na categoria sexual masculina, como “prostituto” e “pegador”, revelam nuances importantes sobre como a sexualidade masculina é representada. Além de descrever comportamentos ou papéis sexuais, muitos desses termos refletem estereótipos sociais que associam a masculinidade à virilidade e à conquista. No entanto, vale destacar que essa categoria também inclui termos com conotações pejorativas e homofóbicas, como “gay”, “bicha” e “baitola”.

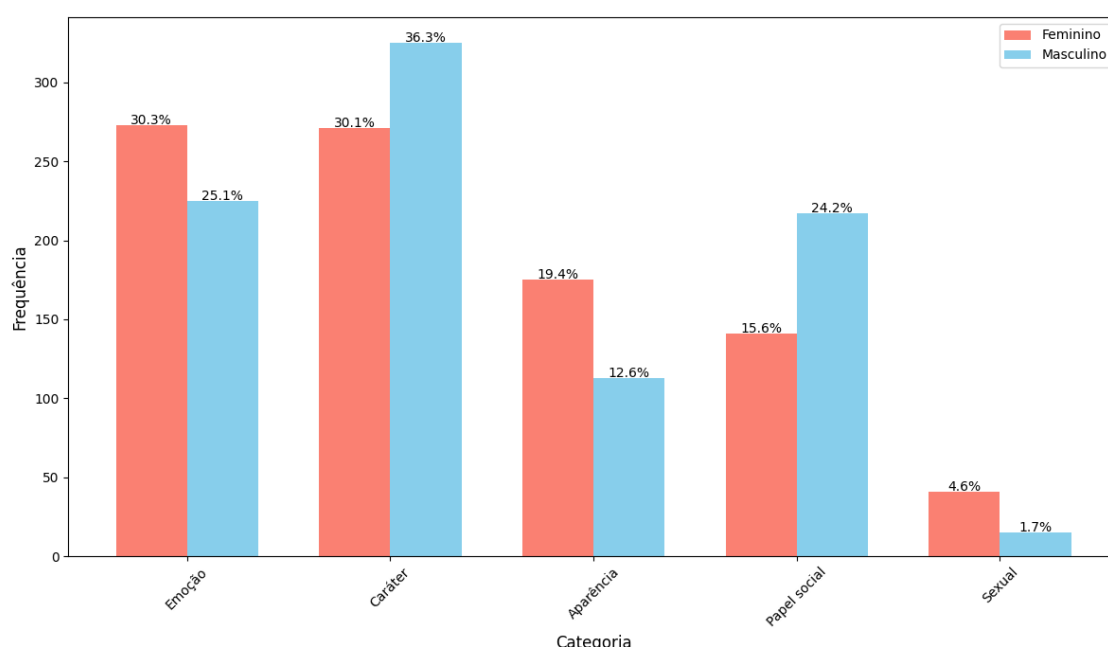


Figura 5. Frequência categorias masculino e feminino

Para as mulheres, observa-se que as categorias “Emoção” com 30,4% e “Caráter” com 30% apresentam uma frequência próxima, sugerindo uma forte ênfase tanto nas descrições emocionais quanto nas de personalidade. A categoria “Aparência” aparece com 19,4%, evidenciando aspectos físicos, como “bonita” e “novinha”. Já “Papel Social” com 15,6%, e associado mulheres a papéis como “mãe” e “dona”. A categoria “Sexual” tem uma representação de 4,6%, significativamente maior que no caso masculino, destacando termos como “safada” e “virgem”.

Observa-se que, no masculino, a ênfase está nas categorias “Caráter” e “Emoção”, sugerindo uma representação mais focada em aspectos internos. Já no feminino, “Emoção” e “Caráter” também dominam, mas a categoria “Aparência” tem uma presença relativamente mais marcante quando comparada ao masculino.

A representação sexual, embora não seja a mais expressiva em ambos os grupos, é

Categoria	Feminino	Masculino
Aparência	bonita, linda, novinha, gostosa, feia, bela, nova, cheia, flor, top	forte, bonito, velho, branco, grande, novo, fraco, lindo, gostoso, preto
Caráter	louca, boa, diferente, maluca, doida, pura, bandida, capaz, namoradeira, ingrata	bom, capaz, louco, diferente, fiel, valente, livre, ruim, guerreiro, certo
Emoção	feliz, sozinha, triste, amor, ciumenta, carente, má, carinhosa, doce, braba	feliz, triste, amor, sozinho, coração, infeliz, distante, carente, cansado, contente
Papel Social	dona, mãe, casada, solteira, brasileira, rica, criança, mamãe, filha, princesa	pobre, casado, rico, pai, mano, amigo, filho, dono, irmão, solteiro
Sexual	safada, virgem, fogosa, puta, safadinha, sapeca, assanhada, assanhadinha, tarada, danadinha	virgem, gay, prostituto, safadinho, pegador, bicha, estuprador, tesão, baitola, brocha

Tabela 2. Top 10 termos mais frequentes para Feminino e Masculino por categoria

mais presente nos textos femininos. A análise dos termos masculinos na categoria sexual revela uma carga homofóbica em muitos deles, com palavras pejorativas como “bicha”, “baitola”, e “brocha”, reforçando estigmas e preconceitos em relação à masculinidade, virilidade e sexualidade.

No feminino, os termos possuem uma forte conotação objetificadora. Palavras como “safada”, “virgem”, “puta” e “piranha” surgem em várias formas, refletindo uma visão estereotipada e hipersexualizada das mulheres. O uso de “sapequinha”, “danadinha”, “ninfeta”, reforça a infantilização e trivialização da mulher em contextos sexuais. Termos como “sapatão”, “mulher-homem”, “travesti” e “lésbica” também refletem preconceitos referente à diversidade sexual.

Nas Figuras 6 e 7 são apresentadas as nuvem de predicadores femininos e masculinos encontrados nas músicas dos corpos, respectivamente.

6.2. Os adjetivos empregados para caracterizar mulheres são utilizados de forma positiva ou negativa?

A análise de sentimentos realizada sobre as sentenças de músicas revelou diferenças significativas no tratamento dado a mulheres e homens. Para o gênero feminino, foram identificadas 6.169 sentenças com sentimento positivo, 4.617 sentenças com sentimento neutro e 4.437 sentenças com sentimento negativo, totalizando 15.223 sentenças. No caso do gênero masculino, foram encontradas 5.370 sentenças positivas, 5.878 sentenças neutras e 3.833 sentenças negativas, com um total de 15.081 sentenças.

O sentimento positivo foi mais prevalente nas sentenças relacionadas ao gênero feminino do que o masculino. Já as sentenças com sentimento neutro foram mais fre-



Figura 6. Nuvem de predicadores femininos



Figura 7. Nuvem de predicadores masculinos

quentes para o gênero masculino, em comparação o feminino. Em relação às sentenças negativas, o gênero feminino também se destacou.

Essas diferenças na distribuição de sentimentos entre os gêneros sugerem que as músicas analisadas tendem a retratar as mulheres de maneira mais emocionalmente carregada, tanto em aspectos positivos quanto negativos, enquanto os homens são frequentemente descritos de forma mais neutra.

A Figura 8 ilustra essas diferenças na frequência dos sentimentos entre os dois gêneros, destacando a maior prevalência de sentimentos positivos e negativos para o gênero feminino, em contraste com a predominância de sentimentos neutros para o masculino.

Realizamos ainda uma leitura mais próxima das sentenças a fim de conhecer o sentimento dos predicadores mais utilizados para cada categoria. Para isso, utilizamos os termos mais frequentes descritos na Tabela 2. Para cada adjetivo, foram buscadas as sentenças em que eles apareceram. Em seguida, essas sentenças foram analisadas com o nosso modelo de classificação de sentimentos. O resultado é apresentado nas subseções a

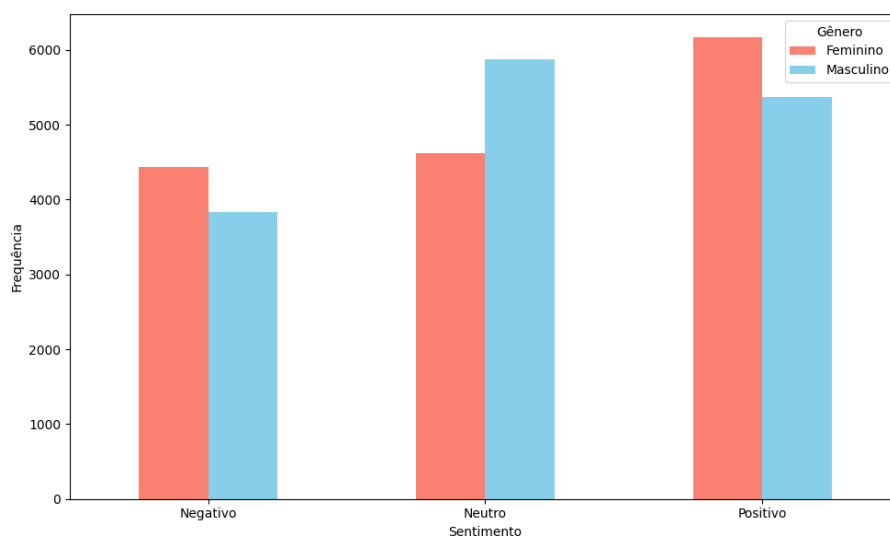


Figura 8. Análise de Sentimentos de todas as sentenças predicadoras

seguir.

6.2.1. Aparência

Os dados revelam que os adjetivos associados à aparência feminina tendem a ter uma carga emocional majoritariamente positiva. Como é possível notar na Figura 9, termos como “bonita” e “linda” têm uma clara conotação positiva. Porém, esses adjetivos podem reduzir a mulher a seu valor estético ou associá-la a padrões de beleza rígidos e idealizados.

Termos como “feia”, “gostosa”, e “novinha” aparecem frequentemente com uma conotação pejorativa ou objetificadora. Embora esses termos negativos estejam presentes, sua proporção é menor em comparação com os adjetivos positivos, sugerindo que, ao descrever a aparência feminina, há uma tendência maior em elogiar do que criticar.

Os termos neutros na aparência feminina são raros, o que sugere que, quando se trata de descrever a aparência das mulheres, as palavras utilizadas tendem a carregar um julgamento — seja positivo ou negativo — sem espaço para descrições imparciais ou neutras.

No caso dos homens (Figura 10), os adjetivos ligados à aparência mostram um equilíbrio maior entre termos positivos e neutros, com uma presença menos frequente de termos negativos em comparação com as mulheres. Palavras como “forte”, “bonito” e “grande” são vistas predominantemente de forma positiva, muitas vezes destacando aspectos físicos que são tradicionalmente valorizados na masculinidade, como força e robustez. O termo “forte”, por exemplo, possui uma conotação bastante positiva, enfatizando a virilidade e o poder físico, que são atributos socialmente valorizados nos homens.

Adjetivos negativos, como “fraco”, são menos frequentes na descrição da aparência masculina, sugerindo que os homens são menos julgados ou criticados com base em sua aparência física. Além disso, termos neutros, como “preto” e “branco”, que

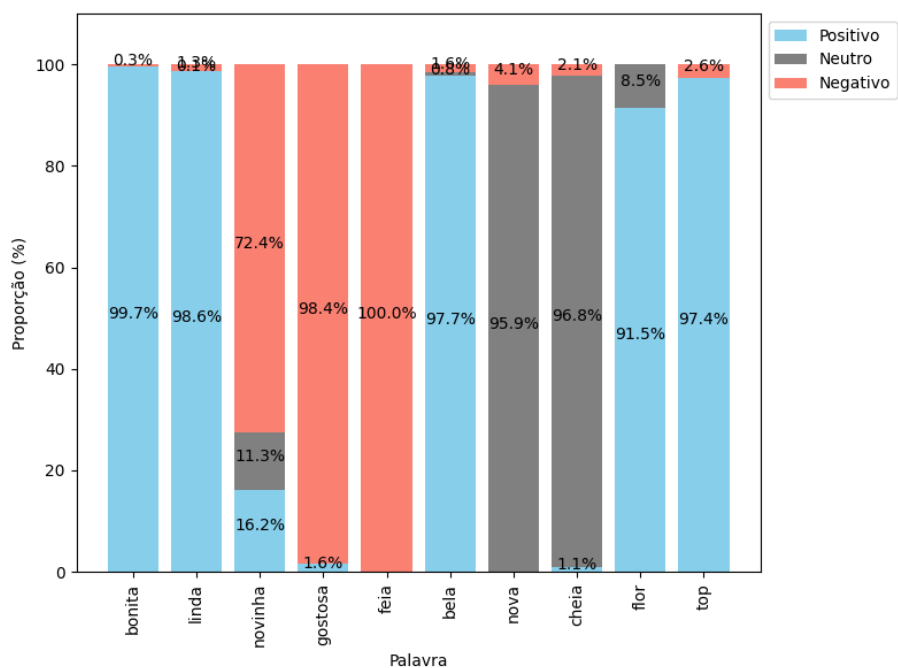


Figura 9. Conotação das palavras relacionadas ao gênero feminino mais relacionadas a aparência

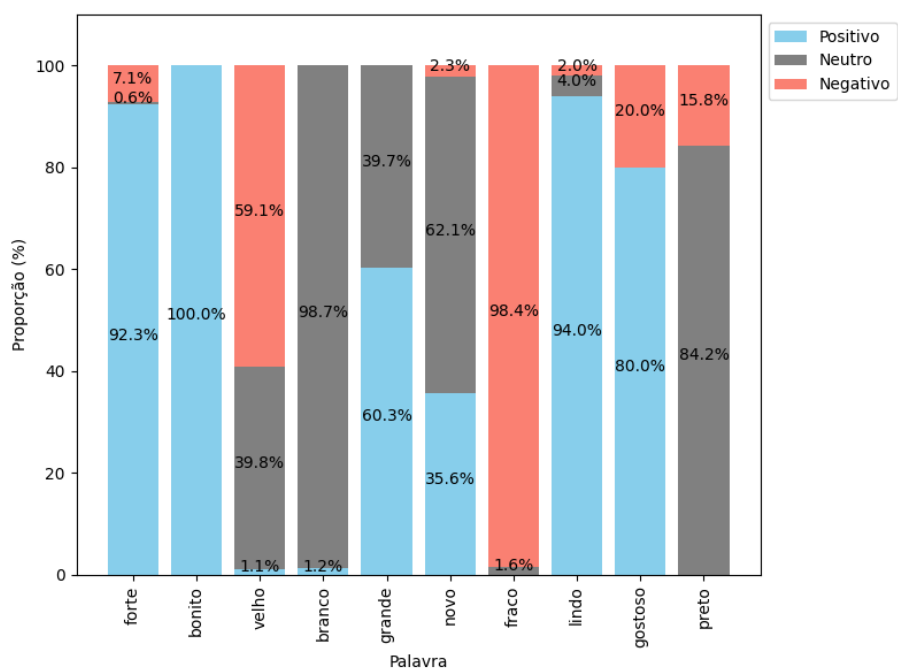


Figura 10. Conotação das palavras relacionadas ao gênero masculino mais relacionadas a aparência

descrevem características físicas sem uma valoração explícita, estão mais presentes na descrição dos homens, indicando que a aparência masculina é tratada com uma objetividade maior em comparação à das mulheres.

6.2.2. Caráter

Como observado na Figura 11, os adjetivos ligados ao caráter das mulheres tendem a refletir predominantemente características negativas, muitas vezes sugerindo fraqueza ou inadequação. Esses termos negativos, comumente associados a estereótipos de instabilidade emocional ou comportamento irracional, refletem uma tendência histórica de descrever o caráter feminino de forma crítica, especialmente quando mulheres demonstram assertividade ou independência. Embora os adjetivos positivos estejam presentes, o uso desses termos depreciativos sugere que o caráter das mulheres é frequentemente interpretado com uma carga emocional negativa.

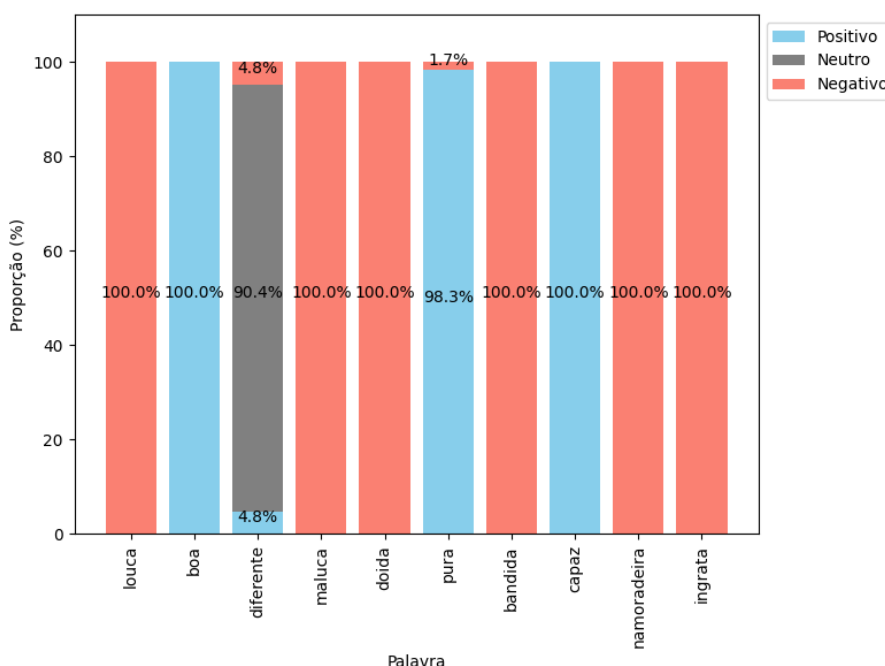


Figura 11. Conotação das palavras relacionadas ao gênero feminino mais relacionadas a caráter

Termos como “boa” e “capaz” têm uma clara conotação positiva, destacando aspectos de competência e confiabilidade. No entanto, há uma forte presença de adjetivos com conotação negativa, como “louca”, “doida” e “maluca”, que são frequentemente usados para descrever mulheres de forma pejorativa ou para diminuir a legitimidade de suas ações e comportamentos.

Além disso, adjetivos neutros ou imparciais são raramente usados para descrever o caráter feminino. As palavras selecionadas tendem a carregar um julgamento, seja ele positivo ou negativo, reforçando a percepção de que as mulheres são constantemente avaliadas e categorizadas em termos de suas características de personalidade.

Por outro lado, na Figura 12, os adjetivos empregados para descrever o caráter masculino são majoritariamente positivos, com destaque para termos como “bom”, “capaz” e “guerreiro”. Esses adjetivos são fortemente associados a qualidades de força, competência e determinação — características tradicionalmente valorizadas na masculinidade.

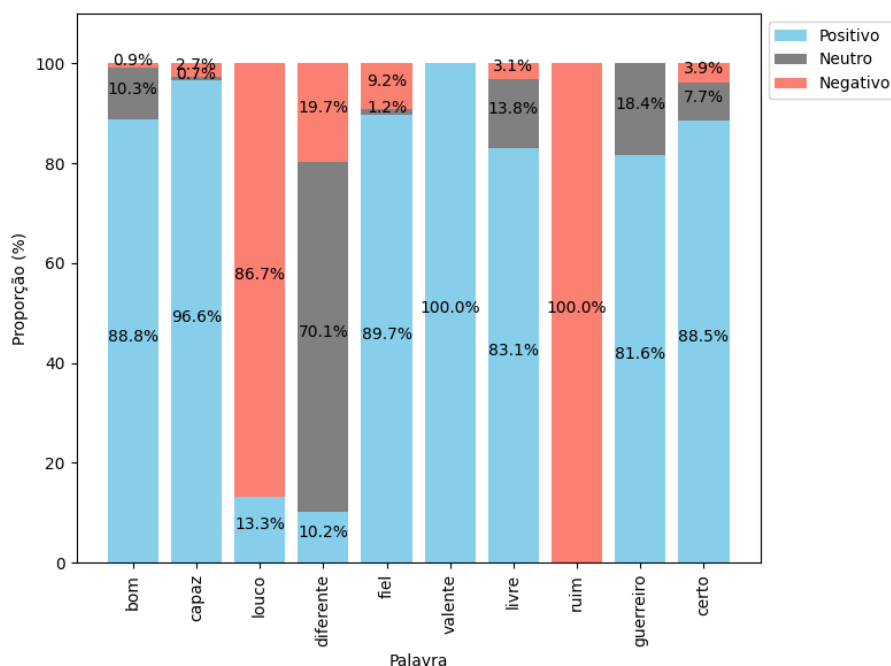


Figura 12. Conotação das palavras relacionadas ao gênero masculino mais relacionadas a caráter

Embora existam alguns termos negativos, como “ruim” e “louco”, sua presença é significativamente menor em comparação com o feminino. Isso indica que, enquanto o caráter feminino é muitas vezes descrito de forma pejorativa ou estigmatizada, o caráter masculino é tratado com maior positividade, com uma ênfase clara em traços que reforçam a ideia de estabilidade, força e confiança.

Adjetivos neutros também são pouco usados na descrição do caráter masculino, reforçando a tendência de que o caráter, em geral, seja avaliado com um viés, seja ele positivo ou negativo. No entanto, as descrições masculinas tendem a ser mais indulgentes e frequentemente associadas a qualidades socialmente desejáveis.

6.2.3. Emoção

Nota-se na Figura 13 que os predicadores ligados às emoções das mulheres são predominantemente negativos. Adjetivos como “triste”, “sozinha” e “ciumenta” aparecem com frequência, refletindo uma visão estereotipada de que as mulheres são emocionalmente frágeis ou excessivamente sensíveis.

Por outro lado, adjetivos positivos como “feliz” e “carinhosa” também têm uma presença significativa, sugerindo uma dualidade na forma como as mulheres são vistas no âmbito emocional. Elas são tanto idealizadas como cuidadoras quanto retratadas como emocionalmente instáveis. Isso reflete um padrão de expectativas emocionais que limita o papel da mulher às esferas de cuidado e sensibilidade, perpetuando a visão de que as emoções femininas são centrais para sua identidade.

Adjetivos neutros são praticamente inexistentes na descrição emocional das mu-

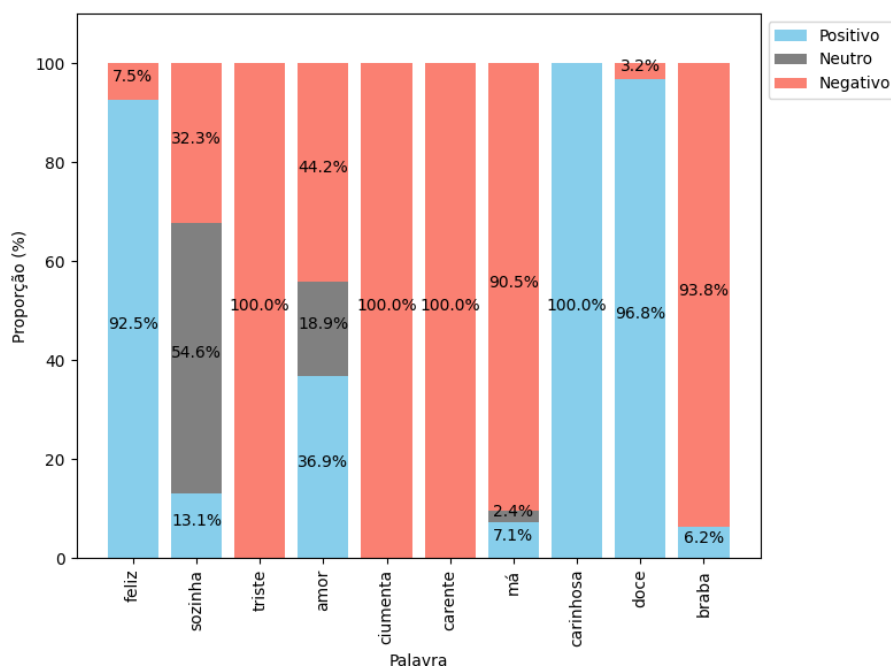


Figura 13. Conotação das palavras relacionadas ao gênero feminino mais relacionadas a emoção

Iheres, o que indica que, quando o caráter emocional feminino é abordado, ele é fortemente carregado de julgamentos — seja para idealizá-las ou para criticá-las.

A Figura 14 revela que, para os homens, a categoria “Emoção” é representada de maneira mais negativa do que positiva. Adjetivos como “triste”, “sozinho”, “infeliz”, “cansado”, “distante”, aparecem com maior frequência, sugerindo que as emoções masculinas são frequentemente associadas a sentimentos de solidão e insatisfação. Esses adjetivos reforçam estereótipos de que os homens, embora menos frequentemente descritos em termos emocionais, são retratados com uma carga negativa quando esse aspecto é abordado.

Apesar da presença de predicções positivas como “feliz” e “amor”, a prevalência de termos negativos é notável, sugerindo uma representação emocional mais sombria para o masculino. Isso pode indicar que, quando os homens são descritos em termos emocionais, suas emoções são associadas a vulnerabilidades internas que não são tão abertamente expressas, mas que carregam uma conotação de sofrimento ou insatisfação.

6.2.4. Papel Social

Na Figura 15, vemos que os adjetivos empregados para descrever o papel social das mulheres estão fortemente associados a suas funções familiares e de cuidado. Termos como “mãe”, “dona”, “casada” e “filha” são dominantes, refletindo uma visão tradicional que coloca a mulher no centro da estrutura familiar. O termo “dona”, em particular, está mais frequentemente associado ao conceito de “dona de casa”, reforçando o papel tradicional da mulher como responsável pelas atividades domésticas e pelos cuidados com a casa. Esses adjetivos têm uma carga emocional neutra ou ligeiramente positiva, uma vez que

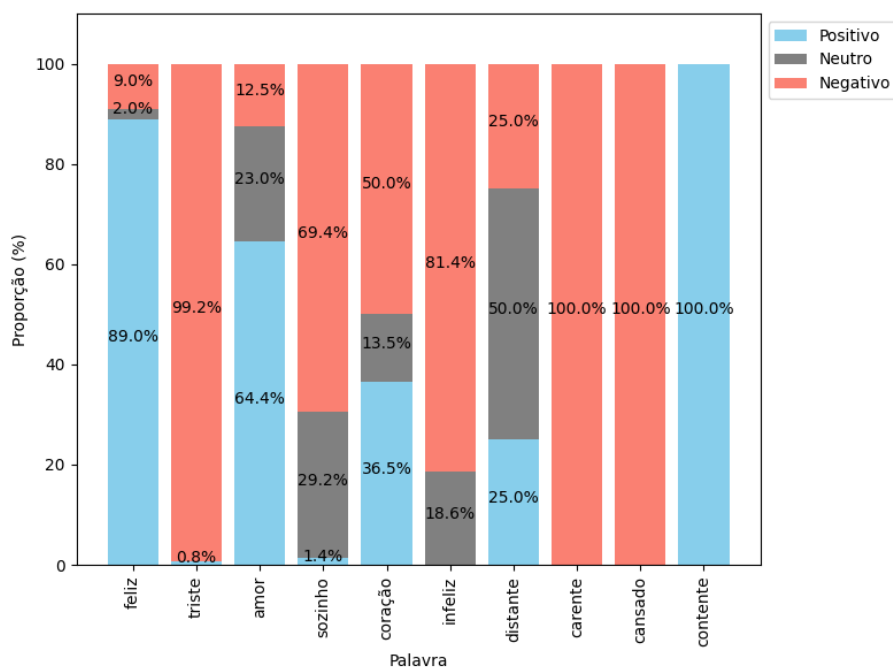


Figura 14. Conotação das palavras relacionadas ao gênero masculino mais relacionadas a emoção

ser mãe ou casada é culturalmente valorizado e visto como um papel social importante.

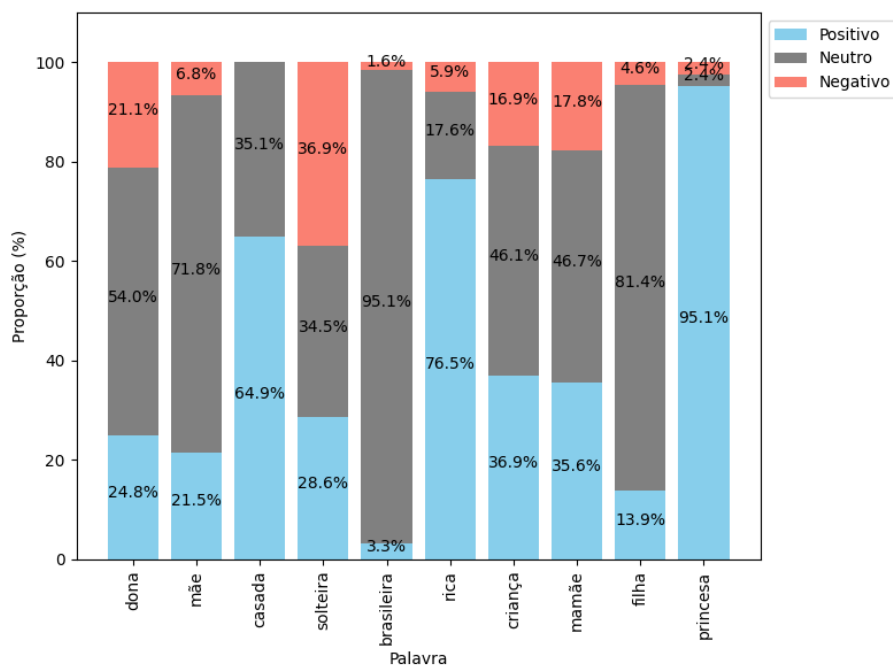


Figura 15. Conotação das palavras relacionadas ao gênero feminino mais relacionadas a papéis sociais

No entanto, a falta de diversidade nos adjetivos que descrevem o papel social feminino sugere que as mulheres ainda são frequentemente enquadradas em papéis limitados.

A ênfase nos termos relacionados à vida familiar ou ao estado conjugal, como “casada” e “mamãe”, reforça estereótipos tradicionais de gênero, nos quais a mulher é vista como cuidadora primária e subordinada às dinâmicas familiares. Isso limita a amplitude de possibilidades para além dessas funções estabelecidas, sugerindo que a mulher é muitas vezes definida pela sua relação com outros (marido, filhos, pais), ao invés de sua individualidade ou realizações fora desse contexto.

Por outro lado, a Figura 16 revela que os adjetivos utilizados para descrever o papel social masculino são mais diversificados, mas igualmente tradicionais. Termos como “pai” e “casado” aparecem com destaque, indicando que os homens também são frequentemente descritos em relação às suas funções familiares. No entanto, há uma diferença significativa: os adjetivos associados ao papel social masculino também enfatizam liderança e autoridade.

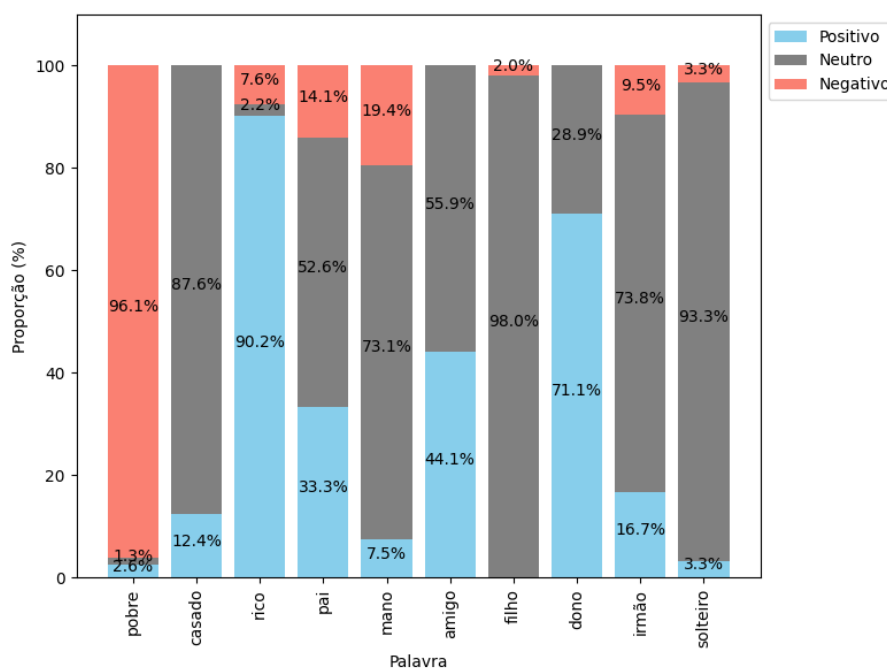


Figura 16. Conotação das palavras relacionadas ao gênero masculino mais relacionadas a papéis sociais

Termos como “dono” sugerem uma posição de poder e controle, reforçando estereótipos de masculinidade que colocam os homens em posições de comando, tanto dentro quanto fora da esfera familiar.

O papel social masculino também inclui termos que refletem uma maior independência e foco na identidade individual, como “amigo”, “mano” e “irmão”. Esses termos sugerem que os homens são vistos como indivíduos com papéis mais variados que vão além das funções familiares ou conjugais. Eles têm espaço para desempenhar papéis fora dessas esferas, o que lhes confere mais autonomia social.

6.2.5. Sexual

Na Figura 17, observamos que os adjetivos ligados à sexualidade feminina têm, em grande parte, uma conotação negativa e objetificadora. Termos como “safada”, “puta”, e “piranha” aparecem com frequência, reforçando estereótipos de hipersexualização e degradação moral das mulheres. Esses adjetivos sugerem que a sexualidade feminina, quando expressa de forma independente ou fora dos padrões normativos, é vista como algo a ser censurado e depreciado.

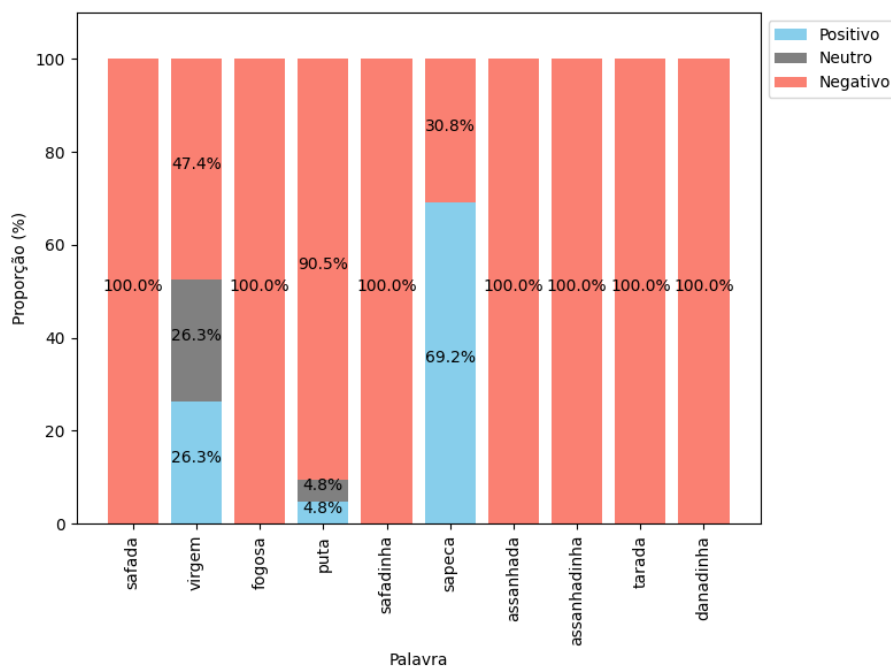


Figura 17. Conotação das palavras relacionadas ao gênero feminino mais relacionadas a sexualidade

Além disso, há um forte uso de diminutivos, como “sapequinha”, “danadinha” e “ninfeta”, que além de hipersexualizarem, também infantilizam a mulher, refletindo uma visão que trivializa a sexualidade feminina. A associação de termos sexualizados com características infantis sugere um padrão perigoso de normalização da exploração sexual de mulheres jovens, moldando suas identidades dentro de uma perspectiva de submissão.

Outro ponto importante é o uso de termos que reduzem as mulheres a seu valor sexual, ignorando outros aspectos de suas identidades. A presença de termos como “virgem” também reflete uma visão binária da sexualidade feminina, onde a mulher é julgada pela pureza ou falta dela, mantendo o foco no controle sobre seu corpo e comportamento sexual.

Termos relacionados à diversidade sexual feminina, como “sapatão” ou “lésbica”, aparecem com menos frequência, mas carregam uma carga preconceituosa e pejorativa quando utilizados de forma depreciativa. Isso reflete uma marginalização das mulheres que não se conformam às normas heteronormativas, reforçando estigmas em torno da sexualidade feminina.

Na Figura 18, observamos que os adjetivos ligados à sexualidade masculina são

frequentemente utilizados de forma negativa e, em muitos casos, carregam uma forte carga homofóbica. Termos como “bicha”, “baitola” e “brocha” são usados para depreciar a masculinidade, sugerindo que qualquer desvio do comportamento sexual ativo e dominante esperado dos homens é motivo de crítica e estigma. O termo “brocha”, por exemplo, ataca diretamente a virilidade masculina, associando a falha sexual à perda de valor e poder, reforçando a pressão social sobre o desempenho sexual como parte central da identidade masculina.

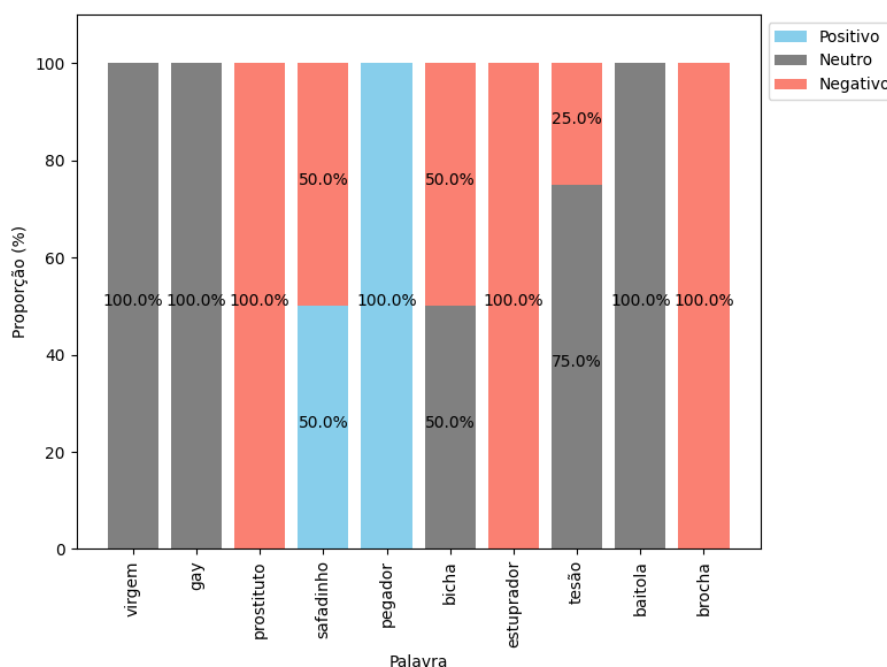


Figura 18. Conotação das palavras relacionadas ao gênero masculino mais relacionadas a sexualidade

A pressão social sobre os homens para se conformarem a esses padrões rígidos de virilidade é evidenciada pelo uso pejorativo de termos que criticam qualquer comportamento que fuja desse modelo, principalmente através de insultos homofóbicos e a ideia de “falha” sexual.

Essa discrepância revela um duplo padrão: enquanto a sexualidade masculina é muitas vezes associada a conquista e virilidade, qualquer desvio desse padrão é fortemente punido socialmente, com termos depreciativos que questionam a masculinidade de maneira direta.

6.3. Quais são as ocupações mais citadas para cada gênero?

Foram identificadas 281 sentenças descrevendo profissões no masculino e 131 no feminino. Após a busca, as profissões foram ranqueadas de acordo com sua frequência, com os resultados apresentados na Tabela 3.

Os dados revelam que as principais ocupações associadas ao gênero masculino incluem “doutor”, “presidente” e “capitão”. Já no gênero feminino, as profissões mais comuns são “atriz”, “professora” e “modelo”. Essa diferença sugere uma divisão de papéis

Feminino		Masculino	
Profissão	Frequência	Profissão	Frequência
Atriz	32	Doutor	31
Professora	10	Presidente	28
Modelo	10	Capitão	14
Chefe	6	Artista	13
Babá	5	Professor	13
Prostituta	4	Cantor	12
Lavadeira	4	Caçador	12
Enfermeira	4	Delegado	11
Costureira	3	Locutor	11
Artista	3	Soldado	11
Total:	131	Total:	281

Tabela 3. Ranking das dez profissões mais frequentes por gênero

profissionais entre os gêneros, refletindo expectativas culturais e sociais presentes nas letras musicais, além de destacar o impacto dessas percepções nas escolhas profissionais e nas oportunidades de carreira para homens e mulheres.

A divisão observada também reflete a perpetuação de estereótipos de gênero em ocupações profissionais. Enquanto os homens são predominantemente associados a papéis de autoridade, as mulheres são representadas em profissões relacionadas ao entretenimento, ensino e cuidado. Esse contraste aponta para uma continuidade de representações tradicionais de gênero, nas quais os homens são vistos como líderes ou figuras de poder, enquanto as mulheres são associadas a papéis ligados à aparência e cuidados.

O [IBGE 2024] disponibiliza dados que evidenciam essa segregação ocupacional. No setor de “transporte, armazenagem e correio”, por exemplo, 80,5% dos cargos gerenciais são ocupados por homens, enquanto as mulheres representam apenas 19,5%. Da mesma forma, na área de “construção”, 85,9% dos cargos gerenciais são masculinos, comparado a apenas 14,1% ocupados por mulheres. Por outro lado, em setores tradicionalmente associados às mulheres, como “saúde humana e serviços sociais”, 63,5% dos cargos gerenciais são ocupados por mulheres, enquanto os homens ocupam 36,5% dessas funções. Esses dados corroboram os estereótipos de gênero, onde os homens predominam em áreas técnicas e de gestão de infraestrutura, enquanto as mulheres são mais presentes em setores ligados ao cuidado e à educação.

No que tange à ocupação de cargos de poder e gerência de forma geral, os homens ocupam uma porcentagem significativamente maior que as mulheres. Os homens são maioria em cargos de gerência em 75% dos grupos, enquanto as mulheres são maioria em apenas 25%. Esse panorama pode ser interpretado como um reflexo de vieses históricos profundamente enraizados, onde as representações midiáticas e culturais reforçam expectativas e comportamentos sociais ligados aos papéis tradicionais de gênero.

6.4. Qual proporção das músicas analisadas apresentam sexismo?

Das 146.612 canções analisadas, foi verificado o uso de frases adjetivas associadas aos gêneros feminino e masculino, com especial atenção para aquelas de conotação negativa. Foram identificadas 15.223 frases adjetivas relacionadas ao feminino e 15.081 ao mascu-

lino. Observou-se uma diferença na carga emocional dos adjetivos: 4.437 frases adjetivas negativas foram direcionadas ao feminino, enquanto 3.833 foram atribuídas ao masculino.

Além disso, ao analisarmos a frequência com que essas frases negativas aparecem, notamos que 3.870 músicas contêm sentenças negativas direcionadas ao feminino e 3.543 músicas contêm sentenças negativas direcionadas ao masculino. Em termos proporcionais, a quantidade de adjetivos negativos direcionados ao feminino representa aproximadamente 2,63% das músicas analisadas, enquanto os negativos para o masculino correspondem a aproximadamente 2,41%.

A análise das músicas que contêm maior incidência de adjetivos negativos revelou padrões específicos entre os intérpretes. Entre as 10 músicas com maior quantidade de adjetivos negativos associados ao feminino, apenas uma foi interpretada por uma mulher, enquanto as outras nove foram cantadas por homens. No caso das frases negativas direcionadas ao masculino, dos 10 intérpretes mais recorrentes, 9 eram homens, e apenas uma dupla mista (homem e mulher) apareceu.

Nas músicas que contêm adjetivos de conotação negativa, foram analisados os dez gêneros musicais mais frequentes. A Figura 19, ilustra a distribuição desses adjetivos nos gêneros associados a figuras femininas, enquanto a Figura 20 apresenta a distribuição correspondente para figuras masculinas.

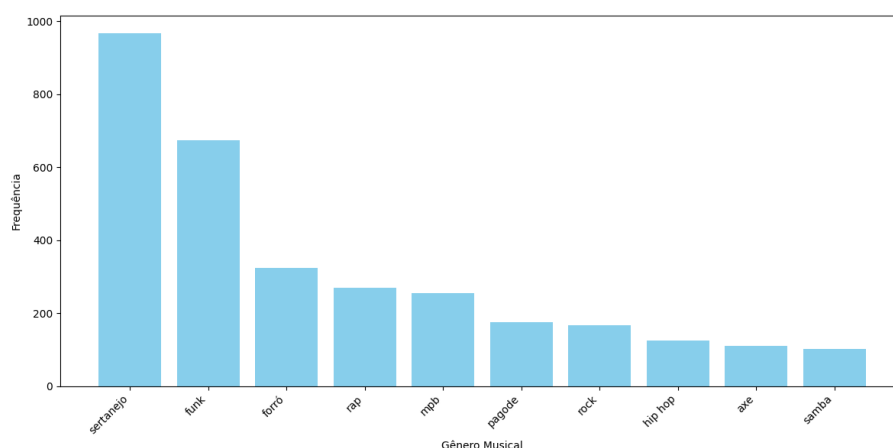


Figura 19. Gêneros Musicais mais frequentes nas Sentenças Negativas para Figuras Femininas

Nota-se que apesar da MBP ser o gênero musical mais representado em nosso corpus, ele está listado como o quinto com mais músicas sexistas. Os estilos Sertanejo e Funk, que representam os segundos e terceiros grupos musicais mais presentes no corpus, por outro lado, figuram como os ritmos com mais estereótipos dentro de nossa análise.

É relevante destacar que os resultados apresentados refletem um recorte específico de dados e, portanto, não devem ser considerados como representativos da totalidade da música brasileira. Assim, evitar generalizações, como afirmar que determinados gêneros são os principais responsáveis pela estereotipização feminina, é essencial para uma interpretação criteriosa e livre de vieses culturais dos achados.

A seguir, uma análise mais detalhada dos estilos musicais é realizada a fim de comparar os predicadores frequentemente utilizados por cada um.

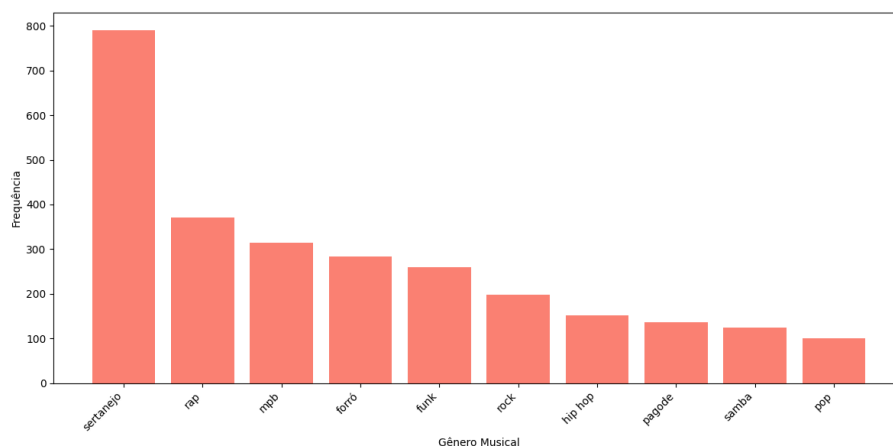


Figura 20. Gêneros Musicais mais frequentes nas Sentenças Negativas para Figuras Masculinas

6.5. Os estereótipos de gênero mudam de acordo com o estilo musical?

Nesta seção, será apresentada a análise das proporções de sentimentos associados a adjetivos em diferentes gêneros musicais, divididos entre masculino e feminino. Para isso, as sentenças, onde os 10 predicadores mais comuns de cada estilo musical aparecem, foram analisadas para determinar o tom emocional das mensagens.

A partir da Figura 21, pode-se notar que a rotulação presente nas canções sertanejas em que os predicadores femininos mais frequentes aparecem é majoritariamente positiva, com 5 adjetivos das categorias “Aparência” (“bonita” e “linda”), “Caráter” (“feliz” e “boa”) e “Papel Social” (“casada”). No entanto, 3 adjetivos apresentam conotação pejorativa: dois relacionados à “Aparência” (“feia” e “gostosa”) e um relacionado ao “Caráter” (“louca”).

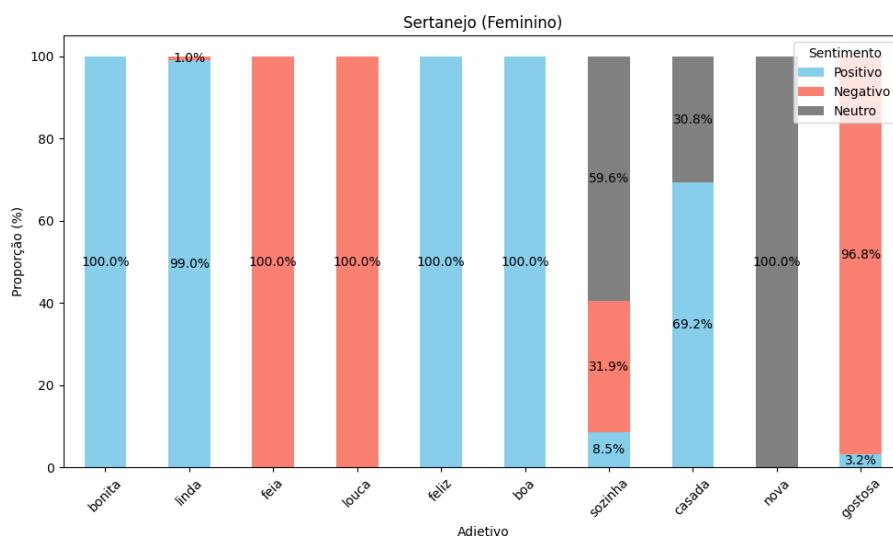


Figura 21. Adjetivos Mais Frequentes no Gênero Musical Sertanejo com Proporção dos Sentimentos para o Feminino

Os predicadores masculinos mais frequentemente empregados nas canções sertanejas

nejas também têm, na maioria das vezes, conotação positiva, com ênfase no “Caráter”: “feliz”, “forte”, “bom”, “capaz” e “rico”. Os 3 adjetivos masculinos empregados de maneira ofensiva – a dizer “pobre”, “triste” e “louco” – referem-se à categorias “Papal Social”, “Emoção” e “Caráter”, respectivamente.

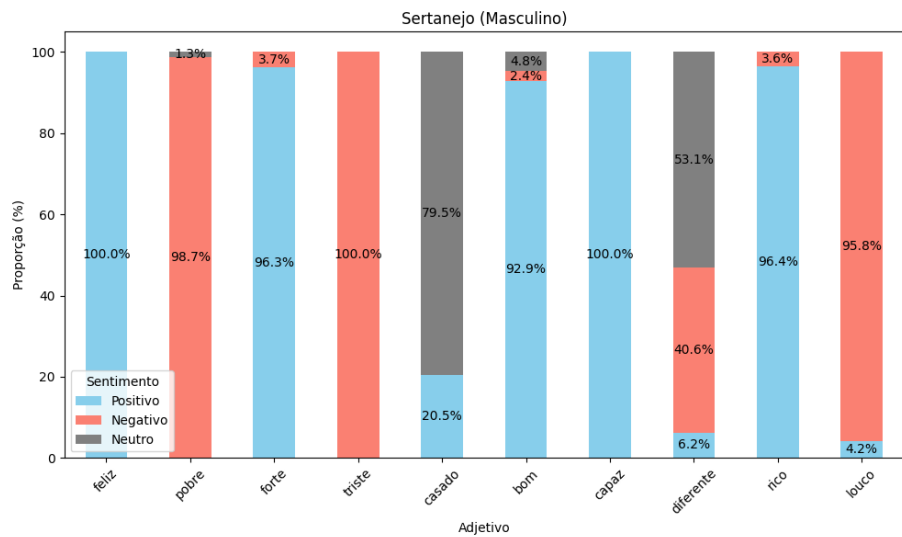


Figura 22. Adjetivos Mais Frequentes no Gênero Musical Sertanejo com Proporção dos Sentimentos para o Masculino

No Funk, como pode-se notar a partir da Figura 23, a intensidade do sexismo feminino aumenta: 5 dos 10 predicadores mais utilizados são pejorativos. Entre os 10 adjetivos, 5 dizem respeito à “Aparência” (“linda”, “gostosa”, “top”, “novinha”, “bonita”) e 4, ao “Caráter” (“louca”, “maluca”, “brava” e “boa”).

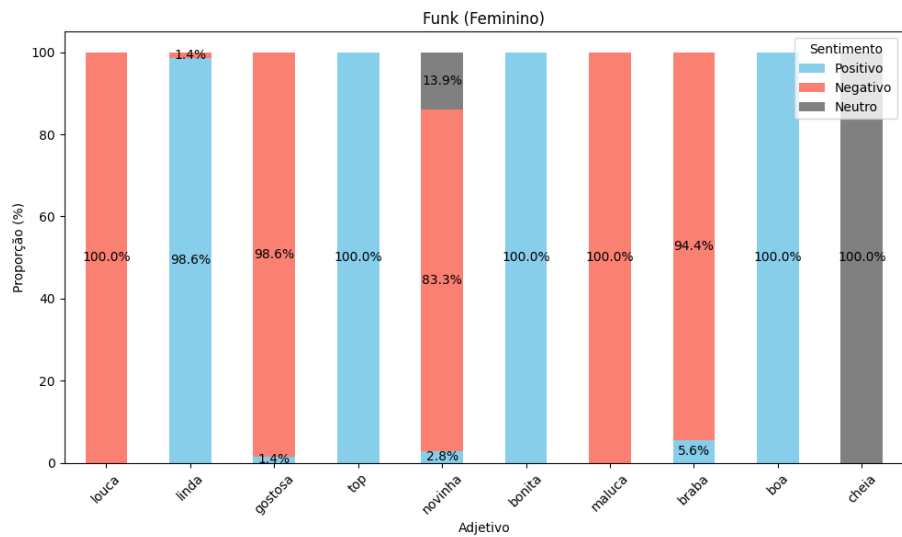


Figura 23. Adjetivos Mais Frequentes no Gênero Musical Funk com Proporção dos Sentimentos para o Feminino

Os adjetivos masculinos mais frequentes no gênero musical Funk apresentam, na maioria das vezes, conotação positiva que exalta virtudes, tais como bondade e fidelidade

(Figura 24.

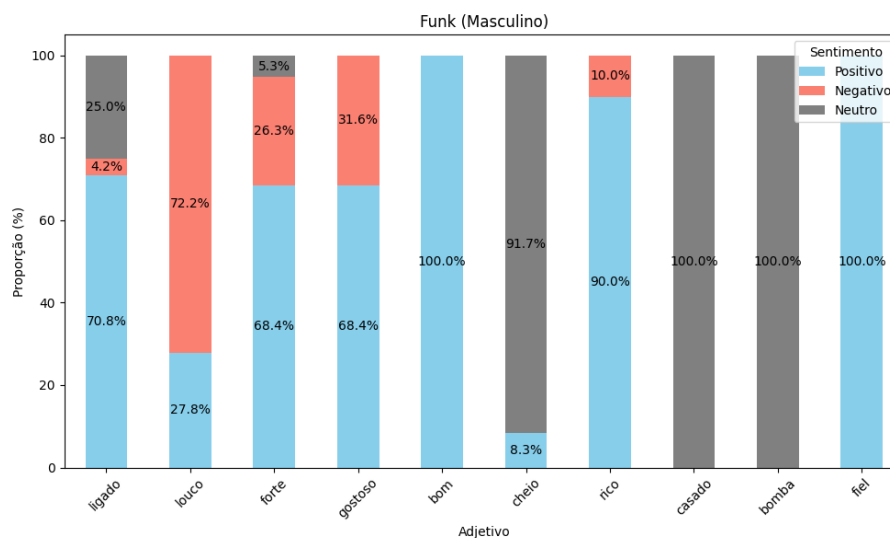


Figura 24. Adjetivos Mais Frequentes no Gênero Musical Funk com Proporção dos Sentimentos para o Masculino

Na MPB, dos 10 termos mais frequentemente utilizados, 2 fazem referências negativas substancialmente negativas: um à “Emoção” (“triste”) e outro ao “Caráter” (“louca”). Entre as referências positivas há menções principalmente à “Aparência” (“bonita”, “linda”, “bela”). Esses dados são descritos na Figura 25.

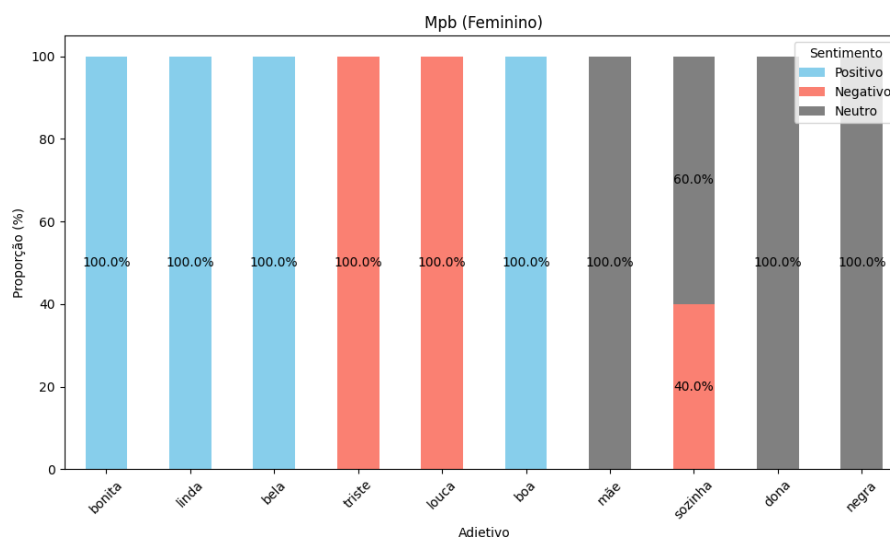


Figura 25. Adjetivos Mais Frequentes no Gênero Musical MPB com Proporção dos Sentimentos para o Feminino

Na Figura 26 são apresentados os sentimentos dos adjetivos masculinos mais frequentes nas canções de MPB. Destacam-se os termos negativos “pobre”, “velho” e “triste”.

As Figuras 27, 28, 29, 30, 31, 32, 33 ilustram as conotações dos termos mais frequentes nas canções de Samba, Rock, Rap, Pagode, Infantil, Hip Hop e Forró, respec-

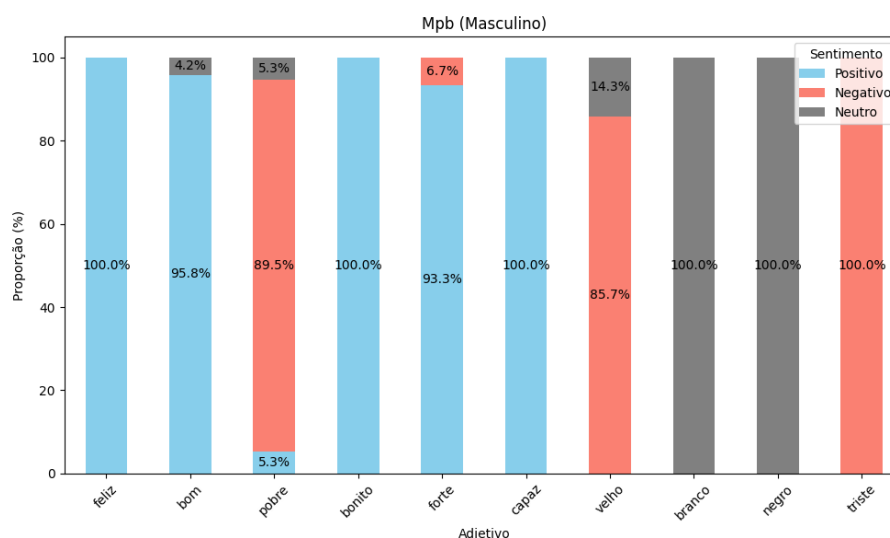


Figura 26. Adjetivos Mais Frequentes no Gênero Musical MPB com Proporção dos Sentimentos para o Masculino

tivamente. Nota-se que os gêneros Rock e Pagode também apresentam alta carga sexista contra mulheres nos dados analisados. No outro extremo, os estilos Hip Hop, Rap e Infantil apresentam menos conotação pejorativa. Destaque para o último gênero que mesmo direcionado a crianças apresenta dois adjetivos negativos frequentemente utilizados para descrever meninas (“teimosa” e “doente”).

No Samba, nota-se uma abertura para a discussão do papel social das mulheres (“dona”, “mãe”). Por fim, a palavra *louca* está presente em 8 dos 10 gêneros musicais analisados como uma das palavras mais utilizadas nas canções.

Pode-se dizer então que a intensidade dos estereótipos de gênero varia de acordo com os estilos musicais, mas o teor da tipificação é centrado na aparência física e emocional das mulheres.

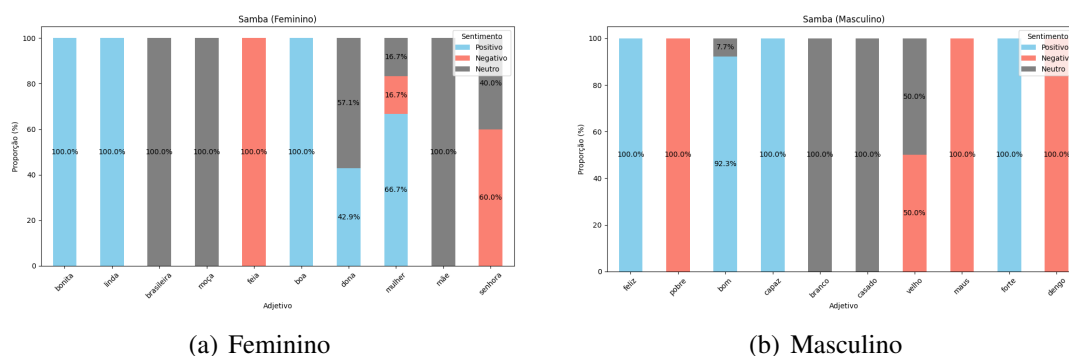
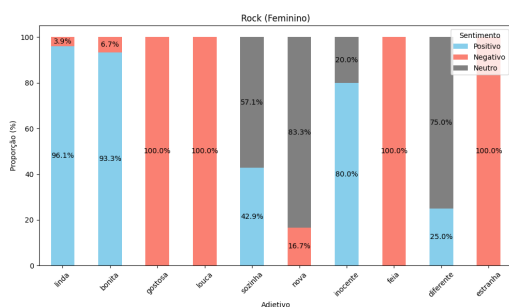


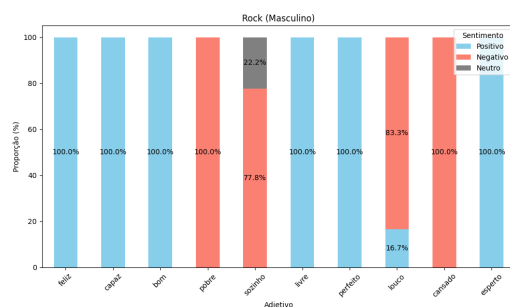
Figura 27. Adjetivos Mais Frequentes no Gênero Musical Samba com Proporção dos Sentimentos

7. Estereótipos da sociedade e a discriminação algorítmica

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

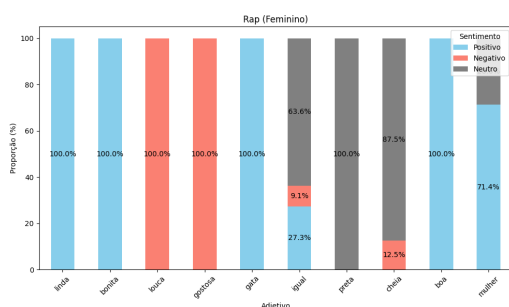


(a) Feminino

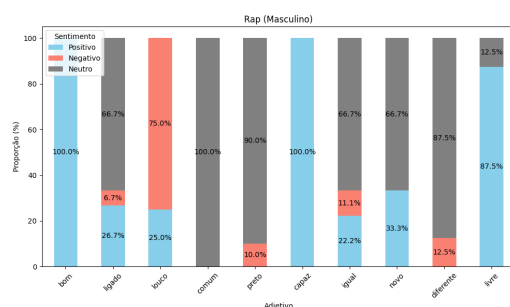


(b) Masculino

Figura 28. Adjetivos Mais Frequentes no Gênero Musical Rock com Proporção dos Sentimentos

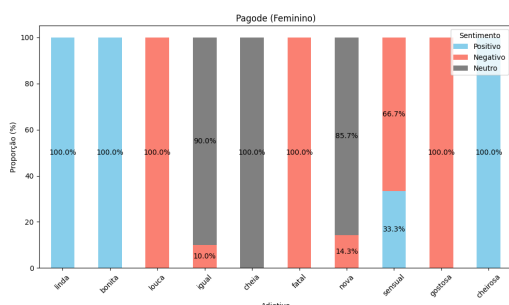


(a) Feminino

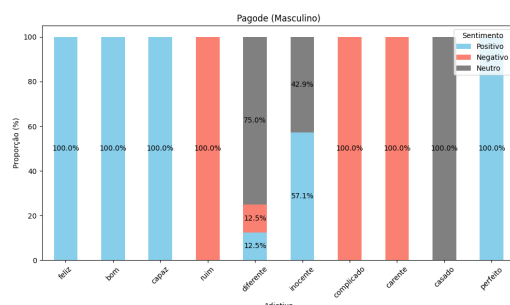


(b) Masculino

Figura 29. Adjetivos Mais Frequentes no Gênero Musical Rap com Proporção dos Sentimentos

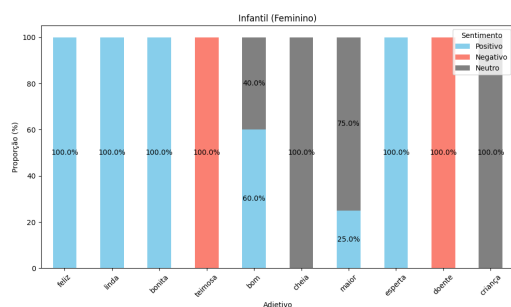


(a) Feminino

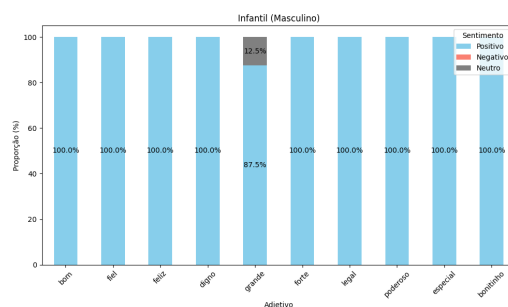


(b) Masculino

Figura 30. Adjetivos Mais Frequentes no Gênero Musical Pagode com Proporção dos Sentimentos

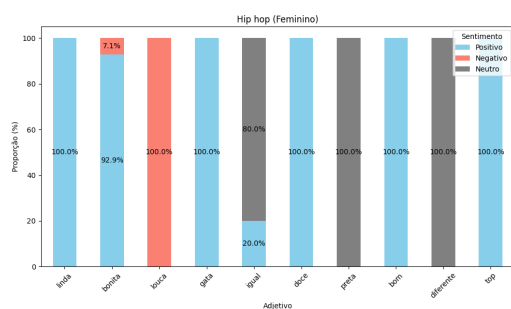


(a) Feminino

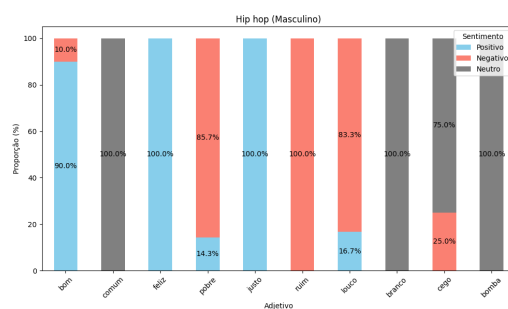


(b) Masculino

Figura 31. Adjetivos Mais Frequentes no Gênero Musical Infantil com Proporção dos Sentimentos

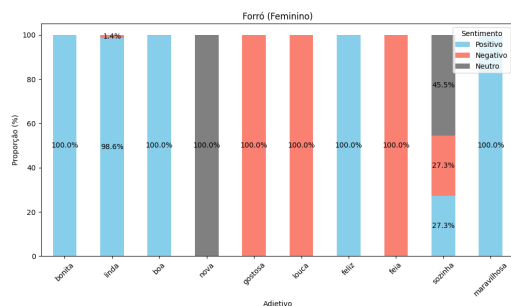


(a) Feminino

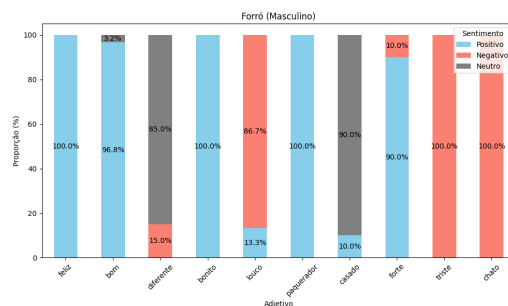


(b) Masculino

Figura 32. Adjetivos Mais Frequentes no Gênero Musical Hip Hop com Proporção dos Sentimentos



(a) Feminino



(b) Masculino

Figura 33. Adjetivos Mais Frequentes no Gênero Musical Forró com Proporção dos Sentimentos

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.

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.

[Moss-Racusin et al. 2012], demonstra o preconceito de gênero como uma preferência ou preconceito em relação a um gênero em detrimento de outro. [Sun et al. 2019] categoriza o preconceito de gênero em tarefas de Processamento de Linguagem Natural (PNL), em quatro formas: “denigração”, que se refere ao uso de termos cultural ou historicamente depreciativos; “estereotipagem”, que reforça estereótipos sociais existentes; “reconhecimento”, que envolve a imprecisão de algoritmos em tarefas de reconhecimento; e “sub-representação”, que descreve a baixa representação de um grupo específico.

A replicação de tendências discriminatórias, como a sub-representação ou estereotipagem das mulheres, pode agravar ainda mais o cenário de desigualdade de gênero. A ausência de um monitoramento rigoroso e de ajustes precisos nesses modelos pode resultar em ainda mais perdas de oportunidades para as mulheres, reforçando ciclos de exclusão e perpetuando a desigualdade.

A presença de vieses de gênero em modelos LLM (do inglês *Large Language Models*), como o GPT-3, é demonstrado por [Lucy and Bamman 2021], onde personagens femininos de histórias eram frequentemente associados com família e aparência, enquanto personagens masculinos relacionados a temas de poder e ação. Além disso. Em [Wan et al. 2023] com cartas de recomendação geradas evidenciaram uma forte disparidade na associação de adjetivos usados para descrever homens e mulheres. Enquanto as cartas para homens frequentemente descreviam qualidades de liderança e competência, das mulheres tendiam a destacar características mais calorosas e de apoio. No contexto brasileiro, [Soares et al. 2023] investigaram a tradução automática realizada pelo GPT-3.5 turbo, observando a presença de viés de gênero na tradução de profissões do inglês para o português. Em sentenças ambíguas em termos de gênero, a tradução para o português teve a tendência de atribuir ao gênero masculino profissões como “médico”, enquanto profissões como “enfermeira” receberam uma tradução feminina.

No projeto “Desvelar – Justiça Racial na Inteligência Artificial e TICs” [?], que mantém uma plataforma reunindo artigos e notícias sobre discriminação algorítmica, é apresentado uma linha do tempo que permite a visualização do impacto de incidentes envolvendo IA em diferentes contextos sociais. Entre os exemplos, pode-se citar situações como: recomendações de plataformas de streaming, artistas do gênero masculino rece-

bem mais de 55% das recomendações, contra apenas 20% das mulheres; um gerador de imagens que inclui armas em *prompts* relacionados a mulheres negras em favelas; um sistema para projetar risco de violência doméstica que descartou a proteção a uma mulher, que acabou assassinada em seguida, com outros 50 casos semelhantes identificados; o teclado do Google sugeriu termos sexuais para a palavra “neguinha”.

A utilização de letras de músicas em modelos de aprendizado de máquina pode apresentar riscos significativos quando não há uma curadoria cuidadosa e diversificada de dados. Em gêneros que frequentemente utilizam descrições pejorativas para retratar mulheres, os algoritmos podem aprender e perpetuar tais estereótipos, refletindo e amplificando essas representações enviesadas. Dessa forma, a presença reiterada de representações estereotipadas pode levar à propagação dessas concepções em tarefas subsequentes, influenciando desde a classificação de conteúdos até a geração automática de textos e imagens.

Para mitigar esses efeitos, torna-se crucial que os datasets utilizados incluam uma diversidade representativa da cultura popular, ao mesmo tempo que sejam cuidadosamente examinados para evitar que determinados padrões perpetuem preconceitos. Essa seleção criteriosa auxilia no desenvolvimento de modelos que respeitem a pluralidade cultural e, simultaneamente, mantenham-se atentos às nuances de representações sociais. Em última análise, o treinamento com conjuntos de dados culturalmente informados, mas balanceados, permite que a tecnologia reflita uma visão mais ética e inclusiva, evitando a propagação inadvertida de estereótipos e favorecendo uma interpretação mais equitativa dos conteúdos provenientes da cultura popular.

8. Trabalhos futuros

Embora este estudo tenha avançado na compreensão da caracterização de mulheres na música brasileira, ainda há muitas questões a serem exploradas.

8.1. O sexismo ao longo do tempo e em diferentes manifestações artísticas

Não foi possível dimensionar a evolução da estereotipação feminina neste trabalho, pois o corpus União não apresenta de maneira consistente e confiável informações a respeito da data de lançamento das músicas. Contudo, experimentos iniciais podem ser feitos com o MusicSpotsBr, explorando características exclusivas desse conjunto de dados.

Além disso, seria relevante investigar as interações entre gravadoras, compositores e intérpretes, analisando como essas relações influenciam a escolha dos temas e a forma como os papéis de gênero são representados nas letras das músicas. Ampliar a análise para incluir outras formas de mídia, tais como cinema, televisão e redes sociais, permitiria verificar se os padrões de objetificação e estereótipos de gênero observados nas músicas também estão presentes em outros contextos midiáticos. As novas análises poderiam fornecer uma compreensão mais abrangente das representações de gênero na cultura popular e sua influência na formação de estereótipos.

8.2. A caracterização de mulheres por meio de suas ações

A partir dos verbos mais frequentemente utilizados por personagens femininos nas músicas, é possível verificar quais as ações mais frequentes para cada gênero. A hipótese

é de que os verbos utilizados no contexto masculino sugerem papéis discursivos ou intelectuais, enquanto verbos utilizados no contexto feminino revelam passividade, foco em expressões de desejo ou aspiração. Vislumbra-se ainda o estudo do emprego de métricas de controles emocionais para qualificar as ações descritas nas canções.

8.3. A sexualização de crianças e adolescentes

A presença de termos como “novinha” nas letras musicais brasileiras aponta para uma preocupante hipersexualização de crianças e adolescentes. Em decorrência, destacamos também a temática da banalização da exploração sexual de menores, que frequentemente aparece velada nas composições. A repetição desses discursos nas músicas tende a legitimar comportamentos de consumo e de aceitação passiva de imagens sexualizadas de adolescentes, o que reforça uma cultura permissiva em relação à vulnerabilidade desses grupos. Como trabalho futuro, propomos uma análise da ocorrência deste tipo de predador nos gêneros musicais.

8.4. Modelo para classificação de sentenças sexistas

Propõe-se a construção de um modelo de classificação de sentenças sexistas, utilizando tanto as anotações de sentimentos quanto as anotações de categorização de sentenças em diferentes dimensões (aparência, caráter, sexual, emoção e papel social). Esses subcorpus anotados servirão para treinar o modelo, permitindo uma análise mais precisa e automatizada do potencial preconceituoso em qualquer conjunto de dados. Com esse modelo, será possível identificar, com maior rigor, padrões linguísticos que perpetuam estereótipos e preconceitos de gênero, evitando que dados problemáticos sejam utilizados para treinar novos modelos. Essa abordagem contribuirá significativamente para o desenvolvimento de sistemas mais éticos e inclusivos, ao permitir uma triagem de datasets para reduzir vieses prejudiciais e promover um ambiente de aprendizado de máquina mais equitativo.

8.5. Multidisciplinaridade na análise de dados

Por fim, destaca-se que a análise dos resultados obtidos não deve ser restrita a profissionais da Computação. Uma colaboração interdisciplinar é fundamental para o avanço de uma compreensão mais ampla e crítica da representação de gênero na cultura popular. Popularizar o conhecimento sobre PLN e IA, assim como torná-las acessíveis a especialistas de outras áreas é crucial para que sociólogos, antropólogos, linguistas e outros estudiosos possam examinar os dados com maior profundidade e em seus respectivos domínios de interesse.

9. Conclusões

O objetivo deste trabalho foi analisar a presença de vieses de gênero na música brasileira. Para isso, foi criado um grande corpus contendo 146.612 obras nacionais, distribuídas em 19 gêneros musicais. Através de uma leitura distante apoiada por técnicas de PLN e anotações manuais, o estudo mostrou que mulheres são frequentemente retratadas por sua aparência e estado emocional. De maneira geral, as descrições utilizadas para descrever o gênero feminino têm conotação positiva, mas quando a interpretação do conteúdo é aproximada por predicativos mais frequentes em cada categoria de descrição, nota-se que muitos termos são ofensivos, principalmente quando fazem referência a caráter e

emoção. Quando os termos são relativos a sexualidade, as mulheres são rotineiramente retratadas por palavras que remetem à hipersexualização.

A limitação do papel social da mulher na sociedade também ficou evidente nas músicas. A análise das profissões e dos agrupamentos de termos nas letras de músicas brasileiras revelou padrões significativos que reforçam estereótipos de gênero. As profissões associadas ao poder e à autoridade, são predominantemente masculinas, enquanto as ocupações vinculadas ao cuidado, ensino e entretenimento, como “atriz”, são majoritariamente atribuídas ao gênero feminino. Isso demonstra uma divisão de papéis profissionais e perspectiva de liderança e gestão, refletindo questões culturais sobre os limites de atuação entre os gêneros apontado pelo IBGE.

Nossos resultados demonstram que os estereótipos de gênero variam em intensidade conforme o estilo musical, mas tendem a ser centralizados na aparência física das mulheres. A análise das sentenças mais populares (ou das músicas do top 10) em cada gênero, utilizando um classificador de sentimento, permitiu identificar nuances emocionais que elucidam ou questionam papéis de gênero tradicionais.

Por fim, nossos resultados mostram que o sexismo está presente em uma pequena parcela das músicas: 2, 29% para mulheres e 2, 17% para homens. No entanto, é importante destacar a intensidade do teor das músicas consideradas sexistas, principalmente no cenário brasileiro em que mulheres são frequentemente representadas de forma estereotipada, reforçando desigualdades de gênero e perpetuando ideias limitantes sobre o papel feminino na sociedade, conforme dados do IBGE e evidenciado neste trabalho. Além disso, há de se enfatizar a possibilidade dessas canções serem utilizadas no treinamento de grandes modelos de linguagem que estão sendo amplamente adotadas pela sociedade.

Nossos resultados contribuem e vão ao encontro daqueles obtidos por [Freitas and Martins 2023], mesmo utilizando um corpus com domínio, linguagem e sazonalidade diferentes. Nossas conclusões também corroboram [Firmino et al. 2024], confirmando a representação estereotipada das mulheres na sociedade através da música, e incorporando outros diversos aspectos.

Este trabalho apresenta limitações relacionadas à biblioteca de PLN utilizada, cujos algoritmos podem influenciar a qualidade da análise das letras das músicas. Esse problema torna-se ainda mais pronunciado devido às estruturas linguísticas não convencionais das canções, o que pode comprometer a precisão da extração de informações. Para mitigar esses problemas, nos comprometemos a realizar uma análise manual adicional, garantindo maior rigor e confiabilidade nos resultados. Além disso, a definição de sexismo, por sua natureza subjetiva, representa um desafio, pois envolve interpretações que podem variar conforme o contexto cultural e individual. Este estudo procura oferecer uma análise fundamentada, ainda que consciente das nuances e desafios que envolvem o tema do sexismo e sua identificação em contextos.

Referências

- Aghajanyan, A., Zettlemoyer, L., and Gupta, S. (2020). Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *CoRR*, abs/2012.13255.
- Assis, C. L. (2014). “entre tapas e beijos”: representações sociais sobre a violência de gênero para adolescentes. 2:229–242.

- Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3):671–732.
- Benatti, R., Severi, F., Avila, S., and Colombini, E. L. (2024). Gender bias detection in court decisions: A brazilian case study. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 746–763, New York, NY, USA. Association for Computing Machinery.
- Betti, L., Abrate, C., and Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1).
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.
- Caseli, H. M. and Nunes, M. G. V., editors (2024). *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2 edition.
- 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.
- Firmino, V., Lopes, J., and Reis, V. (2024). Identificando padrões de sexismo na música brasileira através do processamento de linguagem natural. In *Anais do V Workshop sobre as Implicações da Computação na Sociedade*, pages 59–69, Brasília, DF, Brasil. SBC.
- 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.
- Howard, J. and Ruder, S. (2018). Fine-tuned language models for text classification. *CoRR*, abs/1801.06146.
- Huang, J. (2022). Changing semantics of gendered insults in music lyrics. (THESIS).
- IBGE (2024). Estatísticas de gênero: Indicadores sociais das mulheres no brasil. Acesso em: 25/08/2024.
- 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.
- Lucy, L. and Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. In Akoury, N., Brahman, F., Chaturvedi, S., Clark, E., Iyyer, M., and Martin, L. J., editors, *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.
- Moretti, F. (2008). *A literatura vista de longe*. Editora Arquipélago.
- Moss-Racusin, C., Dovidio, Brescoll, V., Graham, M., and Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109:16474–16479.

- 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.
- Sanches, C. D. L. N. P. L. (2009). A construção do eu feminino na música popular brasileira. *Caderno Espaço Feminino*, 21:181–205.
- Schlösser, A. and Fantin, G. (2022). Percepção de mulheres acerca de conteúdos sexuais na música popular do Brasil. *Ayvu: Revista de Psicologia*, 9.
- Soares, T., Gumiel, Y., Junqueira, R., Gomes, T., and Pagano, A. (2023). Viés de gênero na tradução automática do gpt-3.5 turbo: avaliando o par linguístico inglês-português. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 167–176, Porto Alegre, RS, Brasil. SBC.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). *BERTimbau: Pretrained BERT Models for Brazilian Portuguese*, pages 403–417.
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.
- 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.
- Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. (2023). “kelly is a warm person, joseph is a role model”: Gender biases in LLM-generated reference letters. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.