LLMusic: Modelagem de tópicos em letras de músicas combinando LLM, Engenharia de Prompt e BERTopic

Jesus Daniel Yepez Rojas [1], **Karin Becker** [1]

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

{jesus.rojas, karin.becker}@inf.ufrgs.br

Nível: Mestrado
Ano de ingresso do programa: 2023
Época esperada de conclusão: Maio 2025
Etapas concluídas: Revisão bibliográfica, desenvolvimento do framework, estudo de caso.
Etapas futuras: Aprimoramento do framework, generalização do framework para todos os gêneros musicais.

Abstract. *Song lyrics impose additional challenges to topic modeling, as the discourse is often implicit and must be understood within its context, using figurative and poetic language, and slangs. This paper proposes LLMusic, a new topic modeling approach that leverages Large Language Models (LLMs) to analyze lyrics. We use LLMs and prompting to summarize song excerpts into central themes in an iterative and unsupervised process applied to a corpus representative of the genre. These themes are grouped into a lean, coherent set of topics using BERTopic. Through zero-shot prompts, one can classify new lyrics based on these topics. In the case study, LLMusic captures the social phenomena at the base of Brazilian funk, demonstrating its potential for large-scale analysis.*

Resumo. *Letras de músicas impõem desafios adicionais à modelagem de tópicos, já que o discurso nem sempre é explícito, devendo ser compreendido dentro de seu contexto, sua linguagem figurativa e poética, gírias, etc. Este artigo propõe o LLMusic, uma nova abordagem de modelagem de tópicos que explora o potencial de Grandes Modelos de Linguagem (LLMs) para analisar letras de música. LLMs e prompting são usados para resumir trechos de músicas em temas centrais, em um processo iterativo e não supervisionado aplicado a um corpus representativo do gênero. Esses temas são agrupados em um conjunto enxuto e coerente de tópicos usando BERTopic. Através de prompts zero-shot, pode-se classificar novos trechos de letras com base nesses tópicos. No estudo de caso desenvolvido, LLMusic capturar os fenômenos sociais à base do funk brasileiro, mostrando seu potencial para análise em larga escala.*

Palavras-chaves: LLMs, engenharia de prompt, modelagem de tópicos, BERTopic.

1. Introdução

O estudo da música proporciona uma visão multidimensional para as relações entre cultura e sociedade, oferecendo uma oportunidade de desenvolvimento de análises interdisciplinares e descoberta de conhecimento. Músicas refletem tradições e valores de um determinado povo, espelhando suas condições sociais. Em muitos gêneros musicais (e.g. *blues*, *folk* e *rap*), as letras musicais retratam as realidades sociais de uma época.

Diferentes trabalhos já mergulharam nos aspectos sociais e culturais das letras de música. Em [Smiler et al. 2017], 11 pesquisadores avaliaram a presença de relações amorosas e sexuais em uma base de 1250 músicas para uma análise estatística. No Brasil, o funk tem inspirada análises como o discurso sobre as mulheres [Peres 2023] ou o racismo velado da sociedade brasileira [Lopes and Facina 2012]. Tais trabalhos requerem um número grande de anotadores ou limitam-se a pequenos conjuntos de letras. Assim, ferramentas computacionais podem auxiliar estudos qualitativos, permitindo, em larga escala, a identificação, classificação e sumarização dos discursos presentes nestes textos.

Implementações de técnicas computacionais para análises de letras de músicas foram realizadas em trabalhos abordando viés de gênero [Betti et al. 2023], letras e áudio [Calcina 2022], recorrência de temas em gêneros musicais [Junior et al. 2019] e classificação de sentimentos [Devi and Saharia 2020]. Boa parte desses trabalhos exploram Modelagem de Tópicos (MT), tais como LDA ou BERTopic, como ferramenta de extração de informação em textos. Contudo, letras musicais contém uma grande subjetividade na expressão das ideias. Elementos como figuras de linguagem, poesia e gírias, diversificam as palavras utilizadas para representação de um mesmo contexto. Assim, agrupamentos gerados por essas técnicas podem carecer de relevância discursiva.

Grandes Modelos de Linguagem (Large Language Models - LLMs) vêm sendo aplicados com sucesso em uma diversidade de tarefas de processamento de linguagem natural (PLN). Uma das formas de se interagir com LLMs é através de *prompts*, cujo design através da Engenharia de Prompts (EP) influencia os resultados. Dada a configuração correta dos parâmetros e *prompts*, LLMs têm demostrado capacidade de sumarização de textos semelhantes às feitas por humanos [Zhang et al. 2024]. Isto sugere que a utilização de LLMs para superar desafios de MT pode ser uma abordagem promissora para a análise de letras musicais. LLMs conseguem captar nuances contextuais e semânticas que métodos tradicionais de MTs podem não alcançar. Isso ocorre pois estes modelos são treinados em vastas quantidades de texto, desenvolvendo uma compreensão mais profunda das estruturas e significados linguísticos. A presente pesquisa soma-se aos esforços para combinar o potencial de EP e LLMs para MT [Pham et al. 2024].

Este trabalho propõe LLMusic, um *framework* para extração de tópicos em um *corpus* de letras de músicas. LLMusic combina o poder de PE e LLMs, com técnicas de MT avançadas como o BERTopic, visando assim contribuir às limitações de métodos de MT tradicionais para identificação de tópicos subjetivamente representados em letras musicais. O método requer como entrada um *corpus* de letras representativas do gênero musical. Exploramos a EP e LLMs de duas formas: identificação de temas e atribuição de tópicos. Primeiramente, usamos *prompts* e LLMs para extrair *temas* expressos nos trechos de músicas, explorando a capacidade generativa de LLMs para sumarizar trechos de músicas em temas. Para criação de uma distribuição de temas robusta e representativa, combinamos aleatoriamente trechos de músicas de um *corpus* referência em múltiplas iterações. Usamos o BERTopic para resumir essa distribuição em uma lista não redun-

dante de tópicos representativos. Também usamos PE sobre LLMs para atribuir de forma não supervisionada (*prompts zero shot*) os tópicos identificados a trechos de música para análise em larga escala. Nossos resultados preliminares mostram a capacidade de LLMusic capturar os fenômenos sociais à base do funk.

Sobre o restante do texto: A Seção 2 sumariza os trabalhos relacionados. A Seção 3 detalha o *framework* LLMusic. A Seção 4 discute a aplicação do framework num *corpus* de músicas do gênero funk brasileiro. A Seção 5 discute as direções futuras.

2. Trabalhos relacionados

Os estudos que ligam técnicas computacionais e análise musical vêm aumentando, permitindo extrair *insights* a partir da extensa quantidade de músicas disponíveis [Oramas et al. 2018]. Eles destacam como a música tem sido cada vez mais vista como um sistema linguístico, podendo se beneficiar de técnicas de PLN de ponta, como LLMs.

[Betti et al. 2023] explora aprendizagem supervisionada através de ajustes de modelos BERT para identificação de sexismo e viés de gênero em larga escala. Por ser uma análise supervisionada, sua reprodução é restrita à existência de uma base anotada.

Nos estudos de música que utilizam técnicas não supervisionadas, LDA é a técnica prevalente. Em [Devi and Saharia 2020] é aplicada para classificação de sentimentos. [Junior et al. 2019] investiga como termos específicos (e.g., álcool e relacionamentos) são empregados nas letras do gênero sertanejo. Como o LDA extrai relações de coocorrências das palavras, a geração de tópicos está atrelada a uma recorrência dos artistas em utilizar as mesmas palavras para representar as mesmas situações. Em [Calcina 2022], BERTopic é utilizado para descobrir semelhanças letras de diferentes gêneros musicais. Embora capaz de identificar tópicos com base em similaridade semântica de textos, tem limitações para agrupar discursos implicitamente representados.

Trabalhos recentes destacam o potencial de LLMs em tarefas de sumarização de textos. Em [Zhang et al. 2024], é feita uma avaliação humana de dez diferentes LLMs. O trabalho aponta que a qualidade de sumarização dos LLMs é equivalente à humana. O uso de LLMs explora o paradigma de *prompt learning*. A estratégia (e.g. *zero-shot*, *few-shot*) escolhida na criação do *prompt* é fundamental, uma vez que impacta diretamente o resultado do modelo [Pengfei Liu and Neubig 2023]. Alguns dos desafios para o uso de LLMs incluem alucinações e a natureza não determinística dos LLMs.

Estudos abordaram o potencial de EP em MT de forma geral [Pengfei Liu and Neubig 2023, Pham et al. 2024]. Em [Pham et al. 2024] propõe-se um *framework* onde uma distribuição de tópicos é gerada por EP considerando artigos de Wikipédia e resumos de faturas do congresso americano. A metodologia foi superior a modelos tradicionais de MT como LDA e BERTopic. No entanto, o refinamento dos tópicos exige intervenção manual para análise e definição dos critérios de fusão e exclusão de tópicos. Além disso, a etapa de atribuição de tópicos requer a elaboração de *prompts* contendo exemplos de cada tópico (*few shot*), características que dificultam a implementação em comparação com *frameworks* tradicionais de MT.

Nesse contexto, propomos LLMusic como um framework capaz de lidar com a a natureza figurativa e subjetiva das letras musicais, especialmente em gêneros como o funk brasileiro, por meio e combinada de técnicas avançadas de MT e o uso de LLMs com EP numa abordagem de modelagem multi-tópicos não-supervisionada.

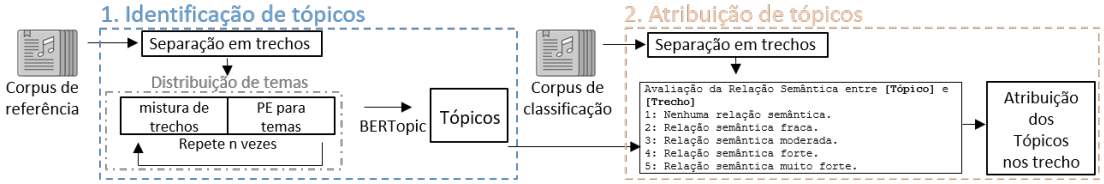

Figura 1. Visão geral do Framework LLMusic

3. LLMusic: Modelagem de tópicos em músicas com Prompt Engineering e BERTopic

Neste trabalho é proposto LLMusic, uma abordagem para identificação de tópicos em letras de músicas. LLMusic combina a capacidade de interpretação das LLMs com a capacidade de identificação de tópicos do BERTopic baseado em similaridade semântica. A Figura 1 resume o *framework* LLMusic, que aborda duas tarefas principais: *Identificação de Tópicos* e *Atribuição de Tópicos*.

A *Identificação de Tópicos* requer como entrada um *corpus* não rotulado de letras divididas em trechos. É realizada em duas fases: a) criação de uma distribuição de *temas* contidos nas músicas usando PE, e b) agrupamento dos temas em um conjunto reduzido, significativo e não redundante de *tópicos*. Para criar uma distribuição confiável, embaralhamos iterativamente os trechos para formar grupos aleatórios usados como entrada no modelo de PE, repetindo este processo várias vezes e agregando os resultados.

A mistura aleatória de trechos aumenta a diversidade de temas identificados através do PE, criando uma distribuição confiável. Ainda, o BERTopic, baseado em agrupamento por densidade, ajuda a eliminar *outliers*, priorizando os temas mais frequentes. Ao descrever o *corpus* musical a partir desses temas, minimizamos o problemas de alucinações e de respostas pouco significativas, comuns em aplicações de LLMs.

A segunda tarefa é a *Atribuição de Tópicos* aos trechos de músicas. Essa atribuição é feita por um estrutura de PE *zero-shot* utilizando uma LLM, onde cada requisição compara individualmente cada tópico com cada trecho de música em uma tarefa de classificação de relacionamento semântico, ilustrado na Figura 1.

4. Resultados preliminares

Nesta seção são apresentados os resultados preliminares do framework LLMusic para extração de tópicos em letras de músicas, discutidos a traves da resultados de implementação do framework em um *corpus* de músicas do gênero funk brasileiro.

Corpus referência: Adotamos as 18 músicas analisadas em [Peres 2023], que estuda a expressão da masculinidade em letras de funk. As 18 músicas do *corpus* foram separadas em 174 trechos.

Distribuição de temas: os 174 trechos foram divididos em 20 grupos, resultando em aproximadamente 9 trechos por grupo. Como LLM, adotamos o Sabiá-2-medium [Ramon Pires and Nogueira 2023]. A arquitetura de *prompting* utilizada foi a *zero-shot* Com 20 repetições, geramos 1.191 temas. A distribuição é rica mas redundante. Por exemplo, entre os 10 temas mais frequentes estão *relacionamentos amorosos*; *amor*; *amor e romance*; *amor e paixão*.

Tabela 1. ID dos Tópicos com descrição e tema representativo

Tópicos	Descrição	Tema Representativo
1	Arrependimento	arrependimento e pedido de perdão aos pais
2	Relações Familiares	feição e relações familiares
3	Relacionamentos	Relacionamentos e comunicação
4	Aparência e comportamento das mulheres	Estereótipos de aparência e comportamento das mulheres
5	Dinâmicas de sedução, sexo, desejo sexual	Comportamento sexual e atração física
6	Festas e bailes Funk	Cultura do funk e dança
7	Consumo e ostentação	Estilo de vida e consumo de produtos de marcas famosas
8	Tráfico e consumo de substâncias	Consumo de álcool e drogas
9	Violência e crime	Violência e linguagem vulgar na música
10	Vida nas favelas	Localização geográfica e referências a bairros e favelas
11	Reflexões sobre a vida e superação de desafios	Reflexão sobre a vida e superação de obstáculos

Geração de tópicos: Usamos BERTopic para agrupar os temas em um conjunto menor de tópicos representativos, de modo a capturar sua similaridade semântica, e tratar temas infrequentes como *outliers*. O resultado foi a identificação de 13 tópicos, onde 260 temas foram considerados *outliers*. Todos os tópicos foram avaliados por uma antropóloga especialista em funk e seu impacto social, que recomendou o descarte de dois por irrelevância de discurso (Apologia ao funk, Apresentação de MCs). O resultado é sumarizado na Tabela 1, descritos por um identificador do tópico, o rótulo definido pela especialista e o tema mais representativo dentre os agrupados no tópico.

Atribuição de tópicos: Os trechos foram classificados relacionando a um ou mais tópicos usando uma estratégia de PE *zero-shots*. A estrutura de *prompt* escolhida é ilustrada na Figura 1, i.e. dado um tópico e um trecho, a LLM quantifica de 1 a 5 a relação semântica entre ambos. Para essa tarefa, o LLM GPT4 teve melhor desempenho comparado ao Sabiá, que apresentou comportamento inconsistente nas respostas do modelo.

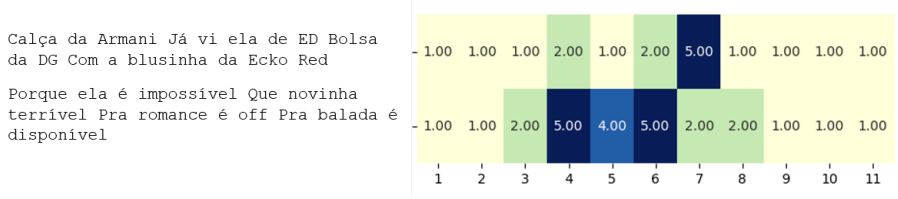

Figura 2. Resultados de classificação de similaridade semântica em trechos selecionados da música *"Novinha Terrível"*

A Figura 2 ilustra os tópicos atribuídos para dois trechos da musica *"Novinha Terrível"*, que descreve uma jovem da favela sob a ótica masculina. O primeiro fala sobre marcas de roupa, o modelo relacionou com tópico 7 (*Consumismo e ostentação*) com nota máxima. O segundo trecho com nota alta nos tópicos 4 (*Aparência e comportamento das mulheres*), 5 (*Dinâmicas de sedução, sexo, desejo sexual*) e 6 (*festas, bailes Funk*), pois critica que ela prefere ir a festas ao invés de envolver-se amorosamente.

Avaliação: construímos uma base de teste para mensurar o desempenho da classificação *zero-shot*. Selecionamos músicas conhecidas contendo tópicos identificados na Tabela 1, divididas em trechos. Três anotadores foram instruídos a anotar para cada trecho presença/ausência dos tópicos da Tabela 1. Selecionamos os trechos onde pelo menos dois anotadores concordaram, resultando em 131 trechos anotados. Os trechos passaram pelo *prompt* de cada um dos 11 tópicos. O tópico foi considerado presente no trecho

Tabela 2. Desempenho da atribuição de tópicas por tópico

.

Tópico	1	2	3	4	5	6	7	8	9	10	11
Revocação ponderada	92,40%	73,93%	75,14%	77,94%	91,95%	69,37%	92,98%	89,84%	86,36%	78,92%	72,52%
Precisão ponderada	87,70%	81,35%	80,24%	78,35%	92,31%	84,45%	88,85%	87,52%	88,13%	81,64%	86,72%
F1 ponderada	81.65%	80.61%	78.84%	77.40%	92.04%	82.74%	84.53%	84.83%	87.64%	81.54%	86.87%
Acurácia	85,50%	83,97%	82,44%	78,63%	92,37%	81,68%	87,02%	86,26%	88,55%	82,44%	87,02%

usando o limiar $>= 3$ em relação à saída do *prompt*.

A Tabela 2 mostra o desempenho do método *zero-shot*, usando precisão, revocação e F1. Para cada tópico, os trechos anotados com o respectivo tópico foram considerados da classe positiva, e os demais, da classe negativa. Os resultados apresentados correspondem à média ponderada das classes positiva/negativa, onde o peso foi a proporção de casos em cada classe.

Observamos um bom desempenho, com bom compromisso entre precisão e revocação, e com equilíbrio entre as classes positiva e negativa. A análise dos resultados trecho a trecho sugerem uma dificuldade do *prompt* na tarefa de atribuição em trechos onde o tópico é expresso de forma mais subjetiva, o que explica a revocação mais baixa.

5. Conclusões e próximos passos

Apresentamos o LLMusic, um *framework* para a identificação e classificação não supervisionada de tópicos em letras de músicas, que explora o potencial de LLMs através da EP para vencer as limitações de métodos tradicionais de MTs. O estudo de caso no corpus de músicas funk apresentou bom desempenho. Contudo, os valores de revocação e a avaliação qualitativa dos trechos sugerem que existe uma margem de melhoria na etapa de atribuição dos tópicos. Assim, a continuação do trabalho explorará arquiteturas mais complexas de PE como *few-shots* e *chain of thougt* na a etapa de atribuição. Ainda, não se descarta a possibilidade de treinamento de modelos pré-treinados BERT via ajuste fino para a tarefa de classificação.

O framework apresenta vários pontos de configuração, tais como o critério de separação e embaralhamento de trechos; o LLM escolhido para geração de temas e para a atribuição de tópicos; o número de iterações necessárias para a geração da distribuição de temas; a arquitetura do PE e a técnica de agrupamento de temas. Todos esse parâmetros foram definidos experimental no contexto específico do estudo de caso, considerando um *corpus* pequeno de músicas apenas do gênero funk. Com o objetivo de consolidar o framework como uma ferramenta de análises musical em grande escala, os próximos passos focarão na implementação do LLMusic em *corpus* de letras maiores, e com outros gêneros musicais. Isto proporcionará um entendimento melhor do impacto dos parâmetros do framework no resultado final, ao mesmo tempo que firmará a capacidade do framework na tarefa de extração de tópicos em letras musicais.

Finalmente, a qualidade dos tópicos gerados pelo LLMusic no estudo de caso foi demostrada através da análise de uma antropóloga especialista em funk, bem como análise qualitativa de um subconjunto de trechos. Para consolidar LLMusic como uma metodologia de modelagem de tópicos, mesmo que dentro do contexto específico da análise musical, é preciso desenvolver comparações com técnicas convencionais de Modelagem de Tópicos (e.g., LDA, BERTopic) utilizando métricas específicas da tarefa, tais como coerência e diversidade dos tópicos.

Agradecimentos: O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). (131387/2023-5).

Referências

Betti, L., Abrate, C., and Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1):10.

Calcina, Erik e Novak, E. (2022). Measuring the similarity of song artists using topic modelling. In *Proc. of the 25th Intl. Multiconference Information Society - Data Mining and Data Warehouses (SiKDD)*, page 103–106.

Devi, M. D. and Saharia, N. (2020). Exploiting topic modelling to classify sentiment from lyrics. In *Proc. of the 2nd Intl. Conferemce on Machine Learning, Image Processing, Network Security and Data Sciences (MIND)*, pages 411–423.

Junior, J. S., Rossi, R., and Lobato, F. (2019). Uma abordagem baseada em letras para a descoberta de conhecimento da música brasileira: o sertanejo como um estudo de caso. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 949–960, Porto Alegre, RS, Brasil. SBC.

Lopes, A. C. and Facina, A. (2012). Cidade do funk: expressões da diáspora negra nas favelas cariocas. *Revista do Arquivo Geral da Cidade do Rio de Janeiro*, 6:193–206.

Oramas, S., Espinosa-Anke, L., Gómez, F., and Serra, X. (2018). Natural language processing for music knowledge discovery. *Journal of New Music Research*, 47:365–382.

Pengfei Liu, Weizhe Yuan, J. F. Z. J. H. H. and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey ofprompting methods in natural language processing. *ACMCom-put.*, 55(9):35.

Peres, F. C. (2023). Puta ou santa: as relações com mulheres enquanto elemento constituinte das masculinidades do funk brasileiro? In *IV Encontro Anual de Antropologia do Mercosul*.

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., and Iyyer, M. (2024). Topicgpt: A prompt-based topic modeling framework. `https://doi.org/10.48550/arXiv.2311.01449`.

Ramon Pires, Hugo Abonizio, T. S. A. and Nogueira, R. (2023). Sabía: Portuguese large language models. *Anais da XII Brazilian Conference on Intelligent Systems*, 12(1):15.

Smiler, A. P., Shewmaker, J. W., and Hearon, B. (2017). From "i want to hold your hand" to "promiscuous": Sexual stereotypes in popular music lyrics, 1960–2008. *Sexuality & Culture*, 21(4):1083–1105.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.