Trabalho de Inteligência artificial
Mestrado

Janaina Nogueira de Souza Lopes[1]

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

Abstract. *This work investigates patterns of gender representation in Brazilian song lyrics, focusing on the comparison between different unsupervised machine learning. Methods such as LDA, BERTopic and KMeans are applied to automatically cluster sentences based on thematic and semantic similarities. From these clusters, we seek to reveal the main topics associated with women and men in musical compositions, identifying patterns that may reinforce gender stereotypes. The proposal expands previous studies by adopting an automated and comparative approach, contributing to the understanding of gender inequalities in Brazilian musical culture.*

Resumo. *Este trabalho investiga padrões de representação de gênero em letras de músicas brasileiras, com foco na comparação entre diferentes técnicas de aprendizado de máquina não supervisionado. São aplicados métodos como LDA, BERTopic e KMeans para agrupar automaticamente frases com base em similaridades temáticas e semânticas. A partir desses clusters, busca-se revelar os principais tópicos associados a mulheres e homens nas composições musicais, identificando padrões que podem reforçar estereótipos de gênero. A proposta amplia estudos anteriores ao adotar uma abordagem automatizada e comparativa, contribuindo para a compreensão das desigualdades de gênero na cultura musical brasileira.*

1. Introdução

A desigualdade de gênero é um problema estrutural que se manifesta em diversos aspectos da sociedade brasileira, como mostram os dados do IBGE sobre distribuição de cargos e salários[IBGE 2024]. A mídia e a cultura, embora possam contribuir para a mudança desse cenário, ainda reforçam estereótipos — especialmente na forma como as mulheres são retratadas[Salles and Pappa 2021, Feijó and Macedo 2013]. A literatura e a música, por exemplo, frequentemente destacam atributos físicos e emocionais femininos, em contraste com características sociais ou morais atribuídas aos homens[Freitas and Martins 2023, Firmino et al. 2024].

Neste contexto, a música brasileira se apresenta como um campo rico para análise crítica da linguagem usada para representar os gêneros. Estudos anteriores indicam que as mulheres são, em grande parte, descritas com foco em aparência e sensualidade, enquanto os homens são associados a força, caráter e ação[Firmino et al. 2024, Lopes et al. 2025]. Embora análises manuais já tenham apontado esses padrões, são raras as abordagens automatizadas que investigam essas construções em larga escala no português.

A pergunta que orienta este estudo é: "Os agrupamentos gerados por diferentes técnicas de aprendizado de máquina não supervisionado são capazes de revelar padrões

de representação de gênero presentes nas letras de músicas brasileiras?". A partir dessa questão, a hipótese central da pesquisa é que os métodos de aprendizado de máquina, ao agruparem frases com base em suas similaridades temáticas e semânticas, conseguirão identificar padrões linguísticos distintos para homens e mulheres, refletindo os estereótipos de gênero já descritos na literatura.

Este trabalho tem como foco comparar os resultados obtidos por meio do Processamento de Linguagem Natural (PLN), utilizando diferentes técnicas de aprendizado de máquina não supervisionado, como LDA, BERTopic e KMeans, aplicadas à tarefa de agrupar automaticamente frases com base em similaridades temáticas e semânticas. A partir desses clusters, busca-se revelar quais são os principais tópicos associados a mulheres e homens.

2. Trabalhos Relacionados

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 predicação e organização quantitativa das ocorrências.

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. Adicionalmente, em [Lopes et al. 2025], as autoras ampliaram a análise anterior e o estudo apontou padrões históricos de sexismo. Também foi identificada uma segmentação de gênero nas profissões mencionadas nas letras, com mulheres vinculadas a funções de cuidado e homens a posições de autoridade.

Referente à comparação de técnicas de aprendizado de máquina não supervisionadas, [Ogunleye et al. 2023] realizou um estudo aplicado ao setor bancário que avaliou diferentes métodos de modelagem de tópicos, como LDA, KMeans e BERTopic. [Chen et al. 2025] utilizam o BERTopic para agrupar mais de 537 mil letras de músicas em tópicos temáticos, analisando sua evolução temporal e distribuição entre gêneros musicais. A modelagem revelou padrões de viés de gênero, como a crescente sexualização de mulheres nas letras, ressaltando a importância de uma análise contextualizada por tema e gênero musical.

3. Metodologia

Todas as análises descritas neste trabalho são realizadas a partir de um corpus extenso de letras de músicas brasileiras, contendo sentenças que qualificam homens e mulheres. A construção deste corpus seguiu o procedimento descrito por [Firmino et al. 2024, Lopes et al. 2025]. A identificação das sentenças predicadoras foi realizada por meio da

aplicação de padrões léxico-sintáticos utilizando a ferramenta SpaCy Matcher, permitindo extrair automaticamente frases com predicações relevantes para análise de gênero[1].

Esta seção descreve a metodologia adotada para a comparação entre diferentes algoritmos de aprendizado de máquina não supervisionado, com foco na modelagem de tópicos e agrupamento semântico. São abordadas em detalhe as aplicações dos seguintes algoritmos: Latent Dirichlet Allocation (LDA), BERTopic e KMeans.

3.1. Seleção de frases & Geração de Embeddings com BERTimbau

Para a realização das tarefas de agrupamento e modelagem de tópicos, foram utilizadas **10.000 frases** extraídas de um corpus de letras de músicas brasileiras, previamente anotadas. Inicialmente, o arquivo CSV contendo as frases foi carregado, e as linhas com valores nulos na coluna foram removidas. Em seguida, os valores dessa coluna foram normalizados para letras minúsculas. Após essa limpeza, foi aplicado um filtro para selecionar apenas frases associadas aos rótulos *female* e *male*. De cada uma dessas categorias, foi realizada uma amostragem aleatória de **5.000 frases**, totalizando **10.000 frases** ao todo. Esse processo garantiu um equilíbrio entre os dois grupos, mantendo a representatividade e a diversidade do conteúdo analisado.

Para a representação vetorial das frases, utilizamos o modelo pré-treinado BERTimbau Base[2]. O tokenizador e o modelo foram carregados a partir da biblioteca transformers. As frases foram convertidas em embeddings através da extração do vetor médio da última camada oculta do modelo. Esse procedimento garantiu a obtenção de representações densas e contextualizadas das sentenças, apropriadas para as tarefas subsequentes de agrupamento e modelagem de tópicos.

3.2. LDA

Para aplicação do LDA[3], primeiramente as frases foram vetorizadas utilizando a técnica *TF-IDF* (Term Frequency-Inverse Document Frequency), com o limite de `max_features=3000` para restringir a dimensionalidade do vocabulário e evitar a dispersão dos tópicos.

Foi aplicada a LDA com `n_components=30`, definindo a extração de 30 tópicos. O parâmetro `random_state=42` foi utilizado para garantir reprodutibilidade dos resultados. Após o ajuste do modelo, cada frase foi atribuída ao tópico com maior probabilidade, e os rótulos foram registrados em uma nova coluna do `DataFrame`. Isso possibilitou uma análise qualitativa dos temas predominantes em diferentes subconjuntos do corpus.

3.3. BERTopic

Foi utilizada a configuração multilíngue do BERTopic[4], utilizando os embeddings previamente gerados com o modelo BERTimbau. Essa abordagem permitiu capturar a semântica das frases em língua portuguesa de forma robusta e contextualizada. O algoritmo identificou automaticamente os tópicos latentes ao agrupar sentenças com base em similaridades

[1]Dados disponíveis em: `https://drive.google.com/drive/folders/1UvoWSO3CAh6BUZtoUpT9tgDrA1ulNdnT?usp=drive_link`

[2]https://huggingface.co/neuralmind/bert-base-portuguese-cased

[3]https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

[4]https://huggingface.co/docs/hub/bertopic

vetoriais. Os tópicos atribuídos a cada frase foram registrados em uma nova coluna do `DataFrame`, possibilitando análises qualitativas dos temas extraídos e uma segmentação coerente e significativa do corpus.

3.4. KMeans

Para identificar agrupamentos de frases com características semelhantes, aplicamos o algoritmo de clusterização *KMeans*[5], com `n_clusters=30`. A clusterização foi realizada sobre os embeddings gerados previamente pelo modelo BERTimbau, de forma a agrupar frases semanticamente próximas em um mesmo grupo. Foi utilizado o parâmetro `random_state=42` para garantir reprodutibilidade.

4. Resultados

Nesta seção, abordamos os resultados[6] das três abordagens distintas para a modelagem de tópicos e agrupamento: LDA, BERTopic e KMeans. As visualizações foram geradas com a técnica UMAP para redução de dimensionalidade, permitindo observar a distribuição dos tópicos ou clusters em duas dimensões. Além das observações visuais, foi realizada uma análise qualitativa dos resultados, considerando a coerência temática dos tópicos.

4.1. LDA

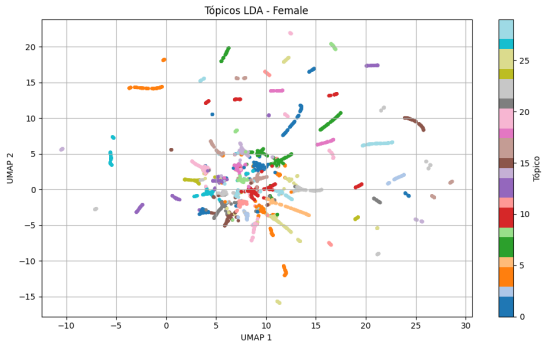

Figura 1. Projeção UMAP dos tópicos LDA com TF-IDF — categoria *female*.

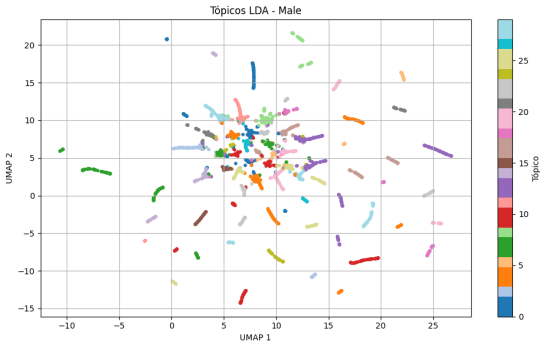

Figura 2. Projeção UMAP dos tópicos LDA com TF-IDF — categoria *male*.

A Figura 1 apresenta a projeção UMAP dos tópicos gerados pelo modelo LDA com base em representações TF-IDF para frases associadas ao padrão *female*. A Figura 2, por sua vez, mostra os tópicos extraídos para frases da categoria *male*.

No caso das frases anotadas como *female*, alguns dos tópicos mais coerentes identificados incluem:

- **Tópico 5**: *"olhos de mel"*, *"cabelos dourados"*, *"sorriso encantador"*, *"corpo de violão"*, *"beleza que hipnotiza"* e *"pele macia feito seda"*, refletindo uma ênfase na estética corporal e atributos físicos, frequentemente idealizados.

[5]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[6]Disponivel em:`https://drive.google.com/drive/folders/1UvoWSO3CAh6BUZtoUpT9tgDrA1ulNdnT?usp=drive_link`

- **Tópico 1**:*“te quero na minha cama”*, *“seu cheiro me enlouquece”*, *“me acende só de olhar”*, *“você me provoca demais”*, *“só de te ver, já me perco”* e *“seu toque me deixa sem chão”*, revelando um discurso centrado no desejo, erotização e atração física.
- **Tópico 3**: *“flor mais linda do jardim”*, *“musa da minha canção”*, *“anjo que caiu do céu”*, *“raio de sol na minha vida”*, *“estrela que ilumina meu céu”* e *“poesia que inspira meu viver”*, caracterizando idealizações românticas e imagens poéticas da figura feminina.

Já na categoria *male*, alguns agrupamentos se destacam:

- **Tópico 3**: *“sou o dono do jogo”*, *“ela me pertence”*, *“mando no que é meu”*, *“ninguém mexe com ela”*, *“comigo ela tá segura”* e *“minha palavra é lei”*, reforçando a ideia de domínio, controle e proteção possessiva.
- **Tópico 7**: *“ela me deixou sozinho”*, *“cansei de te esperar”*, *“coração partido de novo”*, *“fui só mais um pra você”*, *“você brincou comigo”* e *“me iludi com teu amor”*, expressando frustração afetiva e sofrimento emocional.
- **Tópico 1**: *“nasci pra vencer”*, *“sou fera indomável”*, *“homem não chora”*, *“sigo firme na luta”*, *“tenho sangue nos olhos”* e *“vencer é meu destino”*, enfatizando autoafirmação, força e resistência como traços valorizados da masculinidade.

O LDA, demonstrou bom desempenho na identificação de agrupamentos tematicamente coerentes tanto para frases associadas à categoria *female* quanto *male*. A projeção UMAP facilitou a visualização da segmentação dos tópicos, revelando padrões semânticos distintos entre os gêneros. Os resultados indicam que o LDA foi eficaz na captura de estruturas discursivas recorrentes nas letras de música. A separação clara entre os temas emergentes de cada grupo reforça a utilidade do modelo como ferramenta de análise exploratória em estudos culturais e de viés linguístico.

4.2. BERTopic

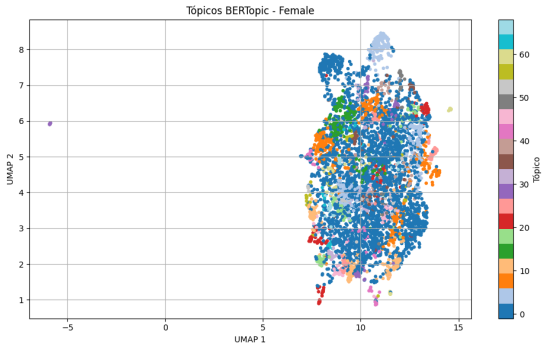

Figura 3. Projeção UMAP dos tópicos BERTopic — categoria *female*.

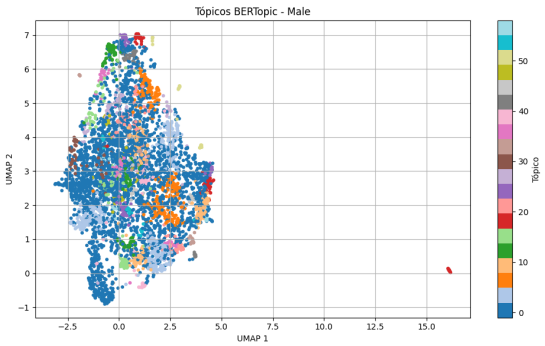

Figura 4. Projeção UMAP dos tópicos BERTopic — categoria *male*.

A Figura 3 apresenta a projeção UMAP dos tópicos gerados pelo modelo BERTopic para frases associadas ao padrão *female*. A Figura 4 mostra a projeção dos tópicos extraídos para frases da categoria *male*.

O modelo BERTopic demonstrou sensibilidade contextual ao identificar agrupamentos semanticamente coesos. No conjunto *female*, por exemplo:

- **Tópico 2**: *"Me reinvento a cada dor"*, *"Levanto mais forte depois de cair"*, *"Sou feita de pétalas e aço"*, *"A beleza está na coragem de continuar"*, *"Florescem as que resistem"* e *"Minha delicadeza também é força"*, indicando um foco na resiliência e beleza interior.
- **Tópico 3**: *"Seu toque é poesia em mim"*, *"Somos dança e silêncio"*, *"O amor em mim floresce com você"*, *"A suavidade do teu beijo me prende"*, *"Te amar é como respirar"* e *"Tudo em você me inspira arte"*, sugerindo um viés estético e sensual.
- **Tópico 4**: *"Ela encanta sem pedir licença"*, *"Beleza que transborda essência"*, *"O mundo para quando ela passa"*, *"Luz própria, brilho único"*, *"A presença dela é calmaria"* e *"Ela é feita de poesia e fogo"*, enfatizando admiração estética e força emocional.

Para o conjunto *male*, o modelo evidenciou agrupamentos como:

- **Tópico 1**: *"Carrego o mundo nas costas e sigo"*, *"Aprendi com a dor a ser mais forte"*, *"Meu silêncio também é resistência"*, *"A vida me bateu, mas tô em pé"*, *"Sou o que levanto depois de cair"* e *"Não corro da luta, corro pro desafio"*, caracterizando discursos de superação e resiliência.
- **Tópico 2**: *"Minha voz é firme, meu coração é leal"*, *"Batalho por quem amo"*, *"Nada me para quando tenho propósito"*, *"Ser homem é proteger e sentir"*, *"Minhas cicatrizes contam histórias"* e *"Respeito é meu escudo"*, reforçando temas de honra, lealdade e sensibilidade.
- **Tópico 4**: *"Ninguém me ensinou a ser forte, aprendi na marra"*, *"A vida forjou meu caráter no fogo"*, *"Levo porrada, mas não caio"*, *"Homem de verdade encara o espelho"*, *"Caminho só, mas nunca em vão"* e *"Não sou perfeito, sou de verdade"*, reforçando uma perspectiva de masculinidade.

O BERTopic apresentou desempenho expressivo na extração de tópicos com maior sensibilidade semântica e contextual. A projeção UMAP evidenciou agrupamentos bem definidos, cujos conteúdos demonstraram coesão temática e nuances emocionais. Adicionalmente, o modelo atribuiu um número considerável de frases ao **tópico -1**, indicando que essas sentenças não se encaixaram de forma clara em nenhum dos clusters definidos.

4.3. KMeans

A Figura 5 apresenta a visualização dos clusters formados pelo modelo KMeans para frases associadas ao padrão *female*. Já a Figura 6 exibe a projeção dos agrupamentos obtidos para sentenças da categoria *male*. A seguir, destacam-se os tópicos mais representativos em cada conjunto, com exemplos ilustrativos extraídos de cada grupo.

- **Tópico 1**: *"Me visto como quero, e daí?"*, *"O corpo é meu, a escolha também"*, *"Não sou sua princesa, sou meu próprio conto"*, *"A saia é curta, minha coragem é longa"*, *"Falo alto porque me calaram demais"* e *"Não me visto pra você, me visto pra mim"*, evidenciando uma estética de empoderamento feminino e autonomia.
- **Tópico 3**: *"Você é luz até nos dias nublados"*, *"Teu sorriso me acalma a alma"*, *"Do teu lado, tudo é abrigo"*, *"Em você encontrei o que nem sabia que procurava"*, *"Tua existência melhora a minha"* e *"Amar você é minha revolução diária"*, caracterizando discursos afetivos e idealizações românticas.

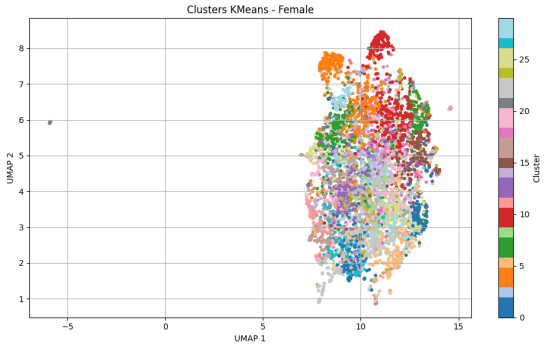

Figura 5. Visualização dos clusters com KMeans — categoria *female*.

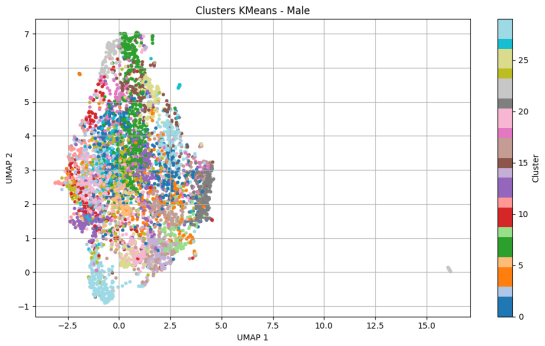

Figura 6. Visualização dos clusters com KMeans — categoria *male*.

- **Tópico 5**: *"Minha luta é diária, mas não me canso"*, *"Mesmo cansada, sigo firme"*, *"Levanto, sacudo e continuo"*, *"Choro escondido, mas sorrio em público"*, *"Ser mulher é resistir em silêncio"* e *"Faço da dor minha motivação"*, expressando temas de resiliência e dor emocional.

O *k-means* demonstrou para *male*:

- **Tópico 2**: *"Minha voz é firme, meu coração é leal"*, *"Batalho por quem amo"*, *"Nada me para quando tenho propósito"*, *"Ser homem é proteger e sentir"*, *"Minhas cicatrizes contam histórias"* e *"Respeito é meu escudo"*, reforçando temas de honra, lealdade e sensibilidade.
- **Tópico 4**: *"De carrão importado eu chego causando"*, *"Relógio caro, whisky na mão, estilo patrão"*, *"Ela vem porque sabe que eu banco tudo"*, *"Tô de corrente de ouro e roupa de grife"*, *"No camarote é só champanhe voando"* e *"Faço dinheiro fácil e gasto como quiser"*, ressaltando ostentação.
- **Tópico 5**: *"Mando e ela obedece sem pensar duas vezes"*, *"Eu que digo como vai ser, ela só segue"*, *"Ela é minha e sabe disso"*, *"Comigo é do meu jeito ou nada feito"*, *"Ela aprende rapidinho quem manda"* e *"Só fica se aceitar minhas regras"*, reforçando atitudes de dominação e controle.

O KMeans apresentou desempenho consistente ao agrupar sentenças com base em semelhanças lexicais, revelando padrões discursivos relevantes tanto na categoria female quanto male. Apesar de sua simplicidade algorítmica, o KMeans conseguiu segmentar com clareza frases semanticamente próximas, especialmente quando os temas são marcados por vocabulário direto e repetitivo. Os resultados indicam que o KMeans é uma ferramenta útil para análises iniciais de agrupamento em conjuntos textuais marcados por padrões discursivos explícitos.

5. Conclusão

Este estudo empregou três técnicas de modelagem de tópicos e agrupamento não supervisionado[7] — LDA, BERTopic e KMeans — aplicadas à análise de sentenças extraídas de letras de músicas brasileiras. As três abordagens foram capazes de identificar padrões discursivos relevantes, cada uma com diferentes níveis de granularidade e coerência temática.

[7]https://youtu.be/7iKqgJE7bxg

Além da análise entre os modelos, o estudo revelou diferenças marcantes nos discursos associados às categorias de gênero. Frases rotuladas como *female* tendem a enfatizar atributos físicos, idealizações românticas e, em alguns grupos, vozes de resistência, autoestima e empoderamento. Já as sentenças categorizadas como *male* frequentemente evocam temas de domínio, força, superação e honra, mas também incluem expressões de vulnerabilidade afetiva e discursos de ostentação. Essas distinções refletem estereótipos de gênero enraizados na cultura popular brasileira e indicam como letras de músicas podem reproduzir, reforçar ou, em alguns casos, tensionar expectativas sociais sobre feminilidade e masculinidade. Portanto, a modelagem de tópicos, além de oferecer uma visão técnica, também se mostra uma ferramenta relevante para revelar padrões discursivos relacionados às construções sociais de gênero em produtos culturais.

Como trabalho futuro, pretende-se explorar outras estratégias de agrupamento e modelagem, como o uso do HDBSCAN isoladamente ou em combinação com outras representações vetoriais, além da aplicação de técnicas como NMF (Non-negative Matrix Factorization) e Spectral Clustering. Essas abordagens podem oferecer diferentes perspectivas na segmentação semântica dos dados e contribuir para uma análise ainda mais refinada dos conteúdos textuais.

Referências

Chen, D., Satish, A., Khanbayov, R., Schuster, C. M., and Groh, G. (2025). Tuning into bias: A computational study of gender bias in song lyrics.

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.

IBGE (2024). Estatísticas de Gênero: Indicadores sociais das mulheres no Brasil. Available in: `https://biblioteca.ibge.gov.br/visualizacao/livros/liv102066\_informativo.pdf`. Last access: 23 May 2025.

Lopes, J. N. d. S., Firmino, V. P., and Reis, V. Q. d. (2025). Muses or stereotypes? identifying historical patterns of sexism in a corpus of brazilian lyrics. *Journal on Interactive Systems*, 16(1):369–380.

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., and Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. *Applied Sciences*, 13(2):797.

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.