Semantic Macro-Themes in Sertanejo Universitário Lyrics: Topic Modeling with BERTopic

Amanda Schmieleski Cossa Manfredini[1], Karin Becker[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS), Brazil
a3schmieleski@gmail.com, karin.becker@inf.ufrgs.br

Abstract. Music lyrics provide a rich textual record of affective, social, and cultural themes, whose organization can be explored at scale through natural language processing. This paper investigates the internal thematic organization of 5,037 stanzas extracted from 959 Brazilian country songs released between 2005 and 2025 (*sertanejo universitário*). The corpus was collected from *letras.mus.br* and analyzed with BERTopic at the stanza level, allowing thematic variation within individual songs to be captured. Five embedding models were compared as a methodological control, with `ptbr-e5-small` adopted as the main configuration. The model identified 44 topics, while 2,066 stanzas (41.02%) were classified as outliers, reflecting the challenge of identifying recurrent semantic patterns in short and often figurative poetic texts. Hierarchical consolidation at a 0.90 conceptual cut yielded 21 macro-topics. Human evaluation of the semantic labels assigned to these macro-topics by four annotators validated 18 of them, with a mean score of 4.23 on a five-point scale. The results reveal a predominantly affective-relational thematic structure, centered on passion, romantic suffering, breakup, and reconciliation, articulated with spirituality, drinking, communication, and sociability. These findings show that stanza-level semantic topic modeling can move beyond identifying isolated recurring themes to characterize how affective and social dimensions are organized within *sertanejo universitário* lyrics.

CCS Concepts: • **Information systems** → **Data mining**; • **Computing methodologies** → **Natural language processing**.

Keywords: topic modeling, music lyrics, BERTopic, embeddings, natural language processing

1. INTRODUÇÃO

Letras musicais podem ser analisadas como documentos culturais e textuais, pois registram temas recorrentes associados a afetos, valores e formas de sociabilidade. [Requena 2016] situa a emergência do chamado "sertanejo universitário" em 2005, período de renovação do gênero sertanejo. Segundo esse autor, mais do que uma categoria musical rigidamente definida, o termo "universitário" funciona como um rótulo mercadológico, onde as letras articulam narrativas sobre amor, término, desejo, sofrimento, bebida, festa e reconciliação [Requena 2016; Schlösser et al. 2016].

Do ponto de vista de Processamento de Linguagem Natural (PLN) e mineração de textos, o sertanejo universitário constitui um objeto relevante por combinar ampla circulação social, repetição lexical, oralidade, metáforas e forte recorrência de temas afetivos. A análise computacional desse gênero permite investigar, em larga escala, como padrões culturais, emocionais e sociais se organizam nas letras, indo além de leituras impressionísticas sobre a presença de amor, sofrimento, bebida ou festa nas canções.

Embora letras musicais já tenham sido exploradas por abordagens computacionais, ainda há lacunas na análise temática não supervisionada de gêneros musicais brasileiros específicos. Trabalhos anteriores

Pesquisa parcialmente apoiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, CNPq (308246/2026-8), e INCT TILD-IAR (Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação) (CNPq/SECTICS/CAPES/FAPs #408490/2024-1).

investigaram relações entre temas musicais e vieses de gênero [Chen et al. 2025], exploração de subgêneros do sertanejo por diferenças linguísticas [Junior et al. 2019], e modelagem de tópicos em outros recortes musicais brasileiros [Yepez Rojas and Becker 2027; Piceni et al. 2025]. No entanto, ainda é pouco explorada uma análise voltada à organização semântica interna das letras de sertanejo universitário por meio de modelagem não supervisionada de tópicos em nível de estrofe.

O presente trabalho aplica técnicas de PLN a letras de sertanejo universitário para investigar duas questões de pesquisa: (i) que estrutura temática interna emerge da modelagem de tópicos em estrofes de sertanejo universitário? e (ii) como diferentes modelos de *embeddings* afetam a coerência, a cobertura e a interpretabilidade dos tópicos quando aplicados ao mesmo corpus? Para desenvolver esse estudo, adotamos BERTopic [Grootendorst 2022], um framework de modelagem de tópicos que explora a similaridade semântica dos textos, e comparamos diferentes modelos de embeddings para representação das estrofes musicais. Para validação, utilizamos métricas de interpretabilidade de tópicos, bem como conduzimos uma avaliação humana sobre os tópicos encontrados e seus significados.

Nossos achados mostram que, no recorte estudado, as estrofes de sertanejo universitário se organizam em termos de um núcleo afetivo-relacional dominante, associado a paixão, sofrimento amoroso, término e reconciliação, articulado a dimensões de religiosidade, comunicação, bebidas e sociabilidade.

O restante desse trabalho está estruturado como segue. A Seção 2 apresenta o referencial teórico, e a Seção 3, os trabalhos relacionados. A Seção 4 apresenta a metodologia adotada, e a Seção 5 discute os resultados. A Seção 6 apresenta as conclusões e os trabalhos futuros.

2. REFERENCIAL TEÓRICO

2.1 Música, cultura e sertanejo universitário

Letras musicais podem ser tratadas como textos culturais porque condensam formas de expressão afetiva, práticas sociais e valores compartilhados. No sertanejo universitário, estudos qualitativos em um pequeno conjunto curado de músicas, apontam a recorrência de temas ligados a relacionamentos amorosos, sofrimento, desejo, festa, bebida e superação [Requena 2016; Schlösser et al. 2016]. Como esses temas podem alternar dentro de uma mesma música, este trabalho adota a estrofe como unidade de análise para observar como eles se agrupam semanticamente dentro das letras.

2.2 Modelagem de tópicos

A modelagem de tópicos busca identificar temas latentes em um corpus sem rótulos prévios. O BERTopic é um framework de modelagem de tópicos configurável que combina *embeddings* de sentenças, redução de dimensionalidade, agrupamento e representação lexical dos tópicos por c-TF-IDF [Grootendorst 2022]. O framework também permite agrupamento hierárquico de tópicos que permite organizar os tópicos identificados em diferentes níveis de abstração, possibilitando a análise de relações semânticas entre tópicos e a obtenção de uma estrutura temática hierarquizada.

A aplicação do BERTopic em letras musicais exige cuidado interpretativo, pois estrofes são textos curtos, repetitivos, metafóricos e marcados por oralidade. Por isso, os resultados foram avaliados tanto por métricas quantitativas, quanto por leitura qualitativa dos tópicos e validação humana dos rótulos atribuídos. Como métricas adotamos: coerência *CV* (avalia a coerência entre as palavras representativas de um tópico), *NPMI* (mede a força de associação normalizada entre essas palavras), *diversidade* (grau de não repetição das palavras entre tópicos), e o RBO invertido (mede a distinção entre seus rankings de palavras) [Röder et al. 2015; Webber et al. 2010].

3. TRABALHOS RELACIONADOS

A análise computacional de letras musicais tem sido utilizada para investigar vieses sociais, associações semânticas e organização temática explorando métodos de modelagem de tópicos. O estudo em [Chen

et al. 2025] integra BERTopic e SC-WEAT para investigar relações entre temas musicais e vieses de gênero em diferentes gêneros musicais. Embora relevante para o uso de PLN em letras musicais, esse trabalho tem foco distinto: analisa viés e representação social, explorando se algum tópicos está mais associado a viés.

No contexto brasileiro, tópicos de letras de funk são extraídos usando LLMusic [Yepez Rojas and Becker 2027], um método de modelagem de tópicos em nível de discurso que combina LLMs, Engenharia de Prompt e BERTopic. Letras relacionadas à ditatura militar no Brasil são analisadas combinando BERTopic e GSDMM [Piceni et al. 2025]. Técnicas de mineração de texto, incluindo LDA, são empregadas em [Junior et al. 2019] para identificar diferenças linguísticas entre subgêneros do sertanejo.

Este trabalho se distingue por investigar a organização temática do sertanejo universitário em nível de estrofe, utilizando modelagem de tópicos baseada em representações semânticas, comparando diferentes modelos de *embeddings* e combinando métricas quantitativas com avaliação humana.

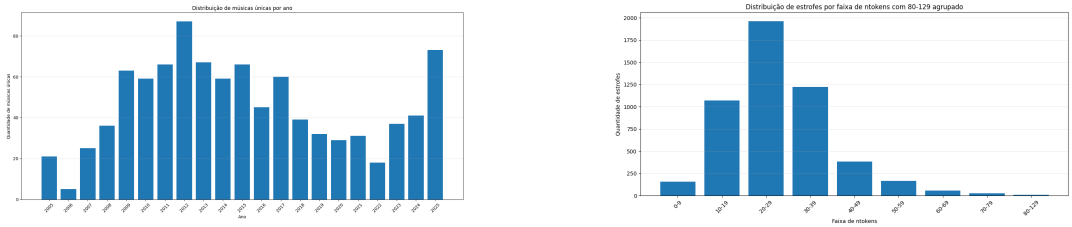
4. METODOLOGIA

A metodologia foi organizada em sete etapas: coleta das músicas por *web scraping*, filtragem e preparação do corpus, segmentação das letras em estrofes, aplicação do BERTopic, comparação entre modelos de *embeddings*, consolidação hierárquica e validação humana dos rótulos dos macro-tópicos mais frequentes. Dados e códigos estão disponíveis em repositório público[1].

4.1 Coleta, composição e pré-processamento

O corpus foi construído a partir de letras de músicas sertanejas coletadas do site *letras.mus.br* por meio de *web scraping*. O critério de coleta foi a classificação da música como do gênero *sertanejo* pelo sítio e a gravação entre 2005 e 2025. A coleta resultou em 959 músicas, correspondetes a 114 artista. Como a coleta usou uma única fonte, o corpus deve ser entendido como um recorte exploratório, sujeito a vieses de cobertura, classificação de gênero musical e inconsistências de metadados.

A unidade de análise adotada foi a estrofe, como proposto em [Yepez Rojas and Becker 2027], já que uma mesma música pode alternar entre sofrimento amoroso, desejo, reconciliação, bebida, festa ou religiosidade. A divisão em estrofes corresponde à separação de versos tal como apresentada no sítio de coleta. O *pipeline* incluiu separação das músicas em estrofes, remoção de músicas em outros idiomas e de estrofes duplicadas, exclusão de estrofes com menos de quatro tokens, limpeza textual e filtro adicional de comprimento, mantendo textos com pelo menos 20 caracteres. O resultado foram 5.037 estrofes únicas após pré-processamento. A Figura 1 mostra a distribuição temporal das músicas e o tamanho das estrofes. A maior parte das estrofes concentra-se entre 20 e 39 tokens, reforçando a necessidade de um método adequado a textos curtos, como o BERTopic.



(a) Músicas por ano. (b) Tokens por estrofe.

Fig. 1: Distribuições do corpus de sertanejo universitário.

[1] `https://github.com/amandaschmieleski/bertopic-sertanejo-universitario`

Table I: Configuração principal do BERTopic.

Etapa	**Configuração**
Embeddings	`jmbrito/ptbr-similarity-e5-small`, via SentenceTransformers
Redução	UMAP: *metric = cosine*, *n_ neighbors = 25*, *n_ components = 5*, *min_ dist = 0.0*, *random_ state = 42*
Agrupamento	HDBSCAN: *min_ cluster_ size = 25*, *min_ samples = 1*, *metric = euclidean*, *cluster_ selection_ method = eom*, *prediction_ data = True*
Representação lexical	CountVectorizer: *ngram_ range = (1, 2)*, *min_ df = 2*, *max_ df = 0.85*, com *stopwords* personalizadas

4.2 Configuração base do BERTopic

A configuração principal usou o modelo ptbr-e5-small, correspondente a jmbrito/ptbr-similarity-e5-small. A configuração do BERTopic foi definida por testes exploratórios. Foram testados diferentes valores de *n_ neighbors*, *n_ components* e *min_ cluster_ size*. A escolha priorizou a minimização do número de *outliers* e o número de tópicos equilibrando representatividade e interpretabilide. Os melhores resultados foram obtidos com os parâmetros resumidos na Tabela I.

4.3 Comparação de modelos de *embeddings* e avaliação dos tópicos

Foram comparados cinco modelos de *embeddings* aplicados ao mesmo conjunto de 5.037 estrofes válidas: `MPNet-multilingual`, `BERT-ASSIN2`, `ptbr-e5-small`, `multilingual-e5-base` e `Serafim-900M`. Em todos os casos, mantiveram-se os mesmos parâmetros de UMAP, HDBSCAN e CountVectorizer; assim, a variação experimental concentrou-se no modelo de *embeddings*.

A avaliação combinou a leitura qualitativa das palavras e das estrofes associadas aos tópicos com quatro métricas quantitativas. *C_ V* e *NPMI* foram calculados no Gensim a partir das dez palavras principais de cada tópico válido e das estrofes tokenizadas. A *diversidade* é calculada como a proporção de palavras únicas entre as listas dos tópicos, enquanto o *RBO invertido* foi calculado como 1 menos a similaridade RBO média entre pares de rankings, com parâmetro $p = 0{,}9$. Em todas as métricas, valores maiores indicam resultados preferíveis.

4.4 Consolidação hierárquica

As diferentes configurações avaliadas geraram de 34 a 50 tópicos. Para reduzir esse número a um conjunto interpretável, dada a configuração final escolhida, foi construída uma hierarquia de macro-tópicos por meio da função `hierarchical_topics(docs)`. O corte conceitual usa a `Distance` dessa hierarquia para agrupar tópicos semanticamente próximos: fusões com `Distance <= dist_cut` são mantidas em uma estrutura *Union-Find*. Fizemos testes exploratórios com cortes entre 0,70 e 0,99, avaliando o ponto que ofereceu o melhor equilíbrio qualitativo entre a redução da fragmentação e a preservação de conceitos interpretáveis.

4.5 Validação humana dos rótulos dos macro-tópicos

Finalmente, o conjunto de macro-tópicos resultantes do corte conceitual foi analisado, e um rótulo interpretativo foi atribuído por uma das autoras a partir das palavras mais relevantes dos tópicos (c-TF-IDF). Para reduzir a subjetividade da nomeação dos macro-tópicos, esses rótulos interpretativos foram avaliados por quatro anotadores humanos. Os avaliadores são alunos de pós-graduação em computação. Os anotadores receberam uma planilha contendo os rótulos propostos para os macro-tópicos, as palavras representativas, bem como um conjunto de estrofes selecionadas pelas autoras dentre as agrupadas no macro-tópicos consideradas como representativas. Eles deveriam avaliar com notas de 1 a 5 o grau de concordância com o rótulo interpretativo, sendo 1 discordância total e 5, a concordância total. Um rótulo foi considerado validado quando apresentou média igual ou superior a 4,0 e pelo menos três das quatro notas iguais ou superiores a 4.

5. EXPERIMENTOS E RESULTADOS

5.1 Comparação e avaliação dos modelos de embeddings e clusterings

A Tabela II apresenta a comparação quantitativa dos tópicos utilizando a mesma configuração base para UMAP, HDSCAN e CountVectorizer (Tabela I), variando os modelos de embeddings. A Tabela III apresenta as métricas de avaliação dos clusterings obtidos. A comparação deve ser interpretada de forma relativa, pois não há padrão-ouro numérico universal para modelagem de tópicos, especialmente em corpus de letras musicais curtas.

Table II: Comparação quantitativa dos modelos aplicados ao mesmo corpus.

Modelo	Tópicos	Textos agrupados	Média/tópico	Outliers	Outliers (%)
`MPNet-multilingual`	45	2.874	63,9	2.163	42,94
`BERT-ASSIN2`	34	**3.019**	88,8	**2.018**	**40,06**
`ptbr-e5-small`	44	2.971	67,5	2.066	41,02
`multilingual-e5-base`	41	2.946	71,9	2.091	41,51
`Serafim-900M`	**50**	2.682	53,6	2.355	46,75

Table III: Métricas de avaliação dos modelos de *embeddings*.

Modelo	Coerência C_V	NPMI	Diversidade	RBO invertido
`MPNet-multilingual`	0,339	-0,230	0,896	0,998
`BERT-ASSIN2`	0,382	-0,255	0,900	0,997
`ptbr-e5-small`	**0,389**	**-0,202**	0,900	0,998
`multilingual-e5-base`	0,323	-0,232	**0,912**	**0,998**
`Serafim-900M`	0,338	-0,248	0,890	0,997

A comparação mostra um *trade-off* entre aproveitamento do corpus e coerência semântica. `BERT-ASSIN2` teve menor taxa de *outliers* e agrupou mais estrofes, enquanto `ptbr-e5-small` obteve os melhores valores de C_V e NPMI. Assim, a escolha do modelo principal não se baseou apenas na quantidade de estrofes agrupadas, mas no equilíbrio entre aproveitamento do corpus e coerência semântica dos tópicos. `Serafim-900M`, embora tenha gerado mais tópicos, apresentou maior proporção de *outliers*, o que indica maior fragmentação. Esses resultados sustentam a configuração usada nas análises seguintes.

5.2 Tópicos gerados pelo modelo principal e Análise de Macro-tópicos

Dada a configuração da Tabela I, foram identificados 44 tópicos válidos, que agrupam 2.971 estrofes (58,98% do total), detalhados na Figura 2. Na consolidação hierárquica, foi escolhido o corte conceitual de 0,9, indicado por uma linha pontilhada na Figura 2, resultando em 21 macro-tópicos.

A Tabela IV apresenta os dez macro-tópicos com o maior número de estrofes agrupadas, indicando na coluna *Tópicos* as fusões considerando os tópicos originais da Figura 2. Os dez macro-tópicos correspondem a 84,52% dos trechos agrupados (2.511). A coluna *Val.* apresenta a média ± desvio-padrão da concordância humana dos rótulos. Por restrições de espaço, a interpretação dos demais 11 macro-tópicos, com tamanho de 25 a 72 documentos, correspondente a 15,48% do total de trechos agrupados, é detalhada no repositório público.

Table IV: 10 Macro-tópicos mais frequentes após o corte conceitual 0,90.

ID	Nome interpretativo	Palavras representativas	Tamanho	Tópicos	Val.
A	Paixão, sofrimento amoroso e término	amo, amar, paixão, tempo	814	(19+6) + (15+13) + ((4+0) + 9)	4,60 ± 0,55
B	Corpo, desejo e afeto	corpo, acordar, dose, sinto	523	21 + ((2+1) + 8)	4,75 ± 0,50
C	Imagens poéticas de natureza, tempo e céu	lua, sol, céu, estrelas	246	(3+22) + (32+37)	4,00 ± 0,82
D	Reconciliação, volta e incerteza afetiva	volta, será, povo, ninguém	243	(31 + (10+11)) + 29	4,50 ± 0,58
E	Espiritualidade, cuidado e proteção	deus, coloca, senhor, cuidado	172	25+5	5,00 ± 0,00
F	Jogos afetivos e disputa amorosa	fora, jogo, jogado, fogo	128	35+7	4,00 ± 0,82
G	Comunicação, recaída e recomeço	celular, internet, mensagem, fila	120	12+28	4,75 ± 0,50
H	Bar, família e sociabilidade	bar, família, virou, homem	95	24+18	4,00 ± 1,41
I	Ex-relacionamento, curtição e superação	ex, acabou, valorizando, amor	92	16+30	5,00 ± 0,00
J	Perda, decisão e mudança de vida	perdeu, dar, jeito, fugida	78	20+33	4,75 ± 0,50

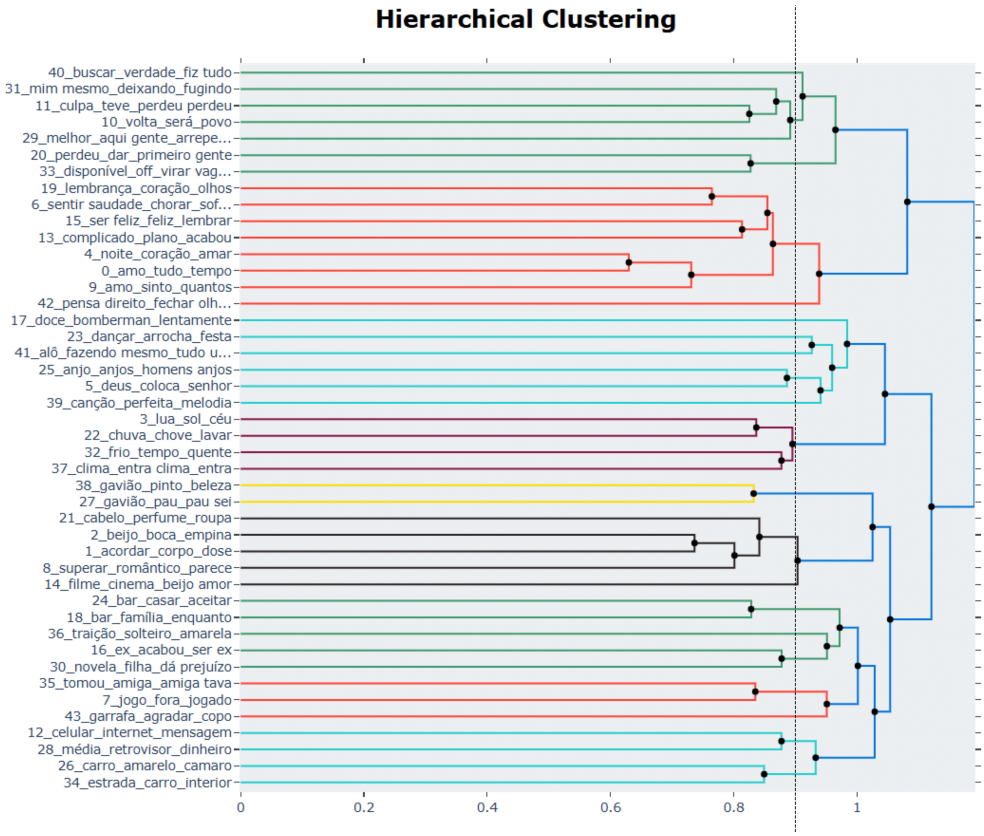


Fig. 2: Hierarquia dos tópicos com corte conceitual adotado *dist-cut* = 0,9.

Os resultados indicam predominância de macrotemas afetivos e relacionais no corpus. O macro-tópico mais frequente (A), interpretado como *paixão, sofrimento amoroso e término*, reúne termos representativos como "amo", "amar", "paixão" e "tempo", indicando a centralidade de vínculos afetivos intensos e perda amorosa. Outros macrotemas, como *reconciliação*, *espiritualidade*, *comunicação digital*, *bar e sociabilidade*, mostram que as letras combinam dimensões afetivas, sociais e cotidianas. Os 11 macro-tópicos de menor frequência, detalhados no material suplementar, ampliam esse repertório com temas mais específicos, como ostentação, festa, traição e bebidas, com menor representatividade.

A consolidação hierárquica também revela concentração temática. Note-se que os dois primeiros tópicos (A e B) agrupam 1.337 estrofes (45,0% do total), sendo 814 estrofes agrupadas (27,4%) no macro-tópico A somente. Considerando os macro-tópicos diretamente associados a relações afetivas, desejo, disputa, ex-relacionamento, perda e mudança de vida (IDs A, B, D, F, I e J), observa-se um total de 1.878 estrofes, equivalente a 63,2% das estrofes agrupadas. Assim, o resultado não mostra apenas a presença de temas afetivos já conhecidos na literatura sobre sertanejo universitário; ele evidencia que esses temas ocupam posição estrutural dominante quando as letras são analisadas em nível de estrofe.

Também é possível distinguir três camadas temáticas entre os macro-tópicos mais frequentes. A primeira corresponde ao núcleo afetivo-relacional, formado por paixão, sofrimento, desejo, reconciliação, disputa amorosa, ex-relacionamento e perda (macro-tópicos A, B, D, F, I e J). A segunda, representada pelo macro-tópico C, reúne elementos simbólicos e expressivos, como imagens de natureza, do tempo e do céu, que funcionam como recursos poéticos recorrentes na literatura. A terceira (macro-tópicos E, G e H) agrega dimensões cotidianas e sociais, como espiritualidade, comunicação digital, bar, família e sociabilidade, indicando que o sofrimento amoroso e a reconciliação aparecem frequentemente articulados a práticas sociais e espaços de convivência.

A validação humana reforçou a adequação da maioria dos rótulos atribuídos aos macro-tópicos. Considerando os 10 macro-tópicos mais frequentes (Table IV), os rótulos avaliados atingiram o critério adotado, com média geral de 4,54. As menores médias foram 4,00, ainda dentro do limiar de validação, e os rótulos "Espiritualidade, cuidado e proteção" e "Ex-relacionamento, curtição e superação" obtiveram média 5,00. O maior desvio-padrão ocorreu no macro-tópico "Bar, família e sociabilidade" ($4{,}00 \pm 1{,}41$). Já a tarefa de nomear os 11 agrupamentos listados no material suplementar, com menor quantidade de estrofes e maior heterogeneidade semântica, foi mais desafiadora, resultando em menor concordância dos avaliadores. Apenas 7 dos 11 tópicos tiveram concordância média superior a 4, variando entre 4,25 e 5. No geral, os resultados indicam uma avaliação positiva dos anotadores quanto à correspondência entre os nomes interpretativos e os agrupamentos temáticos gerados pelo modelo, ao mesmo tempo em que apontam quais rótulos são mais amplos e requerem maior interpretação qualitativa.

Em conclusão, a validade dos macro-tópicos não se apoiou em um único critério, mas na convergência entre evidências quantitativas, inspeção qualitativa das palavras e estrofes associadas, consolidação hierárquica, análise dos *outliers* e avaliação humana dos rótulos. Essa triangulação é especialmente importante na modelagem de tópicos não supervisionada, na qual não há um gabarito externo único para definir a interpretação correta dos agrupamentos.

5.3 Ilustração : Macro-tópicos nas músicas mais tocadas

A Figura 3 ilustra a distribuição dos dez macro-tópicos mais frequentes nas dez músicas mais ouvidas presentes no nosso *corpus*. Usamos o *ranking* do sítio *Letra.Mus.br* e consideramos somente letras que atenderam ao critério de possuir menos de 40% das estrofes classificadas como *outliers*. Para cada música, calculamos a proporção de cada macro-tópico tópico dividindo o número de estrofes da música atribuídas ao respectivo macro-tópico pelo número total de estrofes da música.

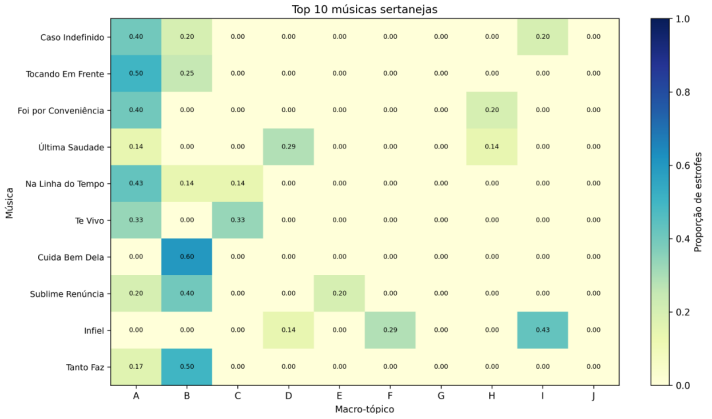


Fig. 3: Distribuição dos dez macro-tópicos no top 10 músicas sertanejas mais ouvidas

A Figura 3 evidencia que uma mesma música pode combinar diferentes macro-tópicos, reforçando a escolha da estrofe como unidade de análise. O macro-tópico A, relacionado a paixão, sofrimento amoroso e término, aparece em sete das dez músicas selecionadas, enquanto o macro-tópico B, associado a corpo, desejo e afeto, também apresenta presença recorrente e alcança sua maior proporção na música "Cuida Bem Dela" (0,60).

5.4 Análise dos outliers

No modelo principal `ptbr-e5-small`, 2.066 das 5.037 estrofes (41,02%) foram classificadas como *outliers*. Essa proporção pode refletir uma característica particular da linguagem das letras musicais: a formação de agrupamentos depende da existência de similaridades semânticas suficientemente explícitas entre os textos, enquanto a expressão poética frequentemente comunica significados de forma

indireta, por meio de metáforas, imagens, expressões figuradas e outras construções que podem representar um mesmo tema sem compartilhar pistas lexicais ou semânticas evidentes. Assim, uma estrofe pode expressar claramente um tema para um leitor humano sem apresentar similaridade suficiente com outras estrofes para formar um agrupamento denso. Os *outliers*, portanto, não devem ser interpretados necessariamente como ausência de conteúdo temático, mas como estrofes para as quais não foram encontradas evidências suficientemente recorrentes de similaridade no espaço semântico utilizado pelo modelo.

6. CONCLUSÃO

Este trabalho investigou a organização temática interna do sertanejo universitário por meio de modelagem de tópicos baseada em similaridade semântica (BERTopic) em nível de estrofe. A comparação de *embeddings* demonstrou que o modelo `ptbr-e5-small` forneceu a representação semântica mais equilibrada. A consolidação hierárquica via BERTopic evidenciou a existência de uma arquitetura temática com um núcleo relacional dominante, bem como recursos poéticos da natureza e referências a práticas cotidianas de sociabilidade, bebida e comunicação digital. A validação com anotadores humanos corroborou a interpretabilidade semântica.

Do ponto de vista metodológico, o volume substancial de *outliers* (41,02%) reflete a complexidade estilística das letras musicais. Como limitações, ressaltam-se a extração a partir de fonte única, a atribuição rígida a um único tópico por estrofe e a ausência de processo de validação composto por especialistas no gênero musical. Baseado em pistas lexicais similares, limitações do BERTopic para capturar os tópicos em níveis de discurso foram observadas.

Para trabalhos futuros, sugere-se a aplicação de abordagens alternativas de modelagem de tópicos, tais como LLMusic [Yepez Rojas and Becker 2027], a análise da evolução temporal dos macro-tópicos entre 2005 e 2025, e o cruzamento dos tópicos com dimensões discursivas de gênero.

Declaração de uso de IA generativa

Foi utilizado o ChatGPT, versão GPT-5.5 Thinking, como apoio à geração/depuração de código e à revisão textual. Todos os códigos, parâmetros, resultados, rótulos e interpretações finais foram revisados, validados e assumidos integralmente como responsabilidade das autoras.

REFERENCES

Chen, D., Satish, A., Khanbayov, R., Schuster, C., and Groh, G. Tuning into bias: A computational study of gender bias in song lyrics. In *Proc. of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. ACL, New Mexico, pp. 117–129, 2025.

Grootendorst, M. BERTopic: Leveraging bert and topic modeling for efficient document clustering. `https://maartengr.github.io/BERTopic/index.html`, 2022.

Junior, J. S., Rossi, R., and Lobato, F. A lyrics-based approach to knowledge discovery in brazilian music: Sertanejo as a case study. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. pp. 949–960, 2019.

Piceni, H. R., Alexandre, P. V., and Balreira, D. G. A música brasileira na ditadura militar: uma análise de tópicos com bertopic e gsdmm. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. SBC, Porto Alegre, RS, Brasil, pp. 349–360, 2025.

Requena, B. H. d. A. F. *A universidade do sertão: o novo retrato cultural da música sertaneja*. M.S. thesis, Universidade de São Paulo, São Paulo, 2016.

Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, pp. 399–408, 2015.

Schlösser, A. et al. Representações sociais de término de relacionamentos amorosos em músicas do sertanejo universitário. *Psicologia em Revista* 22 (2): 407–427, 2016.

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

Yepez Rojas, J. D. and Becker, K. Unsupervised topic modeling of song lyrics with large language models: A zero-shot framework for discourse analysis. *Expert Systems with Applications* vol. 333, pp. 133782, 2027.