INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS PICOS
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

THÁVYNE KEROLLY DIAS RIBEIRO

COMPOSITOR E OUVINTE: ANÁLISE COMPUTACIONAL DE EMOÇÕES NA MÚSICA POPULAR BRASILEIRA UTILIZANDO BERTIMBAU

PICOS, PIAUÍ
2026

THÁVYNE KEROLLY DIAS RIBEIRO

COMPOSITOR E OUVINTE: ANÁLISE COMPUTACIONAL DE EMOÇÕES NA MÚSICA POPULAR BRASILEIRA UTILIZANDO BERTIMBAU

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Picos.

Orientador: Prof. M^{e}. João Paulo Lima do Nascimento

Coorientador: Prof. Dr. Rogerio Figueredo de Sousa

PICOS, PIAUÍ
2026

COMPOSITOR E OUVINTE: ANÁLISE COMPUTACIONAL DE EMOÇÕES NA MÚSICA POPULAR BRASILEIRA UTILIZANDO BERTIMBAU

Thávyne Kerolly Dias Ribeiro[1]
Prof. M[e]. João Paulo Lima do Nascimento[2]
Prof. Dr. Rogério Figueredo de Sousa[3]

RESUMO

Este trabalho apresenta uma abordagem para análise computacional de emoções na Música Popular Brasileira (MPB) utilizando dois modelos BERTimbau ajustados por *fine-tuning* para letras de músicas e comentários do YouTube. Foram analisadas 8.400 letras de 40 artistas da MPB e 48.820 comentários. Os comentários foram rotulados por *weak supervision*, com Snorkel e Gemini, e as letras por meio de um léxico emocional expandido. Os modelos alcançaram acurácia de 79,89% nos comentários e 80,48% nas letras. A comparação entre as emoções expressas nas letras e percebidas pelos ouvintes apresentou convergência de 26,3%. Também foram realizadas análises temporais e modelagem de tópicos. Os resultados evidenciam diferenças entre a emoção expressa nas letras e a percepção dos ouvintes.

Palavras-chave: Música Popular Brasileira; Processamento de Linguagem Natural; BERTimbau; Classificação de Emoções; Análise Compositor-Ouvinte; *Weak Supervision*.

Abstract

This study presents an approach for computational emotion analysis in Brazilian Popular Music (MPB) using two BERTimbau models fine-tuned for song lyrics and YouTube comments. A total of 8,400 lyrics from 40 MPB artists and 48,820 comments were analyzed. Comments were labeled using weak supervision with Snorkel and Gemini, while lyrics were analyzed using an expanded emotion lexicon. The models achieved accuracies of 79.89% for comments and 80.48% for lyrics. The comparison between emotions expressed in lyrics and perceived by listeners showed a convergence of 26.3%. Temporal analysis and topic modeling were also performed. The results indicate differences between emotions expressed in lyrics and those perceived by listeners.

Keywords: Brazilian Popular Music; Natural Language Processing; BERTimbau; Emotion Classification; Composer–Listener Analysis; *Weak Supervision*.

[1] Graduanda em Análise e Desenvolvimento de Sistemas pelo IFPI, Campus Picos. thavynekerolly126@gmail.com
[2] Mestre em Engenharia de Produção. Professor EBTT de Informática do IFPI, campus Picos. joao.nascimento@ifpi.edu.br
[3] Lattes: http://lattes.cnpq.br/9346694618502913. IFPI. rogerio.sousa@ifpi.edu.br

Data de Aprovação: 10 de agosto de 2026

1 INTRODUÇÃO

A música é uma das principais formas de expressão emocional da humanidade. Por meio da combinação entre melodia, ritmo e letra, ela comunica sentimentos, experiências e aspectos culturais que podem ser interpretados de diferentes maneiras pelos ouvintes. Nesse contexto, as letras das músicas exercem papel importante na construção do significado das composições, funcionando como um meio de expressar emoções, memórias e narrativas. Na Música Popular Brasileira (MPB), essa característica ganha ainda mais relevância, pois o gênero acompanha transformações sociais, culturais e históricas do país, refletindo valores, experiências e emoções ao longo das décadas (Juslin e Sloboda, 2010; Galvão, 1972).

Com o crescimento das plataformas digitais, tornou-se possível acessar não apenas as letras das músicas, mas também milhares de comentários publicados pelos próprios ouvintes. Ao mesmo tempo, os avanços do Processamento de Linguagem Natural (PLN) e do aprendizado profundo ampliaram as possibilidades de análise automática desses grandes volumes de texto. Apesar desse avanço, a maior parte das pesquisas ainda está concentrada na língua inglesa. No português brasileiro, especialmente no contexto musical, ainda existem poucos estudos e conjuntos de dados anotados para treinamento e avaliação de modelos de classificação emocional (Gama, 2024; Aparecido et al., 2025).

Embora pesquisas recentes tenham aplicado técnicas de PLN à MPB, ainda existem limitações importantes na compreensão da relação entre a emoção expressa nas letras e a emoção percebida pelos ouvintes. Em especial, a literatura consultada apresenta poucos estudos que integrem essas duas perspectivas em uma mesma metodologia, utilizando modelos especializados para diferentes domínios textuais. A maior parte das abordagens existentes limita-se a analisar apenas uma dessas perspectivas ou desconsidera as particularidades estilísticas e contextuais que diferenciam letras poéticas de comentários espontâneos em plataformas digitais. É justamente essa necessidade de articular e contrapor essas duas dimensões emocionais que motiva o presente trabalho.

Uma questão ainda pouco explorada consiste em investigar em que medida a emoção expressa na letra corresponde à emoção percebida pelo ouvinte. Neste trabalho, a expressão “emoção expressa” refere-se exclusivamente à emoção identificada no conteúdo textual da letra, não representando necessariamente a intenção emocional do autor ou intérprete da obra. Essa distinção é importante porque a intenção do autor pode diferir da emoção identificada no texto, especialmente considerando que a análise é realizada sobre a letra da música. Embora a letra represente uma importante fonte

de informação emocional, sua interpretação pode ser influenciada por fatores como experiências pessoais, contexto sociocultural, memória afetiva e momento da escuta. Assim, uma mesma música pode despertar emoções diferentes em pessoas distintas. Estudos recentes descrevem essa diferença como uma lacuna semântico-perceptual, indicando que a emoção identificada no texto nem sempre coincide com a emoção percebida pelo ouvinte. No contexto da MPB, essa relação ainda apresenta espaço para investigação, especialmente por meio da integração entre a análise das letras e das manifestações espontâneas dos ouvintes em plataformas digitais.

Diante desse cenário, este trabalho propõe uma abordagem baseada no modelo BERTimbau, adaptado via *fine-tuning* para dois domínios textuais distintos. O primeiro modelo é dedicado à classificação das letras musicais (capturando a intenção do compositor), enquanto o segundo processa os comentários do YouTube (refletindo a percepção dos ouvintes). A especialização de cada classificador permite considerar as características linguísticas próprias de cada contexto textual, reduzindo a influência de diferenças de vocabulário, estrutura e estilo na comparação entre essas duas perspectivas.

Outro desafio deste trabalho foi a construção de um conjunto de dados para treinamento dos modelos. Diante da inexistência de um *corpus* de comentários da MPB previamente anotado em larga escala, adotou-se uma estratégia de *weak supervision* (supervisão fraca), combinando heurísticas e modelos de linguagem para a geração automática de rótulos *silver*. Embora essa abordagem reduza drasticamente o custo e o tempo da rotulação manual, ela apresenta como desvantagem a susceptibilidade a ruídos e imprecisões nas anotações, demandando mecanismos de agregação probabilística e validação de concordância para mitigar impactos no desempenho do classificador. Os detalhes dessa estratégia e as etapas de controle de qualidade são apresentados na Seção 3.

Na literatura consultada, foram identificados estudos sobre emoções em letras de músicas e sobre a percepção emocional dos ouvintes. Entretanto, entre os trabalhos analisados, não foram identificados estudos que comparassem diretamente essas perspectivas na MPB por meio de modelos especializados por domínio. Assim, este trabalho analisa as emoções presentes nas letras e as compara com as emoções identificadas em comentários do YouTube. Para isso, foram utilizados dois modelos BERTimbau pré-treinados e especializados por *fine-tuning*. As contribuições incluem: (i) a construção de *corpora* pareados; (ii) a especialização dos modelos por domínio; (iii) o emprego de *weak supervision* com modelos de linguagem para geração de rótulos *silver*[4]; e (iv) a análise da convergência emocional, temporal e de tópicos.

[4] No contexto de Aprendizado de Máquina e PLN, rótulos *silver* (ou anotações de prata) referem-se a dados rotulados automaticamente por meio de regras, heurísticas ou modelos de linguagem, sujeitos a ruídos inerentes, em oposição aos rótulos *gold* (padrão-ouro), anotados manualmente por especialistas humanos.

2 REFERENCIAL TEÓRICO

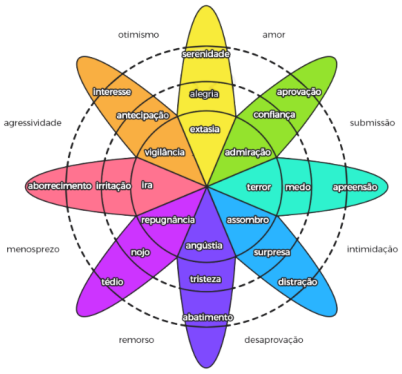
2.1. EMOÇÕES E A RODA DE PLUTCHIK

As emoções humanas podem ser representadas por diferentes modelos teóricos na literatura psicológica e computacional. Entre as abordagens dimensionais, destaca-se o Modelo Circumplexo de Russell (1980), que projeta os estados afetivos em eixos contínuos de valência (positiva/negativa) e ativação (arousal). Embora robustos para análise psicofisiológica, modelos contínuos dificultam a formulação de tarefas supervisionadas de classificação de texto em categorias semânticas bem delimitadas. Por outro lado, modelos categóricos discretos clássicos, como o de Ekman (1992), focado em seis emoções básicas universais, mostram-se mais rígidos para capturar a complexidade da linguagem artística, pois tratam as classes de forma isolada e não contemplam sentimentos compostos ou culturalmente situados.

Nesse contexto, destaca-se a Roda das Emoções proposta por Plutchik (1980), que organiza diferentes estados emocionais considerando relações de proximidade, intensidade e combinação entre emoções.

A Figura 1 apresenta a representação da Roda das Emoções utilizada como referência conceitual neste trabalho.

Figura 1 – Roda das Emoções de Plutchik



Fonte: Adaptada de Plutchik (1980).

Segundo Plutchik (1980), as emoções possuem função adaptativa e podem variar em intensidade ou combinar-se para formar estados emocionais mais complexos. Essas características tornam o modelo relevante para pesquisas de reconhecimento de emoções. Estudos recentes também empregam a Roda de Plutchik como referência em tarefas de Processamento de Linguagem Natural. Aparecido *et al.* (2025), por exemplo, utilizaram esse modelo como referência em um benchmark de reconhecimento de emoções em corpus brasileiro e verificaram desempenho superior do BERTimbau

entre os métodos avaliados. Esse resultado reforça a relevância de modelos baseados em transformadores para tarefas de classificação automática de emoções em língua portuguesa.

2.2. ANÁLISE DE SENTIMENTOS E CLASSIFICAÇÃO DE EMOÇÕES EM TEXTO

A Análise de Sentimentos é uma área do Processamento de Linguagem Natural (PLN) voltada para a identificação de informações subjetivas presentes em textos, como opiniões, avaliações e sentimentos (Liu, 2012). Tradicionalmente, essa tarefa busca identificar a polaridade do texto, classificando-o como positivo, negativo ou neutro. Já o reconhecimento de emoções procura identificar emoções específicas, permitindo uma análise mais detalhada do conteúdo emocional (Nogueira e Capanema, 2023).

As primeiras abordagens para reconhecimento de emoções utilizavam métodos baseados em léxicos, nos quais palavras eram associadas previamente a categorias emocionais. Um dos recursos mais conhecidos é o NRC *Emotion Lexicon* (Mohammad e Turney, 2013). Apesar de simples e fáceis de interpretar, abordagens baseadas em léxicos apresentam limitações quando comparadas a modelos de linguagem em tarefas de detecção de emoções, especialmente em contextos que exigem maior compreensão textual (Aparecido *et al.*, 2025).

Nos últimos anos, modelos baseados na arquitetura Transformer passaram a apresentar melhores resultados em diversas tarefas de PLN, incluindo a classificação de emoções Devlin *et al.* (2019), Hammes e Freitas (2021).

No contexto das letras de músicas, Edmonds e Sedoc (2021) observaram que modelos treinados especificamente nesse domínio apresentam melhor desempenho do que modelos ajustados em textos de redes sociais ou diálogos. Segundo os autores, isso ocorre porque as letras possuem características próprias, como metáforas, linguagem poética e elevada carga emocional, que diferem de outros tipos de texto.

Outra questão importante discutida na literatura é a distinção entre a emoção expressa na letra e a emoção percebida pelo ouvinte. Conforme discutido em Hu *et al.* (2026), essas duas perspectivas podem não coincidir. Por constituir um dos fundamentos da abordagem proposta neste trabalho, essa temática será aprofundada na Seção 2.5, onde são apresentados os principais fatores envolvidos nessa diferença.

2.3. BERTIMBAU E PLN EM LÍNGUA PORTUGUESA

O BERT (*Bidirectional Encoder Representations from Transformers*), proposto por Devlin *et al.* (2019), foi um marco para o Processamento de Linguagem Natural ao introduzir um modelo capaz de compreender palavras considerando o contexto em que aparecem. Baseado na arquitetura Transformer Vaswani *et al.* (2017), o BERT é

inicialmente pré-treinado em grandes volumes de texto e, posteriormente, pode ser ajustado por meio de *fine-tuning* para diferentes tarefas, como classificação de textos, análise de sentimentos e reconhecimento de entidades.

Para o português brasileiro, Souza *et al.* (2020) desenvolveram o BERTimbau, uma versão do BERT pré-treinada em um amplo conjunto de textos em língua portuguesa.

2.4. SUPERVISÃO FRACA E MODELOS DE LINGUAGEM COMO FONTES DE ROTULAÇÃO

Segundo Ratner *et al.* (2017), a obtenção de dados rotulados é um dos principais desafios em tarefas de Aprendizado de Máquina, especialmente quando não existem bases de dados previamente anotadas. A rotulação manual exige tempo, conhecimento especializado e recursos, o que pode dificultar a construção de conjuntos de dados em larga escala. Como alternativa, a *weak supervision* permite gerar automaticamente rótulos aproximados (*silver labels*) a partir de diferentes fontes de informação, reduzindo a necessidade de anotação manual (Ratner *et al.*, 2017).

Uma das ferramentas mais utilizadas para esse tipo de abordagem é o Snorkel (Ratner *et al.*, 2017). Nesse *framework*, a rotulação é realizada por meio de *Labeling Functions* (LFs), que são regras criadas pelo pesquisador para atribuir um rótulo a um texto ou optar por não classificá-lo. Essas funções podem utilizar palavras-chave, léxicos, expressões regulares, regras heurísticas ou outras informações relacionadas ao problema.

Como nem todas as LFs possuem a mesma acurácia, o Snorkel utiliza um modelo probabilístico denominado *Label Model*, que estima a confiabilidade de cada função com base nos seus padrões de concordância e discordância mútua. Por exemplo, diante de um comentário como *"Essa música me lembra a infância no interior, que saudade!"*, uma regra baseada em marcadores temporais pode sugerir o rótulo *Nostalgia*, enquanto uma regra léxica focada no termo explícito sugere *Saudade*. Em vez de aplicar uma votação majoritária simples, o *Label Model* pondera o peso estatístico de cada fonte para inferir a distribuição de probabilidade final da amostra. A partir dessa combinação, é gerado o rótulo consolidado (*silver label*), empregado subsequentemente no treinamento dos classificadores supervisionados.

Nos últimos anos, modelos de linguagem de grande porte (*Large Language Models*, LLMs) também passaram a ser utilizados como fontes adicionais de rotulação. Diferentemente das regras heurísticas determinísticas, esses modelos analisam o texto considerando seu contexto semântico amplo. Smith *et al.* (2024) demonstraram que a integração de LLMs como funções de rotulação dentro de um arranjo de *ensemble* probabilístico no Snorkel produz resultados superiores aos obtidos por regras isoladas ou pelo modelo de linguagem de forma avulsa. Essa abordagem atua como uma

agregação de múltiplos preditores fracos (*weak learners*), onde o *Label Model* pondera a confiança de cada fonte sem a necessidade de anotações manuais de referência.

2.5. EMOÇÃO EXPRESSA E EMOÇÃO PERCEBIDA NA MÚSICA

A psicologia da música distingue dois conceitos importantes relacionados à experiência musical: a emoção expressa, representada neste trabalho pela emoção presente na letra da música, e a emoção percebida, que corresponde à forma como essa obra é interpretada pelo ouvinte (Juslin e Sloboda, 2010). Embora estejam relacionadas, essas duas perspectivas nem sempre coincidem, pois a percepção emocional depende não apenas da música, mas também das experiências e características de cada pessoa.

Entre os fatores que influenciam essa percepção estão as experiências pessoais, a memória afetiva, a identificação com o artista, o contexto social e cultural, o momento da escuta e as preferências musicais do ouvinte. Por esse motivo, uma mesma música pode despertar emoções diferentes em pessoas distintas, mesmo quando sua mensagem permanece a mesma.

Essa diferença tem recebido atenção crescente nas pesquisas sobre reconhecimento computacional de emoções. Hu *et al.* (2026) denominam esse fenômeno de lacuna semântico-perceptual, destacando que a emoção identificada a partir da letra representa apenas uma parte da experiência emocional associada à música. A emoção percebida pelo ouvinte também é influenciada por elementos que não estão presentes no texto, como a melodia, a interpretação vocal, o contexto de escuta e as experiências individuais.

Nesse contexto, comentários publicados espontaneamente em plataformas digitais, como o YouTube, tornam-se uma fonte importante para compreender a emoção percebida pelos ouvintes. Diferentemente das letras, esses comentários registram as reações emocionais do público após o contato com a música, permitindo comparar a emoção expressa na letra com a emoção percebida pelos ouvintes. Essa distinção é um dos fundamentos da abordagem proposta neste trabalho.

2.6. TRABALHOS RELACIONADOS

A utilização de modelos baseados em BERT para o reconhecimento de emoções em língua portuguesa tem sido investigada em diferentes contextos. Hammes e Freitas (2021) realizaram o *fine-tuning* dos modelos BERTimbau-base e BERTimbau-large para a classificação de 27 emoções em sentenças, utilizando o conjunto de dados GoEmotions traduzido para o português. Os autores também empregaram uma estratégia de balanceamento das classes e observaram ganho de desempenho em relação aos resultados originais do conjunto de dados.

Em outro contexto, Oliveira e Sichman (2025) desenvolveram um modelo de detecção de emoções em português brasileiro utilizando o BERTimbau, aplicado a respostas de usuários no Twitter relacionadas a notícias sobre a pandemia de COVID-19. O modelo foi ajustado para identificar as seis emoções básicas de Ekman e a categoria neutro, demonstrando a aplicação do BERTimbau em textos de redes sociais em português brasileiro.

No contexto da Música Popular Brasileira, Gama (2024) analisaram 12.542 músicas de 1.318 artistas, abrangendo o período de 1960 a 2020, utilizando técnicas de Processamento de Linguagem Natural, incluindo modelagem de tópicos com LDA e classificação de emoções com BERTimbau. O estudo apresenta uma abordagem diretamente relacionada à análise emocional de letras da MPB.

No contexto internacional, Edmonds e Sedoc (2021) investigaram a classificação de múltiplas emoções em letras de músicas e criaram um conjunto de dados anotado a partir da perspectiva dos leitores. Os autores observaram que modelos BERT ajustados com dados fora do domínio apresentaram menor capacidade de generalização para letras de músicas, enquanto modelos treinados com conjuntos de dados específicos desse domínio obtiveram desempenho ligeiramente superior. Esse resultado evidencia a relevância da adaptação dos modelos às características do domínio textual.

Aparecido *et al.* (2025) compararam o desempenho do BERTimbau, de um modelo de linguagem de grande porte e de um método baseado em léxico na detecção de emoções em português brasileiro. O estudo avaliou quatro corpora e verificou que os modelos baseados em transformadores superaram o método baseado em léxico em diferentes avaliações, embora os resultados variassem conforme o corpus e a métrica considerada.

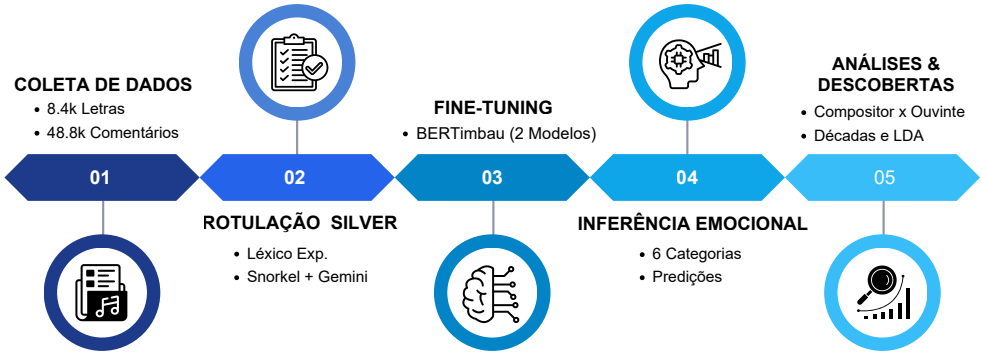
Entre os trabalhos analisados, observa-se que estudos anteriores exploram a classificação de emoções em português, em redes sociais ou em letras de músicas, mas não integram essas duas perspectivas no contexto da Música Popular Brasileira. O presente trabalho diferencia-se ao utilizar dois modelos BERTimbau especializados separadamente para letras e comentários do YouTube, permitindo comparar as emoções identificadas nas letras com aquelas percebidas pelos ouvintes.

3 METODOLOGIA

A metodologia foi organizada em cinco etapas principais, abrangendo a coleta dos dados, a rotulação dos corpora, a especialização dos modelos BERTimbau, a inferência das categorias emocionais e as análises realizadas.

A Figura 2 apresenta uma visão geral do procedimento adotado, enquanto as subseções seguintes detalham cada uma dessas etapas.

Figura 2 – Fluxograma da metodologia proposta.



Fonte: Elaborada pela autora (2026).

3.1. CONSTRUÇÃO DOS CORPORA

A investigação sistemática acerca da convergência e divergência afetiva na Música Popular Brasileira (MPB) demandou a criação de duas bases de dados textuais independentes e complementares: um *corpus* voltado à perspectiva do compositor, composto integralmente por letras musicais, e outro direcionado à percepção dos ouvintes, estruturado a partir de comentários espontâneos publicados na plataforma YouTube sobre as respectivas obras.

O *corpus* lírico foi obtido por meio de rotinas automatizadas de raspagem de dados (*web scraping*) direcionadas ao portal Letras.mus.br[5], utilizando *scripts* desenvolvidos na linguagem Python com o suporte das bibliotecas *Requests* e *BeautifulSoup*. O pipeline de extração realizou consultas sequenciais às discografias de 40 artistas consagrados da MPB, efetuando requisições HTTP e processando as árvores sintáticas DOM em HTML para isolar títulos, estrofes e versos integrais de cada canção. Adicionalmente, integraram-se metadados cronológicos por meio de consultas automatizadas à API MusicBrainz, viabilizando o resgate do ano de lançamento original das faixas para fins de análise temporal. Inicialmente, foram coletadas 8.493 composições líricas. A aplicação de filtros preliminares de higienização voltados à eliminação de duplicatas textuais e ao descarte de registros com extensão inferior a 50 caracteres consolidou a base lírica definitiva em 8.400 letras.

O *corpus* de comentários foi construído a partir de vídeos associados às canções selecionadas no YouTube, empregando a biblioteca *Selenium* para contornar os desafios de renderização dinâmica da interface e permitir a rolagem contínua dos elementos. A estratégia de busca priorizou sistematicamente as faixas oficiais catalogadas pelo YouTube Music por meio do qualificador de canal *"topic"*. Foram descartados vídeos com regravações (*covers*), faixas instrumentais (*playbacks*), registros informais ao vivo e vídeos de reação (*reacts*), minimizando ruídos conversacionais direcionados à performance cênica dos intérpretes e preservando o foco estrito na reação afetiva despertada pela composição. A extração bruta consolidou 49.096 comentários atrelados a 4.479 faixas. Após a remoção de duplicatas e comentários estritamente curtos ou desprovidos de valor semântico, o conjunto final de percepção reuniu 48.820 registros.

A decisão de desacoplar os dois *corpora* constituiu uma escolha de desenho metodológico indispensável para acomodar a assimetria linguística existente entre a escrita poética estruturada e a comunicação informal da *web*. Tal separação permitiu o posterior ajuste fino de classificadores dedicados a cada domínio textual, mitigando vieses estilísticos durante a mensuração da lacuna emocional entre compositores e ouvintes.

[5] Disponível em: <https://www.letras.mus.br>. Acesso em: 2026.

3.2. PRÉ-PROCESSAMENTO TEXTUAL

Antes da etapa de rotulação, ambos os *corpora* foram submetidos a um processo de pré-processamento textual com o objetivo de padronizar os dados, reduzir ruídos e prepará-los para as etapas de rotulação automática e treinamento dos modelos.

Inicialmente, foram aplicadas etapas comuns aos dois *corpora*, compreendendo: (i) conversão de todos os caracteres para letras minúsculas; (ii) remoção de URLs; (iii) substituição de quebras de linha por espaços; (iv) remoção de marcações estruturais presentes nas letras, como indicações de versos e refrões entre colchetes; (v) remoção de caracteres especiais, preservando a acentuação da língua portuguesa; e (vi) normalização de espaços em branco consecutivos.

No corpus de letras, foram excluídas músicas com menos de 50 caracteres após o pré-processamento, registros duplicados com base no texto limpo e letras identificadas automaticamente como pertencentes a idiomas diferentes do português. Essas etapas tiveram como objetivo aumentar a homogeneidade linguística do conjunto de dados e eliminar registros incompletos ou inconsistentes.

No corpus de comentários, foram removidos registros duplicados com base no texto limpo e comentários com até 10 caracteres após a etapa de limpeza textual. Durante a tokenização, comentários que excediam o limite máximo de entrada do BERTimbau foram automaticamente truncados pelo *tokenizer*, garantindo compatibilidade com o modelo durante o treinamento.

Ao final do pré-processamento, ambos os *corpora* encontravam-se preparados para as etapas de rotulação automática e treinamento dos modelos, após a remoção de registros duplicados, de conteúdos textuais insuficientes e a padronização dos textos.

3.3. DEFINIÇÃO DAS CATEGORIAS EMOCIONAIS

A partir da Roda das Emoções de Plutchik, foi definida uma taxonomia própria, adaptada às características emocionais e linguísticas observadas no corpus de Música Popular Brasileira. Foram consideradas seis categorias emocionais: alegria, tristeza, nostalgia, saudade, amor composto e neutro.

Embora Plutchik (1980) proponha oito emoções primárias, esta pesquisa não adotou diretamente sua taxonomia original, mas uma adaptação voltada ao contexto da MPB. As categorias foram selecionadas de acordo com o objetivo de representar manifestações emocionais mais adequadas ao contexto da MPB. As categorias alegria e tristeza foram mantidas por representarem emoções básicas, enquanto saudade e nostalgia foram incorporadas por constituírem estados afetivos recorrentes no contexto cultural brasileiro e nas letras analisadas. A categoria amor composto foi adotada para representar manifestações afetivas relacionadas ao amor e aos relacionamentos,

enquanto a categoria neutro foi incluída para representar letras sem predominância emocional identificável segundo os critérios adotados.

Dessa forma, categorias presentes na proposta original de Plutchik, como medo, raiva, aversão, surpresa, antecipação e confiança, não foram contempladas, pois não constituíam o foco desta investigação. A taxonomia adotada foi utilizada como base para a rotulação e classificação das letras.

3.4. ROTULAÇÃO *SILVER* DAS LETRAS VIA LÉXICO EXPANDIDO

O corpus de letras foi rotulado automaticamente por meio de um léxico expandido, gerando um conjunto de rótulos *silver* utilizado posteriormente no treinamento do modelo BERTimbau especializado em letras. Foram consideradas cinco categorias emocionais: alegria, tristeza, nostalgia, saudade e amor composto. A categoria neutro foi atribuída às letras que não apresentaram correspondência com termos pertencentes a nenhuma dessas categorias.

O léxico expandido foi construído manualmente a partir de um conjunto inicial de palavras representativas de cada emoção, definido com base na Roda das Emoções de Plutchik (Plutchik, 1980) e em palavras e expressões recorrentes em letras da Música Popular Brasileira (MPB). Posteriormente, esse conjunto foi ampliado com a inclusão de sinônimos e expressões semanticamente relacionadas a cada categoria emocional, totalizando 176 termos distribuídos entre as cinco emoções consideradas. Cada categoria foi representada por um conjunto específico de palavras e expressões associadas ao seu significado emocional. Para reduzir possíveis vieses decorrentes de diferenças no tamanho dessas listas, a pontuação de cada categoria foi calculada pela razão entre o número de termos encontrados na letra e a quantidade total de termos do respectivo léxico. Dessa forma, categorias com maior número de termos não eram favorecidas durante o processo de classificação. Além de palavras isoladas, o léxico também incorporou expressões compostas recorrentes nas letras da MPB, ampliando sua capacidade de representar as diferentes manifestações linguísticas de cada categoria emocional.

Nos casos em que duas ou mais categorias apresentavam a mesma pontuação máxima, foi adotado um critério de desempate baseado na especificidade semântica atribuída a cada categoria emocional, seguindo a seguinte ordem de prioridade: saudade, nostalgia, amor composto, tristeza e alegria. Esse critério foi definido para reduzir ambiguidades durante a atribuição dos rótulos.

Optou-se pela utilização de um léxico expandido para o corpus de letras por ser uma abordagem compatível com as características linguísticas desse domínio textual, possibilitando a geração automática de rótulos *silver* para o treinamento do modelo.

Em contraste, os comentários do YouTube apresentam linguagem mais informal,

com uso frequente de gírias, abreviações, emojis, ironias e referências contextuais, fatores que reduzem a eficácia de abordagens exclusivamente lexicais. Por essa razão, optou-se por utilizar uma estratégia de supervisão fraca baseada no *framework* Snorkel apenas para esse corpus, conforme descrito na Seção 3.4.

3.5. ROTULAÇÃO *SILVER* DOS COMENTÁRIOS VIA *WEAK SUPERVISION* COM SNORKEL E GEMINI

Diferentemente do corpus de letras, os comentários publicados no YouTube apresentam maior informalidade, menor extensão textual e forte dependência de contexto, características que dificultam a utilização de métodos exclusivamente baseados em léxico. Além disso, a utilização de uma única regra de rotulação pode limitar a capacidade de representar a diversidade linguística presente nesse tipo de texto. Para lidar com esse cenário, adotou-se uma estratégia de *weak supervision* (supervisão fraca), técnica que permite gerar rótulos automaticamente a partir da combinação de múltiplas fontes de anotação, reduzindo a necessidade de rotulação manual em larga escala.

Para implementar essa estratégia foi utilizado o Snorkel, um *framework* desenvolvido para supervisão fraca que combina diferentes funções de rotulação (*Labeling Functions*) estimou automaticamente a confiabilidade de cada uma delas e inferiu um único rótulo probabilístico para cada instância.

Foram implementadas 19 *Labeling Functions* (LFs), responsáveis por realizar a rotulação automática dos comentários. Cada LF consiste em uma regra heurística que atribui um rótulo quando identifica determinado padrão previamente definido no texto, podendo também optar por não emitir classificação (*abstain*) quando não há evidências suficientes. As funções foram organizadas em sete grupos complementares: (i) palavras-chave fortemente associadas às categorias emocionais; (ii) padrões envolvendo o verbo *lembrar* combinados com referências temporais ou pessoais, utilizados para diferenciar nostalgia de saudade; (iii) padrões estruturais dos comentários, combinando sinais de pontuação e termos emocionais; (iv) referências explícitas à perda, luto e morte, utilizadas como indicadores de saudade; (v) características estruturais dos comentários, utilizando o comprimento do texto como critério auxiliar combinado à presença de expressões positivas; (vi) palavras e expressões semanticamente relacionadas às categorias emocionais, utilizadas para ampliar a cobertura da rotulação automática; e (vii) uma função baseada no modelo de linguagem *Gemini 2.5 Flash* (Google DeepMind, 2025), acessado por meio da plataforma Vertex AI, serviço da Google para execução de modelos de inteligência artificial.

A utilização do Gemini teve como objetivo incorporar uma fonte adicional de conhecimento semântico ao processo de rotulação. Enquanto as demais *Labeling Functions* foram construídas a partir de regras heurísticas, o modelo de linguagem analisa o

texto considerando seu contexto, fornecendo evidências complementares durante a atribuição dos rótulos. Conforme discutido por Smith *et al.* (2024), a combinação entre regras heurísticas e modelos de linguagem torna o processo de supervisão fraca mais robusto do que a utilização isolada de apenas uma dessas estratégias.

Para cada um dos 48.820 comentários, o Gemini recebeu o *prompt* apresentado no Quadro 1„ instruindo o modelo a classificar o comentário em exatamente uma das seis categorias emocionais consideradas neste estudo: alegria, tristeza, nostalgia, saudade, amor composto e neutro. O modelo foi utilizado em configuração *zero-shot*, abordagem em que realiza a classificação sem receber exemplos previamente rotulados, utilizando apenas as instruções presentes no *prompt*, a descrição das categorias emocionais e o texto do comentário. Quando a resposta correspondia à categoria neutro, a *Labeling Function* retornava *ABSTAIN* (abstenção), indicando que optou por não emitir um rótulo para aquela instância. Essa decisão era posteriormente combinada às demais *Labeling Functions* pelo *Label Model* do Snorkel para inferir o rótulo probabilístico final.

Quadro 1 – Prompt utilizado pelo Gemini 2.5 Flash para classificação emocional dos comentários

```
Você é especialista em análise de emoções em comentários sobre Música
Popular Brasileira (MPB).

Classifique a emoção principal do comentário em exatamente uma das
categorias abaixo:

Alegria:  felicidade, contentamento, animação e entusiasmo.
Tristeza:  tristeza, dor, sofrimento e melancolia.
Nostalgia:  lembrança de época passada, memória de geração e velhos
tempos.
Saudade:  falta de alguém específico ou desejo de reviver alguém ou algum
momento.
Amor composto:  amor romântico, paixão, carinho e admiração intensa.
Neutro:  ausência de emoção predominante ou mistura de emoções sem
predominância clara.

Responda utilizando apenas o nome da categoria.

Comentário:
{texto}

Emoção:
```

Fonte: Elaborado pela autora (2026).

As 19 *Labeling Functions* exploraram diferentes evidências presentes nos comentários, incluindo padrões lexicais, contextuais e estruturais, além de uma função baseada no modelo Gemini 2.5 Flash. Cada função atribuía um rótulo de forma independente para a instância ou optava por *ABSTAIN* quando não havia evidências suficientes para a classificação.

Após a execução dessas *Labeling Functions*, o *Label Model*, componente do Snorkel responsável por combinar automaticamente as diferentes fontes de anotação, estimou a confiabilidade de cada função com base nos padrões de concordância e discordância observados entre elas, inferindo probabilisticamente um único rótulo *silver* para cada comentário.

3.6. ARQUITETURA DE MODELOS DESACOPLADOS POR DOMÍNIO

A principal decisão metodológica deste trabalho foi utilizar dois modelos BERTimbau ajustados por meio de *fine-tuning*, cada um ajustado para um domínio textual distinto. Embora letras de músicas e comentários do YouTube estejam relacionados às mesmas obras musicais, esses *corpora* apresentam características linguísticas diferentes. Enquanto as letras possuem linguagem predominantemente poética, maior extensão textual e organização sintática mais estável, os comentários são mais curtos e apresentam elevada informalidade, com uso frequente de gírias, abreviações, emojis e referências contextuais.

Essas diferenças fazem com que os *corpora* apresentem distribuições lexicais e semânticas distintas. Em tarefas de classificação de textos, modelos treinados em dados do próprio domínio tendem a apresentar melhor desempenho do que modelos aplicados diretamente a domínios diferentes (Edmonds e Sedoc, 2021). Por esse motivo, optou-se pelo treinamento de dois modelos independentes, permitindo que cada um aprendesse os padrões linguísticos característicos do corpus para o qual foi desenvolvido.

Além disso, os comentários não foram utilizados como fonte de rotulação das letras. Embora ambos estejam associados à mesma música, eles representam perspectivas emocionais diferentes. As letras expressam a emoção presente na composição, enquanto os comentários refletem a interpretação dos ouvintes. Utilizar os comentários para supervisionar o modelo de letras reduziria a independência entre essas duas perspectivas e poderia influenciar a comparação proposta neste trabalho.

O modelo destinado aos comentários foi treinado exclusivamente com os rótulos produzidos pela estratégia de supervisão fraca baseada em Snorkel, enquanto o modelo das letras utilizou apenas os rótulos gerados pelo léxico expandido. Essa escolha foi motivada pelas características distintas dos dois domínios textuais. As letras de músicas apresentam estrutura mais estável e vocabulário mais elaborado, favorecendo a utilização de um léxico expandido para a rotulação automática. Em contrapartida, os comentários do YouTube apresentam linguagem mais informal, textos mais curtos e maior dependência do contexto, justificando a utilização de múltiplas *Labeling Functions* combinadas pelo Snorkel para gerar rótulos mais robustos. Em ambos os casos, os rótulos *silver* foram utilizados exclusivamente durante o treinamento

dos respectivos modelos. A comparação entre a emoção expressa pelo compositor e a emoção percebida pelos ouvintes foi realizada apenas a partir das predições produzidas pelos modelos treinados.

3.7. TREINAMENTO DOS MODELOS

Após a definição da arquitetura descrita na seção anterior, realizou-se o treinamento dos dois modelos especializados. Ambos foram inicializados a partir do modelo pré-treinado *neuralmind/bert-base-portuguese-cased*, sendo posteriormente ajustados por meio de *fine-tuning* para a tarefa de classificação emocional em seus respectivos *corpora*.

Considerando as diferenças entre os domínios textuais, o modelo de comentários utilizou comprimento máximo de sequência de 128 *tokens*, valor suficiente para acomodar a grande maioria dos comentários coletados sem necessidade de truncamento. Já o modelo de letras utilizou 512 *tokens*, correspondente ao limite máximo suportado pelo BERTimbau, permitindo preservar uma parcela maior do conteúdo das músicas e reduzir a perda de contexto causada pelo truncamento das sequências.

Para ambos os modelos foram utilizados os mesmos hiperparâmetros de treinamento: taxa de aprendizado de 2×10^{-5}, *weight decay* de 0,01, *warmup ratio* de 0,10 e agendador de taxa de aprendizado do tipo *Cosine Scheduler*. Esses hiperparâmetros correspondem a configurações amplamente utilizadas no ajuste fino de modelos BERT, proporcionando treinamento estável e boa capacidade de generalização.

O tamanho do lote (*batch size*) foi definido em 16 para o modelo de comentários e 8 para o modelo de letras. Essa diferença deve-se ao maior comprimento das sequências de letras, que aumenta significativamente o consumo de memória durante o treinamento, exigindo a redução do *batch size* para manter a viabilidade computacional.

Os dados foram inicialmente divididos em 80% para treinamento e validação e 20% para teste, utilizando estratificação por classe para preservar a distribuição das categorias emocionais. Em seguida, o conjunto de treinamento e validação foi novamente dividido de forma estratificada em 90% para treinamento e 10% para validação, resultando em aproximadamente 72% das amostras para treinamento, 8% para validação e 20% para teste. O conjunto de validação foi utilizado durante o ajuste dos modelos e na seleção automática do melhor *checkpoint*, enquanto o conjunto de teste permaneceu completamente separado e foi utilizado apenas na avaliação final do desempenho dos modelos.

O modelo de comentários foi treinado por duas épocas, número definido a partir de experimentos preliminares que indicaram não haver ganhos de desempenho com treinamentos mais longos. Para o modelo de letras, foram realizadas quatro épocas de treinamento para monitorar a evolução da perda no conjunto de validação ao longo do

processo de ajuste fino. Em ambos os casos, foi utilizada a opção de carregamento automático do melhor *checkpoint*, selecionado com base na menor perda observada no conjunto de validação (*validation loss*). Dessa forma, embora o treinamento do modelo de letras tenha sido executado por quatro épocas, o *checkpoint* correspondente à segunda época foi automaticamente selecionado como modelo final por apresentar a menor perda de validação.

3.8. VALIDAÇÃO DAS ROTULAÇÕES *SILVER*

Como os *corpora* utilizados neste estudo não dispõem de anotações humanas de referência (*gold labels*), a qualidade das rotulações *silver* foi avaliada de forma indireta. Para isso, cada modelo BERTimbau foi avaliado independentemente em seu respectivo conjunto de teste, previamente separado durante a etapa de treinamento. Em seguida, calculou-se o coeficiente Kappa de Cohen entre as predições produzidas por cada modelo e os respectivos rótulos *silver*, utilizados como referência durante a avaliação de cada modelo. Esse procedimento permitiu avaliar o grau de concordância entre as predições dos modelos e as rotulações automáticas empregadas em cada domínio textual.

Os conjuntos de teste permaneceram completamente separados do treinamento e foram obtidos por meio de divisão estratificada, garantindo que nenhuma amostra avaliada tivesse sido utilizada durante o ajuste dos modelos. No *corpus* de comentários, a rotulação *silver* foi produzida pelo *Label Model* do Snorkel, responsável por combinar as decisões emitidas pelas 19 *Labeling Functions*, incluindo a função baseada no Gemini 2.5 Flash, inferindo probabilisticamente um único rótulo para cada comentário. Dessa forma, os rótulos utilizados no treinamento resultam da agregação das informações fornecidas pelas múltiplas *Labeling Functions*, e não da decisão de uma única regra (Smith *et al.*, 2024).

Os valores do coeficiente Kappa de Cohen foram interpretados como uma medida da consistência entre as predições dos modelos e os respectivos rótulos *silver*, não como uma medida absoluta da qualidade da rotulação automática. A interpretação seguiu a classificação proposta por Landis e Koch (1977), segundo a qual valores entre 0,61 e 0,80 indicam concordância substancial. Neste estudo, foram obtidos coeficientes Kappa de Cohen de 0,734 para o modelo de comentários e de 0,747 para o modelo de letras, indicando concordância substancial entre as predições dos modelos e os respectivos rótulos *silver* utilizados como referência durante a avaliação de cada modelo em seu respectivo domínio textual.

3.9. INFERÊNCIA NAS LETRAS E ANÁLISE COMPARATIVA

Após o treinamento, os dois modelos BERTimbau foram aplicados aos respectivos *corpora* completos para obtenção das predições emocionais utilizadas nas análises deste trabalho.

Como diversas letras ultrapassam o comprimento máximo de entrada suportado pelo BERTimbau, a inferência do modelo de letras foi realizada por meio de uma estratégia de janela deslizante (*sliding window*). Cada letra foi segmentada em janelas de até 512 palavras, com sobreposição de 256 palavras entre segmentos consecutivos. Em seguida, cada segmento foi tokenizado e convertido para o formato de entrada do modelo, respeitando o limite máximo de 512 *tokens* suportado pelo BERTimbau. A utilização de janelas sobrepostas teve como objetivo reduzir a perda de contexto nas regiões de transição entre segmentos consecutivos, preservando informações que poderiam ficar distribuídas entre janelas adjacentes. Cada segmento foi classificado individualmente pelo modelo, e a emoção final da música foi determinada por votação majoritária entre as predições obtidas para todos os segmentos. Essa estratégia foi adotada por sintetizar as predições dos diferentes segmentos da mesma letra em uma única classificação representativa da obra.

O modelo de comentários foi aplicado ao *corpus* completo, classificando individualmente cada comentário com base nos padrões aprendidos durante o processo de *fine-tuning*.

Para representar a perspectiva dos ouvintes, foi atribuída a cada música a emoção predominante entre os comentários classificados pelo modelo de comentários.A perspectiva da letra foi representada pela emoção predita pelo modelo de letras para a respectiva composição. O cruzamento dessas informações permitiu comparar diretamente a emoção expressa na letra com a emoção percebida pelos ouvintes.

A análise comparativa foi realizada sobre um conjunto de 4.479 músicas presentes simultaneamente nos *corpora* de letras e de comentários. Para representar a percepção dos ouvintes em cada música, foi considerada a emoção mais frequente entre os comentários associados àquela obra. Essas informações foram então combinadas por meio do cruzamento entre título e artista com a emoção prevista pelo modelo de letras, representando a emoção atribuída ao compositor. A partir desse conjunto, foram calculadas a taxa global de convergência emocional entre compositor e ouvinte e o coeficiente Kappa de Cohen, permitindo quantificar o grau de concordância entre essas duas perspectivas. Além da análise global, também foram realizadas análises por categoria emocional do compositor, por artista, por década de lançamento e dos principais padrões de convergência e divergência emocional observados.

4 RESULTADOS E DISCUSSÃO

4.1. DESEMPENHO DO MODELO DE COMENTÁRIOS

O modelo de comentários foi treinado sobre um conjunto de treinamento contendo 35.150 comentários, enquanto 3.906 comentários foram destinados à validação, utilizada durante o processo de ajuste fino para monitorar a perda de validação e selecionar automaticamente o melhor *checkpoint*. A avaliação final foi realizada em um conjunto de teste independente composto por 9.764 comentários. O modelo alcançou acurácia global de 79,89% e F1-score macro de 0,744. A Tabela 1 apresenta o desempenho detalhado por categoria emocional.

Tabela 1 – Desempenho do Modelo A (BERTimbau *fine-tuning* em comentários)

Emoção	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Neutro	0,393	0,569	0,465	167
Alegria	0,856	0,784	0,819	3.297
Tristeza	0,785	0,812	0,798	1.304
Nostalgia	0,718	0,809	0,761	1.039
Amor composto	0,808	0,814	0,811	3.039
Saudade	0,809	0,813	0,811	918
Média macro	0,728	0,767	0,744	9.764
Média ponderada	0,805	0,799	0,801	9.764

Fonte: Elaborada pela autora (2026).

Com exceção da categoria neutro, as demais categorias emocionais apresentaram desempenho consistente, com valores de F1-Score variando entre 0,761 e 0,819, indicando boa capacidade de discriminação entre as diferentes categorias emocionais presentes nos comentários. A categoria alegria apresentou o maior F1-Score (0,819), seguida por amor composto (0,811) e saudade (0,811). As categorias tristeza e nostalgia também apresentaram desempenho satisfatório, com equilíbrio entre precisão e revocação, embora a nostalgia tenha apresentado o menor F1-Score entre as categorias emocionais, desconsiderando a classe neutro.

A categoria neutro apresentou o menor desempenho (F1 = 0,465). Esse resultado é compatível com o reduzido número de exemplos disponíveis no conjunto de teste (167 comentários, aproximadamente 1,7% do corpus). Além da baixa representatividade, essa categoria reúne comentários com pouca informação emocional explícita, dificultando sua distinção em relação às demais emoções.

A análise da matriz de confusão evidenciou que a maior parte dos erros ocorreu entre as categorias alegria e amor composto, emoções semanticamente próximas

segundo a Roda das Emoções de Plutchik. Em diversos comentários, manifestações de felicidade, admiração e afeto aparecem simultaneamente, o que contribui para explicar parte da confusão observada entre essas categorias.

De forma geral, os resultados de acurácia (79,89%), F1-score macro (0,744), precisão, revocação e F1-score por categoria evidenciam desempenho satisfatório do modelo, embora a classe neutro tenha apresentado desempenho inferior às demais categorias emocionais. Complementando essa avaliação, a consistência da rotulação *silver* foi analisada por meio do coeficiente Kappa de Cohen, obtendo-se $\kappa = 0,734$, valor classificado como concordância substancial, conforme a escala proposta por Landis e Koch (1977). Considerando que o modelo foi treinado a partir de rótulos gerados automaticamente, esse resultado indica uma concordância substancial entre as predições do modelo e a rotulação utilizada durante o treinamento.

Durante o desenvolvimento da metodologia foram realizados experimentos preliminares utilizando apenas regras heurísticas para a rotulação dos comentários. Nessa configuração, o coeficiente Kappa apresentou valores superiores a 0,93. Esse comportamento era esperado, pois todas as fontes de anotação eram baseadas em regras determinísticas desenvolvidas para o mesmo domínio, resultando em elevada concordância entre elas, mas com menor capacidade de representar padrões linguísticos mais complexos.

A introdução do Gemini 2.5 Flash como fonte adicional de supervisão, integrada ao processo de agregação probabilística do Snorkel, tornou a rotulação mais robusta ao combinar regras heurísticas com inferências semânticas produzidas por um modelo de linguagem. Essa estratégia está alinhada à abordagem de supervisão fraca proposta por Smith *et al.* (2024). Nesse contexto, o valor de $\kappa = 0,734$ representa um cenário mais realista de supervisão fraca, no qual os rótulos resultam da combinação de múltiplas fontes complementares de supervisão. O desempenho obtido é compatível com um modelo capaz de aprender padrões linguísticos mais gerais e menos dependentes das regras heurísticas utilizadas durante a rotulação.

4.2. DESEMPENHO DO MODELO DE LETRAS

O modelo de letras foi treinado sobre um conjunto de treinamento contendo 6.048 letras, enquanto 672 letras foram destinadas à validação, utilizada durante o processo de ajuste fino para monitorar a perda de validação e selecionar automaticamente o melhor *checkpoint*. A avaliação final foi realizada em um conjunto de teste independente composto por 1.680 letras. O modelo alcançou acurácia global de 80,48% e F1-score macro de 0,764.

A Tabela 2 apresenta o desempenho detalhado por categoria emocional, incluindo a média ponderada das métricas obtidas.

Tabela 2 – Desempenho do Modelo B (BERTimbau *fine-tuning* em letras)

Emoção	**Precisão**	**Revocação**	**F1-Score**	**Suporte**
Neutro	0,770	0,910	0,834	210
Alegria	0,837	0,635	0,722	340
Tristeza	0,736	0,869	0,797	336
Nostalgia	0,746	0,539	0,626	76
Amor composto	0,900	0,850	0,874	614
Saudade	0,634	0,865	0,732	104
Média macro	0,770	0,778	0,764	1.680
Média ponderada	0,815	0,805	0,803	1.680

Fonte: Elaborada pela autora (2026).

Os resultados demonstram desempenho satisfatório na maior parte das categorias emocionais. A classe amor composto apresentou o maior F1-Score (0,874), acompanhado de elevada precisão (0,900) e revocação (0,850), indicando boa capacidade do modelo em identificar padrões afetivos recorrentes nas letras da MPB. As categorias tristeza (F1 = 0,797), saudade (F1 = 0,732) e alegria (F1 = 0,722) também apresentaram desempenho satisfatório, embora algumas delas tenham apresentado diferenças mais acentuadas entre precisão e revocação.

A categoria nostalgia apresentou o menor desempenho (F1 = 0,626), resultado compatível com o reduzido número de exemplos disponíveis no conjunto de teste ($n = 76$) e com a elevada proximidade semântica entre nostalgia e saudade. Em diversos contextos, ambas compartilham vocabulário e construções linguísticas semelhantes, dificultando sua distinção automática.

A categoria neutro apresentou F1-Score de 0,834 e elevada revocação (0,910), indicando boa capacidade do modelo em identificar letras sem predominância de vocabulário emocional, conforme o critério de rotulação adotado.

A consistência da rotulação *silver* foi avaliada por meio do coeficiente Kappa de Cohen, obtendo-se $\kappa = 0,747$, indicando concordância substancial segundo os critérios de Landis e Koch (1977). Esse resultado indica elevada concordância entre os rótulos utilizados durante o treinamento e as predições produzidas pelo modelo no conjunto de teste.

Comparando-se os resultados obtidos com os do modelo de comentários, observa-se que o modelo treinado sobre letras apresentou desempenho ligeiramente superior, alcançando acurácia de 80,48% contra 79,89% e F1-score macro de 0,764 contra 0,744. Essa diferença pode ser atribuída a fatores complementares. Em primeiro lugar, os dois modelos foram treinados a partir de estratégias distintas de geração dos rótulos *silver*. Enquanto o corpus de comentários utilizou supervisão fraca baseada

no Snorkel, combinando múltiplas *Labeling Functions* e um modelo de linguagem, o corpus de letras foi rotulado por meio de um léxico expandido, o que pode ter influenciado os rótulos *silver* utilizados no treinamento de cada modelo. Além disso, o comportamento observado também é compatível com as características linguísticas dos corpora utilizados. As letras da MPB apresentam maior extensão, organização sintática mais regular e vocabulário emocional relativamente estável, favorecendo a aprendizagem de padrões linguísticos. Em contraste, os comentários publicados no YouTube são mais curtos, informais e frequentemente contêm gírias, abreviações, ironias, emojis e referências contextuais, tornando a tarefa de classificação emocional mais desafiadora.

A evolução das curvas de treinamento reforça essa interpretação. A menor perda de validação foi obtida na segunda época (*Validation Loss* = 0,636), enquanto nas épocas subsequentes esse valor passou a aumentar, mesmo com a redução da perda de treinamento. Esse comportamento caracteriza o início do sobreajuste (*overfitting*), justificando a seleção automática do *checkpoint* com menor perda de validação para compor o modelo final.

4.3. ANÁLISE EMOCIONAL DAS LETRAS DA MPB

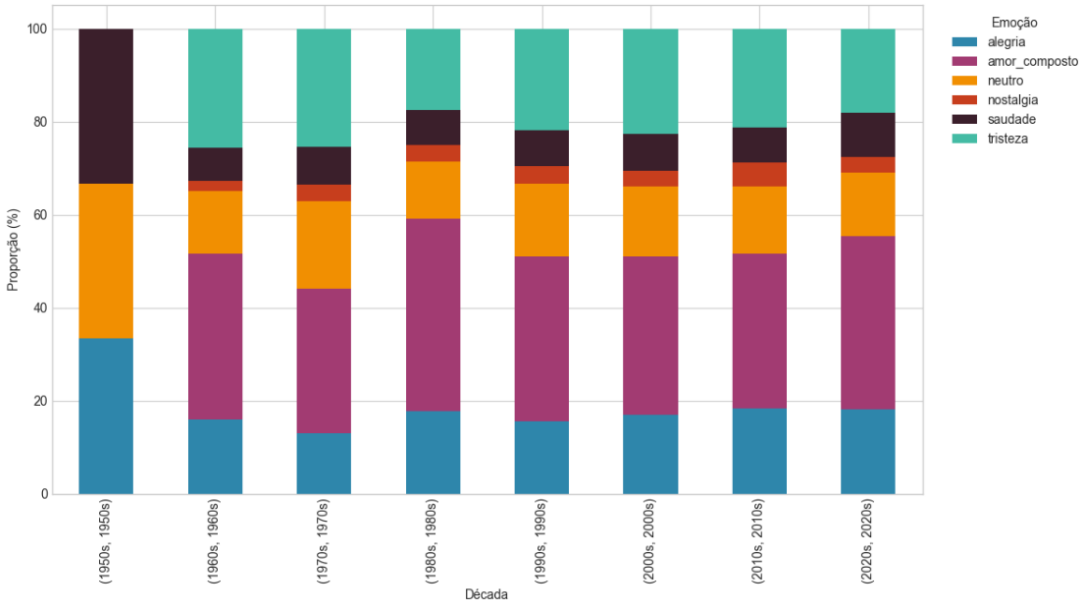
O modelo de letras foi aplicado ao corpus completo de 8.400 letras, permitindo identificar a emoção predominante em cada composição. A emoção amor composto predominou em 35,0% das letras analisadas (2.940 letras), seguida por tristeza (21,6%), alegria (16,8%), neutro (15,0%), saudade (7,9%) e nostalgia (3,7%). Esses resultados indicam que emoções associadas ao afeto e aos relacionamentos ocupam posição de destaque no corpus analisado. Essa predominância é consistente com a elevada frequência da categoria amor composto observada tanto na análise por década quanto no perfil emocional dos artistas, apresentados nas Figuras 3 e 4.

Esse comportamento é coerente com o achado central de Gama (2024), que identificou o amor como tema predominante em todas as décadas da MPB. Diferentemente daquele estudo, baseado em análise temática, neste trabalho essa predominância foi identificada por meio de um modelo BERTimbau ajustado por *fine-tuning* e validado quantitativamente para classificação emocional.

A análise por década revelou que, desconsiderando a década de 1950, representada por apenas seis composições, a categoria amor composto permaneceu como a emoção predominante em todos os períodos compreendidos entre as décadas de 1960 e 2020. Observam-se variações nas proporções das demais categorias emocionais ao longo do tempo, com maior participação relativa da categoria tristeza nas décadas iniciais e aumento da frequência da categoria alegria nas décadas mais recentes. A categoria nostalgia permaneceu pouco frequente durante todo o período analisado.

A Figura 3 apresenta a distribuição das categorias emocionais predominantes nas letras da MPB ao longo das décadas analisadas.

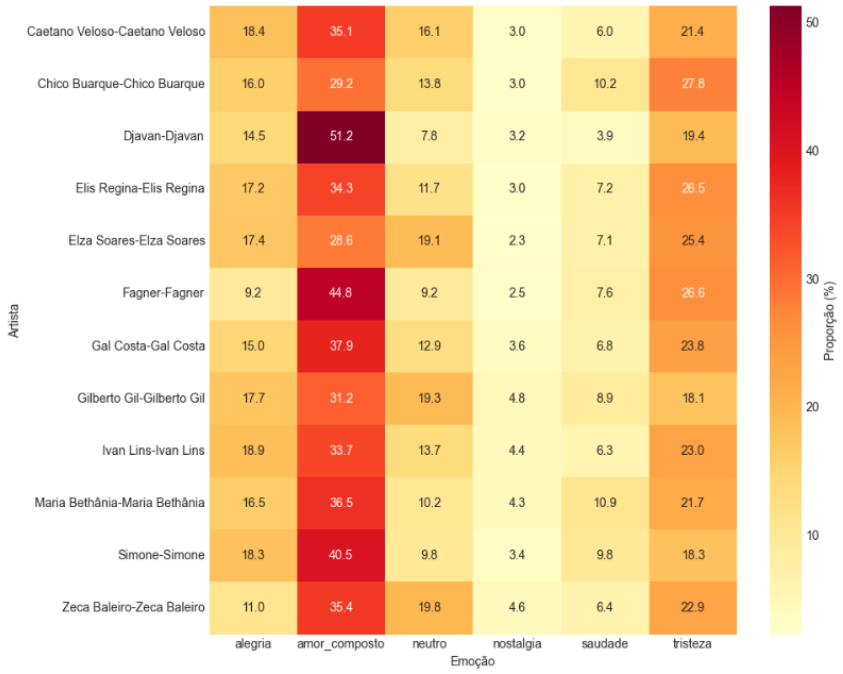
Figura 3 – Distribuição emocional por década



Fonte: Elaborada pela autora (2026).

A Figura 4 detalha o perfil emocional dos 12 artistas com maior volume de letras no *corpus*.

Figura 4 – Perfil emocional dos artistas da MPB



Fonte: Elaborada pela autora (2026).

Conforme observado na Figura 4, a categoria *amor composto* permaneceu como a emoção predominante entre os principais artistas analisados, embora sua concentração varie consideravelmente entre os repertórios. Djavan apresentou a maior concentração nessa classe (51,2%), seguido por Fagner (44,8%), evidenciando uma forte inclinação temática voltada às relações afetivas. Em contrapartida, Gilberto Gil exibiu uma distribuição mais balanceada entre as classes, sem hegemonia acentuada de uma única emoção. Além disso, as categorias *nostalgia* e *saudade* mantiveram-se pouco frequentes em todos os perfis destacados, enquanto *tristeza* e *alegria* assumiram proporções intermediárias conforme a estética lírica de cada intérprete.

4.4. ANÁLISE COMPARATIVA: COMPOSITOR VERSUS OUVINTE

A principal contribuição deste trabalho consiste na comparação entre a emoção expressa nas letras, representando a perspectiva da composição (inferida pelo modelo treinado sobre letras), e a emoção percebida nos comentários do YouTube, representando a perspectiva do ouvinte (inferida a partir da emoção predominante nos comentários classificados pelo modelo de comentários). Essa comparação foi realizada sobre um conjunto de 4.479 músicas obtido por meio do cruzamento entre os corpora de letras e de comentários, utilizando título e artista como chaves de correspondência e mantendo apenas as músicas presentes em ambas as bases.

A utilização de modelos especializados para cada domínio textual, aliada à estratégia de supervisão fraca empregada na rotulação dos comentários e enriquecida pelo julgamento semântico do Gemini 2.5 Flash, permitiu comparar duas perspectivas emocionais distintas de uma mesma obra musical. Essa abordagem difere da adotada em estudos anteriores, que normalmente empregavam um único modelo para ambos os tipos de texto.

A taxa global de convergência emocional foi de 26,3% ($\kappa = 0,053$), indicando que a emoção prevista pelo modelo de letras coincidiu com a emoção predominante observada nos comentários da mesma música, classificados pelo modelo de comentários. Em contrapartida, 73,7% das músicas apresentaram divergência entre essas duas perspectivas.

Diferentemente do coeficiente Kappa empregado anteriormente para avaliar a concordância entre as predições dos modelos e seus respectivos rótulos *silver*, nesta análise o coeficiente foi utilizado para medir a concordância entre duas perspectivas distintas: a emoção atribuída às letras pelo modelo especializado e a emoção predominante observada nos comentários dos ouvintes. Assim, esse valor não representa o desempenho dos modelos de classificação, mas o grau de convergência emocional entre compositor e ouvinte.

A Tabela 3 apresenta a taxa de convergência por categoria emocional do compositor,

enquanto a Figura 5 ilustra a distribuição completa das convergências e divergências entre as emoções atribuídas pelo compositor e pelo ouvinte.

Tabela 3 – Taxa de convergência emocional por categoria do compositor

Emoção do Compositor	**Taxa de Convergência (%)**
Amor composto	41,7
Alegria	40,2
Nostalgia	14,8
Tristeza	13,4
Saudade	11,6
Neutro	1,9
Global	26,3

Fonte: Elaborada pela autora (2026).

Os resultados evidenciam que a convergência emocional não ocorre de maneira uniforme entre as categorias avaliadas. Enquanto emoções de valência positiva e afetuosa apresentaram maiores taxas de concordância como *amor composto* (41,7%) e *alegria* (40,2%), categorias de valência negativa ou reflexiva exibiram índices expressivamente menores, como *nostalgia* (14,8%), *tristeza* (13,4%) e *saudade* (11,6%).

A concentração acentuada de predições dos ouvintes nas classes *alegria* e *amor composto* reflete um fenômeno duplo, metodológico e comportamental. Sob a ótica comportamental, as plataformas digitais como o YouTube apresentam um viés de autoseleção e positividade inerente: os usuários tendem a interagir e registrar comentários impulsionados por sentimentos de admiração, fruição estética e carinho pelo intérprete ou pela obra (expressos por meio de elogios como *"música maravilhosa"* ou *"amo essa canção"*). Dessa forma, mesmo diante de composições com lírica estritamente triste ou melancólica, a manifestação discursiva do ouvinte na seção de comentários canaliza-se frequentemente como uma reação afetiva positiva.

Sob a perspectiva da psicologia da música (JUSLIN; SLOBODA, 2010), esse padrão dialoga diretamente com a distinção entre a emoção transmitida pelo texto e o afeto despertado no ouvinte durante a fruição da obra. A assimetria observada, portanto, não denota viés algorítmico dos modelos, mas evidencia empiricamente a lacuna semântico-perceptual apontada por Hu *et al.* (2026), comprovando que a recepção pública tende a ressignificar o conteúdo lírico em termos de valor afetivo e exaltação do artista.

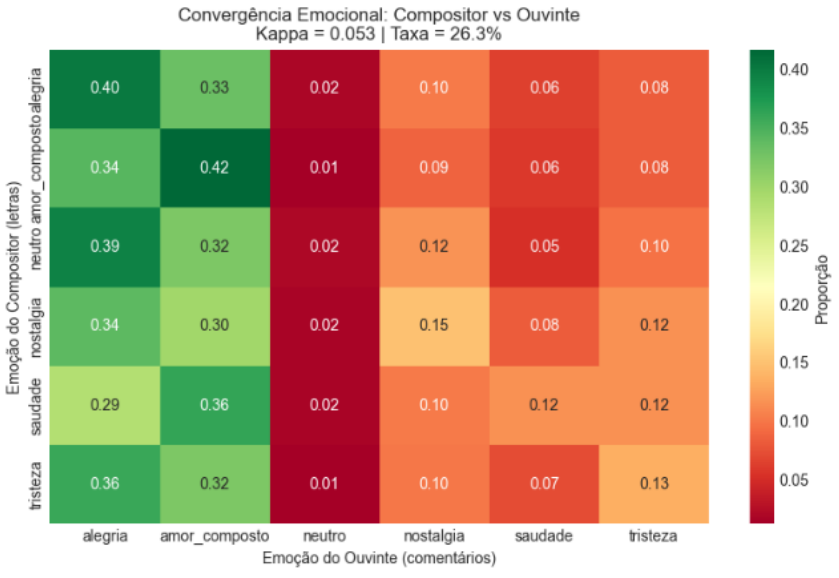
O baixo valor do coeficiente Kappa não indica falha dos modelos nem inconsistência dos dados, mas reflete a diferença entre duas perspectivas distintas: a emoção expressa pelo compositor na letra e a emoção percebida pelo ouvinte durante a recepção da obra. Conforme discutido na Seção 2.5, a literatura distingue essas duas perspecti-

vas, fenômeno descrito por (Hu *et al.*, 2026) como uma lacuna entre a emoção expressa e a emoção percebida e amplamente reconhecido na psicologia da música (Juslin e Sloboda, 2010). Nesse contexto, os resultados obtidos corroboram essa perspectiva ao evidenciarem que a emoção predominante identificada nas letras nem sempre coincide com a emoção predominante percebida pelos ouvintes nos comentários da mesma música.

O comportamento observado na Tabela 3 revela um padrão interessante. Quando o compositor expressa amor composto, aproximadamente quatro em cada dez músicas apresentam a mesma emoção predominante na percepção dos ouvintes (41,7%). Para alegria, essa concordância foi de 40,2%. Em contraste, emoções como nostalgia (14,8%), tristeza (13,4%), saudade (11,6%) e, principalmente, neutro (1,9%) apresentaram taxas de convergência consideravelmente menores. Esses resultados sugerem que parte das emoções negativas expressas nas letras tende a ser percebida pelos ouvintes como emoções de valência mais positiva durante a recepção das obras.

Esse comportamento reforça a hipótese de que a emoção percebida pelo ouvinte não depende exclusivamente do conteúdo textual da letra, mas também da interpretação individual, do contexto sociocultural, das experiências pessoais e de aspectos musicais não analisados neste trabalho, como melodia, harmonia, timbre e interpretação vocal.

Figura 5 – Matriz de convergência emocional entre compositor e ouvinte



Fonte: Elaborada pela autora (2026).

A análise da matriz de convergência permite identificar dois padrões principais de divergência. O primeiro ocorre entre emoções semanticamente próximas, especialmente entre amor composto e alegria. As combinações amor composto $\rightarrow$ alegria (546 ocorrências), alegria $\rightarrow$ amor composto (251 ocorrências) e amor composto $\rightarrow$ nostal-

gia (140 ocorrências) constituíram alguns dos pares de divergência mais frequentes, refletindo a proximidade semântica existente entre essas categorias.

Também se destacaram as combinações alegria $\rightarrow$ amor composto (251 ocorrências) e amor composto $\rightarrow$ nostalgia (140 ocorrências), reforçando a existência de diferentes padrões de interpretação emocional entre compositor e ouvinte.

O segundo padrão envolve mudanças entre emoções de valências distintas. Destacam-se as combinações tristeza $\rightarrow$ alegria (343 ocorrências), tristeza $\rightarrow$ amor composto (306 ocorrências) e saudade $\rightarrow$ amor composto (131 ocorrências). Em conjunto, esses resultados sugerem um padrão recorrente de percepção de emoções de valência mais positiva pelos ouvintes em relação às emoções expressas nas letras.

Exemplos ilustrativos desse comportamento incluem “Graffitis”, de Adriana Calcanhotto, classificada pelo modelo de letras como amor composto e percebida predominantemente como nostalgia pelos ouvintes, e “Luz dos Olhos”, de Cássia Eller, classificada como alegria nas letras e percebida como saudade nos comentários. Também foram observados casos como “O Homem Que Eu Amo”, de Fafá de Belém, classificada como amor composto nas letras e como tristeza pelos ouvintes. Esses exemplos ilustram como a emoção percebida pelos ouvintes pode diferir da emoção expressa na letra da música.

Cabe ressaltar que os comentários do YouTube constituem uma amostra de conveniência e podem apresentar viés de seleção, uma vez que usuários tendem a comentar músicas com as quais possuem maior identificação ou envolvimento emocional. Essa característica deve ser considerada na interpretação dos resultados. Ainda assim, a recorrência do padrão observado em 4.479 músicas, distribuídas entre 40 artistas, sugere que a assimetria identificada não ocorreu de forma isolada, constituindo uma hipótese consistente sobre a relação entre a emoção expressa nas letras e a emoção percebida pelos ouvintes no contexto da MPB.

4.5. MODELAGEM DE TÓPICOS POR EMOÇÃO

A modelagem de tópicos baseada em *Latent Dirichlet Allocation* (LDA) foi aplicada separadamente às letras classificadas em cada categoria emocional pelo modelo de letras. O LDA é uma técnica não supervisionada utilizado para identificar automaticamente tópicos recorrentes em um conjunto de documentos, agrupando palavras que frequentemente ocorrem em conjunto. Diferentemente de Gama (2024), que organizou a modelagem por décadas, neste trabalho os tópicos foram extraídos a partir das categorias emocionais identificadas pelo modelo de classificação.

Em todas as categorias emocionais, o modelo LDA foi configurado para extrair três tópicos ($k = 3$), executando 15 iterações de treinamento sobre o *corpus*. O vocabulário foi previamente filtrado para desconsiderar termos com frequência inferior

a dois documentos ou presentes em mais de 90% dos textos, visando mitigar ruídos causados por termos excessivamente raros ou hiperfrequentes.

A qualidade e a interpretabilidade dos tópicos gerados foram avaliadas por meio da métrica de Coerência de Tópicos C_v (Röder, Both e Hinneburg, 2015), que quantifica o grau de consistência semântica entre os termos mais representativos de cada agrupamento com base em janelas de coocorrência e similaridade de cosseno. Valores superiores de C_v indicam representações temáticas mais nítidas e semanticamente coesas. Os escores médios obtidos variaram entre 0,285 (*nostalgia*) e 0,403 (*tristeza*), conforme detalhado na Tabela 4.

Os valores moderados de C_v obtidos são característicos da modelagem de tópicos em textos artísticos e poéticos, cuja polissemia, linguagem figurada e extensão reduzida impõem desafios adicionais à coocorrência estrita de termos. Ademais, o menor volume de amostras em subcoleções como *nostalgia* ($n = 314$) e *saudade* ($n = 666$) influencia diretamente a densidade estatística das matrizes de coocorrência. Esse comportamento alinha-se aos achados de Gama (2024), que também reportou escores moderados de coerência semântica ao aplicar LDA sobre subconjuntos de letras da MPB. Não obstante as variações numéricas de C_v, a inspeção qualitativa confirmou que os agrupamentos lexicais gerados mantiveram interpretabilidade temática consistente com as respectivas categorias afetivas.

A Tabela 4 apresenta o número de letras utilizado em cada modelo e os respectivos valores de coerência Cv obtidos.

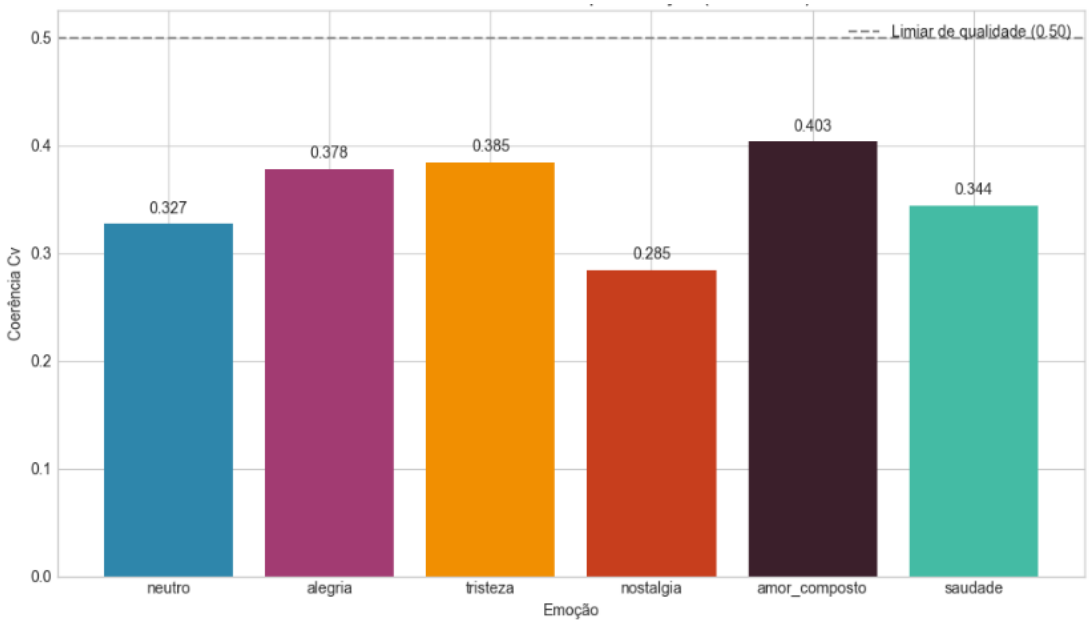
Tabela 4 – Coerência dos modelos LDA por categoria emocional

Emoção	**Número de letras (n)**	**Coerência Cv**
Amor composto	2.940	0,403
Tristeza	1.813	0,385
Alegria	1.410	0,378
Saudade	666	0,344
Neutro	1.257	0,327
Nostalgia	314	0,285

Fonte: Elaborada pela autora (2026).

A Figura 6 apresenta a comparação dos valores de coerência Cv obtidos para os modelos LDA em cada categoria emocional, permitindo visualizar as diferenças de desempenho entre os modelos gerados para cada emoção.

Figura 6 – Coerência dos modelos LDA por categoria emocional



Fonte: Elaborada pela autora (2026).

Na categoria amor composto, que apresentou o maior valor de coerência (Cv = 0,403), os tópicos foram compostos predominantemente por termos relacionados às relações afetivas, como amor, coração e vida, reforçando a predominância dessa temática nas letras analisadas. Esse resultado é consistente com a distribuição emocional apresentada anteriormente e corrobora o estudo de Gama (2024), que identificou o amor como tema recorrente na MPB.

Na categoria tristeza, observaram-se tópicos compostos por termos recorrentes como vida, dia, mar, sol, amor e quero. Embora também contenham palavras de alta frequência na língua portuguesa, os tópicos apresentam padrões lexicais compatíveis com o conjunto de letras classificadas nessa emoção.

Os tópicos da categoria saudade apresentaram recorrência de termos como saudade, Deus, amor, mar e tempo, evidenciando associações frequentes entre lembrança, espiritualidade e relações afetivas nas letras classificadas nessa emoção.

Na categoria nostalgia, que apresentou a menor coerência (Cv = 0,285), observaram-se termos relacionados à passagem do tempo, como tempo, anos e dia. Apesar da maior diversidade lexical, esses tópicos sugerem associação predominante com referências temporais e memórias do passado.

Na categoria alegria, destacaram-se palavras como vida, amor, vem, vou e bom, indicando maior frequência de vocabulário associado a experiências positivas, dinamismo e relações afetivas.

Em conjunto, os resultados da modelagem de tópicos complementam a análise quantitativa realizada pelo modelo de classificação emocional, permitindo identificar os

principais conjuntos lexicais associados a cada emoção presente nas letras da MPB.

4.6. AMEAÇAS À VALIDADE

Os resultados deste trabalho devem ser interpretados considerando algumas limitações decorrentes das escolhas metodológicas adotadas.

A primeira ameaça refere-se às diferenças entre os processos de geração dos rótulos *silver* utilizados no treinamento dos modelos. Enquanto o modelo de letras foi treinado com rótulos produzidos por um léxico expandido, o modelo de comentários utilizou supervisão fraca baseada no Snorkel, combinando múltiplas *Labeling Functions* e inferências produzidas pelo Gemini 2.5 Flash. Dessa forma, parte das divergências observadas entre as emoções atribuídas às letras e aos comentários pode decorrer de diferenças sistemáticas entre os processos de supervisão adotados, e não exclusivamente de diferenças emocionais entre compositor e ouvinte. Assim, a comparação entre as duas perspectivas deve ser interpretada como uma análise exploratória, e não como uma medida definitiva da divergência emocional entre ambas. A utilização de um conjunto de referência anotado manualmente para ambos os domínios textuais constitui uma oportunidade para trabalhos futuros.

A segunda ameaça está relacionada ao viés de seleção presente nos comentários do YouTube. Usuários tendem a comentar principalmente músicas com as quais possuem maior envolvimento ou identificação emocional, o que pode favorecer a manifestação de emoções positivas, como alegria e amor composto, na perspectiva do ouvinte. Consequentemente, os resultados referentes à comparação entre compositor e ouvinte devem ser interpretados considerando essa característica da plataforma. Além disso, por terem sido coletados exclusivamente do YouTube, os comentários refletem o perfil dos usuários e a dinâmica de interação dessa plataforma, não sendo possível assumir que os mesmos padrões emocionais ocorreriam em outros ambientes digitais ou públicos distintos.

A terceira ameaça refere-se ao emprego do modelo de linguagem *Gemini 2.5 Flash* como uma das fontes de rotulação durante a supervisão fraca. Por se tratar de um modelo proprietário, seu comportamento pode sofrer alterações entre versões, além de não disponibilizar informações detalhadas sobre seu processo interno de decisão. Embora suas respostas tenham sido utilizadas apenas como uma das fontes agregadas pelo *Label Model*, essa dependência constitui uma limitação quanto à reprodutibilidade completa do processo de anotação.

A quarta ameaça está associada à delimitação do corpus. O corpus final utilizado neste estudo foi composto por 8.400 letras pertencentes a 40 artistas da Música Popular Brasileira. A seleção dos artistas considerou sua relevância e popularidade no contexto da MPB, buscando contemplar diferentes períodos e gerações da música brasileira.

Outra limitação refere-se à origem das letras utilizadas na construção do corpus. Embora a plataforma *Letras.mus.br* disponibilize um amplo acervo de músicas da Música Popular Brasileira, o conjunto analisado depende das obras disponíveis na plataforma e dos critérios de seleção adotados nesta pesquisa. Assim, artistas, composições ou períodos não contemplados podem apresentar distribuições emocionais distintas das observadas neste estudo.

Além disso, a classificação emocional foi restrita às seis categorias adotadas neste trabalho (alegria, tristeza, amor composto, saudade, nostalgia e neutro). Embora essas categorias tenham sido suficientes para atender aos objetivos da pesquisa, elas não contemplam toda a complexidade das emoções humanas. Emoções como medo, raiva, surpresa, esperança ou culpa, por exemplo, não foram consideradas durante o processo de classificação. Consequentemente, algumas nuances emocionais presentes nas letras e nos comentários podem não ter sido capturadas pelos modelos desenvolvidos.

Por fim, os modelos de tópicos baseados em LDA apresentaram valores moderados de coerência Cv, especialmente na categoria nostalgia (Cv = 0,285). Embora os tópicos obtidos tenham possibilitado interpretações coerentes com os conjuntos lexicais observados, os valores de coerência indicam que esses modelos devem ser interpretados com cautela, particularmente nas categorias com menor número de documentos.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem para análise computacional de emoções na Música Popular Brasileira (MPB) por meio de uma arquitetura composta por dois modelos BERTimbau ajustados por *fine-tuning* e especializados em diferentes domínios textuais. O modelo destinado à classificação de comentários foi treinado utilizando rótulos obtidos por supervisão fraca, combinando o *framework* Snorkel e o modelo de linguagem *Gemini 2.5 Flash*, enquanto o modelo de letras foi treinado a partir de rótulos gerados por um léxico emocional expandido. Os resultados obtidos evidenciaram desempenho satisfatório em ambos os cenários, alcançando acurácia de 79,89% e coeficiente Kappa de Cohen de 0,734 para o modelo de comentários, e acurácia de 80,48% e Kappa de 0,747 para o modelo de letras, indicando concordância substancial entre as predições dos modelos e os respectivos rótulos *silver* utilizados como referência durante a avaliação.

A principal contribuição desta pesquisa consistiu em investigar, de forma integrada, a relação entre a emoção expressa nas letras da MPB e a emoção percebida pelos ouvintes a partir dos comentários publicados no YouTube. Diferentemente de abordagens que analisam apenas o conteúdo das letras ou apenas as reações do público, este trabalho reuniu ambas as perspectivas em um mesmo processo analítico, permitindo observar como uma mesma obra pode despertar interpretações emocionais distintas durante sua recepção.

A análise realizada sobre 4.479 músicas revelou uma taxa global de convergência emocional de 26,3%, sugerindo que apenas aproximadamente um quarto das músicas apresentou coincidência entre a emoção inferida para a letra e a emoção predominante observada nos comentários. Além disso, verificou-se que emoções de valência positiva apresentaram maiores taxas de convergência do que emoções de valência negativa, enquanto categorias como tristeza e saudade foram frequentemente reinterpretadas pelos ouvintes como amor composto ou alegria. Esses resultados corroboram estudos da psicologia da música que distinguem a emoção expressa pela obra da emoção percebida durante sua escuta, reforçando que a experiência emocional da música é influenciada não apenas pelo conteúdo da letra, mas também pela subjetividade do indivíduo e pelo contexto de recepção.

A modelagem de tópicos baseada em LDA complementou a classificação automática ao identificar padrões lexicais característicos para cada categoria emocional. Embora os modelos tenham apresentado valores moderados de coerência, especialmente nas categorias com menor quantidade de documentos, os tópicos extraídos mostraram-se semanticamente interpretáveis, oferecendo uma perspectiva qualitativa complementar que auxiliou na interpretação dos padrões emocionais identificados pelos modelos de classificação.

Do ponto de vista metodológico, este trabalho também contribui ao demonstrar a viabilidade da utilização de supervisão fraca para construção de conjuntos de treinamento em português, reduzindo a necessidade de grandes volumes de dados anotados manualmente. A combinação entre funções de rotulação, modelos de linguagem, validação por meio do coeficiente Kappa de Cohen e modelos especializados para diferentes domínios textuais constitui uma estratégia reproduzível que pode ser adaptada para outras tarefas de classificação emocional em Processamento de Linguagem Natural.

As limitações da pesquisa incluem o emprego de rótulos *silver* em substituição a anotações humanas, a utilização de estratégias distintas para geração dos rótulos *silver* dos corpora de letras e comentários, o viés inerente aos comentários publicados no YouTube, a dependência parcial de um modelo de linguagem proprietário durante o processo de supervisão fraca, a delimitação do corpus à Música Popular Brasileira e a restrição da classificação às categorias emocionais definidas neste estudo. Embora essas limitações não invalidem os resultados obtidos, elas delimitam o escopo das conclusões apresentadas. Em particular, a comparação entre as emoções inferidas para as letras e para os comentários deve ser interpretada como uma análise exploratória, uma vez que os dois modelos foram treinados com estratégias distintas de geração de rótulos *silver*. Essas limitações também indicam oportunidades para investigações futuras.

Como perspectivas para trabalhos futuros, destacam-se a construção de um corpus de referência anotado manualmente para validação das rotulações automáticas, a

ampliação da abordagem para outros gêneros musicais brasileiros, a incorporação de diferentes plataformas digitais para representar a percepção dos ouvintes, a utilização de múltiplos modelos de linguagem durante o processo de supervisão fraca e a exploração de abordagens multimodais que integrem informações provenientes das letras, do áudio e de outros elementos musicais, permitindo uma compreensão mais abrangente da experiência emocional associada às obras. Também se destaca como perspectiva futura a investigação de arquiteturas mais recentes baseadas em transformadores e grandes modelos de linguagem para apoiar as etapas de rotulação e classificação emocional.

Em síntese, esta pesquisa demonstra que a integração entre modelos especializados por domínio textual, técnicas de supervisão fraca e análise comparativa entre compositor e ouvinte constitui uma metodologia reproduzível para investigar emoções em músicas brasileiras. Para assegurar a reprodutibilidade integral dos experimentos, os *scripts* de coleta, as 19 *Labeling Functions*, o léxico afetivo com 176 termos, os *datasets* consolidados e as especificações de ambiente foram disponibilizados em repositório público de acesso aberto[6].

[6] Disponível em: <https://github.com/thavyne-KDR/mpb-emotion-analysis>. Acesso em: 20 ago. 2026.

REFERÊNCIAS

APARECIDO, T. D. D. *et al.* Benchmarking psychological lexicons and large language models for emotion detection in Brazilian Portuguese. **AI**, v. 6, n. 10, p. 249, 2025. Citado 3 vezes nas páginas 5, 6 e 9.

DEVLIN, J. *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)**. Minneapolis: [s.n.], 2019. p. 4171–4186. Citado na página 6.

EDMONDS, A.; SEDOC, J. Multi-emotion classification for song lyrics. In: **Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)**. [S.l.: s.n.], 2021. p. 221–235. Citado 3 vezes nas páginas 6, 9 e 16.

GAMA, J. M. N. **Análise de Emoções e Temas Presentes na Música Popular Brasileira com Processamento de Linguagem Natural**. Dissertação (Trabalho de Conclusão de Curso) — Universidade Federal de Campina Grande, Campina Grande, 2024. Citado 5 vezes nas páginas 9, 23, 28, 29 e 30.

Google DeepMind. **Gemini 2.5: Our Most Intelligent AI Model**. 2025. Acesso em: 6 jul. 2026. Disponível em: <https://deepmind.google/technologies/gemini/>. Citado na página 14.

HAMMES, K. R.; FREITAS, L. A. d. Classificação de emoções em textos no português brasileiro: Uma abordagem com bertimbau. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)**. Porto Alegre, RS, Brasil: SBC, 2021. p. 189–198. Citado 2 vezes nas páginas 6 e 8.

HU, Q. Q. H. *et al.* Revisiting the role of lyrics in music emotion recognition: A critical analysis of the semantic-perceived emotion gap and methodological challenges. **Information Fusion**, 2026. Citado 4 vezes nas páginas 6, 8, 26 e 27.

JUSLIN, P. N.; SLOBODA, J. A. **Handbook of Music and Emotion: Theory, Research, Applications**. Oxford: Oxford University Press, 2010. Citado 3 vezes nas páginas 8, 26 e 27.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977. Citado 2 vezes nas páginas 18 e 22.

LIU, B. **Sentiment Analysis and Opinion Mining**. San Rafael: Morgan & Claypool Publishers, 2012. Citado na página 6.

MOHAMMAD, S. M.; TURNEY, P. D. Crowdsourcing a word-emotion association lexicon. **Computational Intelligence**, v. 29, n. 3, p. 436–465, 2013. Citado na página 6.

NOGUEIRA, R. A. G.; CAPANEMA, D. O. **Análise de Sentimentos para a Recomendação de Músicas Baseada nos Sentimentos do Usuário**. Dissertação (Trabalho de Conclusão de Curso) — Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, 2023. Citado na página 6.

OLIVEIRA, F. B.; SICHMAN, J. S. Portuguese emotion detection model using BERTimbau applied to COVID-19 news and replies. In: **Intelligent Systems: BRACIS 2024**. [S.l.]: Springer, 2025, (Lecture Notes in Computer Science, v. 15413). p. 274–289. Citado na página 9.

PLUTCHIK, R. **Emotion: A Psychoevolutionary Synthesis**. New York: Harper & Row, 1980. Citado 3 vezes nas páginas 5, 12 e 13.

RATNER, A. *et al.* Snorkel: Rapid training data creation with weak supervision. **Proceedings of the VLDB Endowment**, v. 11, n. 3, p. 269–282, 2017. Citado na página 7.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: **Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)**. [S.l.: s.n.], 2015. p. 399–408. Citado na página 29.

SMITH, R. *et al.* Language models in the loop: Incorporating prompting into weak supervision. **ACM/IMS Journal of Data Science**, v. 1, n. 2, p. 1–30, 2024. Citado 4 vezes nas páginas 7, 15, 18 e 21.

SOUZA, F. *et al.* BERTimbau: Pretrained BERT models for brazilian portuguese. In: **Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.]: Springer, 2020. p. 403–417. Citado na página 7.

VASWANI, A. *et al.* Attention is all you need. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. Citado na página 6.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, presença constante em cada passo desta caminhada. Nos dias de medo e de cansaço, foi Sua mão que me sustentou e Sua voz silenciosa que me lembrou que eu nunca estive sozinha. Foi Ele quem transformou incertezas em fé e abriu, no tempo certo, os caminhos que eu não conseguia enxergar. Toda honra e toda glória pertencem a Ti.

À minha avó Elenice, que sempre foi minha mãe. Não existem palavras suficientes para expressar a gratidão que sinto por tudo o que você fez e faz por mim. Obrigada por nunca medir esforços para que eu pudesse estudar. Sei que a vida nunca foi fácil para nós; cada sacrifício que você fez, cada renúncia silenciosa, cada preocupação e cada oração me trouxeram até este momento. Esta conquista não é apenas minha, ela é nossa.

À minha avó Luzemildes, minha eterna saudade. Obrigada por ter segurado minhas mãos quando eu ainda dava os primeiros passos na vida. Mesmo que hoje você não esteja aqui para compartilhar este momento, sei que uma parte de você permanece viva em tudo aquilo que sou. Levo comigo o seu carinho e o seu amor. Sentirei sua falta para sempre.

À minha mãe, agradeço por sempre acreditar em mim. Obrigada por cada demonstração de carinho. Saber que você acreditava em mim me deu forças para continuar.

Aos meus tios, Kaelles e Kelson, agradeço por todo o apoio, por toda a ajuda e por estarem presentes sempre que precisei. O carinho e a dedicação de vocês fizeram diferença em muitos momentos desta caminhada.

Aos meus irmãos, Messi, Alípio e Silas, agradeço por fazerem parte da minha vida. Vocês representam uma parte muito importante de quem eu sou. Que Deus continue cuidando de cada um de vocês.

À Livia, agradeço de forma muito especial por estar comigo nos momentos mais difíceis desta graduação, entre risos, lágrimas, apresentações, trabalhos e inseguranças. Obrigada por nunca me deixar enfrentar essa caminhada sozinha.

Ao Filipe, agradeço por sua presença constante, por seu carinho e por sua dedicação a mim ao longo desta caminhada.

Ao meu orientador, Professor João Paulo Lima do Nascimento, agradeço pela paciência, pela confiança e pelos ensinamentos que levarei por toda a minha vida acadêmica e profissional.

Ao meu coorientador, Professor Rogério Figueiredo de Sousa, agradeço pela disponibilidade e pelos ensinamentos transmitidos durante o curso.

Por fim, agradeço a todas as pessoas que, de alguma forma, fizeram parte desta caminhada. A todos, uma vida longa e próspera.