UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Instituto de Informática
STIL - 2025

BRACIS 2025

A Música Brasileira na Ditadura Militar: uma Análise de Tópicos com BERTopic e GSDMM

Henry R. Piceni, **Pedro V. Alexandre** e **Dennis G. Balreira**

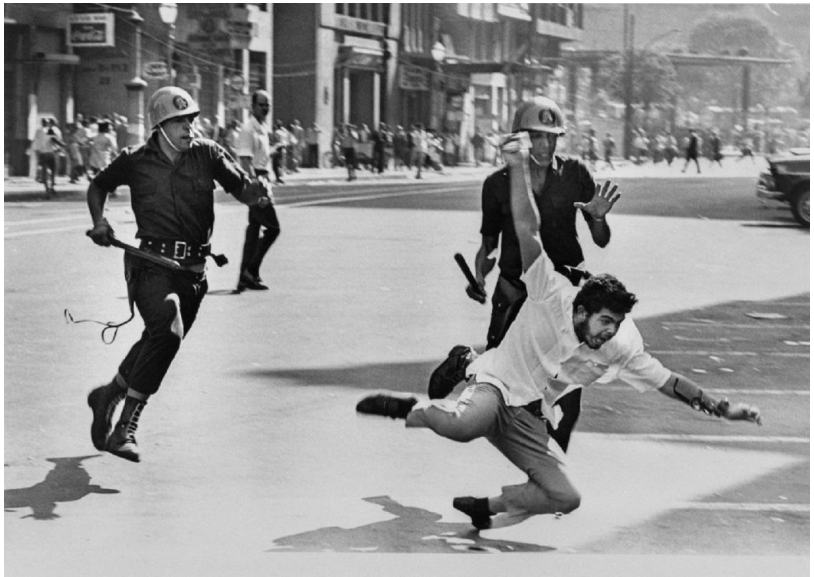
{henry.piceni,pvalexandre,dgbalreira}@inf.ufrgs.br

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf INSTITUTO DE INFORMÁTICA UFRGS

Motivação e Objetivos

- A Ditadura Militar (1964-1985) no Brasil foi marcada por censura, repressão e moralismo.
  - Muitos artistas usaram a **música** como forma de protesto, com metáforas e poesia para driblar a censura e expressar ideais democráticos.



Estudante sofrendo repressão por militares durante a ditadura. (Foto: Evandro Teixeira)

- Buscamos investigar temas presentes em letras de música.
  - Identificar **aspectos sociais, políticos e culturais** do período.

Trabalhos relacionados

- Poucos estudos usam Modelagem de Tópicos em músicas para análise **histórico-sociopolítica**.

- Períodos complexos (ex.: Ditadura Militar) são pouco explorados, muitas vezes pela falta de **corpora estruturados**.

Inspiração Metodológica:

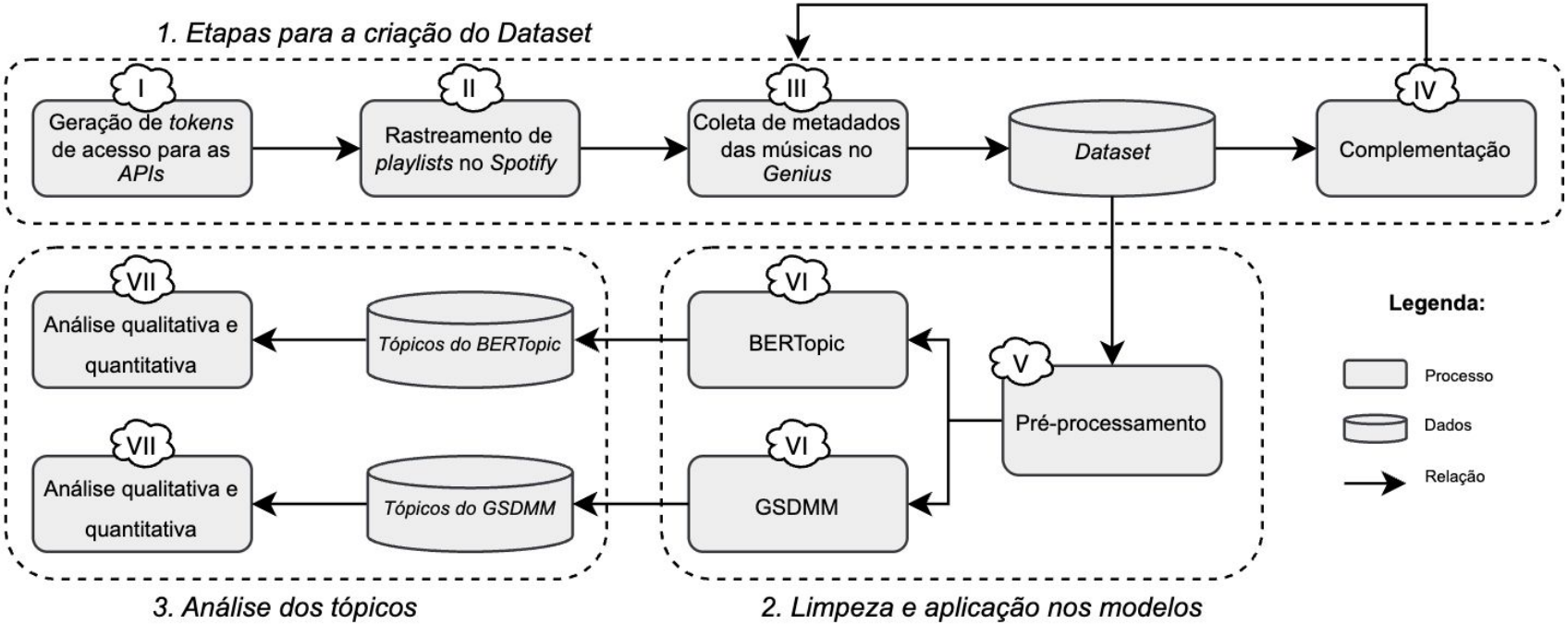
Amorim et al. (2022):

- Modelagem de Tópicos em 45.097 tweets.
- Técnicas Avaliadas: **GSDMM**, **BERTopic**, entre outras.

Contribuição para este trabalho:

- Baseou a escolha das métricas de avaliação (coerência e diversidade).
- Inspirou a seleção preliminar de hiperparâmetros dos algoritmos.

Metodologia

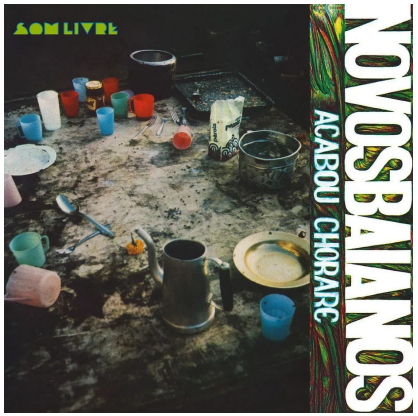


Metodologia: Geração do *Dataset*

- Pesquisa de *playlists* no Spotify por uso de palavras-chave:
  - **Ex:** ditadura, censura, MPB, militar...

Metodologia: Geração do *Dataset*

- Álbuns adicionados manualmente devido à sua relevância para o período.



Acabou Chorare (1972)

Novos Baianos

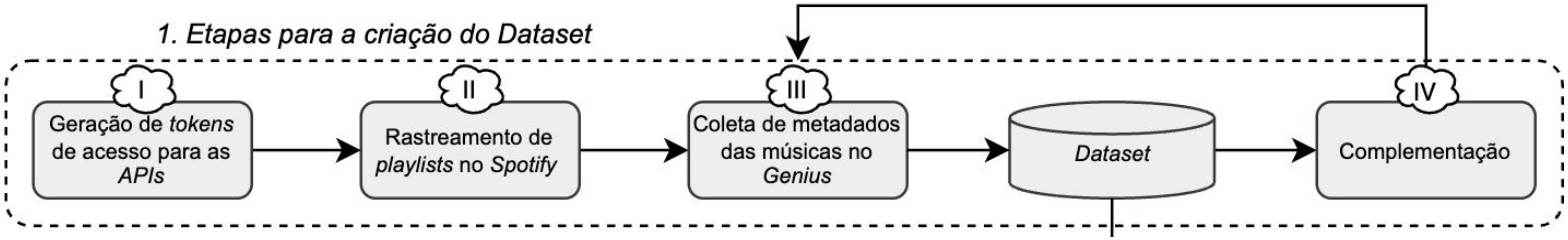
Clube da Esquina (1972)

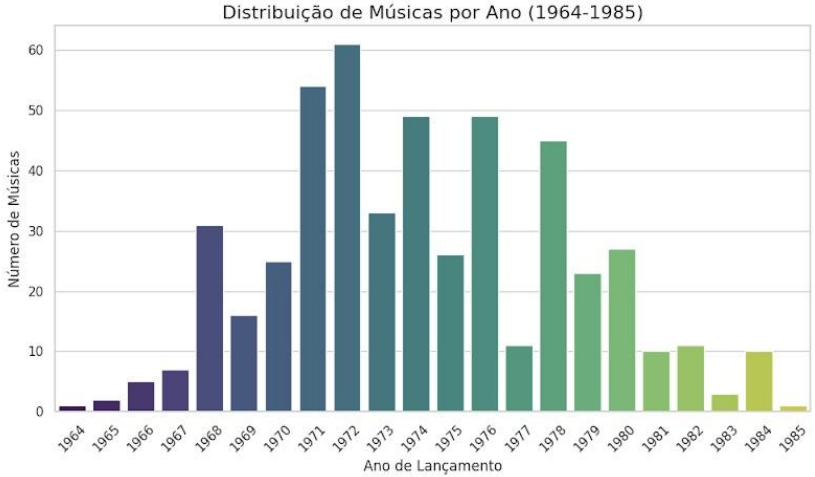
Milton Nascimento e Lô Borges

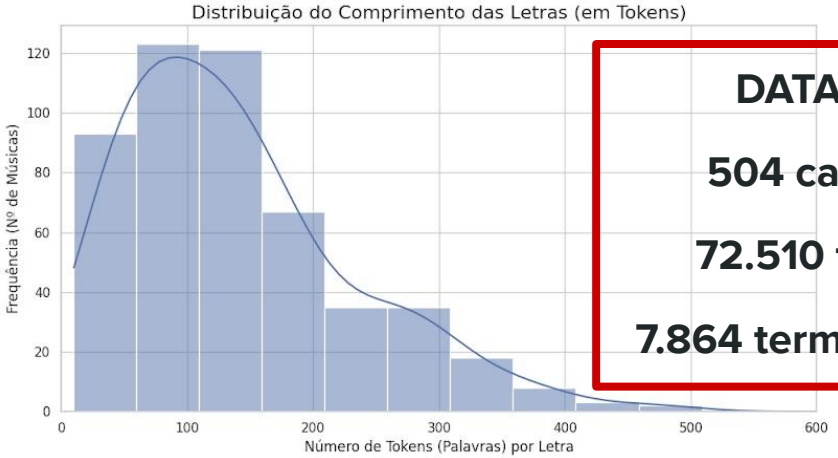
Rita Lee (1980)

Rita Lee

Metodologia: Geração do *Dataset*







DATASET:

504 canções

72.510 tokens

7.864 termos únicos

Pré-processamento

Problemas?

Cinco, quatro, três, dois
Parem, esperem aí
Onde é que vocês pensam que vão?
Ahn, ahn

Plunct plact zum
Não vai a lugar nenhum
Plunct plact zum
Não vai a lugar nenhum

...

Carimbador Maluco (1983)
Raul Seixas

He he he hey boy
O teu cabelo tá bonito hey boy
Tua caranga até assusta hey boy
(Tchu aa uu)
Vai passear na rua Augusta tá

He he he hey boy
Teu pai já deu tua mesada hey boy
A tua mina tá gamada hey boy (Tchu aa uu)

...

Hey Boy (1970)
Os Mutantes

Ooh-ooh-ooh, ooh-ooh-ooh
Ooh-ooh-ooh, ooh-ooh-ooh, ooh

Eu sou apenas um rapaz
latino-americano
Sem dinheiro no banco
Sem parentes importantes
E vindo do interior

...

Apenas um Rapaz Latino-Americano(1976)
Belchior

Pré-processamento

Cinco, quatro, três, dois
Parem, esperem aí
Onde é que vocês pensam que vão?
Ahn, ahn

Plunct plact zum
Não vai a lugar nenhum
Plunct plact zum
Não vai a lugar nenhum

...

He he he hey boy
O teu cabelo tá bonito hey boy
Tua caranga até assusta hey boy
(Tchu aa uu)
Vai passear na rua Augusta tá

He he he hey boy
Teu pai já deu tua mesada hey boy
A tua mina tá gamada hey boy
(Tchu aa uu)

...

Problemas?

Ooh-ooh-ooh, ooh-ooh-ooh
Ooh-ooh-ooh, ooh-ooh-ooh, ooh

Eu sou apenas um rapaz
latino-americano
Sem dinheiro no banco
Sem parentes importantes
E vindo do interior

...

- Desafio inerente ao domínio de composições:
  - Muitas palavras e termos sem **valor semântico** no contexto: interjeições, marcas de oralidade, gírias, palavras de outros idiomas, etc...
- Podem gerar tópicos **pouco descritivos** ao introduzirem ruídos no modelo
  - Como reduzir os impactos disso?

Metodologia: Pré-processamento

Desafio: Letras são textos curtos, abstratos e com ruído (interjeições, gírias)

Estratégia: Dois *pipelines* distintos, adequados a cada algoritmo.

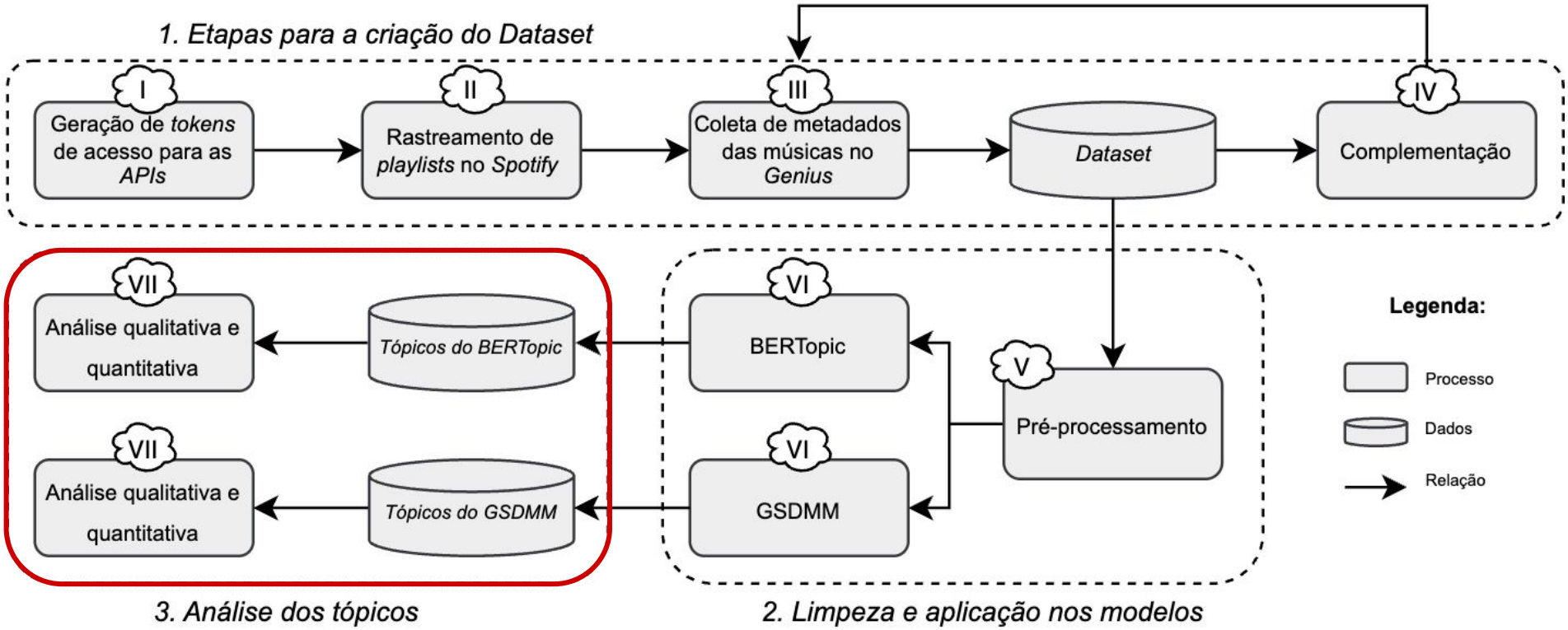
BERTopic	GSDMM
Pré-processamento Mínimo Preserva contexto para o BERT	**Pré-processamento Extensivo** Reduz ruído para o modelo probabilístico
• Conversão para minúsculas • Remoção de pontuação • Lematização • **Manter *stopwords* e acentos** • Remover docs com <10 tokens	• Conversão para minúsculas • Remoção de pontuação • Lematização • **Remover *stopwords*, interjeições e palavras de baixo semântico** • Aplicação de TF-IDF (*min_df = 0.01, max_df = 150*)

Metodologia: Hiperparâmetros

Objetivo: Distribuição uniforme de tópicos e identificação de nichos temáticos.

<table>
<tr><th>BERTopic</th><th>GSDMM</th></tr>
<tr><td>Modelo de Embedding:
paraphrase-multilingual-MiniLM-L12-v2</td><td rowspan="5">Hiperparâmetros do Modelo:
K = 15 (tópicos)
α = 0.5
β = 1.0
Iterações = 100</td></tr>
<tr><td>UMAP (Redução de Dimensão):
n_components=5, n_neighbors=3, metric='cosine'</td></tr>
<tr><td>HDBSCAN (Clustering): min_cluster_size=15</td></tr>
<tr><td>CountVectorizer: min_df=3, max_df=0.6</td></tr>
<tr><td>BERTopic: min_topic_size=15, top_n_words=10</td></tr>
</table>

Metodologia: Análise dos tópicos



Análise de Tópicos

Análise Qualitativa:

- Interpretação do significado e a coerência dos temas gerados.
- **Foco:** Entender o que os agrupamentos de palavras representavam no contexto das músicas.

Análise Quantitativa:

- Utilização de métricas de mercado para medir a performance técnica dos modelos.
- **Foco:** Medir a Coerência (sentido interno dos tópicos) e a Diversidade (diferença entre os tópicos).

Análise Qualitativa: *Prompt* de Nomeação

“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.”

Análise Qualitativa: 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

Análise Qualitativa: GSDMM

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

Análise Qualitativa

BERTopic

- Linguagem popular e temas do cotidiano

Crônicas Urbanas (1: “papo, amigo, botar, guarda, bloco”)

- Resistência implícita

Consciência Social (6: “sociedade, cidadão, coragem, lei”) e Jogos de Poder (8: “rei, atento, temer, prova”)

- Escapismo lírico

Alma Cigana (5: “cigano, rosa, roda, corro”) e Aventura Marítima (14: “barco, pirata, navegar, porto, luxo”)

GSDMM

- Segmentação em temas afetivos e cotidianos

Apelos Familiares (6: “doutor, filho, mãe, hoje”) e Jornada Popular (7: “mamãe, povo, rua, tempo”)

- Temas psicológicos: opressão, angústia, espiritualidade

Cenas de Tensão (4: “medo, carro, chorar, mão, entrar”), Luta Corporal (10: “inferno, corpo, suor, gritar”) e Chamado Espiritual (9: “jesus, cristo, pai, porta”)

- Escapismo mais realista

Cultura Noturna (12: “rock, cinema, noite, dinheiro”) e Hedonismo Rock (13: “banda, felicidade, cama, pecado”)

Análise Quantitativa

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

BERTopic

- Tópicos mais variados e com menos sobreposição entre si.
  Média de Diversidade: 0,962

GSDMM

- Tópicos com maior coesão interna e mais fáceis de interpretar.
  Média de Coerência: 0,141

Análise dos Modelos

- **Processo de Otimização e Calibragem**
  - Para alcançar a distribuição de tópicos mais descritiva, foram realizadas mais de 100 execuções. Tempo médio de treinamento foi de **3 min.** para o BERTopic e **1,5 min.** para o GSDMM.
  - A seleção final foi baseada em um processo iterativo de experimentação e análise (característico do **Aprendizado Não Supervisionado**).
- **Alta sensibilidade** dos hiperparâmetros de ambos algoritmos.
  - **BERTopic** se mostrou mais sensível: variações mínimas geraram distribuições de tópicos completamente distintas
- **Pré-processamento adequado** foi um fator crítico de sucesso.
  - A remoção de termos ruidosos (característicos da linguagem poética) foi essencial para garantir que os tópicos tivessem alto valor semântico.

Conclusão

- A **Modelagem de Tópicos** pode ser uma ferramenta eficaz para a análise cultural.
  - Embora a análise de textos curtos e poéticos, como letras de música, apresente um desafio interpretativo considerável devido à sua linguagem subjetiva, metafórica e coloquial.
- Cada modelo se destacou em uma dimensão.
  - **BERTopic** mapeou temas amplos e decodificou as intenções críticas e o viés político-cultural dos artistas.
  - **GSDMM** aprofundou subtemas e capturou o registro emocional e as vivências concretas da população.

Ambos **se complementam** em decodificar linguagem simples e cotidiana, possivelmente como estratégia contra a censura.

Trabalhos Futuros

- Expandir o dataset
  - Colocar mais composições do período.
- Testar novos pré-processamentos
  - Manter/remover stopwords, realizar/não realizar stemming e/ou lematização, etc.
- Explorar outros algoritmos de modelagem de tópicos
  - LDA, HDP ou testar outros *embeddings* para o BERTopic, etc.
- Outras temáticas
  - Nacionais: análise geográfica e regional, discografia de um artista específico, etc.
  - Internacionais: períodos históricos de outros países
- **Usar as letras de músicas como ferramenta de estudo**
  - Outras técnicas de PLN além de Modelagem de Tópicos.

Obrigado!

Henry Ribeiro Piceni
henry.piceni@inf.ufrgs.br

Pedro Vitor Alexandre
pvalexandre@inf.ufrgs.br

Dennis Giovani Balreira
dgbalreira@inf.ufrgs.br

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

BRACIS 2025

.inf
INSTITUTO DE INFORMÁTICA
UFRGS