Retrato_Cantado: Criação e Análise de um Corpus para Representações de Identidade em Letras de Músicas Brasileiras

Vitória P. Firmino[1], Janaina N. de S. Lopes[1], Bruno M. Nogueira[1], Valéria Q. dos Reis[1,2]
[1]Universidade Federal de Mato Grosso do Sul
[2]Leuphana Universität Lüneburg
Correspondence: {vitoria.firmino,janaina.nogueira,bruno.nogueira,valeria.reis}@ufms.br

Resumo

This paper presents the development of Retrato_Cantado, a dataset of sentences extracted from Brazilian song lyrics and manually annotated to identify and categorize predicative constructions that describe individuals. The corpus findings validate the effectiveness of lexical-syntactic patterns for identifying predicative sentences, confirming their suitability for large-scale linguistic annotation tasks. The dataset also serves as a valuable resource for the analysis of textual discourse and the representation of social groups in Brazilian culture. We additionally trained a person-characterization classifier to illustrate the applicability of the dataset to the automatic detection of predicative descriptions, which achieved high accuracy and highlights the potential for creating more specialized models capable of detecting physical and sociocognitive categories, as well as performing sentiment polarity analysis.

1 Introdução

Para Aristóteles, a arte constitui uma forma de representação simbólica da realidade, por meio da qual se expressam experiências, emoções e aspectos essenciais dos fenômenos humanos, muitas vezes de maneira mais perceptível do que na realidade imediata (Aristotle, 1995). Nesse sentido, a música ocupa um papel central ao expressar dimensões sociais, culturais e emocionais vivenciadas por indivíduos e grupos em contextos históricos específicos, funcionando como um reflexo das experiências coletivas e individuais. Conforme argumenta Kong (1995), entretanto, a música não apenas reflete a vida social, mas também contribui para reproduzi-la, atuando como um veículo de circulação de emoções, construções sociais e estereótipos. Essa compreensão dialoga com a concepção de discurso de Foucault (2012), segundo a qual a produção discursiva é regulada e legitimada por relações de poder, sendo autorizada apenas quando compatível com a manutenção das estruturas vigentes. Assim, a música pode ser compreendida como um discurso que participa ativamente da criação, reprodução e manutenção de representações sociais.

Apesar dos avanços recentes, grande parte dos estudos sobre representações sociais na música brasileira ainda se apoia em análises predominantemente manuais (Sanches, 2009; Assis, 2014; Schlösser e Fantin, 2022). Abordagens automatizadas baseadas em Processamento de Linguagem Natural (PLN), capazes de analisar grandes volumes de dados e sem a necessidade de muitos recursos humanos capacitados, têm sido mais exploradas no contexto internacional (Huang, 2022; Betti et al., 2023; Barman et al., 2019; Casanovas-Buliart et al., 2024), permanecendo relativamente escassas no cenário brasileiro. Nesse panorama, os trabalhos de Firmino et al. (2024) e Lopes et al. (2025) destacam-se como iniciativas pioneiras ao empregar expressões regulares para a identificação de sentenças predicativas em letras de músicas, contribuindo para evidenciar manifestações de sexismo nesse domínio. Ainda assim, tais estudos apresentam limitações, especialmente no que se refere à ausência de validação sistemática das regras sintáticas propostas e à pouca consideração do contexto semântico dos predicativos identificados, fatores que podem restringir a precisão e a interpretabilidade das análises.

Ao longo deste artigo, adotamos o termo "sentença predicativa" para designar construções que atribuem alguma forma de caracterização de pessoas dentro de uma oração. Do ponto de vista semântico, adjetivos podem ser interpretados como expressões que atribuem propriedades a indivíduos (Kamp, 1975). Do ponto de vista sintático, tais elementos podem desempenhar funções como adjunto adnominal e predicativo (Cunha e Cintra, 2008). No segundo caso, conforme descrições funcionais do português, diferentes classes de palavras, como adjetivos e substantivos, podem exercer função pre-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 687–696

dicativa ao estabelecer relações de caracterização (Neves, 2000). Nesse sentido, consideramos construções envolvendo verbos de ligação ou copulativos (que estabelecem uma relação entre o sujeito e uma propriedade ou estado, como em *ser* ou *estar*), bem como perífrases verbais que expressam mudança ou progressão de estado (como em *estar ficando* ou *vir a ser*). Também são considerados apostos e adjuntos adnomiais. Estão presentes ainda os predicativos nominais, elementos que atribuem propriedades ou classificações ao sujeito ou ao objeto da oração, e modificadores adverbiais que intensificam ou especificam a predicação. A representação dessas relações sintático-semânticas é particularmente relevante para a modelagem linguística e computacional (Bertoldi e Chishman, 2006). Exemplos representativos incluem: *"Ela está ficando cansada."*, *"Mário é professor."*, *"Ela foi homenageada."* e *"Mariana, a vaca."* e *"Lili sonhadora."*.

O corpus resultante constitui um recurso relevante para o treinamento de modelos compactos voltados à extração automática de descrições de pessoas em textos da língua portuguesa. A partir das sentenças predicativas identificadas, torna-se possível conduzir estudos de análise sociológicas e computacionais sobre representações de grupos sociais ao longo do tempo e em diferentes gêneros musicais, contribuindo tanto para o fortalecimento da análise cultural quanto para a identificação de discriminações reproduzidas no discurso musical.

As análises conduzidas evidenciam que a interpretação automática de letras de músicas é uma tarefa desafiadora, devido à informalidade linguística e à dependência de contexto, mas também indicam que abordagens baseadas em padrões sintáticos são eficazes para a identificação de sentenças predicativas, apresentando elevada concordância entre anotadores e desempenho satisfatório em modelos treinados a partir do corpus. Para a construção do *Retrato_Cantado*, contou-se com a colaboração de 38 anotadores, que rotularam 18.778 sentenças extraídas de 146.612 canções brasileiras lançadas entre 1947 e 2024.

Como principais contribuições, este trabalho disponibiliza (i) um corpus anotado manualmente para a caracterização de pessoas em letras de músicas brasileiras, (ii) uma análise sistemática desse recurso e (iii) um modelo treinado para a detecção automática de sentenças predicativas. Os recursos produzidos são fundamentais para o desenvolvimento de ferramentas de análise sociológica e computacional em língua portuguesa, incluindo aplicações voltadas à identificação de discriminação, alinhando-se aos Objetivos de Desenvolvimento Sustentável 5 e 10 das Nações Unidas, referentes à igualdade de gênero e à redução das desigualdades.

2 Trabalhos correlatos

A literatura relacionada a este trabalho abrange tanto metodologias de criação e anotação de corpora quanto análises linguísticas e computacionais voltadas à caracterização de representações sociais em obras artísticas, em especial na música e na literatura.

A identificação manual de sentenças predicativas é uma tarefa complexa, frequentemente apoiada por ferramentas de análise léxico-sintática. Nesse contexto, Freitas e Martins (2023) empregam analisadores sintáticos em uma abordagem de leitura distante para caracterizar personagens dos gêneros masculino e feminino em obras literárias brasileiras, analisando a distribuição quantitativa de predicações no corpus. De forma semelhante, Firmino et al. (2024) e Lopes et al. (2025) utilizam expressões regulares e padrões sintáticos para identificar predicações e profissões atribuídas a mulheres em letras de músicas brasileiras, evidenciando manifestações de sexismo no discurso musical. Embora esses estudos representem avanços importantes, carecem de validação sistemática das regras sintáticas propostas e não exploram de maneira aprofundada o contexto semântico ou emocional das sentenças identificadas. Neste trabalho, avançamos nesse sentido ao validar manualmente padrões léxico-sintáticos previamente propostos, além de treinar um modelo de classificação automática para a identificação de sentenças predicativas, que são ainda anotadas segundo categorias físicas e sociocognitivas, bem como quanto à sua carga emocional.

No âmbito dos estudos linguísticos, pesquisas recentes têm aprofundado a análise das relações léxico-semânticas (Oliveira et al., 2024) e das propriedades léxico-sintáticas de adjetivos no português (Martinez et al., 2024). Esses trabalhos fornecem fundamentos teóricos relevantes para a compreensão dos mecanismos linguísticos subjacentes à atribuição de características a indivíduos, especialmente no que se refere às estruturas sintáticas responsáveis pela realização da predicação, aspectos centrais para a abordagem adotada neste estudo.

No que diz respeito a análises automatizadas em larga escala, estudos internacionais têm empregado

técnicas de PLN e Aprendizado de Máquina para examinar grandes volumes de letras musicais em busca de padrões discursivos relacionados a gênero e discriminação. Huang (2022), analisa mais de 237 mil músicas em inglês por meio de embeddings dinâmicos baseados em BERT, com o objetivo de investigar mudanças semânticas ao longo do tempo, destacando diferentes usos e conotações do termo *bitch*. De modo complementar, Betti et al. (2023) utilizam modelos supervisionados para analisar aproximadamente 378 mil letras em inglês, evidenciando tendências crescentes de sexismo em músicas de artistas masculinos ao longo de cinco décadas. Já Casanovas-Buliart et al. (2024) investigam a evolução do sexismo em músicas populares espanholas entre 1960 e 2022, indicando um aumento significativo de conteúdos sexistas a partir de 2015.

Apesar da relevância desses estudos, observa-se que análises computacionais de grande escala aplicadas a letras de músicas em língua portuguesa ainda são escassas. Além disso, não há, até onde sabemos, corpora de músicas brasileiras anotados especificamente para a identificação e caracterização de sentenças predicativas. Embora existam corpora voltados à detecção de discurso de ódio em português (Fortuna et al., 2019; Leite et al., 2020; Vargas et al., 2022; Oliveira et al., 2023; Trajano et al., 2023), tais recursos não contemplam a granularidade linguística necessária para a análise de estruturas predicativas. Nesse sentido, a abordagem proposta neste trabalho, ao combinar anotação sintática, categorização semântica e rotulagem emocional, oferece uma alternativa mais refinada para o estudo computacional das representações sociais no discurso musical em língua portuguesa.

3 Método de criação do corpus

Esta seção apresenta os métodos utilizados na criação do *Retrato_Cantado*.

O fluxo metodológico seguido está detalhado no diagrama ilustrado na Figura 1.

3.1 Coleta de dados

Para o desenvolvimento da pesquisa, utilizamos o corpus compilado por Lopes et al. (2025). O recurso apresenta um total de 146.612 canções, distribuídas em 73 (sub)gêneros, lançadas entre os anos de 1.947 e 2.024, contendo as colunas: *Nome da música*, *Artista*, *Gênero da Música* e *Letra da Música*.

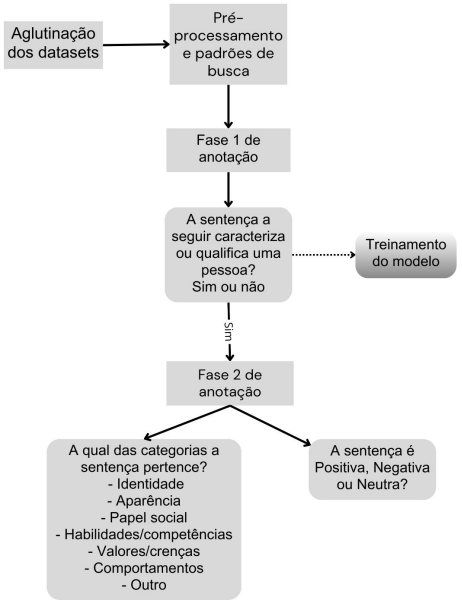

Figura 1: Diagrama do fluxo de criação do *corpus*.

3.2 Pré-processamento e padrões de busca

O pré-processamento das letras de músicas foi realizado com a biblioteca spaCy. Utilizou-se um modelo pré-treinado de grande porte para o português, capaz de fornecer anotações linguísticas detalhadas, como classes gramaticais, lemas, traços morfológicos e dependências sintáticas.

Cada letra foi segmentada em sentenças, que passaram a constituir a unidade básica de análise. O objetivo dessa etapa foi identificar construções predicativas responsáveis por caracterizar pessoas, isto é, sentenças que atribuem propriedades, estados, papéis ou qualidades a sujeitos humanos.

Para essa identificação, foram desenvolvidos padrões de busca léxico-sintáticos inspirados no trabalho de Lopes et al. (2025) e posteriormente expandidos para contemplar uma maior diversidade de estruturas gramaticais observadas no corpus. Esses padrões foram implementados por meio do *Matcher* do spaCy e combinam informações de classe gramatical, traços morfológicos e ordem sequencial dos constituintes na sentença. A utilização direta de relações de dependência sintática também foi considerada. No entanto, experimentos preliminares indicaram que, no domínio de letras de música — caracterizado por construções com maior liberdade estilística — abordagens baseadas exclusivamente em árvores de dependência tendiam a produzir um número elevado de correspondências espúrias. Assim, optou-se por padrões léxico-sintáticos mais controlados, capazes de preservar

maior precisão na identificação das construções predicativas de interesse.

No total, foi definido um conjunto extensivo de padrões-base, cada um deles expandido em múltiplas variações estruturais. Essas variações permitem capturar diferentes formas de realização do sujeito, incluindo sujeitos nominais explícitos, nomes próprios e pronomes pessoais, o que amplia significativamente a cobertura dos padrões sem comprometer sua precisão sintática.

Os padrões contemplam construções envolvendo verbos copulativos, verbos de predicação verbal e perífrases verbais, bem como modificadores adverbiais e predicativos nominais, de modo a refletir a diversidade sintática e estilística presente na língua portuguesa. A descrição completa dos padrões implementados, assim como exemplos representativos das sentenças capturadas, está disponível publicamente[1].

As 18.731 sentenças que correspondiam a pelo menos um dos padrões de busca foram selecionadas para compor o conjunto de dados a ser anotado. O propósito era verificar manualmente a classificação produzida pelos critérios automáticos. Assim, seria possível demonstrar que o filtro baseado em padrões conseguia identificar corretamente construções predicativas. Outras 100 sentenças que não foram classificadas positivamente pelos padrões de busca também foram incluídas no conjunto de anotação, como forma de controle de falsos negativos.

3.3 Anotações de sentenças

As anotações foram realizadas de forma manual, garantindo maior confiabilidade às análises.

Ocorreram duas etapas de anotações. A primeira identificava se uma sentença descrevia uma pessoa. A segunda, caracterizava tal descrição a partir de categorias e da polaridade da sentença predicativa.

3.3.1 Caracterização de pessoas

Na primeira fase de anotação, os colaboradores tinham que responder à seguinte pergunta: *A sentença a seguir caracteriza ou qualifica uma pessoa?*

A resposta consistia em *"sim"* ou *"não"*, permitindo filtrar apenas as sentenças que descreviam uma pessoa, auxiliando, assim, a validar os padrões léxico-sintático usados.

Essa etapa foi essencial para identificar sentenças que não expressavam predicação pessoal, como expressões genéricas ou sentenças sem referência humana explícita.

Somente as sentenças majoritariamente classificadas com a resposta "sim" foram consideradas para a Fase 2.

3.3.2 Classificação das sentenças

Na Fase 2, foram anotadas manualmente 15.312 sentenças em relação à polaridade. As considerações para cada categoria foram:

- **Positiva**: Sentenças que expressam opiniões ou informações favoráveis sobre o gênero. Por exemplo: "*Meninas são muito inteligentes.*".
- **Negativa**: Sentenças que contêm comentários desfavoráveis, degradantes ou pejorativos sobre o gênero. Por exemplo: "*Ele é feio e chato.*".
- **Neutra**: 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.*".

3.3.3 Categorização de predicações

Na Fase 2, para a análise das sentenças predicativas, os anotadores responderam à seguinte pergunta: *A qual das categorias a sentença pertence?*

Considerando o interesse na caracterização humana, foram consideradas seis categorias que refletem os diferentes aspectos da humanidade: *identidade*, *aparência*, *papel social*, *habilidades/competências*, *valores/crenças* e *comportamentos*.

As categorias utilizadas foram elaboradas com base na literatura e validadas por um grupo de estudos em Psicologia e Sociologia, de modo a melhor refletirem diferentes dimensões daquilo que constitui o ser humano (Hall, 2014; Borges, 2019). A seguir, apresenta-se uma breve descrição de cada categoria adotada na análise. O foco é dado nos gêneros masculino e feminino por serem os grupos sociais mais bem representados nas sentenças. Porém, outros grupos poderiam ser protagonistas da análise tais como pessoas negras, com deficiência ou idosos.

Partindo da concepção de gênero como uma categoria relacional, que produz relações de poder entre homens e mulheres, sendo as hierarquias e subordinações daqui resultantes expressas em diversos planos, dentre eles, o simbólico, no qual

[1]Veja link disponibilizado no final do artigo.

enquadramos as músicas. A partir das representações sociais femininas produzidas e reproduzidas nessa forma de discurso, visualizamos prescrições e normatizações direcionadas às mulheres, com a finalidade de construção e conformação de suas identidades subjetivas e também para justificar práticas e discursos que violam seus direitos.

Com isso, existem prescrições do que é ser homem/mulher (*identidade*), quais atributos físicos são desejados e indesejados para cada gênero (*aparência*), quais funções sociais cabem a eles (*papel social*), o que as pessoas podem e devem fazer (*habilidades/competências*), quais valores e normas devem reger suas vidas (*valores/crenças*) e como elas devem agir (*comportamentos*). Dessa forma, elencamos seis categorias como ilustrativas das relações na música brasileira e produtoras de discursos e práticas que buscam a manutenção da ordem.

Para pensar a primeira categoria, *Identidade*, partimos do campo sociológico e antropológico, definindo-a como um processo contínuo, fluido e situacional no qual adquirimos e perdemos elementos compartilhados coletivamente que garantem o sentimento de pertencimento a um grupo e que nos ajudam a definir quem nós somos, quais habilidades devemos desenvolver, quais capacidades nos são cobradas, etc. (Hall, 2014). Dessa forma, a *identidade* é construída e reconstruída socialmente, culturalmente e historicamente. Quando afirmamos que "somos homens" ou "somos mulheres", por exemplo, partimos do que está prescrito em cada sociedade e em cada cultura para indivíduos que adotam essas identidades, ou seja, como ser homem e como ser mulher. No entanto, existe uma ideia essencialista de que homens e mulheres são determinados biologicamente, ou seja, já nascem com atributos e capacidades inerentes ao ser. A visão essencialista e determinista sobre identidade de gênero ainda pauta práticas e discursos cotidianos, reproduzindo estereótipos de gênero e aprisionando homens e mulheres em concepções fechadas, sendo tudo isso expresso em artefatos culturais como as músicas.

"Me desculpem as feias, mas beleza é fundamental". Com essa frase do poeta brasileiro Vinícius de Moraes, introduzimos a segunda categoria, *Aparência*, que reproduz a ideia construída ao longo de séculos e nas mais diversas sociedades de que as mulheres são o belo sexo, ou seja, indivíduos, cujo único atributo é a beleza, já que seriam desprovidas de capacidades mentais. Historicamente, as mulheres são descritas e pensadas a partir de atributos físicos, sendo a beleza do corpo aquela que garantiria êxito e trânsito social a esse grupo social; em contrapartida, poder e inteligência são considerados atributos essencialmente masculinos (Borges, 2019). Dessa forma, firma-se no imaginário social uma série de representações sociais femininas pautadas em adjetivações positivas (linda, perfeita, gostosa etc.) e adjetivações negativas (feia, mocreia, velha etc.) que emergem nas músicas e reproduzem a ideia de que o essencial para as mulheres é a aparência física.

Por fim, temos as quatro categorias restantes (*Papel Social*, *Habilidades/Competências*, *Valores/Crenças*, *Comportamentos*) que estão interligadas no imaginário social produzido pela lógica patriarcal. Historicamente, construíram-se as noções de feminino e feminilidade atreladas a um conjunto de características, atributos, papeis sociais e comportamentos que seriam inerentes às mulheres, expressos nos exemplos a seguir: donas de casa, mães e cuidadoras (*Papel Social*); que cuidam da casa e de todos ao seu redor (*Habilidades/Competências*); que acreditam no amor e são fiéis (*Valores/Crenças*); recatadas, delicadas e amáveis (*Comportamentos*). Portanto, essas categorias nos ajudam a visualizar como os discursos presentes nas músicas contribuem para prescrever e normatizar formas de ser, agir e pensar que seriam corretas para as mulheres dentro da lógica patriarcal.

Além disso, foi criada a categoria *Outro*, destinada a agrupar ocorrências de caracterização que não se encaixavam claramente nas dimensões semânticas definidas no esquema de anotação ou que apresentavam múltiplas possibilidades interpretativas. Trata-se, portanto, de uma classe residual que captura predicações cuja interpretação depende fortemente do contexto ou não corresponde diretamente às categorias estabelecidas. Exemplos incluem frases como *"Se está com ele está sozinha."*, *"Você é outra história."* e *"Filha única do meu amigo."*.

3.4 A interface de anotação

A interface de anotação foi desenvolvida com a ferramenta Streamlit[2], o que possibilitou disponibilizar aos colaboradores um sistema web leve e de fácil acesso. Cada participante possuía suas próprias credenciais individuais de acesso (usuário e

[2] https://streamlit.io/.

senha).

3.5 Os anotadores

Todo o processo de anotação contou com a participação de 38 voluntários, estudantes de graduação da Universidade Federal de Mato Grosso do Sul. Os participantes eram oriundos dos cursos de Ciência da Computação (15), Sistemas de Informação (7), Engenharia de Software (6), Engenharia de Computação (5), Psicologia (3), Fisioterapia (1) e Ciência de Dados (1), com idades entre 18 e 31 anos. Dos participantes, 17 eram do gênero feminino e 21 do gênero masculino. A participação foi voluntária e não remunerada, mas garantiu certificação de 40 e 45 horas de atividades de extensão para os participantes da primeira e segunda fases de rotulação.

Antes do início de cada etapa de anotação, foi realizada uma reunião explicativa com os estudantes colaboradores para esclarecer os critérios de anotação e garantir um entendimento compartilhado entre os participantes. Durante o processo de rotulação, os participantes podiam se comunicar com a equipe de pesquisa via um grupo de aplicativo de mensagens criado especificamente para esse fim.

3.6 Consolidação de anotações

Inicialmente, cada sentença foi atribuída a 3 pessoas anotadoras. No processo de treinamento do modelo de classificação, somente sentenças rotuladas por pelo menos duas pessoas foram consideradas.

Na primeira fase da anotação (*"A sentença a seguir caracteriza ou qualifica uma pessoa?"*), as divergências foram resolvidas com base na classificação majoritária. Em caso de empate, assumiu-se que, por não haver consenso suficiente, a sentença seria classificada negativamente. Ao final desta etapa, obteve-se uma taxa de concordância de 89,31% entre os anotadores.

O tratamento para as respostas sobre o teor emocional das sentenças predicativas foi o mesmo. Já as respostas para a pergunta *"A qual das categorias a sentença pertence?"* permitiam a escolha de múltiplos rótulos.

4 Análise do corpus

Na Tabela 1 são apresentadas algumas estatísticas do *Retrato_Cantado*. As 146.612 letras musicais foram segmentadas em 2.685.695 sentenças que foram filtradas pelos padrões de busca léxico-sintáticos, resultando em 18.731 sentenças supostamente predicativas. Entre as sentenças candidatas à predicação, 7.417 apresentavam padrão masculino e 11.314, padrão feminino. À essas sentenças, juntaram-se 100 outras que não estavam em conformidade com a caracterização de pessoas. O objetivo dessa inclusão foi avaliar a existência de padrões de descrição de pessoas que não foram considerados em nosso estudo.

Item	Quantidade
Número Total Músicas	146.612
Número Sentenças Genéricas	2.685.695
Número Sentenças com Padrão	18.731

Tabela 1: Número de letras musicais, sentenças segmentadas e sentenças encontradas pelos padrões de busca.

Qtd. Sentenças	Polaridade		
Predicativas	Positiva	Negativa	Neutra
15.312	6.095	3.480	5.737

Tabela 2: Padrão de gênero e polaridade das sentenças.

Após anotação, 88% das sentenças candidatas à predicação foram confirmadas com confiabilidade, o que mostra que padrões de busca são instrumentos eficazes na identificação de sentenças predicativas. Entre as sentenças sem predicação inseridas na anotação, 92% foram corretamente identificadas como negativas, indicando a alta precisão do processo de seleção e reforçando a adequação dos critérios adotados para diferenciar sentenças com e sem predicação.

Na Tabela 2 estão apresentadas estatísticas do corpus anotado. São 15.312 sentenças validadas como predicativas: 6.095, 3.480 e 5.737 sentenças classificadas para o sentimento positivo, negativo e neutro, respectivamente. Padrões morfológicos e lexicais foram utilizados para identificar o gênero das predicações. Foram reconhecidas 9.333 sentenças predicativas do padrão feminino e 5.811 do padrão masculino. Na Figura 2 é apresentada a divisão da análise de sentimento de acordo com o gênero da predicação.

Por fim, a categorização das sentenças predicativas é apresentada na Figura 3. A categoria Identidade apresenta o maior número de ocorrências (7.440 sentenças). Em seguida, aparecem as categorias Comportamento (2.766 sentenças) e Aparência (2.549 sentenças). A categoria Outros reúne (1.409 sentenças). As categorias de menor frequência são Valor/Competência (805 sentenças), Papel social

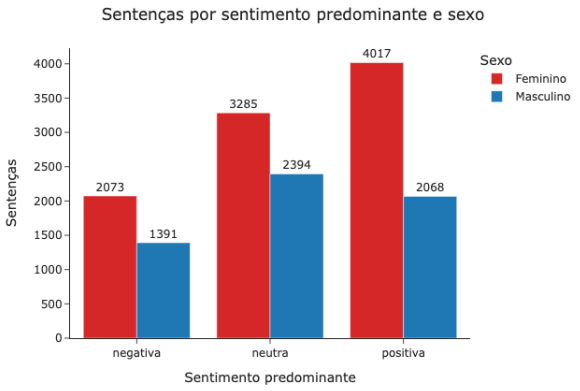

Figura 2: Análise de sentimento das sentenças por gênero identificado.

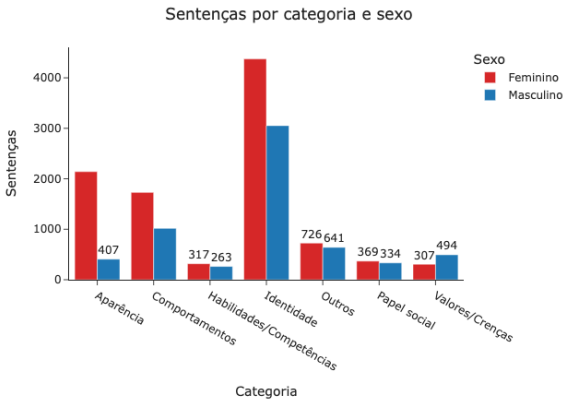

Figura 4: Divisão das sentenças por categorias de acordo com o gênero identificado.

(766 sentenças) e Habilidade/Compreensão (581 sentenças). Na Figura 4, mostra-se a divisão das categorias por gênero de predicação identificado.

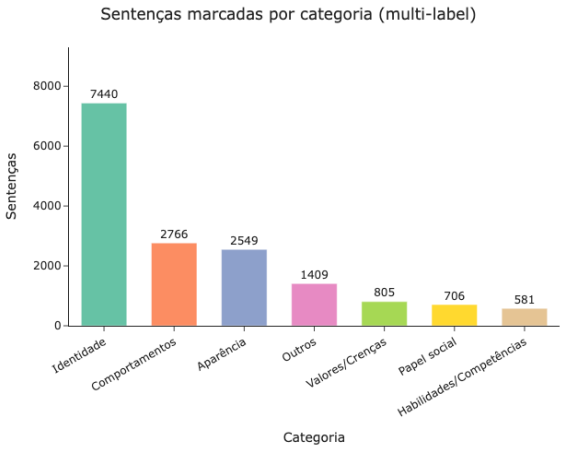

Figura 3: Número de sentenças rotuladas por categoria.

5 Treinamento de classificador

Nesta seção, apresentamos o processo de anotação assistida e o treinamento de um classificador supervisionado para a identificação automática de sentenças predicativas.

5.1 Anotação Semântica assistida por LLM

A estratégia adotada neste trabalho priorizou a anotação humana nos casos que exigem maior precisão, em especial na verificação de falsos positivos oriundos da extração baseada em padrões linguísticos. Frases identificadas automaticamente como contendo caracterização foram, portanto, validadas manualmente, assegurando que as regras definidas não introduzissem rótulos incorretos.

Por outro lado, frases que não acionaram nenhum dos padrões definidos foram tratadas como candidatas a não conter caracterização. Em vez de assumir automaticamente essa condição, delegou-se a um *Large Language Model* (LLM) a identificação de possíveis falsos negativos, isto é, ocorrências de caracterização implícita ou contextualmente expressa que escapam a regras formais. Essa escolha permitiu explorar a capacidade dos LLM de realizar inferência semântica em larga escala, preservando o esforço humano para os casos mais críticos. Cabe destacar que o LLM foi utilizada apenas como ferramenta exploratória para a inspeção de possíveis falsos negativos, não substituindo o processo de anotação humana nem sendo empregada diretamente para rotular exemplos positivos no conjunto de treinamento.

A análise das respostas indicou que parte das frases classificadas como caracterização pelo LLM apresentava formas implícitas ou indiretas de descrição que não eram capturadas pelos padrões linguísticos utilizados. Em geral, tratava-se de construções sem adjetivação explícita ou que expressavam caracterização por meio de metáforas, comparações ou estruturas predicativas menos convencionais. Exemplos desse tipo de ocorrência incluem frases como *"Por ser gaúcho, muito homem e mulherengo"*, *"Você é complicada demais"*, *"Fofoqueiro, linguarudo, sem vergonha"* e *"Mais gostosa do que chocolate"*. Nesses casos, a caracterização emerge de construções comparativas, elípticas ou dependentes de contexto, que não necessariamente correspondem às estruturas sintáticas capturadas pelos padrões linguísticos utilizados na etapa de extração.

Essas limitações são inerentes à abordagem ba-

seada em padrões, que privilegia precisão na identificação de estruturas linguísticas explícitas, mas tende a apresentar menor cobertura para expressões implícitas ou contextuais. Embora tais casos possam, em princípio, motivar a criação de novos padrões linguísticos, a incorporação dessas variações exige análise qualitativa detalhada e ampliação progressiva do conjunto de regras.

Para essa etapa exploratória, foram selecionadas aleatoriamente 6.000 frases não capturadas por padrões linguísticos e submetidas ao modelo GPT-4.1 mini, acessado via *OpenRouter* e configurado com temperatura zero. As respostas foram normalizadas para um julgamento binário, resultando em 5.010 frases classificadas como não contendo caracterização e 990 como contendo caracterização, evidenciando a presença de ocorrências implícitas não detectadas pela abordagem baseada em padrões.

Esses resultados foram integrados ao processo de construção do dataset de forma conservadora. Apenas as frases inicialmente não capturadas por padrões linguísticos e posteriormente classificadas pela LLM como não contendo caracterização foram incorporadas ao conjunto final de treinamento. As frases identificadas pela LLM como contendo caracterização implícita não foram incluídas no conjunto de treino, sendo preservadas para análise posterior e para possível expansão futura dos padrões linguísticos utilizados.

5.2 Treinamento do Classificador

O treinamento do classificador supervisionado foi conduzido utilizando um modelo BERT pré-treinado para o português, especificamente a variante *neuralmind/bert-base-portuguese-cased*, adaptada para a tarefa de classificação binária por meio da adição de uma camada de classificação na saída do modelo. O treinamento foi realizado sobre o conjunto de dados anotado, com a seguinte distribuição de rótulos: 7.153 instâncias rotuladas como *"não"* e 15.013 instâncias rotuladas como *"sim"*.

A tokenização das frases foi realizada com o tokenizer correspondente ao modelo pré-treinado, adotando *padding* dinâmico por *batch* e truncamento com comprimento máximo de 128 *tokens*, de modo a otimizar o uso de memória e o tempo de treinamento. O treinamento foi conduzido em ambiente com GPU, utilizando *mixed precision* para maior eficiência computacional.

Para avaliação do desempenho e da robustez do modelo, foi empregada validação cruzada estratificada com 10 *folds*, garantindo a preservação da distribuição dos rótulos em cada partição. Em cada *fold*, o modelo foi reinicializado e treinado por duas épocas, utilizando o otimizador AdamW com taxa de aprendizado de 1×10^{-5}. A função de perda adotada foi a entropia cruzada ponderada por classe, com pesos calculados a partir da distribuição dos rótulos no conjunto de treinamento de cada *fold*, de modo a mitigar efeitos de desbalanceamento.O treinamento incorporou ainda *gradient clipping*, *learning rate scheduling* com *warmup* linear e atualização dinâmica da taxa de aprendizado ao longo das épocas.

O desempenho do classificador foi avaliado ao final de cada *fold* por meio das métricas de *accuracy*, *precision*, *recall* e *F1-score*, reportadas tanto na forma ponderada quanto na forma macro, permitindo uma análise mais equilibrada entre as classes. Ao final da validação cruzada estratificada, o modelo apresentou desempenho consistente e estável, alcançando *accuracy* média de $0{,}9081 \pm 0{,}0062$ e *F1-score* ponderado de $0{,}9076 \pm 0{,}0064$. Considerando métricas macro, que atribuem peso igual às classes, o classificador obteve *F1-score* médio de $0{,}8937 \pm 0{,}0075$, evidenciando desempenho equilibrado mesmo diante do desbalanceamento do conjunto de dados. Os baixos desvios padrão observados reforçam a robustez do treinamento e a estabilidade do modelo em diferentes partições dos dados.

Após a etapa de validação cruzada, um modelo final foi treinado utilizando todo o conjunto de dados disponível, mantendo as mesmas configurações validadas experimentalmente. Esse modelo foi então exportado juntamente com o tokenizer e o codificador de rótulos, sendo destinado exclusivamente a uso em inferência e aplicações posteriores.

6 Trabalhos futuros

O *Retrato_Cantado* abre diversas possibilidades de investigação. Embora neste trabalho já tenhamos explorado seu potencial para o treinamento de modelos de aprendizado de máquina por meio da construção de um classificador de predicações, futuras investigações podem aprofundar e ampliar essa linha de pesquisa. Entre as possibilidades está o desenvolvimento de modelos complementares, como classificadores voltados à análise de sentimentos associados às caracterizações presentes nas letras.

Considerando o corpus como uma manifestação da cultura popular, tais modelos poderiam ser apli-

cados ao próprio *Retrato_Cantado* para produzir uma leitura distante sobre como diferentes grupos, tal como homens e mulheres, são representados nas músicas. Evidenciar a forma como o gênero feminino é retratado na cultura popular, e de modo particular em distintos estilos musicais, constitui uma contribuição relevante para o campo das humanidades digitais. Além disso, o teor emocional associado às caracterizações permitiria compreender a polaridade presente nessas descrições.

Outra linha promissora de investigação consiste na ampliação dos padrões linguísticos utilizados na etapa de extração. As ocorrências identificadas pelo LLM como possíveis casos de caracterização implícita, mas que não foram capturadas pelos padrões atuais, constituem um conjunto particularmente relevante para análise qualitativa. O uso mostrou-se vantajoso nesse contexto por permitir a identificação, em larga escala, de ocorrências semanticamente plausíveis que escapam a regras linguísticas estritamente formais, funcionando como um mecanismo exploratório para a detecção de possíveis lacunas na cobertura dos padrões. A inspeção sistemática dessas frases poderá orientar a criação de novos padrões linguísticos capazes de abranger construções comparativas, elípticas ou dependentes de contexto, ampliando a cobertura da abordagem baseada em regras sem comprometer sua precisão.

7 Conclusões

Este trabalho apresentou a motivação, o processo de construção e a análise de um corpus de letras de músicas brasileiras, cuidadosamente anotado para identificar descrições de pessoas que vão desde aspectos físicos até dimensões sociais e cognitivas mencionadas nas canções. Mostramos que a utilização de padrões léxico-sintáticos é eficaz para detectar sentenças predicativas. O modelo desenvolvido para identificar predicação a partir deles alcançou bons resultados, constituindo um recurso relevante para a análise em larga escala de outros tipos de textos.

Apesar das contribuições do *Retrato_Cantado*, este trabalho possui algumas limitações. Primeiro, a validação dos padrões léxico-sintáticos não garante cobertura completa das construções predicativas em letras de músicas: predicações implícitas, elipses, metáforas, vocativos, dependência de anáfora e informações distribuídas por múltiplos versos podem não ser capturadas, o que tende a privilegiar caracterizações mais explícitas e localmente realizadas na sentença. Além disso, a anotação foi realizada por voluntários não especialistas, o que pode aumentar a variabilidade interpretativa em casos ambíguos (p.ex., ironia, caracterização indireta e polissemia). Por fim, a etapa assistida por LLM para inspeção de falsos negativos é útil para exploração, mas não substitui a validação humana e pode refletir vieses do próprio sistema.

Uso de IA generativa

O uso de ferramentas de IA generativa na escrita deste trabalho restringiu-se exclusivamente ao aprimoramento linguístico do texto, incluindo reescrita, parafraseamento e lapidação da redação produzida pelos autores. Nenhuma dessas ferramentas foi empregada para sugerir, gerar ou desenvolver novo conteúdo intelectual, limitando-se a funções análogas às de corretores gramaticais, ortográficos ou dicionários.

Disponibilidade dos dados e códigos

Os dados e códigos utilizados neste trabalho estão disponíveis em `https://github.com/firminovitoria/retrato-cantado-dataset/tree/main`.

Referências

Aristotle. 1995. *Poetics*, volume 199 de *Loeb Classical Library*. Harvard University Press, Cambridge, MA.

Cleber Lizardo Assis. 2014. "Entre tapas e beijos": representações sociais sobre a violência de gênero para adolescentes. 2:229–242.

Manash Pratim Barman, Amit Awekar, e Sambhav Kothari. 2019. Decoding the style and bias of song lyrics. *CoRR*, abs/1907.07818.

Anderson Bertoldi e Rove Luiza Oliveira Chishman. 2006. A semântica dos adjetivos e os sistemas de extração de informação na Web. *Letras de Hoje*, 41(2):325–340.

Lorenzo Betti, Carlo Abrate, e Andreas Kaltenbrunner. 2023. Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1).

Maria de Lourdes Borges. 2019. Beleza e gênero. Em Ana Maria Colling e Losandro Antonio Tedeschi, editores, *Dicionário Crítico de Gênero*, páginas 74–80. Editora UFGD, Dourados.

Laura Casanovas-Buliart, Priscila Alvarez-Cueva, e Carlos Castillo. 2024. Evolution over 62 years: an

analysis of sexism in the lyrics of the most-listened-to songs in Spain. *Cogent Arts & Humanities*, 11(1):2436723.

Celso Cunha e Lindley Cintra. 2008. *Nova Gramática do Português Contemporâneo*. Lexikon, Rio de Janeiro.

Vitória Firmino, Janaina Lopes, e Valéria Reis. 2024. Identificando Padrões de Sexismo na Música Brasileira através do Processamento de Linguagem Natural. Em *Anais do V Workshop sobre as Implicações da Computação na Sociedade*, páginas 59–69, Brasília, DF, Brasil. SBC.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, e Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. Em *Proceedings of the Third Workshop on Abusive Language Online*, páginas 94–104, Florence, Italy. Association for Computational Linguistics.

Michel Foucault. 2012. *História da Sexualidade I: A vontade de saber*. Edições Graal, Rio de Janeiro.

Cláudia Freitas e Flávia Martins. 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.

Stuart Hall. 2014. *A identidade cultural na pós-modernidade*. Lamparina, Rio de Janeiro.

Jasmine Huang. 2022. Changing semantics of gendered insults in music lyrics. (THESIS).

Hans Kamp. 1975. *Two Theories About Adjectives*.

Lily Kong. 1995. *Popular Music in Geographical Analyses*, 15th edição. Progress in Human Geography.

João Augusto Leite, Diego Silva, Kalina Bontcheva, e Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. Em *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, páginas 914–924, Suzhou, China. Association for Computational Linguistics.

Janaina N. S. Lopes, Vitória P. Firmino, e Valéria Q. Reis. 2025. Muses or Stereotypes? Identifying Historical Patterns of Sexism in a Corpus of Brazilian Lyrics. *Journal on Interactive Systems*, 16(1):369–380.

Ryan Martinez, Jorge Baptista, e Oto Vale. 2024. Towards a syntactic lexicon of Brazilian Portuguese adjectives. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, páginas 532–538, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Maria Helena de Moura Neves. 2000. *Gramática de usos do português*. Unesp, São Paulo.

Felipe Oliveira, Victoria Reis, e Nelson Ebecken. 2023. TuPy-E: detecting hate speech in Brazilian Portuguese social media with a novel dataset and comprehensive analysis of models. *Preprint*, arXiv:2312.17704.

Hugo Gonçalo Oliveira, Ricardo Rodrigues, Bruno Ferreira, Purificação Silvano, e Sara Carvalho. 2024. BATS-PT: Assessing Portuguese masked language models in lexico-semantic analogy solving and relation completion. Em *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, páginas 207–217, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Chirlei Dutra Lima; Nanci Patrícia Lima Sanches. 2009. A construção do eu feminino na música popular brasileira. *Caderno Espaço Feminino,*, 21:181–205.

Adriano Schlösser e Gabriela Fantin. 2022. Percepção de mulheres acerca de conteúdos sexuais na música popular do Brasil. *Ayvu: Revista de Psicologia*, 9.

Douglas Trajano, Rafael H. Bordini, e Renata Vieira. 2023. OLID-BR: offensive language identification dataset for brazilian portuguese. *Lang. Resour. Eval.*, 58(4):1263–1289.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, e Fabrício Benevenuto. 2022. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. Em *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, páginas 7174–7183, Marseille, France. European Language Resources Association.