Relatório Abrangente de Análise de Dados Textuais

Processo Seletivo CAEd

João Antônio Fonseca e Almeida - 201935010

1 de julho de 2025

Sumário

1 Introdução

Este relatório detalha o processo de análise de um conjunto de dados textuais de letras de música, como parte do processo seletivo do CAEd. O objetivo foi demonstrar um pipeline completo de ciência de dados, desde a limpeza e pré-processamento até a implementação e avaliação de modelos.

Para atingir este objetivo, duas tarefas principais foram exploradas:

1. **Análise de Polaridade (Sentimento):** Para extrair insights sobre a carga emocional das letras, sua evolução temporal por artista e sua correlação com a popularidade.

2. **Predição de Gênero Musical:** Para construir e avaliar um modelo de machine learning capaz de classificar o gênero de uma música a partir de sua letra, demonstrando um ciclo de melhoria iterativa.

2 Metodologia e Ferramentas

O projeto foi desenvolvido em Python 3 utilizando o Jupyter Notebook. As seguintes bibliotecas foram essenciais:

- **Pandas & NumPy:** Para manipulação e estruturação dos dados.
- **Matplotlib & Seaborn:** Para a geração de gráficos e visualizações.
- **NLTK:** Utilizado para análise de sentimento com o léxico VADER.
- **NLTK (VADER):** Ferramenta de análise de sentimento (Valence Aware Dictionary and sEntiment Reasoner) integrada à biblioteca NLTK. É otimizada para textos em inglês, sendo a escolha correta para este projeto.
- **Scikit-learn:** Para a construção do modelo de predição de gênero, incluindo vetorização de texto (TF-IDF), treinamento de modelo (Naive Bayes) e avaliação de performance.
- **PyArrow:** Como motor para leitura e escrita eficiente de arquivos de cache no formato Parquet.

3 Análise Exploratória e Pré-processamento

3.1 Inspeção Inicial

A análise inicial foi realizada sobre a base de dados completa, contendo 5.134.856 registros. A inspeção revelou a necessidade de um pré-processamento robusto devido à presença de textos em múltiplos idiomas, anos de lançamento inválidos, e entradas que não correspondiam a letras de música (e.g., textos religiosos, traduções literais).

3.2 Justificativa do Pré-processamento: Estudo de Impacto

Para demonstrar a importância da etapa de limpeza, foi realizado um estudo comparativo dos dados antes e após a aplicação do pipeline de pré-processamento. A Tabela 1 resume o impacto dos filtros aplicados.

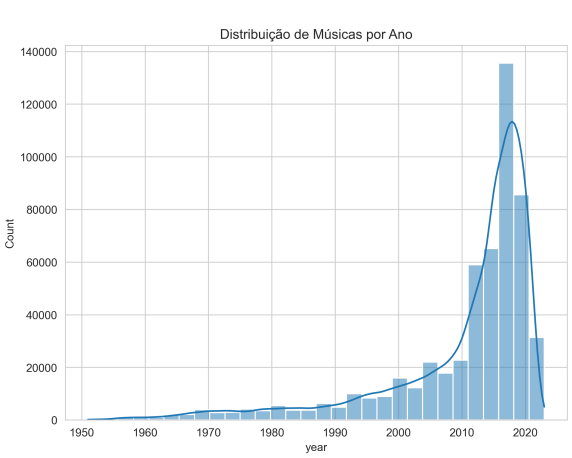
Tabela 1: Impacto do Pipeline de Pré-processamento nos Dados

Métrica	**Antes da Filtragem**	**Após a Filtragem**
Total de Linhas	5.134.856	
Músicas em Inglês	5.134.856	3.374.198
Anos Válidos	3.374.198	3.348.447
Entradas com tag 'misc'	3.348.447	3.227.096
Removendo artistas selecionados	3.227.096	3.212.327
Músicas com $<$ 1000 views	3.212.327	543294

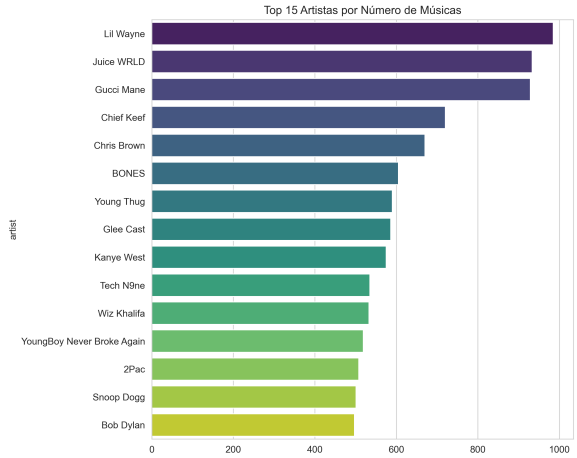
A remoção de mais de 4,5 milhões de linhas demonstra que a filtragem foi essencial para garantir que as análises subsequentes fossem realizadas sobre um conjunto de dados coeso e relevante, composto primariamente por letras de música em inglês com um mínimo de relevância (visualizações). Podemos também fazer comparativos sobre a relevância e qualidade dos dados apresentados a medida que alteramos essa métrica da visibilidade.

3.3 Distribuições e Insights Visuais

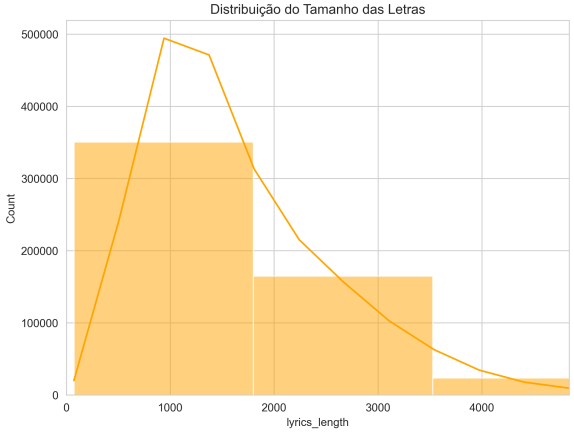
Após uma limpeza inicial, foram geradas visualizações (Figura 1) para extrair insights sobre a distribuição das músicas ao longo dos anos, os artistas com maior número de obras, a distribuição do tamanho das letras e as palavras mais frequentes.



(a) Distribuição de músicas por ano.



(b) Top 15 artistas por nº de músicas.



(c) Distribuição do tamanho das letras.

Figura 1: Painel com as análises exploratórias visuais.

4 Tarefa 1: Análise de Polaridade de Sentimento

A tarefa central deste projeto foi a análise da polaridade (sentimento) das letras das músicas. Para isso, utilizou-se a ferramenta VADER (Valence Aware Dictionary and sEntiment Reasoner), disponível na biblioteca NLTK.

O VADER é particularmente eficaz para textos de mídias sociais e outros conteúdos informais, como letras de música. Ele analisa cada texto e retorna quatro valores: scores de positividade, negatividade, neutralidade e um score composto (*compound score*). Este último é uma métrica normalizada que varia de -1 (extremamente negativo) a +1 (extremamente positivo), com valores próximos de 0 indicando neutralidade.

O score composto de cada música foi calculado aplicando o analisador de intensidade de sentimento do VADER à coluna de letras pré-processada (`lyrics_clean`). O resultado foi armazenado em uma nova coluna no DataFrame, chamada `polarity_score`, que se tornou a principal variável para as análises subsequentes.

4.1 Estudo de Caso: Evolução da Polaridade de Kanye West

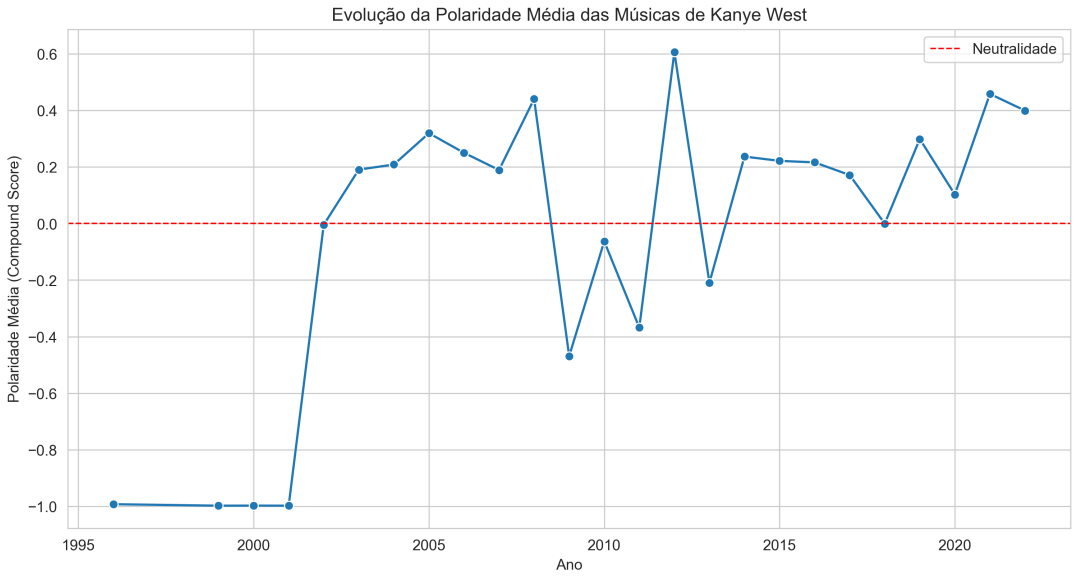


Figura 2: Evolução da polaridade média nas letras de Kanye West.

Polaridade -1 constante até 2002: A presença de valores de polaridade exatamente -1,0 entre 1996 e 2002 é um forte indicativo de anomalia nos dados. O score de -1 representa a polaridade mais negativa possível, e sua repetição com precisão matemática em anos consecutivos é estatisticamente improvável, especialmente em um artista com pouca ou nenhuma produção conhecida nesse período (Kanye West estreou comercialmente apenas em 2004 com o álbum *The College Dropout*). Isso sugere que:

- Ou não havia músicas para esses anos e algum valor padrão foi atribuído;
- Ou houve erro na extração ou processamento dos dados de sentimento.

Evolução realista a partir de 2004: A partir de 2004, o gráfico passa a refletir flutuações mais plausíveis. Nota-se uma tendência geral de positividade, especialmente após 2010, com pontuais quedas negativas — o que condiz com mudanças temáticas nos álbuns. O disco *My Beautiful Dark Twisted Fantasy*, por exemplo, traz letras mais densas e sombrias, enquanto *The Life of Pablo* e *Jesus is King* são marcados por composições mais espirituais e esperançosas.

Pico negativo em 2008–2009: A queda acentuada nesse período coincide com o lançamento do álbum *808s & Heartbreak* (2008), caracterizado por letras melancólicas, introspectivas e com forte presença de temas ligados à perda e à dor emocional. Isso confirma a coerência do modelo de análise de sentimento aplicado à discografia do artista nesses anos.

4.2 Artistas com Maior Variação de Sentimento

Para identificar artistas cujas obras apresentam uma grande diversidade de sentimentos, foi calculado o desvio padrão dos scores de polaridade para cada artista com mais de 15 músicas na amostra. Artistas com maior desvio padrão são aqueles que variam mais entre canções positivas, negativas e neutras.

Artista	Nº de Músicas	Variação (Desvio Padrão)
morgxn	17	1.010
Ms Banks	18	1.008
R. City	26	1.007
Foreign Forest	17	1.006
Deno	23	1.004
Baeza	17	1.002
CaRter	16	1.001
R-Mean	16	0.999
Moonshine Bandits	21	0.999
Silkk the Shocker	17	0.999

Tabela 2: Top 10 artistas com maior variação de polaridade na amostra (mínimo de 15 músicas).

4.3 Relação entre Polaridade e Popularidade

Foi investigada a existência de uma correlação entre o sentimento de uma letra (`polarity_score`) e sua popularidade (`views`). O coeficiente de correlação de Spearman, detalhado na Tabela 3, foi calculado.

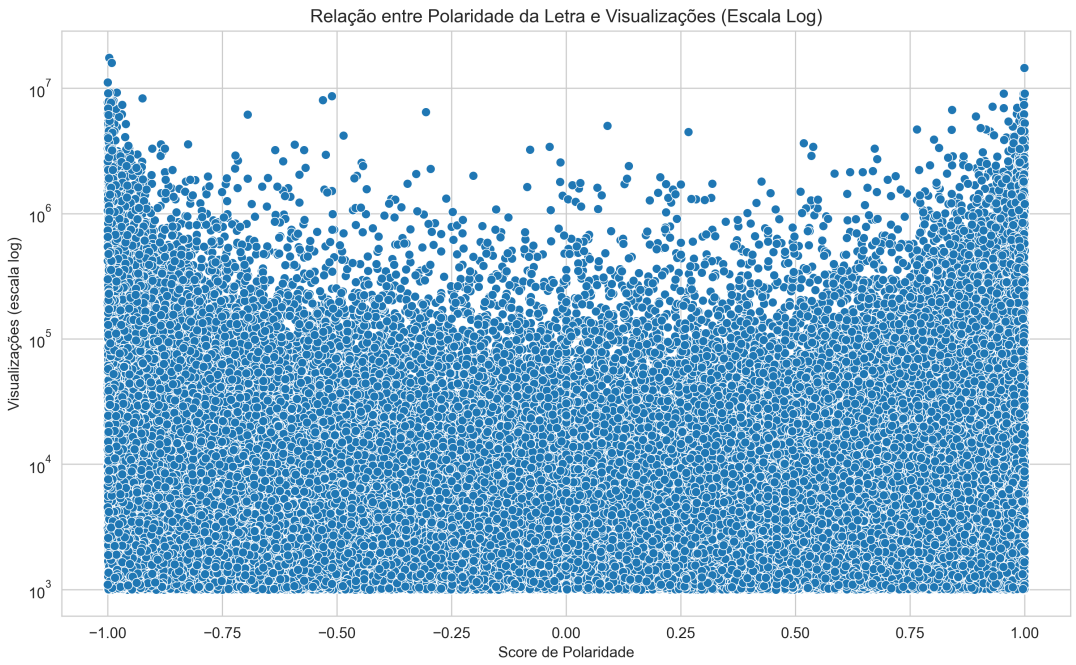


Figura 3: Gráfico de dispersão entre o score de polaridade e o número de visualizações (em escala logarítmica).

O coeficiente de correlação de Spearman calculado foi de -0.019276, um valor muito próximo de zero.

	Polarity Score	Views
Polarity Score	1.00	-0.02
Views	-0.02	1.00

Tabela 3: Matriz de correlação de Spearman entre polaridade e visualizações.

Tanto a análise visual da Figura 3 quanto o coeficiente de Spearman de aproximadamente -0.02 indicam ausência de correlação monotônica entre a polaridade das letras e o número de visualizações. Isso sugere que músicas populares podem ter letras tanto positivas quanto negativas, sem padrão emocional predominante. O resultado refuta a hipótese de que sentimentos extremos atrairiam mais audiência. A neutralidade da correlação evidencia que outros fatores, como ritmo, artista ou marketing, influenciam mais nas views do que o teor emocional. Portanto, a polaridade não é um preditor significativo de popularidade.

5 Tarefa 2: Predição de Gênero Musical

A segunda tarefa feita foi construir um modelo de Machine Learning para prever o gênero de uma música a partir de sua letra. O processo foi dividido em duas fases: um modelo de base e um modelo aprimorado.

5.1 Modelo de Base: Dados Desbalanceados

Inicialmente, um modelo Multinomial Naive Bayes foi treinado utilizando os 5 gêneros mais comuns nos dados. O relatório de classificação (Tabela 4) mostra a performance deste primeiro modelo.

Tabela 4: Relatório de Classificação do Modelo de Base (Dados Desbalanceados)

Gênero	Precision	Recall	F1-Score	Support
country	0.69	0.05	0.10	3078
pop	0.53	0.75	0.62	36420
rap	0.82	0.84	0.83	36398
rb	0.52	0.03	0.06	6906
rock	0.59	0.41	0.49	25808
Macro Avg	**0.63**	**0.42**	**0.42**	**108610**
Accuracy			**0.64**	

A análise revela um bom desempenho para ‘rap‘, mas uma falha quase completa em identificar os gêneros minoritários ‘country‘ e ‘rb‘, como evidenciado pelo baixíssimo ‘recall‘. A ‘macro avg f1-score‘ de 0.42 confirma que o modelo, apesar da acurácia de 0.64, é desequilibrado.

5.2 Melhoria do Modelo com Undersampling

Para corrigir o desequilíbrio, foi aplicada a técnica de Undersampling, onde o conjunto de treino foi reamostrado para que todos os cinco gêneros tivessem o mesmo número de

exemplos da menor classe ('country'). A Tabela 5 mostra o resultado do novo modelo treinado com dados balanceados.

Tabela 5: Relatório de Classificação do Modelo Aprimorado (Undersampling)

Gênero	**Precision**	**Recall**	**F1-Score**	**Support**
country	0.63	0.64	0.64	3078
pop	0.37	0.28	0.32	3078
rap	0.78	0.77	0.78	3078
rb	0.52	0.57	0.55	3079
rock	0.50	0.55	0.52	3078
Macro Avg	**0.56**	**0.56**	**0.56**	**15391**
Accuracy			**0.56**	

5.3 Análise Comparativa

O segundo modelo, embora com uma acurácia geral menor (0.56), é vastamente superior. O 'macro avg f1-score' subiu de 0.42 para 0.56, indicando um modelo muito mais justo e equilibrado. Houve um aumento dramático no 'f1-score' de 'country' (de 0.10 para 0.64) e 'rb' (de 0.06 para 0.55). Isso demonstra que a técnica de undersampling foi bem-sucedida em forçar o modelo a aprender as características das classes minoritárias, ao custo de uma performance reduzida na classe 'pop'.

6 Conclusão Geral

Este projeto executou com sucesso um pipeline completo de análise de dados. A Tarefa 1 revelou que a polaridade das letras não é um preditor forte de popularidade, mas pode traçar a evolução temática de um artista. A Tarefa 2 demonstrou a construção de um modelo de classificação de gênero e, mais importante, um processo iterativo de melhoria, identificando e corrigindo o problema de desbalanceamento de classes através de Undersampling para criar um modelo mais robusto e equitativo.

7 Limitações e Passos Futuros

As principais limitações foram o tempo e o escopo da análise de sentimento, que poderia ser aprimorada com modelos mais complexos (e.g., BERT). Passos futuros incluiriam treinar o modelo de gênero com o dataset completo e testar técnicas de amostragem mais avançadas, como SMOTE (oversampling).