Detecção de tópicos frequentes em músicas virais no Spotify

Jorge H. F. da Silva

Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

jorge.ferreira@dcc.ufmg.br

Abstract. *This project examines the behavior of the music industry, observing the characteristics that make a song go viral. This study analyzed the evolution of hit songs from 2017 to 2021 using natural language processing (NLP) methods such as Latent Dirichlet Allocation (LDA) and BERTopic to identify recurring topics. The data is pre-processed to remove noise and then analyzed to detect predominant patterns and sentiments. The comparison obtained between LDA and BERTopic shows that LDA provides a comprehensive view of the topics, while BERTopic offers more specific and detailed analyses. These findings highlight the diversity of song topics and language over the years and suggest areas for future research, including the influence of culture and society on musical success.*

Resumo. *Este projeto examina o comportamento da indústria musical, observando as características que tornam uma música viral. Este estudo analisou a evolução de músicas hit de 2017 a 2021 usando métodos de processamento de linguagem natural (PLN) como Latent Dirichlet Allocation (LDA) e BERTopic para identificar tópicos que são recorrentes. Os dados são pré-processados para remover ruídos e então analisados para detectar padrões e sentimentos predominantes. A comparação obtida entre LDA e BERTopic mostra que o LDA fornece uma visão abrangente dos tópicos, enquanto BERTopic disponibiliza análises mais específicas e detalhadas. Estas descobertas destacam a diversidade dos tópicos das músicas e da linguagem ao longo dos anos e sugerem áreas para pesquisas futuras, incluindo a influência da cultura e da sociedade no sucesso musical.*

1. Introdução

A indústria da música é um campo dinâmico e complexo, onde a compreensão dos padrões de sucesso e as tendências emergentes desempenham um papel crucial. Nesse viés, o POC I teve como objetivo principal explorar os padrões de sucesso e viralidade na indústria musical, com foco na influência das colaborações entre diferentes gêneros musicais; buscou-se compreender por que algumas músicas se tornam virais e populares, enquanto outras permanecem desconhecidas, e como as combinações entre gêneros podem impactar o sucesso das músicas.

Os resultados obtidos revelaram diferenças significativas nos gêneros musicais populares em mercados regionais, bem como a influência da cultura musical latina no gosto global. A aplicação do algoritmo Apriori permitiu a identificação eficaz de padrões e regras de associação, revelando interações complexas entre gêneros musicais em diversos mercados, destacando a influência da cultura musical latina e as tendências consistentes nos países de língua inglesa, como rap, trap e hip hop.

Partindo desse contexto, damos continuidade a essa investigação, ampliando nosso escopo acerca das características das músicas virais, explorando a temática dos tópicos presentes em suas letras. Através da aplicação de técnicas de Processamento de Linguagem Natural (NLP), como Latent Dirichlet Allocation (LDA) e Bertopic, o objetivo central é identificar e analisar os temas recorrentes nas letras das músicas virais, buscando padrões e correlações que, juntamente com os resultados obtidos no POC I, possam apresentar insights que contribuam para a compreensão do sucesso musical.

O restante deste artigo está organizado para responder a todas essas perguntas. Apresenta-se primeiro trabalhos relacionados na Seção 2; introduz-se a metodologia baseada em técnicas de processamento de linguagem natural, detalha-se a avaliação experimental e são apresentados os resultados na Seção 3. Finalmente, é feito as considerações gerais sobre as descobertas na Seção 4.

2. Referência Teórica

A análise de letras musicais se tornou um campo de pesquisa em crescimento, impulsionado pelo aumento da disponibilidade de dados e pelo avanço das técnicas de Processamento de Linguagem Natural (NLP). O objetivo principal da detecção de tópicos em letras musicais é identificar temas recorrentes e padrões semânticos que caracterizam as músicas. Essa análise proporciona insights valiosos sobre diversos aspectos da música, incluindo temas, emoções, estilos e tendências. Além disso, tem aplicações comerciais importantes, como auxiliar na criação de estratégias de marketing e desenvolvimento de produtos musicais que atendam às preferências dos consumidores.

O NLP oferece uma variedade de técnicas para a análise de texto em letras musicais. Essas técnicas incluem análise lexical para identificar palavras-chave e termos relevantes, análise semântica para compreender o significado das palavras e frases no contexto das letras, e modelagem de tópicos para agrupar palavras e frases em tópicos temáticos relacionados.

Explorando a literatura da área, pode-se destacar alguns artigos relevantes sobre detecção de tópicos em letras musicais, como “Measuring the Similarity of Song Artists using Topic Modelling” (2022) por Calcina, E., Novak, E., na qual apresentam um método que encontra similaridades entre artistas usando modelagem de tópicos em um conjunto de dados contendo letras de músicas. Na mesma medida,"Exploiting Topic Modelling to Classify Sentiment from Lyrics" (2020), Maibam, D., Navanath, S., tendo como objetivo realizar a extração de classes de sentimentos a partir de letras usando técnicas de modelagem de tópicos como Latent Dirichlet Allocation (LDA) e Heuristic Dirichlet Process (HDP).

O artigo de Röder, Both e Hinneburg (2015), "Exploring the Space of Topic Coherence Measures", por sua vez, propõe uma framework para criar e combinar medidas de coerência de tópicos baseadas em palavras, melhorando a avaliação da interpretabilidade dos resultados de modelos de tópicos. A pesquisa mostra que novas combinações de medidas superam as anteriores em correlação com avaliações humanas, oferecendo avanços significativos para a mineração de texto e recuperação de informações.

3. Metodologia e resultados

Esta seção apresenta a metodologia utilizada para encontrar padrões frequentes e excepcionais em músicas virais e de sucesso. Partimos de um conjunto de dados contendo músicas de sucesso e virais em mercados globais e regionais e a letra dessas músicas (3.1). Em seguida, pré-processamos as letras, removendo caracteres especiais, stopwords e contrações, além de realizar lematização e outros tratamentos (Seção 3.2). Aplicamos o Latent Dirichlet Allocation (LDA) e o BERTopic para modelagem de tópicos, identificando temas prevalentes nas letras das músicas (Seção 3.3). Por fim, são apresentados os resultados obtidos, com visualizações dos tópicos identificados e análise desses. (Seção 3.4).

3.1. Dados

Neste estudo, empregamos o Music Genre Dataset (MGD) [Oliveira et al. 2020] para obter dados de sucesso no Spotify de 2017 a 2021. Os dados do Spotify incluem informações diárias das 200 músicas mais reproduzidas em cada país e território, além de um gráfico global consolidado. Este conjunto de dados abrange os gráficos globais e oito dos dez principais mercados musicais conforme a IFPI em 2019: Estados Unidos (#1), Japão (#2), Reino Unido (#3), Alemanha (#4), França (#5), Canadá (#8), Austrália (#9) e Brasil (#10). O MGD fornece as músicas que ingressaram nos gráficos diários para cada mercado e ano, juntamente com características acústicas que descrevem essas músicas (foram consideradas para execução apenas uma aparição de cada música). Adicionalmente, são utilizados as letras extraídas do Genius correspondentes a essas músicas, sendo Genius uma plataforma online que fornece letras de músicas, anotações e informações detalhadas sobre canções e artistas.

3.1. Pré-processamento dos dados

Durante a etapa de pré-processamento dos dados, é estruturado o conjunto de dados a ser utilizado a partir dos dados de músicas e seus rankings no Spotify juntamente com as letras oriundas do Genius. Tratando os dados, processos como conversão das letras para minúsculas, remoção de pontuação e caracteres especiais e palavras e frases entre colchetes, marcadores de seção, são realizados. As stopwords em inglês, espanhol e alemão são excluídas para reduzir ruído, além da remoção da menção dos nomes dos artistas. Contrações comuns são expandidas para suas formas completas para manter a consistência semântica. Por fim, aplica-se a lematização para reduzir as palavras às suas formas base, ou seja, radicais, garantindo que palavras com diferentes terminações sejam tratadas de forma unificada. Este conjunto de passos assegura que o texto final seja limpo e padronizado, pronto para análise de tópicos usando LDA e BERTopic.

3.2. Algoritmos para Modelagem de Tópicos

A modelagem de tópicos visa identificar os principais temas presentes em grandes volumes de texto. Para analisar as letras de músicas e descobrir os tópicos mais relevantes, foram utilizados dois algoritmos principais: Latent Dirichlet Allocation

(LDA) e BERTopic. Abaixo detalha-se a construção, as decisões tomadas e os parâmetros utilizados para cada um desses algoritmos.

3.2.1 Latent Dirichlet Allocation (LDA)

O LDA é um modelo de geração de tópicos que supõe que cada documento é uma mistura de um número finito de tópicos e que cada palavra no documento é atribuída a um desses tópicos.

A implementação do LDA foi feita utilizando a biblioteca scikit-learn em Python, onde o número de tópicos foi definido como 10, um equilíbrio entre granularidade e generalização dos temas, e a randomização foi controlada para garantir reprodutibilidade. Foi utilizado a vetorização TF-IDF com *max_df* de 0.95 e *min_df* de 2 para remover termos muito frequentes e muito raros, respectivamente. Após a vetorização, o LDA foi ajustado e os tópicos foram visualizados utilizando pyLDAvis.

3.2.1 BERTopic

O BERTopic é um algoritmo de modelagem de tópicos baseado em embeddings de linguagem (BERT) que agrupa documentos em tópicos utilizando técnicas de aprendizado não supervisionado. Ele combina a representação semântica de BERT com técnicas de clustering e redução de dimensionalidade.

Utilizou-se um modelo pré-treinado BERT para gerar embeddings de texto, que foram reduzidos em dimensionalidade usando UMAP e, em seguida, agrupados com HDBSCAN. A visualização dos tópicos gerados foi a partir da funcionalidade de visualização do BERTopic.

3.3. Resultados

A análise dos tópicos e sentimentos ao longo dos anos revela uma evolução constante na temática e na diversidade linguística da música popular. De 2017 a 2021, a popularidade de temas românticos em espanhol inicialmente predominou, refletindo um sentimento positivo notável em 2017.

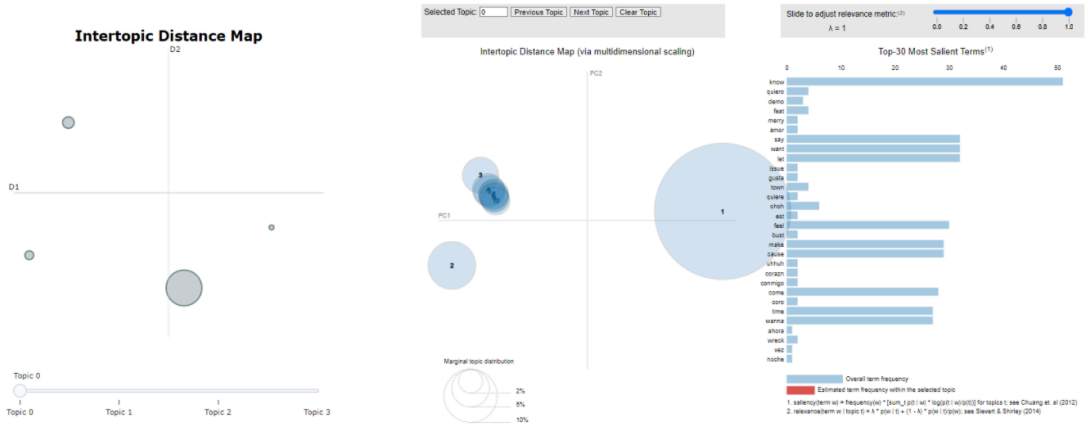

Figura 1. Mapa de distância inter-tópicos pelo BERTopic e LDA com dados de 2017, respectivamente.

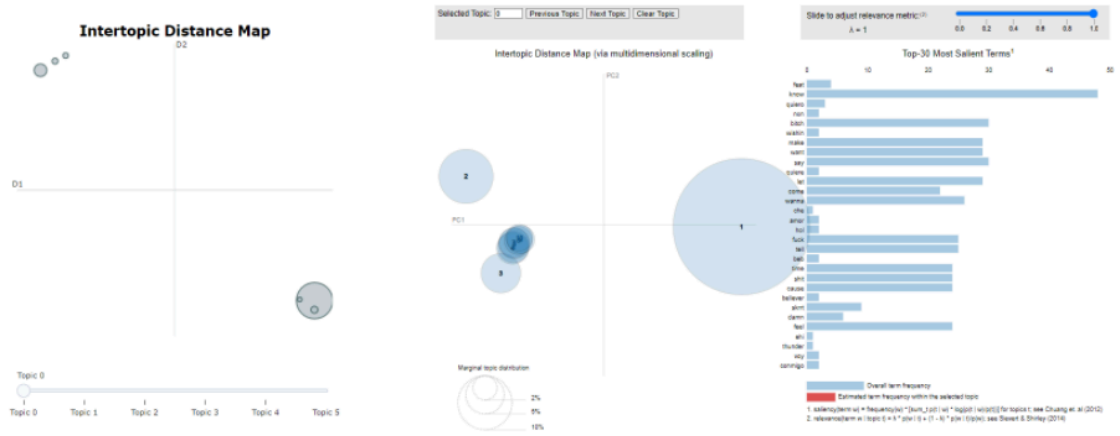

Figura 2. Mapa de distância inter-tópicos pelo BERTopic e LDA com dados de 2018, respectivamente.

Figura 3. Mapa de distância inter-tópicos pelo BERTopic e LDA com dados de 2019, respectivamente.

Figura 4. Mapa de distância inter-tópicos pelo BERTopic e LDA com dados de 2020, respectivamente.

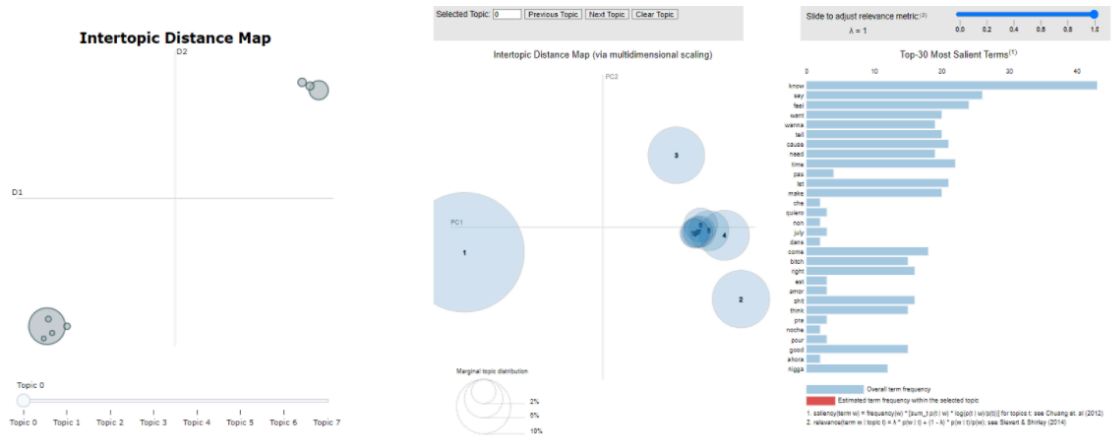

Figura 5. Mapa de distância inter-tópicos pelo BERTopic e LDA com dados de 2021, respectivamente.

Contudo, a partir de 2018, a diversidade linguística começou a se expandir, com a inclusão de idiomas como francês, italiano e alemão, acompanhada por uma leve queda no sentimento médio. Essa tendência de diversificação continuou nos anos seguintes, com a inclusão de temas em turco e uma crescente influência de músicas em diferentes idiomas, como o alemão e o português, em 2021. O sentimento geral tornou-se neutro a partir de 2019, indicando uma mudança em relação ao otimismo inicial. A constância dos temas natalinos ao longo dos anos sugere a continuidade de tradições festivas na música, como pode ser verificado nos gráficos de frequência na figura 6.

Comparando os modelos de análise BERTopic e LDA, BERTopic tende a fornecer uma visão mais focada e específica dos tópicos, enquanto LDA captura uma gama mais ampla de temas dentro de cada tópico. Por exemplo, em 2017, BERTopic identifica um tópico claramente relacionado ao Natal com termos como "christmas", "santa", "merry" e "year", enquanto LDA distribui essas palavras em tópicos variados. Em geral, BERTopic oferece uma análise mais detalhada e específica, enquanto LDA proporciona uma visão mais ampla dos temas abordados, sendo cada modelo útil para diferentes objetivos analíticos.

4. Conclusões e relação de trabalhos futuros

Em síntese, este projeto forneceu insights significativos sobre os temas presentes nas letras de músicas virais e de sucesso, destacando a evolução e diversificação temática ao longo dos anos. A aplicação de técnicas de Processamento de Linguagem Natural, como Latent Dirichlet Allocation (LDA) e BERTopic, permitiu a identificação de tópicos recorrentes e a análise de sentimentos nas letras, revelando padrões importantes que ajudam a compreender o sucesso musical em diferentes mercados.

Os resultados destacaram a predominância inicial de temas românticos em espanhol, seguida por uma diversificação linguística e temática a partir de 2018, com a inclusão de novos idiomas e uma mudança para um sentimento mais neutro. A persistência dos temas natalinos ao longo dos anos reflete a continuidade das tradições festivas na música popular. Além disso, a comparação entre os modelos LDA e BERTopic

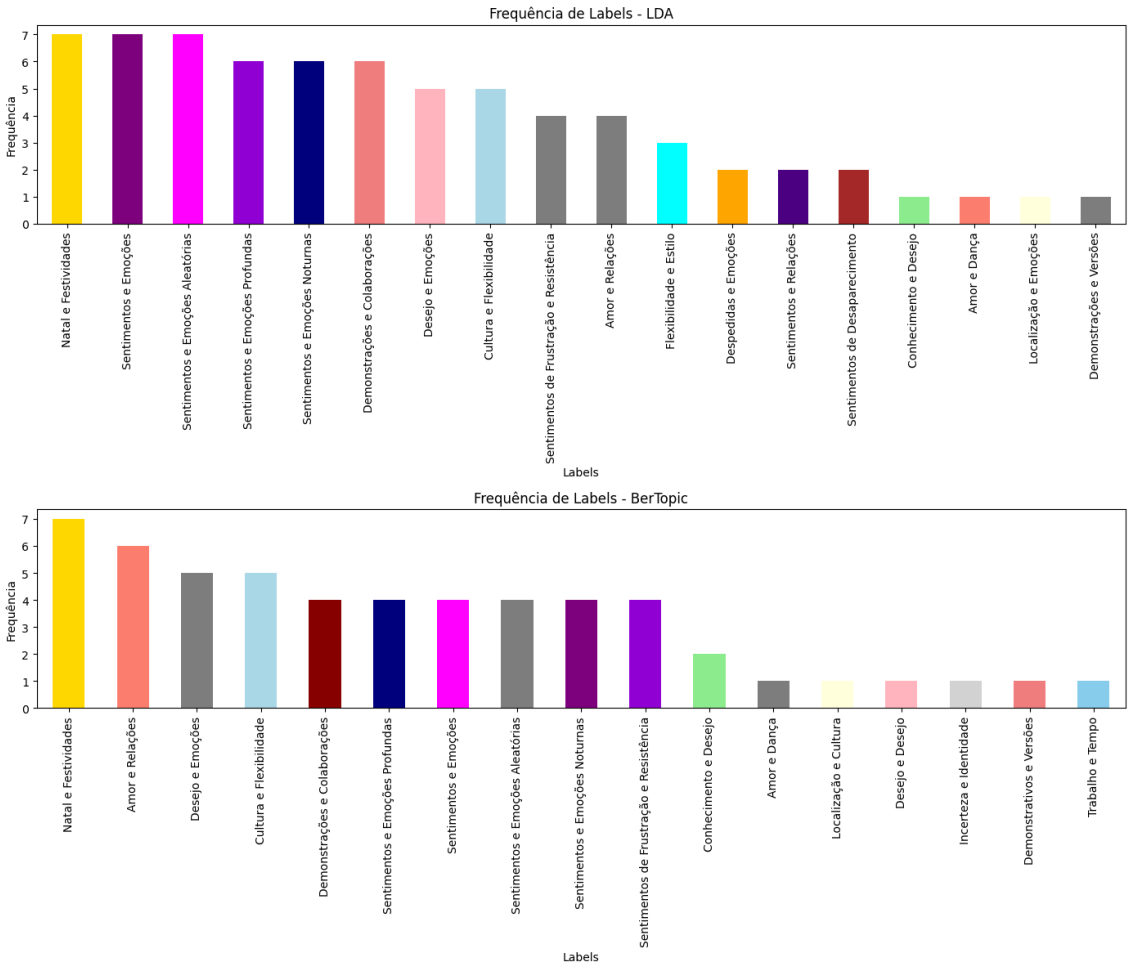

Figura 6. Gráficos de frequências de etiquetas presentes nas letras musicais dentre os anos 2017 a 2021 para BERTopic e LDA.

evidenciou que, enquanto BERTopic oferece uma análise mais específica, LDA proporciona uma visão mais abrangente dos temas, ambos sendo valiosos para diferentes objetivos de análise.

Para trabalhos futuros, recomenda-se uma análise mais aprofundada das variações observadas nos padrões temáticos e sentimentais, levando em consideração fatores culturais, sociais e econômicos. Além disso, explorar a evolução desses padrões ao longo do tempo e em diferentes regiões pode fornecer insights valiosos para estratégias de lançamento e promoção musical. Adicionalmente, a inclusão de novos dados e a aplicação de técnicas de análise mais avançadas podem enriquecer ainda mais a compreensão das dinâmicas da indústria musical global. Este trabalho estabelece uma base sólida para futuras investigações que busquem aprofundar o entendimento das complexas relações entre letras de músicas e seu impacto no sucesso na indústria musical.

Referências

Iloga, S. et al. (2018) “A sequential pattern mining approach to design taxonomies for hierarchical music genre recognition”, Pattern Anal. Appl. 21 (2): 363–380.

Oliveira, G. P. et al. (2020) “Detecting collaboration profiles in success-based music genre networks”, In ISMIR. pp. 726–732.

Oliveira, G. P. et al. (2023) “Mining Exceptional Genre Patterns on Hit Songs”.

Ordanini, A. et al. (2018) “The featuring phenomenon in music: how combining artists of different genres increases a song’s popularity”, Market. Letters vol. 29, pp. 485–499.

Ren, J. and Kauffman, R. J. (2019) “Understanding music track popularity in a social network”, In ECIS. AIS, Atlanta, GA, USA, pp. 374–388.

Rompré, L. et al. (2017) “Using association rules mining for retrieving genre-specific music files. In FLAIRS Conference”, AAAI Press, pp. 706–711.

Shin, S. and Park, J. (2018) “On-chart success dynamics of popular songs”, Adv. in Comp. Systems 21 (3-4): 1850008.

Silva, M. O. et al. (2014) “Collaboration as a driving factor for hit song classification”, In WebMedia. ACM, pp. 66–74, 2022. Zaki, M. J. and Meira Jr., W. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

Calcina, E.and Novak, E. (2022) “Measuring the Similarity of Song Artists using Topic Modelling”.

Maibam, D. and Navanath, S. (2020) “Exploiting Topic Modelling to Classify Sentiment from Lyrics”, International Conference on Machine Learning, Image Processing, Network Security and Data Sciences.

Röder, Both and Hinneburg (2015) “Exploring the Space of Topic Coherence Measures”, WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.

Sievert, C., and Shirley, K. (2014) “LDAvis: A Method for Visualizing and Interpreting Topics”.