A Música Brasileira na Ditadura Militar: uma análise de tópicos com BERTopic e GSDMM

Henry R. Piceni[1], **Pedro V. Alexandre**[1], **Dennis G. Balreira**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{henry.piceni,pvalexandre,dgbalreira}@inf.ufrgs.br

Resumo. *Durante a ditadura militar no Brasil, artistas recorreram à música como forma de expressão, valendo-se de uma linguagem poética e metafórica. Nesse contexto, a análise sociopolítica dessas letras apresenta-se como um desafio interpretativo. Este trabalho busca analisar letras de músicas brasileiras lançadas durante este período, utilizando BERTopic e GSDMM para modelagem de tópicos, a fim de identificar temas-chave que refletem aspectos sociais, políticos e históricos do período. Os modelos analisados revelaram uma linguagem simples e cotidiana, com pouco uso de vocabulário erudito ou político, sugerindo que os artistas optaram por uma expressão direta e emocional, possivelmente para ampliar o diálogo com o público. Particularmente, o BERTopic destacou-se por mapear a diversidade temática com pouca redundância, enquanto o GSDMM segmentou temas dominantes em subtópicos altamente coesos. Este trabalho mostra como a música brasileira, por meio de uma linguagem poética e acessível, retratou o cotidiano da época e revelou o potencial da modelagem de tópicos como ferramenta para a análise cultural.*

Abstract. *During the Brazilian military dictatorship, artists turned to music as a form of expression, employing poetic and metaphorical language. In this context, the sociopolitical analysis of these lyrics presents an interpretive challenge. This study examines Brazilian song lyrics released during this period using BERTopic and GSDMM for topic modeling, with the aim of identifying key themes that reflect the period's social, political, and historical aspects. The analyzed models revealed simple, everyday language, with little use of erudite or overtly political vocabulary, suggesting that artists opted for a direct and emotional form of expression, possibly to broaden their connection with the audience. In particular, BERTopic stood out for mapping thematic diversity with minimal redundancy, while GSDMM segmented dominant themes into highly cohesive subtopics. This work shows how Brazilian music, through poetic and accessible language, portrayed the everyday life of the era and demonstrated the potential of topic modeling as a tool for cultural analysis.*

1. Introdução

A ditadura militar no Brasil (1964–1985) foi um período marcado por repressão, censura e controle ideológico. Durante mais de duas décadas, a preservação dos bons costumes e da moralidade era exigida em documentos, textos e divulgações públicas [Cavalcanti 2018]. Nesse contexto, a música popular desempenhou um papel relevante como forma de expressão crítica, permitindo que artistas, por meio de recursos poéticos e metafóricos,

transmitissem valores democráticos e o anseio por liberdade [Maia 2015]. Assim, ela transcendeu a função artística e se consolidou como instrumento de protesto e resistência, além de servir como importante registro histórico, atuando como vetor de memória coletiva e das lutas vivenciadas pela sociedade [Rosenberg 2013]. As músicas carregam consigo parte da história, funcionando como documentos culturais que permitem estudar o passado e refletir sobre transformações sociais. No entanto, muitas dessas composições, especialmente as de caráter mais poético ou lírico, são marcadas por linguagem subjetiva, textos curtos e pouca explicitação sobre o contexto que as motivou. Essa natureza impõe desafios à análise, uma vez que seus significados podem ser multifacetados e profundamente enraizados em nuances culturais [Tagg 1982].

Nesse cenário, a Modelagem de Tópicos, técnica de Processamento de Linguagem Natural (PLN) que identifica automaticamente temas com base na coocorrência lexical, revela-se útil para análise exploratória de grandes volumes textuais, dispensando anotações manuais. No entanto, ao aplicar métodos tradicionais como o *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003] às letras de música, que marcadas por escassez lexical, uso de interjeições e contextos implícitos, a detecção de padrões temáticos torna-se significativamente mais complexa [Qiang et al. 2019].

O objetivo central deste trabalho é investigar os tópicos gerados pelos modelos BERTopic [Grootendorst 2022] e *Gibbs Sampling Dirichlet Multinomial Mixture* (GSDMM) [Yin and Wang 2014] em um *corpus* de letras de música do período da ditadura militar brasileira. Como principais contribuições, este estudo apresenta: (i) um novo *dataset* com aproximadamente 500 composições do período; (ii) uma metodologia que visa explorar as vantagens de cada abordagem; (iii) a análise das características e resultados de cada modelo na extração de temas em textos curtos e poéticos e (iv) uma discussão sobre os temas socioculturais identificados. O *GitHub*[1] deste trabalho é público e gratuito e conta com o *dataset* originado e demais códigos relacionados.

2. Trabalhos relacionados

As músicas têm sido utilizadas como ferramenta de estudo em diversos contextos no campo do PLN. Em um contexto histórico, [Ribeiro et al. 2023] analisou letras de músicas brasileiras e portuguesas de 1960 a 2020, aplicando LDA [Blei et al. 2003], BERTopic e análise de sentimentos para identificar variações temáticas e emocionais ao longo das décadas. Na esfera do MPB, o estudo de [Dalmora and Tavares 2019] aplicou modelagem de tópicos para classificar letras de músicas populares brasileiras em contextos narrativos, comparando a eficácia de métodos baseados em aprendizado de máquina com a classificação humana. No contexto generativo, o LL-Music [Yepez et al. 2024] combina LLMs e BERTopic para modelagem de tópicos em letras de funk, revelando a presença de discursos periféricos urbanos. Já nas abordagens supervisionadas, [de Araújo Lima et al. 2020] treinou classificadores como *SVM* [Cortes and Vapnik 1995], *Random Forest* [Breiman 2001] e *BiLSTM* [Graves and Schmidhuber 2005] para prever gêneros musicais a partir das letras. Na mesma linha, o estudo de [Fernandes Tavares and José Ayres 2025] propôs um sistema multilíngue (português-inglês) para classificação automática de gêneros musicais a partir de letras, utilizando *Sentence-BERT* (sBERT)[Reimers and Gurevych 2019] para gerar

[1] `https://github.com/henrybeiro/Modelagem-de-Topicos-na-Ditadura`

embeddings que capturam semântica cruzada entre idiomas.

Contudo, na Modelagem de Tópicos, as músicas ainda têm sido pouco utilizadas como objetos de pesquisa, especialmente em contextos históricos e sociopolíticos complexos, como o período da ditadura militar brasileira, possivelmente devido à ausência de dados estruturados e *corpora* organizados disponíveis. Diante da escassez de trabalhos diretamente relacionados, a análise exploratória proposta por este trabalho se volta para domínios que apresentam desafios análogos, como o de textos curtos. Nesse contexto, o estudo de [Amorim et al. 2022] avaliou experimentalmente a Modelagem de Tópicos em um *corpus* de 45.097 *tweets* da rede social X^2 (antigo *Twitter*) utilizando cinco técnicas distintas, dentre elas, o GSDMM e o BERTopic. Portanto, a metodologia quantitativa deste trabalho foi alinhada à de [Amorim et al. 2022], que inspirou tanto no emprego da mesma base de métricas para a avaliação de coerência e diversidade dos tópicos, quanto na escolha preliminar dos hiperparâmetros dos algoritmos utilizados aqui.

3. Metodologia

Para este caso de estudo, foram utilizados dois algoritmos de modelagem de tópicos: o BERTopic [Grootendorst 2022] e o *Gibbs Sampling Dirichlet Multinomial Mixture* (GSDMM) [Yin and Wang 2014]. O BERTopic emprega representações vetoriais baseadas em *embeddings* geradas por modelos pré-treinados, o que permite capturar relações semânticas profundas entre palavras [Grootendorst 2022]. Já o GSDMM segue uma abordagem probabilística fundamentada em frequência de termos e coocorrência, sem levar em conta o contexto semântico [Yin and Wang 2014]. Além disso, enquanto o BERTopic foi desenvolvido para lidar com textos de formatos genéricos, o GSDMM é particularmente indicado para documentos curtos [Amorim et al. 2022]. Cabe destacar que o objetivo deste trabalho não é comparar o desempenho entre os algoritmos, mas sim investigar o potencial de cada um para gerar tópicos coerentes e diversos, respeitando suas limitações e a depender da seleção de hiperparâmetros adotada. A Figura 1 apresenta de forma simples o fluxograma completo deste trabalho:

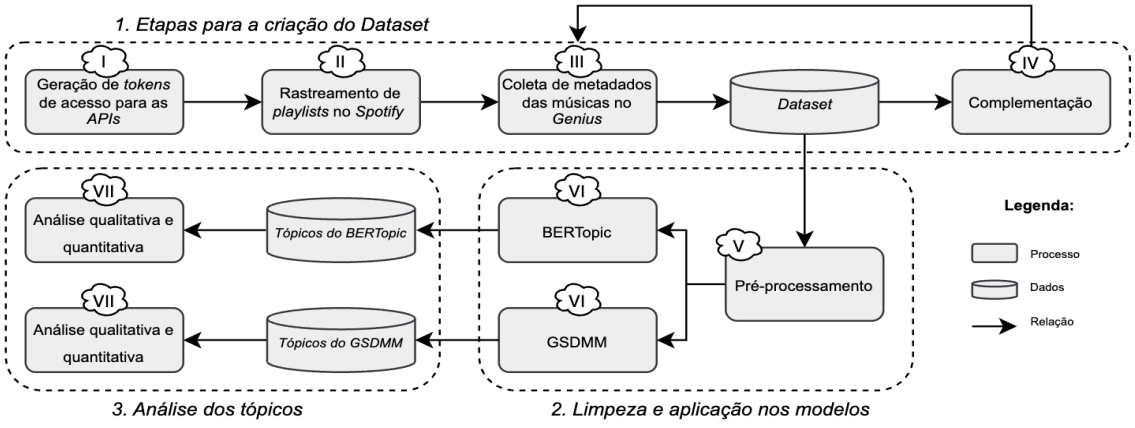

Figura 1. Metodologia adotada neste trabalho, dividida em três etapas: (1) criação do dataset, envolvendo geração de tokens, rastreamento de playlists, coleta de metadados e complementação manual; (2) pré-processamento das letras e aplicação dos modelos BERTopic e GSDMM; e (3) análise qualitativa e quantitativa dos tópicos extraídos.

[2] https://x.com

3.1. Geração do dataset

O processo de criação da base de dados deste trabalho foi estruturado em um pipeline de quatro etapas conceituais. A primeira, de **(i) rastreamento e coleta de músicas**, iniciou-se com a identificação de *playlists* públicas no Spotify[3] com temáticas relacionadas à ditadura militar brasileira. Esse rastreamento foi realizado de forma sistemática por meio de buscas com palavras-chave (ex.: "ditadura", "militar", "Brasil", "MPB") e curadoria manual, com o objetivo de captar composições que, direta ou indiretamente, dialogam com o período entre 1964 e 1985. A *API* do Spotify[4] foi utilizada como fonte primária para a coleta dos metadados oficiais das faixas encontradas, incluindo título, artista(s) e data de lançamento, etc..., de forma semelhante a outros estudos que empregaram a plataforma em análises musicais [Wukkadada 2025].

A partir dos metadados coletados, a segunda etapa consistiu na **(ii) extração das letras** de forma automática, por meio da *API* do Genius[5]. Cada música coletada na etapa anterior foi consultada individualmente, utilizando título e artista como parâmetros de busca, e apenas os textos líricos foram extraídos, sem considerar anotações, traduções ou comentários da plataforma, sendo esta etapa essencial para obter o conteúdo textual a ser analisado [Lim and Benson 2021]. Em seguida, na fase de **(iii) filtragem e validação**, foram definidos critérios específicos de inclusão para garantir a coerência temporal e autoral do *corpus*. Apenas composições com data de lançamento entre 1964 e 1985 e autoria brasileira foram mantidas no conjunto, e letras que não puderam ser encontradas ou que apresentaram erros graves de extração foram descartadas. Essa etapa também envolveu uma verificação pontual dos metadados e do conteúdo textual para assegurar a consistência da base.

Por fim, a etapa de **(iv) complementação** teve como objetivo ampliar a diversidade e representatividade do *corpus*, incluindo obras que não haviam sido captadas nas *playlists* temáticas. Foram inseridos álbuns completos e reconhecidos como marcos do período, como *Acabou Chorare*[6] dos Novos Baianos, *Clube da Esquina*[7] de Milton Nascimento e Lô Borges e *Rita Lee*[8] de Rita Lee, incluídos manualmente com base em sua relevância histórica e artística para o contexto da ditadura militar.

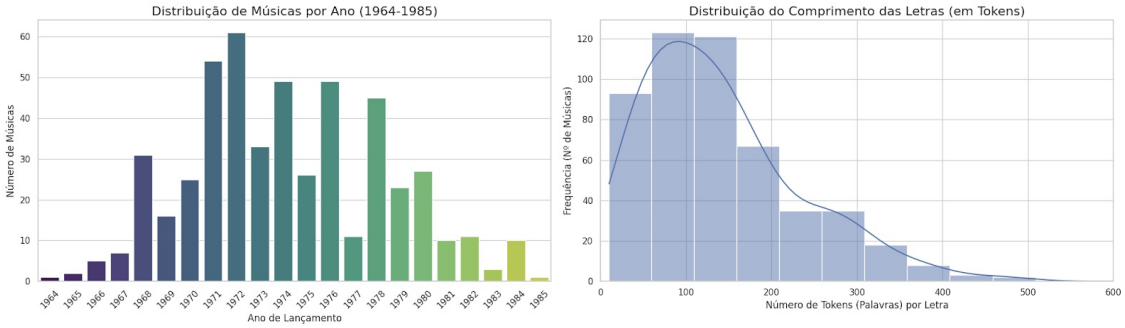

Figura 2. Distribuições Temporal e de Comprimento

[3]https://www.spotify.com
[4]https://developer.spotify.com
[5]https://docs.genius.com
[6]https://pt.wikipedia.org/wiki/Acabou_Chorare
[7]https://pt.wikipedia.org/wiki/Clube_da_Esquina
[8]https://pt.wikipedia.org/wiki/Rita_Lee_(%C3%A1lbum_de_1980)

Deste modo, o *dataset* gerado apresenta **504** canções, totalizando **72.510** *tokens* e um vocabulário de **7.864** termos *únicos*. A Figura 2 ilustra a distribuição temporal das canções e a do comprimento das letras, onde se observa uma tendência na concentração de músicas com aproximadamente **95** *tokens*. Por sua vez, o histograma de distribuição de músicas por ano evidencia uma concentração de lançamentos na década de 1970, com picos notáveis em 1971 e 1972.

3.2. Hiperparâmetros

A escolha dos valores dos hiperparâmetros foi realizada de forma empírica, com base em testes preliminares que buscaram (i) promover uma distribuição mais uniforme dos documentos entre os tópicos, (ii) reduzir a concentração em poucos tópicos dominantes e (iii) preservar a capacidade de identificar nichos temáticos relevantes no conjunto de dados. A seguir, descrevem-se os valores utilizados neste caso de estudo.

3.2.1. Hiperparâmetros do BERTopic

Dentre os vários hiperparâmetros que o BERTopic pode receber, a seguir está a descrição das classes essenciais que foram usadas neste trabalho. Para a redução de dimensionalidade, foi utilizado o *Uniform Manifold Approximation and Projection* (UMAP) [McInnes et al. 2018] com `metric = 'cosine'`, `min_dist = 0.0`, `n_components = 5`, `n_neighbors = 3` e `random_state = 42`. Já para a clusterização, o *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) [McInnes et al. 2017] foi adotado com `min_cluster_size = 15`, `min_samples = 1` e `metric = 'euclidean'`. O CountVectorizer [Pedregosa et al. 2011] com `min_df = 3` e `max_df = 0.6` e `stopwords` recebendo uma lista de stopwords fornecidas pelo NLTK (*Natural Language Tool Kit*) [Bird et al. 2009] foi utilizado para a contagem e a filtragem dos termos nos tópicos. Por fim, para o BERTopic em si, o modelo de *embedding* utilizado foi o *paraphrase-multilingual-MiniLM-L12-v2* [Reimers and Gurevych 2019] junto com `min_topic_size = 15` e `top_n_words = 10`.

3.2.2. Hiperparâmetros do GSDMM

Para o GSDMM, foram utilizados os seguintes hiperparâmetros: $K = 15$ (número de tópicos), $\alpha = 0.5$ e $\beta = 1.0$, com 100 interações.

3.3. Avaliação dos resultados

A avaliação dos resultados obtidos pela modelagem de tópicos foi conduzida por duas abordagens complementares: uma avaliação qualitativa, baseada na interpretação empírica dos agrupamentos gerados, e uma avaliação quantitativa, fundamentada em métricas amplamente utilizadas na literatura. O objetivo dessa etapa foi tanto examinar a relevância e coerência dos temas emergentes quanto comparar o desempenho técnico do BERTopic e do GSDMM frente às características do corpus analisado.

3.3.1. Avaliação qualitativa

A avaliação qualitativa dos resultados foi realizada por meio da interpretação dos tópicos gerados pelos modelos. A análise concentrou-se na observação direta das palavras mais representativas de cada agrupamento, visando verificar coerência semântica e identificar temas predominantes, sabendo se tratarem de letras musicais.

Para nomeação temática, foi utilizado o modelo *Gemini 2.5 Pro* [Google DeepMind 2024] após testes comparativos com reformulação de *prompt* e outros modelos LLM. O Gemini se destacou pela capacidade interpretativa, enquanto outros LLMs frequentemente limitavam-se a combinar palavras do conjunto sem abstrair conceitos unificadores. A cada agrupamento gerado, forneceu-se um *prompt* com a seguinte instrução: *Receba uma lista com 10 palavras que representam um tópico extraído de letras de músicas (curtas, poéticas e metafóricas), em ordem decrescente de frequência. Sua tarefa é propor um título genérico e conciso (até 2 palavras) que capture a ideia central ou a atmosfera sugerida pelo conjunto, buscando transcrever o que os autores podem ter buscado expressar com estas letras.*

Além da interpretação de cada tópico individualmente, também foi conduzida uma discussão geral sobre os principais eixos temáticos identificados, aplicando o contexto da ditadura militar brasileira.

3.3.2. Avaliação quantitativa

A qualidade dos tópicos foi avaliada mediante métricas complementares: (1) coerência, medida por *C_V* e *Normalized Pointwise Mutual Information* (NPMI) [Röder et al. 2015], que combinam coocorrência estatística e similaridade semântica; e (2) diversidade, analisada através da proporção de *tokens* únicos [Dieng et al. 2020] e do *Rank-Biased Overlap* (RBO) [Webber et al. 2010], método que pondera a sobreposição entre rankings de termos com ênfase nos itens mais relevantes. Essa abordagem multidimensional permite avaliar simultaneamente a consistência interna dos tópicos e sua distinção mútua, atendendo aos critérios estabelecidos na literatura para análise de modelos temáticos.

3.4. Preparação dos dados

Um dos maiores desafios neste trabalho é definir o melhor pré-processamento possível para textos curtos, de natureza abstrata e frequentemente marcados por termos ruidosos. Exemplos desses ruídos são interjeições (ex.: “ô”, “ai”, “ah”), marcas de oralidade (ex.: “pô”, “tá ligado?”, “partiu”) ou palavras com baixa carga semântica (ex.: *Batmakumbayêyê*). Embora tenham função expressiva e comunicativa no contexto musical, esses elementos podem interferir na qualidade da modelagem de tópicos — introduzindo ruído nos vetores de representação textual no caso do BERTopic, ou desbalanceando a frequência das palavras no caso do GSDMM. A fim de se obter o melhor desempenho possível de cada abordagem, foram aplicadas duas estratégias de pré-processamento distintas, adequadas às particularidades de cada algoritmo, apresentadas na sequência.

3.4.1. Pré-processamento para o BERTopic

Visto que os modelos BERT [Devlin et al. 2019] foram pré-treinados em linguagem natural completa, optou-se por um pré-processamento mínimo para preservar o contexto original das letras das músicas. Nesta etapa, foram realizadas as etapas de conversão para minúscula, remoção da pontuação e padronização do vocabulário com a lematização. Portanto, as *stopwords* e caracteres com acentuação foram mantidos. Além disso, textos com menos de dez *tokens* e termos que aparecem em menos de 1% dos documentos foram removidos. Nesta etapa, não foi realizado *TF-IDF* [Spärck Jones 1972] no vocabulário pois a classe *CountVectorizer* do BERTopic implementa isso posteriormente [Pedregosa et al. 2011].

3.4.2. Pré-processamento para o GSDMM

Em contraste com a abordagem anterior, o GSDMM é um modelo probabilístico que opera sob o paradigma *Bag of Words* [Harris 1954]. Para este modelo, foi necessário um pré-processamento mais extensivo para reduzir o ruído e destacar os termos semanticamente relevantes. O pipeline consistiu em: (i) conversão para minúsculas; (ii) remoção de pontuação; (iii) remoção de stopwords, verbos de ligação, interjeições, marcas da oralidade e palavras com baixo poder semântico; e (iv) lematização dos termos. Somando a isto, foi aplicado um *TF-IDF* com min_df = 0,01 e max_df = 150 no conjunto de dados a fim de eliminar potenciais ruídos.

4. Resultados e Análise Qualitativa

Nas tabelas que serão apresentadas abaixo, a coluna *Tópico* representa os agrupamentos realizados, a coluna *Nome Proposto* é resultado do *prompt* descrito na seção anterior, *Nº Docs.* é a quantidade de documentos no tópico e *Palavras Representativas* são as palavras mais relevantes do tópico.

4.1. BERTopic

Tabela 1. Tópicos Gerados pelo BERTopic

Tópico	Nome Proposto	Nº Docs.	Palavras Representativas
-1	***Outlier***	15	banho, mato, reino, deserto, par, bandeira, quer, iaiá, estrela, faltar
0	Fé e Devoção	**51**	cristo, jesus, amanhã, salvar, felicidade, tristeza, glória, filha, banda, amigo
1	Crônicas Urbanas	50	papo, amigo, apesar, botar, preto, amanhã, guarda, faltar, bloco, tirar
2	Desejo e Noite	48	lançar, proibir, loucurar, escuro, morena, usar, inferno, carro, cheiro, chuva
3	Rock Cotidiano	41	rock, baby, pessoa, sala, jantar, rolar, viola, coqueiro, puder, canção
4	Samba e Malandragem	38	samba, rodar, modo, banda, música, malandro, gritar, debaixo, puedo, paulo
5	Alma Cigana	36	cabelo, cigano, debaixo, morena, azul, acordar, rosa, amigo, roda, corro
6	Consciência Social	36	fé, costumar, inventar, alternativo, sociedade, rodar, ler, cidadão, coragem, lei
7	Dilemas Afetivos	30	amo, mamãe, mole, engano, turma, fome, tratar, baby, duro, saúde
8	Jogos de Poder	27	abraço, negar, rei, atento, atenção, temer, prova, prato, prata, salvar
9	Conflitos da Alma	23	pai, afastar, vinho, doutor, pecado, santa, baixo, existir, ferir, resto
10	Jornada Interior	22	maluco, viro, conseguir, beleza, vejo, certeza, mistério, nariz, passo, cheguei
11	Destino e Juventude	22	suor, destino, rapaz, certeza, brincadeira, capricho, aviso, baixar, menino, cansado
12	Retratos do Brasil	20	brasil, boi, sambar, morro, dança, passado, joão, pagar, brasileiro, mês
13	Metáforas Elementares	18	voador, fruto, rato, nenhum, quente, defender, cristal, preço, além, raio
14	Aventura Marítima	15	pedra, barco, pirata, baby, navegar, navio, gastar, porto, cigano, luxo

Conforme a Tabela 1, o BERTopic identificou 15 tópicos com ênfase em conceitos abstratos e críticas veladas ao regime, fora o agrupamento de documentos considerados *outliers* [Grootendorst 2022].

Ainda que as críticas ao regime ditatorial não sejam explícitas, é possível notar que os tópicos gerados convergem para uma visão a partir da massa popular brasileira. A exemplo, agrupamentos como *Consciência Social*, *Jogos de Poder* e *Retratos do Brasil* demonstram estratégias de resistência cultural através de metáforas e abstrações, características da produção artística sob censura, com destaque para temas do cotidiano popular e expressões culturais brasileiras. Não há tópicos sobre riqueza ou relacionados à elite, tampouco referências a bens materiais ou ao luxo. Os agrupamentos refletem o cotidiano do cidadão comum, especialmente da classe média-baixa.

Na época, na tentativa de impor os bons costumes e a moralidade, o elemento religioso, sobretudo o católico, foi peça-chave na consolidação de uma visão conservadora de sociedade [Cavalcanti 2018], servindo de respaldo ideológico à censura e à repressão de comportamentos considerados desviantes, como a homossexualidade, a promiscuidade e outras expressões de liberdade individual. Tal contexto pode explicar por que o tópico, *Fé e Devoção*, relacionado à religião aparece como o mais representativo. Tópicos como *Samba e Malandragem*, *Rock Cotidiano* e *Crônicas Urbanas* transportam a manifestação artística para a proximidade do cotidiano popular. Ainda assim, temas de escapismo lírico são demonstradas em *Aventura Marítima* e *Alma Cigana*, onde é desejada uma realidade longe da cinzenta ditadura.

4.2. GSDMM

Tabela 2. Tópicos Gerados pelo GSDMM (Ordenados por Frequência)

Tópico	Nome Proposto	Nº Docs.	Palavras Representativas
0	Intensidade Amorosa	**86**	amor, vir, esperar, louco, deixar, vida, dia, pensar, hora, levar
1	Cores do Amor	73	amor, dia, coração, cantar, inventar, velho, hoje, lançar, dor, cor
2	Canto Coletivo	70	baby, gente, mundo, cantar, bem, rodar, maria, samba, tempo, gostar
3	Devoção Amorosa	41	amor, sol, amar, deus, medo, comigo, deixar, menina, cair, coração
4	Cenas de Tensão	40	medo, ficar, tempo, tirar, carro, entrar, chorar, mão, disco, papo
5	Jornada Existencial	38	viver, vivo, mundo, outro, senhor, falar, bem, volta, dia, rei
6	Apelos Familiares	36	agora, pedir, outro, doutor, filho, amor, mundo, ficar, mãe, hoje
7	Jornada Popular	27	mamãe, voltar, fim, povo, vida, parte, botar, rua, tempo, feliz
8	Destino Grandioso	18	chegar, duro, virar, cidade, salvar, tocar, grande, rio, brasil, mundo
9	Chamado Espiritual	18	vir, chamar, jesus, cristo, pai, ano, bem, correr, porta, passado
10	Luta Corporal	17	nunca, andar, inferno, corpo, mão, gritar, suor, girar, engano, vida
11	Realidade Onírica	16	sonho, boi, abraço, show, sala, sonhar, pessoa, dança, jantar, acabar
12	Cultura Noturna	10	rock, escuro, cinema, usar, perder, cantar, ninguém, dinheiro, noite, rapaz
13	Hedonismo Rock	9	cheio, menina, banda, felicidade, homem, cama, baixo, deixar, pecado, bota
14	Necessidade e Tensão	7	preciso, proibir, jeito, forte, sim, amigo, tempo, atenção, atento, temer

Diferente do BERTopic, a Tabela 2 revela que o GSDMM teve maior dificuldade em gerar tópicos com distribuição uniforme de documentos, mesmo após as diversas configurações de hiperparâmetros testadas. O Tópico 0, *Intensidade Amorosa*, concentra aproximadamente **17%** de todas as composições analisadas; e junto de *Cores do Amor* e *Devoção Amorosa*, somam quase **40%** e evidenciam a centralidade dos temas afetivos que serão explorados a seguir.

Em uma primeira análise, é possível notar que o tema dominante nas canções é o das emoções e relações humanas, o qual o modelo foi capaz de segmentar em diferentes agrupamentos, destacando impactos psicológicos da repressão. Tópicos como *Cenas de Tensão* e *Necessidade e Tensão* capturam o medo generalizado e a ansiedade coletiva. *Luta Corporal* evidencia o sofrimento físico. O GSDMM foi capaz de identificar a presença do clima de opressão e a violência psicológica do período da ditadura.

Por outro lado, *Hedonismo Rock* e *Cultura Noturna* podem significar um escapismo como resposta comportamental à censura, uma visão diferente e realista do escapismo lírico traduzido pelo outro modelo. Os temas *Intensidade Amorosa* e *Devoção Amorosa*, assim como *Cores do Amor*, diferentemente do que foi observado no BERTopics, focam na experiência emocional direta e imediata.

Em conjunto, os modelos fornecem perspectivas complementares sobre o *corpus*: o BERTopic possui um viés político-cultural, decodificando as intenções críticas e estratégias metafóricas dos artistas, enquanto o GSDMM faz um registro emocional, capturando as manifestações concretas da vivência coletiva. A utilização desta dualidade analítica revela o caráter multifacetado da produção cultural durante o regime autoritário, onde a expressão artística simultaneamente contestava e transcendia a opressão através de códigos estéticos e afetivos.

4.3. Análise Quantitativa

Ao observar a Tabela 3, é possível notar um claro *trade-off* entre os dois modelos, onde cada um se destaca em uma das dimensões avaliadas:

Tabela 3. Comparação quantitativa dos modelos

Modelos	Coerência *C_V*	Coerência NPMI	Média das Coerências	Diversidade	Inverted RBO	Média das Diversidades
BERTopic	0,455	-0,426	0,014	0,927	0,997	**0,962**
GSDMM	0,407	-0,125	**0,141**	0,820	0,984	0,902

Na avaliação da coerência, que mede a interpretabilidade dos tópicos, o GSDMM apresentou um desempenho médio superior ao do BERTopic. Essa vantagem foi impulsionada por um resultado significativamente menos negativo na métrica NPMI, embora o BERTopic tenha registrado um valor ligeiramente maior na Coerência *C_V*. No quesito de diversidade, que avalia a distinção entre os tópicos, o BERTopic demonstrou uma superioridade clara e consistente, superando o GSDMM em todas as métricas avaliadas, tanto na média geral quanto nas medidas individuais.

Os resultados quantitativos, portanto, revelam um perfil de desempenho distinto para cada modelo. O BERTopic se destaca na geração de tópicos que tendem a ser únicos e variados, mas ao custo de uma baixa coesão semântica.

As baixas pontuações de coerência NPMI observadas para ambos os modelos podem ser atribuídas a duas características intrínsecas do corpus: (i) seu tamanho reduzido, que limita a robustez estatística da métrica e (ii) a natureza poética e subjetiva das letras. Nestes textos, a riqueza vocabular e o uso de metáforas diminuem a frequência de coocorrência literal de termos semanticamente relacionados. Isso sugere que, para este tipo

de corpus, a avaliação qualitativa e a interpretabilidade humana dos tópicos se tornam um critério de análise mais crucial do que a avaliação puramente quantitativa.

A Tabela 4 sintetiza os resultados quantitativos deste estudo em comparação com os obtidos por [Amorim et al. 2022] em sua análise sobre *tweets*, permitindo observar similaridades e diferenças de desempenho entre os modelos em contextos distintos.

Tabela 4. Comparação quantitativa com o trabalho de [Amorim et al. 2022].

Modelo	Resultados do Presente Trabalho						Resultados de [Amorim et al. 2022]					
	C_V	NPMI	Méd. Coer.	Diver.	Inv. RBO	Méd. Diver.	C_V	NPMI	Méd. Coer.	Diver.	Inv. RBO	Méd. Diver.
BERTopic	0,455	–0,426	0,014	0,927	0,997	**0,962**	0,492	–0,104	0,194	0,983	0,999	**0,991**
GSDMM	0,407	–0,125	**0,141**	0,820	0,984	0,902	0,623	0,209	**0,416**	0,833	0,973	0,903

A análise comparativa com [Amorim et al. 2022] corrobora o comportamento do BERTopic de gerar maior diversidade temática, enquanto o GSDMM tende a formar agrupamentos semanticamente mais homogêneos. Esta tendência persiste mesmo diante da disparidade entre os *corpora*, sendo o de [Amorim et al. 2022] aproximadamente 85 vezes mais extenso.

5. Conclusão

Este trabalho evidencia que letras de música, embora textos curtos, apresentam alta complexidade linguística, marcada por subjetividade, metáforas e traços de oralidade. Um pré-processamento adaptado a essas características permitiu construir um *corpus* coeso, adequado à investigação por modelagem de tópicos. Os modelos BERTopic e GSDMM geraram tópicos majoritariamente interpretáveis, com diferenças significativas: o BERTopic apresentou maior diversidade temática com mínima redundância, enquanto o GSDMM destacou-se na segmentação de temas dominantes em subtópicos coesos. A análise revelou o uso predominante de linguagem simples e cotidiana, sugerindo uma escolha artística por formas de expressão emocional e acessível, possivelmente como estratégia para contornar a censura ao tratar de temas políticos de forma subjetiva.

Finalmente, a principal contribuição deste trabalho reside na aplicação e avaliação de técnicas de modelagem de tópicos em um corpus que impõe desafios específicos à área de Processamento de Linguagem Natural (PLN), como a concisão e a expressividade estilística das letras de música. O ciclo metodológico proposto, que abrange desde a construção de um *dataset* inédito até a comparação entre modelos, contribui para a compreensão das possibilidades da Modelagem de Tópicos em domínios textuais não convencionais e abre diversos caminhos para trabalhos futuros. Nesse sentido, propõe-se a expansão do *dataset*, visando incluir uma maior diversidade de gêneros musicais e um maior equilíbrio na representação de artistas e na distribuição temporal das canções. Adicionalmente, futuras investigações podem explorar o impacto de diferentes estratégias de pré-processamento e testar variações nos hiperparâmetros, aprofundando a análise sobre a sensibilidade dos modelos. A exploração de outros algoritmos de modelagem de tópicos também se apresenta como uma continuação natural, podendo oferecer novas perspectivas sobre a análise qualitativa e permitindo compreender de forma mais profunda como a produção musical brasileira expressou, resistiu e reinterpretou a realidade sob um regime autoritário, preservando na arte a memória e as vozes de seu tempo.

Referências

Amorim, A., Murrugarra-Llerena, N., Silva, V., de Oliveira, D., and Paes, A. (2022). Modelagem de tópicos em textos curtos: uma avaliação experimental. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 254–266. SBC.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc., Sebastopol, CA.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Cavalcanti, I. L. L. (2018). Censura moral e música na ditadura militar no brasil: o regime contra a transgressão da família e dos bons costumes. *Working Paper*, (75). Acesso em: 10 ago. 2023.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dalmora, A. and Tavares, T. (2019). Identifying narrative contexts in brazilian popular music lyrics using sparse topic models: A comparison between human-based and machine-based classification. In *Simpósio Brasileiro de Computação Musical (SBCM)*, pages 17–21. SBC.

de Araújo Lima, R., de Sousa, R. C. C., Lopes, H., and Barbosa, S. D. J. (2020). Brazilian lyrics-based music genre classification using a blstm network. In *International Conference on Artificial Intelligence and Soft Computing*, pages 525–534. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020). Topic modeling in embedding spaces. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*.

Fernandes Tavares, T. and José Ayres, F. (2025). Multi-label cross-lingual automatic music genre classification from lyrics with sentence bert. *arXiv e-prints*, pages arXiv–2501.

Google DeepMind (2024). Gemini 2.5 Pro. Modelo multimodal avançado de linguagem (LLM) com contexto de até 1 milhão de tokens. Disponível via Google AI Studio, Vertex AI e API.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Lim, D.-H. and Benson, A. (2021). Expertise dynamics in online annotation communities: The case of genius.com. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 480–491.

Maia, A. V. (2015). A música popular brasileira e a ditadura militar: vozes de coragem como manifestações de enfrentamento aos instrumentos de repressão.

McInnes, L., Healy, J., Astels, S., et al. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Qiang, J., Zhenyu, Q., Yunhao, Y., and Xindong, W. (2019). Short text topic modeling techniques, applications, and performance: A survey. *arXiv preprint arXiv:1904.07695*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ribeiro, R. D. S. F. M., Ramos, P. d. P. N., et al. (2023). Sentiment analysis and topic modeling of portuguese and brazilian song lyrics through the years. Master's thesis, iscte.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Rosenberg, T. (2013). The soundtrack of revolution: Memory, affect, and the power of protest songs. *Culture Unbound*, 5(2):175–188.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Tagg, P. (1982). Analysing popular music: theory, method and practice. *Popular music*, 2:37–67.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

Wukkadada, S. (2025). Decoding spotify hits: Statistical and predictive analysis of track features driving song popularity. *Academy of Marketing Studies Journal*, 29(1).

Yepez, J., Tavares, B., Peres, F., and Becker, K. (2024). Na batida do funk: modelagem de tópicos combinando llm, engenharia de prompt e bertopic. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 613–625. SBC.

Yin, J. and Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242.