UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOSÉ MATHEUS DO NASCIMENTO GAMA

ANÁLISE DE EMOÇÕES E TEMAS PRESENTES NA MÚSICA POPULAR BRASILEIRA COM PROCESSAMENTO DE LINGUAGEM NATURAL

CAMPINA GRANDE - PB

2024

JOSÉ MATHEUS DO NASCIMENTO GAMA

ANÁLISE DE EMOÇÕES E TEMAS PRESENTES NA MÚSICA POPULAR BRASILEIRA COM PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: MAXWELL GUIMARÃES DE OLIVEIRA

CAMPINA GRANDE - PB

2024

JOSÉ MATHEUS DO NASCIMENTO GAMA

ANÁLISE DE EMOÇÕES E TEMAS PRESENTES NA MÚSICA POPULAR BRASILEIRA COM PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Maxwell Guimarães de Oliveira

Orientador – UASC/CEEI/UFCG

Dalton Dario Serey Guerrero

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 15 de maio de 2024.

CAMPINA GRANDE - PB

RESUMO

Este trabalho consistiu na análise de letras da música popular brasileira ao longo das décadas, utilizando técnicas de Processamento de Linguagem Natural para identificar temas predominantes e emoções subjacentes. A importância desse tema residia na compreensão das transformações culturais e emocionais na sociedade brasileira ao longo do tempo, uma vez que a música reflete as emoções, preocupações e experiências da sociedade. Para produzir o trabalho, foram coletadas e pré-processadas 12.542 músicas distribuídas entre 1.318 artistas, abrangendo as décadas de 1960 a 2020. As principais tecnologias utilizadas incluíram o algoritmo de Alocação Latente de Dirichlet para Modelagem de Tópicos e o modelo BERTimbau para classificação de emoções. Os resultados revelaram temas recorrentes, como o amor, e padrões emocionais ao longo do tempo, fornecendo insights valiosos sobre as mudanças culturais e emocionais na sociedade brasileira. Em suma, o estudo destacou a constância do amor como tema central nas músicas, além das mudanças emocionais refletidas nas letras, contribuindo para uma compreensão mais profunda da mudança cultural no Brasil.

ANALYSIS OF EMOTIONS AND THEMES PRESENT IN BRAZILIAN POPULAR MUSIC WITH NATURAL LANGUAGE PROCESSING

ABSTRACT

This work consisted of analyzing lyrics from Brazilian popular music over the decades, using Natural Language Processing techniques to identify predominant themes and underlying emotions. The importance of this topic lay in understanding the cultural and emotional transformations in Brazilian society over time, as music reflects the emotions, concerns, and experiences of society. To produce the work, 12,542 songs distributed among 1,318 artists were collected and pre-processed, spanning the decades from 1960 to 2020. The main technologies used included the Latent Dirichlet Allocation algorithm for Topic Modeling and the BERTimbau model for emotion classification. The results revealed recurring themes, such as love, and emotional patterns over time, providing valuable insights into cultural and emotional changes in Brazilian society. In summary, the study highlighted the constancy of love as a central theme in the songs, as well as the emotional changes reflected in the lyrics, contributing to a deeper understanding of cultural change in Brazil.

Análise de emoções e temas presentes na Música Popular Brasileira com Processamento de Linguagem Natural

José Matheus do Nascimento Gama
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
jose.gama@ccc.ufcg.edu.br

Maxwell Guimarães de Oliveira
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
maxwell@computacao.ufcg.edu.br

RESUMO

Este trabalho consistiu na análise de letras da música popular brasileira ao longo das décadas, utilizando técnicas de Processamento de Linguagem Natural para identificar temas predominantes e emoções subjacentes. A importância desse tema residia na compreensão das transformações culturais e emocionais na sociedade brasileira ao longo do tempo, uma vez que a música reflete as emoções, preocupações e experiências da sociedade. Para produzir o trabalho, foram coletadas e pré-processadas 12.542 músicas distribuídas entre 1.318 artistas, abrangendo as décadas de 1960 a 2020. As principais ferramentas utilizadas incluíram o algoritmo de Alocação Latente de Dirichlet para Modelagem de Tópicos e o modelo BERTimbau para classificação de emoções. Os resultados revelaram temas recorrentes, como o amor, e padrões emocionais ao longo do tempo, fornecendo insights valiosos sobre as mudanças culturais e emocionais na sociedade brasileira. Em suma, o estudo destacou a constância do amor como tema central nas músicas, além das mudanças emocionais refletidas nas letras, contribuindo para uma compreensão mais profunda da mudança cultural no Brasil.

PALAVRAS-CHAVE

Música popular brasileira. Processamento de linguagem natural. Modelagem de tópicos. Classificação de emoções.

1. INTRODUÇÃO

A falta de compreensão e valorização da identidade cultural de um país pode resultar na desconexão e perda de identidade para os indivíduos que ali vivem, bem como na desvalorização de suas tradições culturais. Além disso, essa falta de reconhecimento pode ter impactos sociais mais amplos, como dificuldades de integração social e participação cívica informada. Por outro lado, é importante destacar que a música possui um papel fundamental na transmissão e preservação da identidade cultural de um país. Através da música, é possível explorar e compreender as diferentes facetas de uma cultura, desde suas raízes históricas até suas expressões contemporâneas.

A Música Popular Brasileira (MPB), sob essa denominação, ganha destaque no Brasil a partir dos anos 60, quando músicos brasileiros passaram a abordar de forma mais enfática as questões sociais, políticas e culturais da época em suas composições. Embora tais temáticas já estivessem presentes na Bossa Nova, a MPB se destacou por ter um maior compromisso em interpretar o mundo corrente (GALVÃO, 1976), especialmente os aspectos cotidianos dos cidadãos brasileiros. Desde questões como desigualdade, pobreza e injustiça, até temas mais universais, como amor, saudade e identidade cultural, a MPB oferece uma narrativa rica e complexa que ressoa com pessoas de todas as origens.

É relevante observar que, ao longo do tempo, os problemas socioculturais no Brasil evoluíram, gerando novos desafios, enquanto muitos dilemas antigos se aprofundaram. Nessa perspectiva, como enfatizado por NEDER (2010), não se pode limitar o conceito de música popular ao que foi definido e produzido na década de origem da MPB, pois isso reduziria fortemente a sua compreensão. Isto é, a música brasileira, por meio de suas letras, tem sido e continuará sendo um reflexo da configuração sociocultural de cada época, seja como uma ferramenta de denúncia ou como expressão dos costumes, valores e sentimentos do povo.

Nesse contexto, uma análise lírica da MPB oferece a oportunidade de identificar os temas explorados nesse estilo musical, desde seu surgimento até os dias atuais, e também permite compreender as diversas emoções expressas pelos músicos brasileiros em suas composições. Ao longo dos anos, a forma como as músicas são distribuídas para o público passou por uma transformação significativa, especialmente com o advento dos serviços de streaming. Estes proporcionam acesso a um vasto catálogo de músicas de muitos gêneros e épocas. A crescente quantidade de músicas disponíveis nestas plataformas é notável, com milhões de faixas acessíveis instantaneamente para os ouvintes.

No contexto da MPB, essa abundância de conteúdo oferece uma rica fonte de dados para entendermos as mudanças nas temáticas exploradas pelos artistas ao longo das décadas. No entanto, devido à vasta quantidade de músicas brasileiras disponíveis, torna-se humanamente inviável realizar uma análise minuciosa, letra por letra, dessas canções, bem como reunir todas as informações pertinentes a elas.

Diante desse problema, este trabalho se propôs a coletar da Web, de maneira automática, uma extensa base de dados contendo informações sobre músicas populares brasileiras. Utilizando técnicas avançadas de Processamento de Linguagem Natural (PLN), como Modelagem de Tópicos e Classificação de Emoções, o objetivo é identificar os temas e emoções presentes nas letras dessas músicas, década por década. Essa análise desempenha uma função social crucial, representando uma tentativa de resgatar parte da identidade cultural do Brasil que foi expressa por meio das composições musicais. O código-fonte,

documentação e artefatos utilizados neste estudo estão disponíveis em um repositório[1] público do GitHub.

2. FUNDAMENTAÇÃO TEÓRICA

A corrente seção apresenta PLN como uma ferramenta essencial para analisar as letras da música popular brasileira. Destaca-se a Modelagem de Tópicos, especialmente a Alocação Latente de Dirichlet (LDA), para identificar temas e a métrica de coerência para avaliar a qualidade dos resultados. Também se discute a relevância da classificação de emoções, onde se utilizou a arquitetura BERT, especialmente o modelo BERTimbau, adaptado para textos em português.

2.1 Processamento de Linguagem Natural

Diante da vasta quantidade de músicas criadas por artistas brasileiros, a busca por temas e emoções nelas sem o auxílio de ferramentas eficazes e precisas é praticamente inviável para os humanos.

No campo da ciência da computação e linguística, o Processamento de Linguagem Natural (PLN) é amplamente definido como "um subcampo da inteligência artificial e da linguística computacional que lida com a interação entre computadores e humanos através da linguagem natural" (JURAFSKY & MARTIN, 2008). Em essência, o PLN capacita os computadores a compreender, interpretar e gerar linguagem humana de maneira semelhante aos seres humanos.

Nesse sentido, o PLN se torna uma ferramenta indispensável para desvendar tópicos latentes e sentimentos presentes nas letras da música popular brasileira, possibilitando uma análise abrangente, precisa e eficiente.

2.2 Modelagem de Tópicos

A Modelagem de Tópicos é uma abordagem poderosa dentro do campo de PLN, empregada para extrair os temas predominantes em grandes volumes de texto, como as letras de músicas.

Essa técnica considera cada documento textual como uma composição de diferentes tópicos, sendo que cada tópico é, por sua vez, uma mistura de palavras. Ao analisar um conjunto de documentos textuais, a Modelagem de Tópicos visa identificar os tópicos subjacentes e as palavras associadas a cada um deles, revelando padrões e estruturas semânticas intrínsecas.

2.2.1 Alocação Latente de Dirichlet

A Alocação Latente de Dirichlet (LDA) é um modelo probabilístico utilizado em PLN para identificar tópicos latentes em um conjunto de documentos textuais.

Proposto por BLEI et al. (2003), o modelo LDA parte de dois pressupostos fundamentais: primeiro, que cada documento em uma coleção é composto por uma mistura de diversos tópicos; segundo, que cada palavra em um documento é atribuída a um dos tópicos presentes no documento. Durante o treinamento, o LDA ajusta as distribuições de tópicos para maximizar a probabilidade de observar os documentos fornecidos.

Para cada documento, o LDA gera uma distribuição de tópicos e, para cada palavra em um documento, atribui um tópico com base na distribuição de tópicos do documento e na distribuição de palavras do tópico. Durante o processo de geração de documentos, o LDA segue um procedimento aleatório, escolhendo aleatoriamente uma distribuição de tópicos para o documento e, em seguida, para cada palavra, escolhendo aleatoriamente um tópico com base na distribuição de tópicos do documento e uma palavra com base na distribuição de palavras do tópico escolhido. Um tópico nada mais é do que uma coleção de palavras-chave dominantes que tipicamente representam algum tema.

O LDA possui parâmetros importantes, como o número de tópicos a serem extraídos e a distribuição de Dirichlet que modela a distribuição de palavras em cada tópico. Uma vez treinado, o LDA pode inferir os tópicos subjacentes em novos documentos textuais, proporcionando informações valiosas acerca dos temas presentes.

2.2.2 Métrica de Coerência

A métrica de coerência é uma ferramenta fundamental no contexto da modelagem de tópicos, especialmente quando se trata de avaliar a qualidade dos tópicos identificados pelo algoritmo do modelo LDA. A métrica de coerência visa quantificar o grau de interpretabilidade e consistência dos tópicos extraídos, oferecendo uma forma de avaliar a relevância e a coesão semântica das palavras associadas a cada tópico.

Dentro das abordagens para calcular a coerência de tópicos, destaca-se, conforme explanado por SYED & SPRUIT (2017), a métrica Cv. Esta métrica é composta por quatro etapas distintas: a segmentação dos dados em pares de palavras, o cálculo das probabilidades associadas a palavras ou pares de palavras, a determinação de uma medida de confirmação que avalia a robustez da relação entre conjuntos de palavras e, por fim, a agregação das medidas individuais de confirmação para gerar uma pontuação de coerência global.

Ao aplicar a métrica de coerência aos resultados da modelagem de tópicos, é possível identificar o número ideal de tópicos a serem extraídos, bem como ajustar os parâmetros do modelo para otimizar a interpretabilidade e a coesão dos tópicos identificados. Em última análise, a métrica de coerência desempenha um papel importante na validação e refinamento dos modelos de tópicos, contribuindo para uma análise mais precisa e significativa de dados textuais, inclusive de letras de músicas.

2.3 Classificação de emoções

No contexto da composição musical, as emoções desempenham um papel fundamental no processo criativo dos compositores. Ao criar música, os compositores frequentemente canalizam suas próprias experiências emocionais, pensamentos e sentimentos para informar a direção e o conteúdo de suas composições. As emoções dos compositores podem servir como uma fonte de inspiração e motivação, levando-os a expressar uma ampla gama de sentimentos, como amor, tristeza, alegria, raiva e esperança, através da música. Além disso, as escolhas musicais feitas durante

[1] Repositório github: https://github.com/mixmaxze/tcc_mpb

o processo de composição, como harmonia, melodia, ritmo e dinâmica, são frequentemente influenciadas pelas emoções que o compositor deseja evocar em sua música. Portanto, as emoções dos compositores desempenham um papel central na criação de obras musicais que ressoam com os ouvintes, transmitindo significado e profundidade emocional.

A classificação de emoções em PLN é uma tarefa que envolve identificar e categorizar as emoções expressas em textos escritos. É frequentemente realizada utilizando técnicas de aprendizado de máquina, onde modelos computacionais são treinados em conjuntos de dados rotulados com exemplos de textos e as emoções associadas a eles. Uma vez treinado, o modelo pode classificar automaticamente as emoções em novos textos com base em padrões identificados durante o treinamento.

2.4 BERT

O BERT (Bidirectional Encoder Representations from Transformers) se baseia na arquitetura Transformer e no conceito de pré-treinamento bidirecional em PLN. Desenvolvido por VASWANI et al. (2017), a arquitetura Transformer permite o processamento eficiente de sequências de texto, utilizando camadas de autoatenção para capturar o contexto de palavras em ambas as direções da sequência.

O BERT inova ao introduzir o pré-treinamento bidirecional, treinando o modelo para prever palavras mascaradas em uma sentença de maneira bidirecional, o que possibilita a captura de relações mais complexas entre as palavras. Essa abordagem, descrita detalhadamente no artigo "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" por DEVLIN et al. (2018), tem levado a resultados de estado-da-arte em diversas tarefas de PLN, tornando o BERT uma ferramenta fundamental na análise e compreensão de texto.

2.4.1 BERTimbau para a Classificação de Emoções

O BERTimbau, desenvolvido por SOUZA et. al (2020), é uma adaptação do modelo BERT para a língua portuguesa brasileira. Essa adaptação envolveu o pré-treinamento do BERTimbau em uma vasta quantidade de dados em português, permitindo que o modelo aprendesse representações contextualizadas das palavras específicas do idioma. Com essa capacidade, o BERTimbau demonstrou um desempenho avançado em diversas tarefas de processamento de linguagem natural, como classificação de texto, tradução automática e análise de sentimentos, tornando-se uma ferramenta valiosa para aplicações relacionadas ao processamento de linguagem em português.

Com base nisso, o estudo realizado por HAMMES & FREITAS (2021) explorou o potencial dos modelos BERTimbau para a tarefa de classificação de emoções em textos na língua portuguesa. Os pesquisadores ajustaram os modelos BERTimbau-base e BERTimbau-large usando o conjunto de dados GoEmotions, que foi traduzido para o português. Eles compararam os resultados obtidos com os originais em inglês e identificaram uma melhoria no desempenho, atribuída a um método de balanceamento aplicado aos dados para lidar com o desbalanceamento das classes de emoções.

Ao realizar o ajuste fino dos modelos BERTimbau, os autores utilizaram métricas como Precisão, Sensibilidade e Medida-F para avaliar o desempenho. Eles também estabeleceram um limiar de confiança para as previsões do modelo, selecionando apenas aquelas com escores iguais ou superiores a 0,3. Essa abordagem permitiu uma análise mais precisa das emoções presentes nos textos em português, contribuindo para avanços na compreensão da análise de emoções em língua portuguesa.

Como resultado desse estudo, foi produzido um modelo capaz de classificar documentos textuais em 27 emoções distintas: amor, alegria, desejo, otimismo, entusiasmo, aprovação, realização, tristeza, decepção, zelo, aborrecimento, confusão, admiração, surpresa, remorso, luto, diversão, alívio, raiva, constrangimento, curiosidade, medo, nervosismo, desaprovação, gratidão, nojo e orgulho, além da emoção neutra.

3. METODOLOGIA

A metodologia adotada neste estudo emprega uma abordagem baseada em PLN para analisar as letras da música popular brasileira, visando identificar temas predominantes e emoções subjacentes. Inicialmente, foi realizada uma coleta extensiva de dados, compilando um corpus representativo de letras de músicas de artistas brasileiros. Em seguida, a técnica de Modelagem de Tópicos foi aplicada para extrair os tópicos latentes presentes nas letras, utilizando o LDA como modelo probabilístico. Posteriormente, foi realizada uma avaliação da qualidade dos tópicos identificados utilizando métricas de coerência. Por fim, foi realizada, com o BERTimbau, a classificação de emoções para compreender as nuances emocionais presentes nas letras, contribuindo para uma análise mais abrangente do conteúdo das músicas. Os detalhes de cada uma dessas etapas estão apresentados nas subseções a seguir.

3.1 Coleta de dados

Para se construir uma base com dados de músicas da MPB, foi criado um raspador de dados, feito com o framework Selenium[2] na linguagem Python. Como local para extração desses dados, foi escolhida uma página que filtra músicas da MPB no Genius[3], site que contém uma enorme coleção de informações musicais públicas. O motivo dessa escolha é que algumas APIs com propósito semelhante, como do LastFM e do Spotify, possuem limites de requisições diárias, dificultando a extração das informações necessárias para a análise.

A partir dessa página, que continha 1.000 músicas de variados artistas, foram extraídos delas os links dos perfis de cada artista principal e participante. Para cada uma das músicas encontradas nesses perfis, foram extraídas as seguintes informações: título, nome do artista, letra, nome do álbum e data de lançamento.

3.2 Pré-processamento dos dados

Inicialmente, foi preciso criar uma coluna derivada para armazenar o ano de lançamento da música, informação presente no nome do álbum ou na data de lançamento, para então criar uma

[2] Selenium: https://www.selenium.dev/pt-br/documentation/
[3] Genius: https://genius.com/tags/mpb/all?page=1

coluna com a década de lançamento. As músicas que não possuíam essa informação foram removidas da base de dados. Também foram retiradas músicas da década de 1950, pois nela havia apenas 7 artistas, o que tornaria a análise enviesada para essa década.

Após essa etapa inicial, foi realizada a remoção de músicas duplicadas, priorizando a ocorrência mais antiga, e a remoção daquelas que continham letras vazias, especialmente as que consistiam apenas de música instrumental. Em seguida, as letras originais foram convertidas para minúsculas, removendo-se partes entre colchetes que indicavam inícios de versos e refrões. Além disso, as quebras de linha foram substituídas por espaços em branco. As letras processadas até esse ponto foram posteriormente utilizadas na análise de emoções.

Adicionalmente, para o uso na modelagem de tópicos, foi criada uma coluna para armazenar as letras musicais resultantes da etapa anterior, porém sem pontuações, caracteres isolados e stopwords. Stopwords são palavras que não contribuem com significado relevante para o texto, por ocorrerem com muita frequência em qualquer texto, como artigos, preposições e pronomes.

Por último, para filtrar as músicas em português, foram utilizadas, para decidirem em unanimidade, duas bibliotecas Python de classificação de idiomas: lingua[4] e langdetect[5]. A escolha de utilizar duas bibliotecas simultaneamente se justifica pelo fato de que, quando usadas de forma isolada, ambas tendem a confundir o idioma português com o idioma espanhol em alguns casos.

Após o término do pré-processamento, a base de dados resultante consistiu em um total de 12.542 músicas de 1.318 artistas distintos, abrangendo sete décadas, de 1960 a 2020 (até janeiro de 2024). É possível visualizar a distribuição de músicas por década na Figura 1.

Já na Figura 2, pode-se visualizar a distribuição de artistas por década. É importante destacar que um mesmo artista pode aparecer em décadas diferentes.

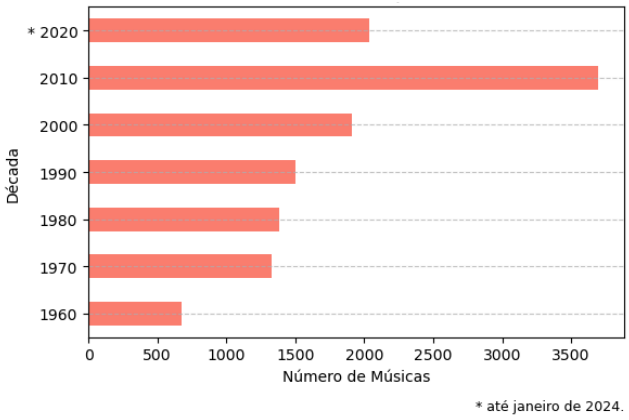

Figura 1: Gráfico de barras horizontais com número de músicas por década

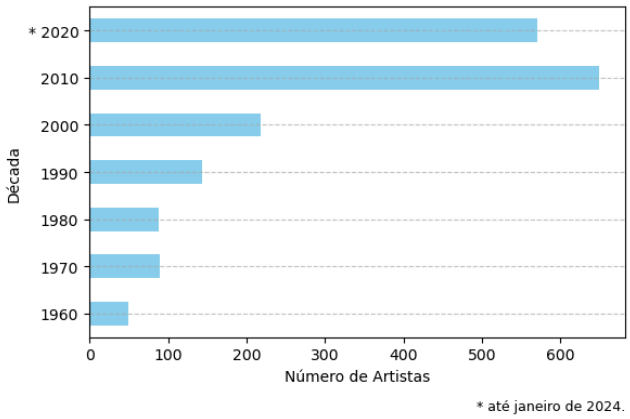

Figura 2: Gráfico de barras horizontais com número de artistas por década

3.3 Descoberta de tópicos

Para a descoberta de tópicos com o modelo LDA e o cálculo da métrica de coerência de cada tópico descoberto, utilizou-se a biblioteca Gensim[6] para a linguagem Python. Essa biblioteca é uma ferramenta poderosa dentro da área de PLN e oferece diversos pacotes com funcionalidades úteis para análise de texto.

Um dos principais módulos do Gensim é o LdaModel[7], que foi utilizado para treinar modelos de tópicos utilizando o algoritmo LDA. Com o LdaModel, é possível identificar automaticamente os principais tópicos presentes em um conjunto de documentos.

Além disso, para avaliar a qualidade dos tópicos gerados pelo modelo LDA, foi empregado o CoherenceModel[8]. Este módulo é responsável por calcular a métrica de coerência Cv dos tópicos descobertos. Essa métrica indica quão semanticamente relacionadas são as palavras dentro de cada tópico. Quanto mais alta a coerência, mais interpretação e relevância os tópicos apresentam. Dessa forma, o CoherenceModel é fundamental para avaliar a consistência e a interpretabilidade dos tópicos identificados pelo modelo LDA, possibilitando ajustes nos parâmetros do modelo para a obtenção de resultados mais significativos.

Assim, após a etapa de pré-processamento, foram agrupadas as letras de músicas por década e, iterando sobre cada década, foram executadas sobre essas letras a seguinte sequência de passos:

1. Transformar as palavras das letras das músicas em tokens atômicos;
2. Construir bigramas com esses tokens, isto é, guardar todas as sequências de dois tokens, nesse caso distintos, que aparecem nas letras das músicas;

[4] lingua: https://pypi.org/project/lingua-language-detector/1.0.1/
[5] langdetect: https://pypi.org/project/langdetect/
[6] Gensim: https://pypi.org/project/gensim/
[7] LdaModel: https://radimrehurek.com/gensim/models/ldamodel.html
[8] CoherenceModel: https://radimrehurek.com/gensim/models/coherencemodel.html

3. Criar um dicionário com esses bigramas, que mapeia cada bigrama para um ID exclusivo;
4. Criar um corpus desses bigramas, que contém as contagens de palavras para cada documento do conjunto de músicas daquela década;
5. Construir modelos LDA com diferentes quantidades de tópicos, com um alcance de 1 até 10 tópicos, utilizando o dicionário e corpus criados anteriormente;
6. Calcular a mediana de coerências dos tópicos de cada modelo LDA gerado, levando em conta os 20 bigramas mais relevantes de cada tópico.

Ao final desse processo, para decidir uma quantidade em comum de tópicos a serem analisados para todas as décadas, foi criado o mapa de calor mostrado na Figura 3. Neste mapa, cada valor representa a mediana de coerências dos tópicos de um modelo LDA de uma década específica (eixo y) e com uma quantidade específica de tópicos (eixo x). Valores mais próximos de 1 indicam que tais tópicos são mais coerentes e interpretáveis, enquanto valores mais próximos a 0 indicam o contrário.

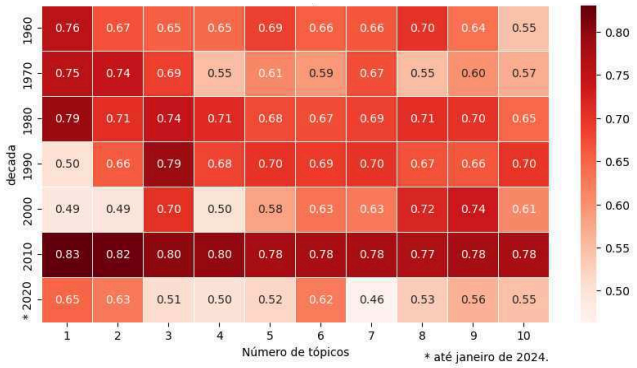

Figura 3: Mapa de calor com medianas das coerências dos tópicos em cada modelo LDA

Com base no mapa apresentado e auxiliado pela visualização na Figura 4, nota-se que o desempenho mediano em relação ao número de tópicos atinge seu ponto mais alto com 1 tópico. No entanto, para uma análise menos generalista, optou-se por selecionar o maior valor seguinte, que neste caso é a mediana da coerência dos modelos LDA com 3 tópicos.

Após o treinamento de todos os modelos LDA, eles foram aplicados para categorizar cada letra de música em um dos 3 tópicos correspondentes à década em que se encaixam. Para uma interpretação mais precisa dos tópicos, foram construídas e analisadas nuvens de palavras contendo os 20 bigramas mais relevantes de cada tópico, além das nuvens de palavras com os 20 bigramas mais frequentes de cada década. O tamanho de cada bigrama exibido nessas nuvens é proporcional à sua relevância no respectivo tópico, facilitando a visualização dos elementos mais importantes.

Além disso, também foram analisados exemplos de letras de músicas aleatórias relacionadas a cada tópico para aprofundar a compreensão dos subtemas presentes em cada um. Isso se tornou necessário, uma vez que, como ilustrado na Figura 3, algumas coerências entre os bigramas não são tão altas e a interpretabilidade não é facilmente obtida apenas observando os bigramas isoladamente. Dessa forma, tornou-se viável elaborar descrições para os tópicos, destacando as subdivisões presentes.

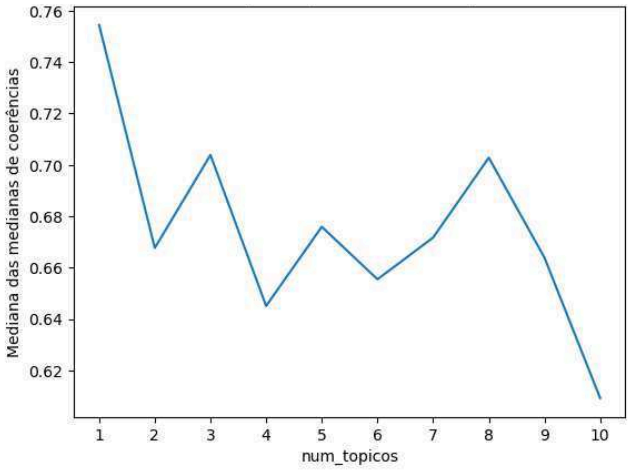

Figura 4: Gráfico de linha com a mediana de medianas de coerências por número de tópicos

3.4 Classificação de emoções

Para a classificação de emoções nas letras das músicas, utilizou-se a biblioteca Transformers[9] da linguagem Python, que permite o carregamento e a aplicação de modelos de classificação pré-treinados, incluindo modelos BERT.

Neste estudo, empregamos o BERTimbau para categorizar as emoções das músicas em 27 emoções distintas, incluindo a emoção neutra. O modelo calcula uma pontuação para cada letra de música de forma proporcional ao conjunto de emoções disponíveis, de modo que a soma das pontuações de todas as emoções resulta sempre em 100%.

Após a classificação das emoções em todas as letras das músicas, considerando que uma música pode evocar emoções diferentes ao longo dos versos, optou-se por remover a emoção neutra e selecionar, dentre as restantes, as 3 emoções com as pontuações mais altas de acordo com o modelo BERTimbau.

4. RESULTADOS

Após a análise dos tópicos de cada década, percebeu-se que o tema romântico é prevalente em grande parte dos tópicos de todas as décadas, apresentando-se com muitas facetas em diversas formas e contextos. Além disso, como será descrito ao decorrer desta seção, cada década se destaca por apresentar, além desse tema principal, seus próprios tópicos particulares.

Complementando esse ponto, conforme evidenciado na Figura 5, a análise das emoções revela um padrão consistente, com o amor mantendo-se como a emoção predominantemente expressa ao longo das décadas. Adicionalmente, é possível observar variações nas composições emocionais de cada período temporal. Por exemplo, uma redução na incidência da emoção de tristeza é perceptível ao longo do tempo, com uma tendência de crescimento em 2020.

[9] Transformers: https://pypi.org/project/transformers/

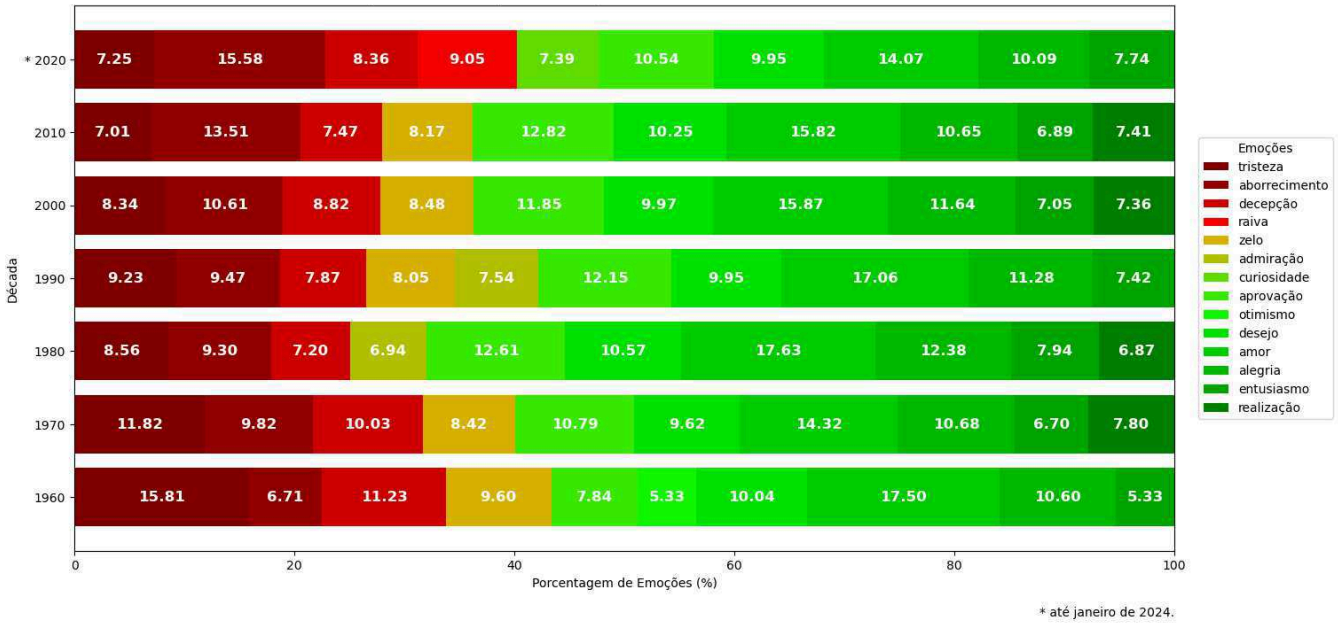

Figura 5: Gráfico de barras horizontais empilhadas com as porcentagens das 10 emoções mais frequentes nas letras de músicas de cada década

No entanto, destaca-se um aumento progressivo na representação da emoção de aborrecimento ao longo das décadas, indicando uma tendência de intensificação. Notavelmente, a emoção de "decepção" exibiu uma significativa relevância nas décadas de 60 e 70, declinando sua presença a partir dos anos 80. Entretanto, uma ressurgência dessa emoção é observada em 2020, particularmente relevante devido à limitada disponibilidade de dados para esta década. Essas nuances oferecem *insights* adicionais sobre as dinâmicas emocionais presentes nas letras de músicas ao longo do tempo, sugerindo possíveis correlações com contextos sociais e culturais específicos.

Na década de 60, conforme ilustrado na Tabela 1 e Figura 6, as músicas populares brasileiras abordam uma ampla gama de temas, desde as nuances dos relacionamentos até as mudanças culturais da época. No entanto, é essencial considerar o contexto político desse período. A censura e as pressões políticas exerciam uma influência marcante na produção artística, moldando de maneira sutil as mensagens transmitidas nas letras das músicas. Apesar dessas restrições, os artistas encontraram formas inventivas de expressar suas emoções e pontos de vista.

Tabela 1: Tópicos da década de 1960

Tópico	Frequência (%)
Amor, saudade, esperança, desilusão, sonho e memória	36,97
Relacionamentos, cultura, identidade, descontentamento	33,58
Rotina diária, nostalgia, mar, festas, amor distante	29,45

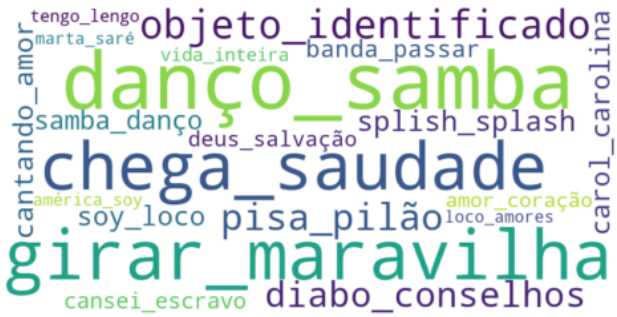

Figura 6: Nuvem com os bigramas mais presentes nas músicas da década de 1960

Assim, as músicas dos anos 60 não apenas refletem as preocupações e aspirações da sociedade, mas também a complexidade do ambiente político e cultural. Emoções como amor, tristeza, decepção e alegria são as que mais permeiam as composições, evocando tanto a melancolia quanto a esperança. Além disso, as letras frequentemente retratam a vida cotidiana com uma certa ingenuidade, refletindo os desafios e a simplicidade enfrentados pelos trabalhadores da época, ao mesmo tempo em que demonstram a habilidade dos artistas de se expressarem de forma criativa, mesmo em tempos de repressão política.

Já na década de 70, conforme evidenciado na Tabela 2 e Figura 7, as músicas refletem uma rica mistura de paixão, natureza, viagens e jornadas, muitas vezes incorporando elementos de introspecção e imaginação. Elas capturam a essência das experiências humanas, desde os relacionamentos românticos até as reflexões sobre a vida e a existência. Essas músicas também são permeadas por sentimentos profundos e imagens poéticas, que ajudam a transmitir uma gama de emoções e pensamentos. A

busca por liberdade é um tema recorrente, refletindo um período de transformação social e política no Brasil. Além disso, as referências culturais brasileiras são frequentes, destacando a riqueza da tradição artística e cultural do país. Essas músicas não apenas refletem as preocupações e aspirações do povo brasileiro, mas também celebram sua identidade e diversidade. O predomínio das emoções de amor, tristeza, aprovação e alegria se mescla com uma busca por liberdade e identidade cultural, refletindo as lutas e aspirações dos brasileiros em meio a um contexto de intensa transformação política e social.

Tabela 2: Tópicos da década de 1970

Tópico	Frequência (%)
Reflexões existenciais, referências culturais brasileiras, busca por liberdade	38,09
Sentimentos profundos, imagens poéticas, expressões artísticas	32,08
Paixão, natureza, viagens, jornada, introspecção, imaginação	29,83

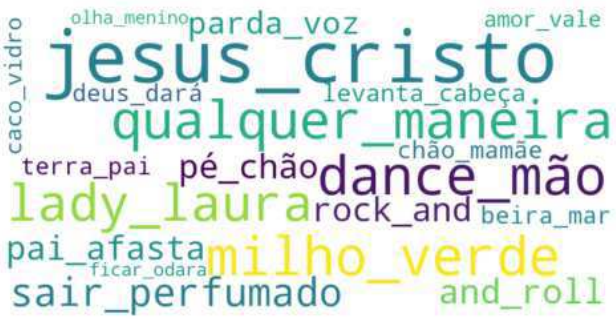

Figura 7: Nuvem com os bigramas mais presentes nas músicas da década de 1970

Em seguida, na década de 80, como destacado na Tabela 3 e Figura 8, as letras de músicas ecoam uma era mais otimista e exuberante, permeada pelas emoções frequentes de amor, aprovação, alegria e desejo. A liberdade recém-conquistada com o fim da ditadura alimentou a criatividade e a positividade nas composições musicais da segunda metade dessa década, contribuindo para um clima festivo e confiante. No âmbito do romance e autenticidade, essas músicas transmitem sentimentos de paixão e esperança ao explorar os relacionamentos humanos. As festividades são celebradas com canções que refletem a alegria e vivacidade da cultura brasileira. Além disso, há uma forte conexão com as tradições locais e regionalidades, destacando a riqueza cultural de diferentes regiões do Brasil. A natureza é uma constante fonte de inspiração, evocando paisagens exuberantes e a relação íntima do povo brasileiro com o ambiente natural. O tema do amor perdido traz uma dimensão de melancolia e reflexão, enquanto as músicas abordam as lutas diárias e os desafios enfrentados pela população, exaltando sua resiliência e força. As paisagens urbanas e rurais oferecem um panorama vívido da vida cotidiana, e as alusões brasileiras destacam a identidade cultural única do país e sua diversidade étnica e regional.

Tabela 3: Tópicos da década de 1980

Tópico	Frequência (%)
Romance, autenticidade, otimismo, festividades, reflexão	38,32
Tradições locais, regionalidades, conexão com natureza, amor perdido	32,18
Cotidiano, resistência, paisagens, relações humanas, alusões brasileiras	29,50

Figura 8: Nuvem com os bigramas mais presentes nas músicas da década de 1980

Prosseguindo, as letras de músicas dos anos 90, conforme destacado na Tabela 4 e Figura 9, revelam uma mistura de melancolia e autenticidade. O amor é retratado de maneira mais sincera e realista, com composições que exploram temas como companheirismo, conflitos internos e identidade pessoal. Há uma marcante valorização das raízes e tradições, evidenciada pela forte conexão com a cultura brasileira. As músicas celebram a diversidade cultural do país, destacando suas influências e expressões regionais. Além disso, as músicas frequentemente abordam questões sociais e políticas, com um caráter mais reflexivo sobre a vida coletiva. Aqui, as emoções predominantes são amor, aprovação, alegria e desejo, que refletem o contexto de transformação e busca por identidade vivenciado pela sociedade nessa década. Essas emoções não apenas refletem as experiências individuais, mas também capturam a essência coletiva de uma época marcada pela busca por autenticidade e pela luta por uma vida mais plena e significativa.

Com o avanço tecnológico e a globalização da cultura pop, as letras de músicas dos anos 2000 refletem um período marcado por uma riqueza emocional e uma profunda conexão com as questões culturais e existenciais do Brasil, como pode ser visto na Tabela 5 e Figura 10. A busca pelo amor e pela conexão humana continua sendo uma constante, mas acompanhada por uma reflexão sobre adversidades e conflitos internos. Os afetos e sentimentos intensos sugerem uma expressão artística que mergulha nas emoções humanas, possivelmente refletindo as mudanças sociais e políticas da época.

Tabela 4: Tópicos da década de 1990

Tópico	Frequência (%)
Relacionamentos, lembranças, espiritualidade, filosofia, criatividade, desafios	40,66
Paixão, diversidade, igualdade, crítica	29,97
Amor, tempo, busca por autenticidade, condição humana, sociedade, memória	29,37

Figura 9: Nuvem com os bigramas mais presentes nas músicas da década de 1990

Tabela 5: Tópicos da década de 2000

Tópico	Frequência (%)
Afetos, sentimentos intensos, brasilidades, questões existenciais, apreciação	36,68
Vínculos pessoais, autoconsciência, herança cultural, sensualidade	32,39
Relacionamentos românticos, autodescoberta, reflexão, celebração, resistência, luta	30,93

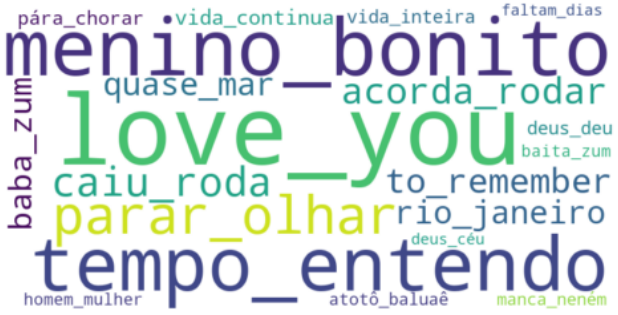

Figura 10: Nuvem com os bigramas mais presentes nas músicas da década de 2000

Ainda nos anos 2000, a apreciação pela cultura brasileira indica uma valorização das tradições, da diversidade e da identidade nacional. Os vínculos pessoais e a autoconsciência sugerem uma busca por identidade em meio à rápida transformação e globalização. A sensualidade representa uma abordagem mais aberta à expressão sexual e uma celebração da beleza brasileira. Os relacionamentos românticos e a autodescoberta refletem um interesse na intimidade e individualidade, enquanto a reflexão, a celebração e a resistência revelam uma consciência social e política na música dos anos 2000. As emoções mais frequentes como amor, aprovação, alegria e aborrecimento mostram uma tentativa de lidar com as adversidades da vida contemporânea, buscando felicidade e significado.

Com o advento das redes sociais e a crescente influência da Internet, as letras de músicas dos anos 2010, como evidenciado na Tabela 6 e Figura 11, refletem uma busca incessante por significado e autenticidade em um mundo cada vez mais interconectado. Especificamente ao analisar os bigramas dessa década e da seguinte, foi necessário remover algumas palavras obscenas que haviam entre eles. No entanto, vale ressaltar que essas palavras não foram ignoradas durante a interpretação do significado dos tópicos destas duas últimas décadas.

Tabela 6: Tópicos da década de 2010

Tópico	Frequência (%)
Amor, libertação, sociedade, política, busca por significado, autoaceitação	35,54
Desejo romântico, perseverança, liberdade, busca por sucesso, separação	33,22
Relacionamentos interpessoais, dificuldades cotidianas, criatividade, mudanças, pertencimento	31,24

Figura 11: Nuvem com os bigramas mais presentes nas músicas da década de 2010

Nos anos 2010, temas de amor, libertação, sociedade, política, busca por significado e autoaceitação se entrelaçam. As canções frequentemente abordam questões sociais como inclusão, igualdade e justiça, além de refletir a crescente conscientização sobre identidade e valores individuais na sociedade brasileira. Ao mesmo tempo, o desejo romântico, a perseverança e a busca pelo

sucesso são narrativas comuns, muitas vezes retratando a superação de obstáculos. A liberdade, tanto pessoal quanto política e social, é um tema recorrente, assim como a separação, principalmente de relacionamentos amorosos. Além disso, as músicas refletem as dificuldades cotidianas e os relacionamentos interpessoais, destacando a criatividade, a busca por mudanças e a necessidade de pertencimento e conexão com os outros e com o mundo ao redor. Entre as emoções mais presentes estão o amor, que permeia as narrativas românticas e de autoaceitação, o aborrecimento, evidenciado nas reflexões sobre as dificuldades cotidianas e os obstáculos sociais, a aprovação, refletida na busca por sucesso e reconhecimento, e a alegria, que surge nas narrativas de superação.

Por fim, até janeiro de 2024, a década de 2020 traz canções que exploram desde questões sociais e políticas até aspectos mais íntimos da experiência humana. Conforme ilustrado na Tabela 7 e Figura 12, o amor é abordado de maneira hedonista, sensual e explícita, com ênfase no prazer e nas conexões íntimas. A celebração da identidade étnica e cultural brasileira emerge como uma temática proeminente, realçando a diversidade e a riqueza das tradições do país, ao mesmo tempo que promove mensagens de inclusão e respeito mútuo.

Tabela 7: Tópicos da década de 2020 (até janeiro de 2024)

Tópico	Frequência (%)
Diversão, sonhos, prazer, euforia, conexão humana	35,61
Celebração, identidade étnica, relações sentimentais, justiça, autoconhecimento	32,71
Sensualidade, empoderamento, origem, identidade cultural, amores complexos	31,68

Figura 12: Nuvem com os bigramas mais presentes nas músicas da década de 2020 (até janeiro de 2024)

Adicionalmente, muitas composições refletem uma consciência aguçada sobre questões de justiça social e empoderamento, tratando de temas como representatividade e servindo como veículos para expressar aspirações individuais e coletivas, além de despertar a consciência sobre problemas sociais prementes. Frequentemente, essas músicas exploram temas como superação, desejo intenso e resistência, refletindo uma busca incessante por autenticidade e empoderamento em meio às incertezas e desafios do século XXI. Em meio a esse contexto, emoções como aborrecimento, amor, aprovação e alegria sugerem uma sensação de desconforto e inquietação em um mundo em constante mutação, onde a curiosidade atua como guia na busca por entendimento, enquanto o amor oferece conforto e esperança diante das incertezas.

5. CONCLUSÕES E TRABALHOS FUTUROS

A análise realizada das letras de músicas ao longo das décadas oferece um vislumbre das transformações temáticas e emocionais que marcaram a segunda metade do século XX e o início do século XXI. Em todos os períodos examinados, o amor emerge como um tema central, apresentando-se de maneiras diversas e multifacetadas. Na década de 60, observamos um amor mais idealizado, evoluindo para uma fase introspectiva nos anos 70, que se transforma em uma expressão mais alegre nos anos 80. Na década de 90, surge uma maior liberdade de crítica, acompanhada, nas primeiras décadas do século XXI, pelo levantamento de questões identitárias e reflexões sobre a identidade no mundo. Isso se entrelaça com uma valorização crescente do amor próprio nos anos 2010, culminando em uma expressão mais hedonista da conexão humana desde 2020 até o presente.

Os resultados deste estudo evidenciam a eficácia da aplicação da Modelagem de Tópicos, baseada no algoritmo LDA, na identificação de subtemas recorrentes nas músicas populares brasileiras, oferecendo valiosos *insights* sobre os temas predominantes ao longo das décadas. Além disso, a classificação das emoções utilizando o BERTimbau também se mostrou útil ao esclarecer as emoções preponderantes em cada período, proporcionando uma compreensão mais profunda de como o contexto cultural era refletido através das emoções mais frequentes.

Adicionalmente, os resultados desta análise também destacam a eficácia da métrica de coerência Cv, ao utilizar os bigramas mais frequentes de palavras distintas, para gerar tópicos mais acessíveis e interpretáveis. No entanto, é importante notar que as maiores coerências foram observadas na década com a maior quantidade de músicas coletadas, como evidenciado nos anos 2010. Isso sugere a possibilidade de que um corpus ainda mais abrangente de músicas poderia proporcionar tópicos mais diversificados e interpretáveis, capturando nuances adicionais.

Como sugestão para próximos passos é válido considerar a utilização de um *corpus* ainda mais amplo de letras de músicas da MPB. Isso poderia proporcionar uma visão mais abrangente e representativa das tendências temáticas ao longo das décadas, permitindo uma análise mais aprofundada das mudanças culturais e emocionais. Além disso, seria promissor explorar o uso de modelos mais avançados de modelagem de tópicos, capazes de interpretar automaticamente os tópicos gerados sem intervenção manual. Essa abordagem poderia facilitar uma análise mais automatizada e precisa das ideias e narrativas presentes nas letras de músicas, tornando possível identificar padrões temáticos de forma ainda mais eficiente.

6. REFERÊNCIAS

[1] JURAFSKY, Daniel, MARTIN, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2008. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>. Acesso em: 04 de abril de 2024.

[2] GALVÃO, Walnice Nogueira. MMPB: uma análise ideológica, 1976. Disponível em: <https://www.novacultura.info/post/2020/10/08/mmpb-uma-analise-ideologica>. Acesso em: 30 de setembro de 2023.

[3] NEDER, Álvaro. O estudo cultural da música popular brasileira: dois problemas e uma contribuição, 2010. Disponível em: <https://doi.org/10.1590/S1517-75992010000200015> Acesso em: 30 de setembro de 2023.

[4] BLEI, David M., NG, Andrew Y., JORDAN, Michael I. Latent Dirichlet Allocation, 2003. Disponível em: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>. Acesso em: 04 de abril de 2024.

[5] SYED, Shaheen, SPRUIT, Marco. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation, 2017. Disponível em: <https://www.researchgate.net/profile/Marco-Spruit/publication/345665781_Full-Text_or_Abstract_Examining_Topic_Coherence_Scores_Using_Latent_Dirichlet_Allocation/links/61435dc627c6bf14579815f6/Full-Text-or-Abstract-Examining-Topic-Coherence-Scores-Using-Latent-Dirichlet-Allocation.pdf>. Acesso em 28 de abril de 2024.

[6] VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, USZKOREIT, Jakob, JONES, Llion, GOMEZ, Aidan N., KAISER, Lukasz, POLOSUKHIN, Illia. Attention is All You Need, 2017. Disponível em: <https://arxiv.org/pdf/1706.03762.pdf> Acesso em: 12 de abril de 2024.

[7] DEVLIN, Jacob, CHANG, Ming-Wei, LEE, Kenton, TOUTANOVA, Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018. Disponível em: <https://arxiv.org/pdf/1810.04805.pdf> Acesso em: 12 de abril de 2024.

[8] SOUZA, Fábio, NOGUEIRA, Rodrigo, LOTUFO, Roberto, 2020. Disponível em: <https://www.researchgate.net/publication/345395208_BERTimbau_Pretrained_BERT_Models_for_Brazilian_Portuguese>. Acesso em: 14 de abril de 2024.

[9] HAMMES, Luiz, FREITAS, Larissa, 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17784/17618>. Acesso em: 14 de abril de 2024.