Classificação de gêneros a partir de letras de músicas em português

Matheus Bastos de Oliveira[1], **João Baptista de Oliveira e Souza Filho**[1]

[1]Programa de Engenharia Elétrica
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brasil

{oliveiraa.maatheus,jbfilho}@poli.ufrj.br

Abstract. *Associating songs with genres is not easy. The subjectivity and diversity in musical works make assigning unequivocal labels a challenging task. However, textual features can aid in characterizing genres. This work proposes a system for classifying song lyrics in Portuguese, exploring Deep Learning models, such as LSTM networks and Transformers. It also includes more straightforward strategies like Logistic Regression classification of TF-IDF-generated embeddings. Experiments demonstrated that the Transformer model achieved the best performance, reaching an accuracy of 61.6% for ten music genres.*

Resumo. *Associar canções a gêneros não é fácil. A subjetividade e diversidade das obras musicais tornam a atribuição de rótulos inequívocos uma tarefa desafiadora. Porém, atributos textuais podem contribuir para a caracterização de gêneros. Este trabalho propõe um sistema para classificar letras de músicas em português, explorando modelos de Aprendizado Profundo, tais como Redes LSTM e Transformers. São também incluídas estratégias mais simples como a classificação por Regressão Logística de representações geradas por TF-IDF. Experimentos demonstraram que o modelo Transformer apresentou o melhor desempenho, alcançando uma acurácia de 61,6% para dez gêneros musicais.*

1. Introdução

A música, enquanto prática artística e cultural, tem como essência a diversidade e a pluralidade. Neste sentido, muitos fatores podem influenciar sua concepção, como, por exemplo, o seu local de origem, seu enquadramento cronológico e o seu contexto etnográfico.

A fim de atender tal diversidade, as obras musicais são comumente categorizadas em gêneros, de acordo com as similaridades observadas entre elas. Entretanto, tal tarefa pode ser desafiadora e demandante, uma vez que os estilos musicais podem inspirar e gerar novas formas de música com características específicas. Entre alguns exemplos têm-se: o "pop-rock", que une características dos gêneros "pop" e "rock"; a "bossa nova", a qual pode ser considerada tanto um gênero musical próprio quanto um subgênero do "samba" ou do "jazz"; e a "pisadinha", que é um gênero recente, surgido em meados dos anos 2000, oriundo do "forró" [Vicente 2022]. Um outro ponto problemático é o fato de uma canção ser uma expressão artística complexa, podendo admitir diferentes interpretações, o que dificulta a sua caracterização plena por rótulos pré-determinados.

Uma vez que uma canção é uma conjugação de dois tipos de linguagens: a verbal e a musical, aspectos específicos podem estar presentes tanto em percepções auditivas, como melodia, ritmo e harmonia, como em elementos textuais [Bonds 2018].

Neste contexto, as técnicas de Processamento de Linguagem Natural (NLP) têm tornado mais factível a automatização da tarefa de identificação de gêneros a partir das letras de músicas. Tal processo pode ser útil para diferentes aplicações, tais como a indexação, recomendação e a distribuição de músicas, cada vez mais relevantes à vida digital.

Desenvolver um modelo de NLP dedicado a esta tarefa presume a coleta de um volume expressivo de letras. Em particular, o português, apesar de ser significativamente falado ao redor do mundo, carece de maiores *corpora* e de modelos específicos. Motivado por tais questões, este trabalho busca apresentar um conjunto de dados balanceado, especialmente elaborado para a classificação de gêneros musicais através de letras de canções em português, e disponibilizado em um repositório público [de Oliveira 2023]. Adicionalmente, é proposto um modelo de classificação automática que opera com gêneros tipicamente brasileiros e pouco observados na literatura, os quais são identificados apenas por meio do conteúdo semântico contido nas letras, portanto de forma independente ao ritmo segundo o qual a música é executada.

Neste estudo, considerando-se diferentes métodos de geração de representações distribuídas das palavras na forma de vetores numéricos, processo conhecido como "geração de *embeddings*", foram avaliados algoritmos promissores para o tratamento do problema, tais como as Redes Neurais do tipo *Long Short-Term Memory* (LSTM) [Hochreiter and Schmidhuber 1997] e os *Transformers* [Vaswani et al. 2017], bem como a classificação por Regressão Logística [Hastie et al. 2009], uma alternativa mais simples, a fim de melhor relacionar a complexidade dos modelos adotados e o desempenho obtido.

Este artigo é estruturado da seguinte maneira: a Seção 2 realiza uma breve revisão bibliográfica sobre o tema de classificação de gêneros musicais; a Seção 3 descreve a base de dados e o seu pré-processamento, bem como aponta algumas características estatísticas de interesse dos dados; a Seção 4 apresenta a metodologia utilizada para o projeto dos modelos, descrevendo as arquiteturas e os processos de geração de *embeddings* empregados. Por fim, a Seção 5 exibe os resultados obtidos, enquanto a Seção 6 discute as conclusões.

2. Trabalhos Relacionados

Trabalhos sobre a classificação automática de músicas a partir de sinais de áudio são frequentes na literatura. Em [Jeong and Lee 2016] foi proposta uma estrutura para o aprendizado de características temporais discriminantes que alcançou 63% de acurácia para a identificação de 10 gêneros, quando combinada com o aprendizado convencional de características espectrais. De forma similar, uma acurácia de 85% é reportada por [da Silva Muniz and de Oliveira e Souza Filho 2023], o qual considerou a geração de 81 atributos específicos. Cabe destacar que um estudo subjetivo descrito em [Gjerdingen and Perrott 2008] sinaliza 70% de acerto para humanos quando ouvidas amostras com uma duração de 3 segundos.

Quando consideradas letras de canções em português, em [Guimarães et al. 2020] foram realizados experimentos considerando 6000 canções e 7 gêneros: forró, gospel, MPB, samba, sertanejo, bossa nova e axé. Cada canção foi transformada num vetor de inteiros em que cada componente é o índice no vocabulário de suas 200 primeiras palavras. Em seguida, foram avaliados seis modelos de classificação: LSTM, FastText, *eXtreme Gradient Boosting*, *Random Forest* (RF), Árvore de Decisão e *Multilayer Perceptron*. O melhor resultado foi obtido pelo modelo LSTM (50% de acurácia), ao considerar uma ca-

mada de *Embedding* de 100 dimensões e treinamento explorando as técnicas de *Dropout* e *Gradient Clipping*. Em [de Araújo Lima et al. 2020], foi apresentado um conjunto de dados com cerca de 138 mil canções brasileiras e 14 gêneros. Os experimentos com ele realizados exploraram os modelos *Support Vector Machine* (SVM), RF e LSTM Bidirecional (BiLSTM), cada um associado a diferentes técnicas para a geração dos *embeddings*. Os melhores resultados foram obtidos a partir da combinação da rede BiLSTM com o modelo *Wang2Vec* pré-treinado em português, que alcançou 48% na média do F1-*score* inferido para cada classe. Em [Pimenta and Pugliesi 2022], foram considerados 3 gêneros: sertanejo, MPB e funk, e 12 mil músicas, realizando-se a vetorização das letras através da técnica TF-IDF. Para a classificação, foram avaliados os modelos *Linear Discriminant Analysis*, *k-Nearest Neighbors*, SVM, Árvore de Decisão, RF e Regressão Logística, o último de melhor eficácia, que atingiu uma acurácia de 80%. Uma tendência natural observada nestes trabalhos é que os modelos com um menor número de gêneros obtiveram um melhor desempenho, visto resolverem tarefas mais simples.

3. Base de Dados

Nesta seção são apresentadas as etapas envolvidas na geração e tratamento dos dados.

3.1. Coleta dos Dados

A base de dados [Neisse 2022] explorada neste trabalho foi coletada no repositório "Kaggle", comunidade virtual de ciência de dados, e integra letras de 379,893 canções de 4,239 artistas, das quais cerca de 41% delas estão em português, e foram portanto consideradas para este trabalho, exceto quando dispunham de campos com dados faltantes.

Esta base se encontra dividida em duas tabelas distintas. A primeira, "*lyrics-data*", relaciona cada letra de música a seu nome e idioma, bem como apresenta os endereços *web* para a letra e para o perfil do artista; a outra, "*artists-data*", relaciona cada artista com os seus gêneros musicais, o número de canções que possui, sua pontuação de popularidade inferida pelo número de acessos ao site e o endereço *web* do seu perfil.

3.2. Pré-processamento e Formação da Base

A preparação da base de dados para a classificação envolveu os seguintes passos: para cada letra presente na *lyrics-data*, foi utilizado o endereço *web* para o perfil do artista como item de busca dos gêneros a ele associados na *artists-data*. Em seguida, mantiveram-se apenas as colunas correspondentes à letra (atributos) e ao primeiro - por ser o mais relevante ao artista - desses gêneros (alvo). A fim de expressar um bom compromisso quanto a variedade de gêneros e o quantitativo de canções disponível por gênero, foram enumerados e selecionados apenas os 10 gêneros com a maior quantidade de dados (forró, funk carioca, gospel, MPB, pagode, pop-rock, rock, romântico, samba e sertanejo). Para balancear a base, foi utilizada a técnica de *undersampling*. Com efeito, notou-se que dentre estes gêneros, o que possuía menos letras associadas contava com 3990 canções, assim foi assumida tal quantidade de letras para todos os demais gêneros. Logo após, foi realizada a divisão da base em dados de treino (para o ajuste dos parâmetros internos dos modelos), validação (para a sintonia de seus hiperparâmetros) e teste (para a avaliação final e comparação entre modelos), de forma aleatória, em proporções de, respectivamente, 72%, 8% e 20% do total de dados. Por fim, foram identificadas e removidas 12 letras cujos conteúdos não representavam canções reais, com conteúdos como "Essa música é

INSTRUMENTAL", "(Mensagem)" e "Introdução". Como tal quantitativo não é substancial, assumiu-se que a base continuou balanceada, dispondo-se de 3986 músicas de pop-rock, 3987 de MPB, 3988 de rock, 3989 de forró, funk e gospel, e 3990 de pagode, romântico, samba e sertanejo.

3.3. Análise Exploratória

A Tabela 1 descreve a quantidade de palavras, versos e estrofes médios por canção, bem como os valores mínimo, médio e máximo de palavras únicas por canção; e a Tabela 2, as 3 palavras mais comuns ao longo das letras, cada uma com a sua respectiva taxa de ocorrência (T_O), dada pela razão entre a quantidade de ocorrências de uma dada palavra ao longo de todas as canções de um gênero pelo total de canções deste gênero.

Cabe destacar que todas as análises não diferenciaram letras maiúsculas de minúsculas. Relativamente à Tabela 2, foram excluídas as *stopwords* indicadas pela biblioteca NLTK [Bird et al. 2009], bem como pontuações e as palavras "ai", "vou", "vai", "vem", "pra", "pro", "tô", "tá", "tão", "mim", "ti", "então", "lá", por não se acreditar que contribuiriam significativamente para a discriminação dos gêneros.

Tabela 1. Comprimentos médios e palavras únicas por canção

Gênero	Palavras	Versos	Estrofes	Mínimo	Média	Máximo
Forró	142,1 ± 89,6	25,1 ± 16,0	5,7 ± 4,3	13	70,1	459
Funk	213,3 ± 131,1	35,9 ± 21,6	7,5 ± 5,1	6	86,2	660
Gospel	130,4 ± 73,6	22,0 ± 11,3	4,7 ± 3,1	11	68,4	571
MPB	126,9 ± 81,3	25,0 ± 16,2	4,5 ± 4,2	4	68,0	643
Pagode	144,6 ± 84,9	25,4 ± 14,7	4,9 ± 3,7	13	73,1	487
Pop-Rock	160,5 ± 82,8	30,2 ± 14,8	6,3 ± 4,2	4	75,3	466
Rock	153,3 ± 73,2	28,3 ± 14,1	6,1 ± 4,3	6	75,2	492
Romântico	128,1 ± 62,8	23,8 ± 11,8	4,8 ± 2,9	5	66,3	378
Samba	122,5 ± 70,0	23,6 ± 13,7	4,5 ± 4,1	11	68,2	309
Sertanejo	158,0 ± 65,8	27,2 ± 12,4	5,8 ± 3,9	15	86,9	307

É interessante observar na Tabela 1 o quanto os gêneros se diferenciam no comprimento médio de suas canções. Neste quesito, cabe destaque ao funk, que apresenta, para as três medidas realizadas, o valor mais expressivo e distinto dos demais. Por outro lado, nota-se uma grande semelhança entre os trios de valores apresentados pelos gêneros MPB, romântico e samba. Além disso, é surpreendente a variabilidade dos gêneros, que contêm alguns exemplos de canções com mais de 300 palavras únicas. Por outro lado, há canções com um número muito pequeno de palavras. Um caso digno de nota é a música do gênero MPB "Tutano" do artista Walter Franco, cuja letra é "Quem tem tutano, tutano tem. Quem não tem tutano, tutano não tem.", formada portanto por apenas 4 palavras únicas.

Pela Tabela 2 constata-se que o gênero gospel é o que possui uma maior disparidade quanto às palavras mais frequentes em suas letras, visto "Deus", "senhor" e "Jesus" não figurarem como resultado de nenhum outro estilo. Além disso, cabe destacar a grande similaridade entre as palavras mais frequentes dos demais estilos, demonstrando que o uso de palavra(s)-chave(s) é ineficaz para uma determinação direta do gênero da música, ressaltando assim a não-trivialidade do problema.

Tabela 2. Palavras mais frequentes por canção

Gênero	Palavra	T_O	Gênero	Palavra	T_O
Forró	amor	1,427	**Funk**	quero	0,788
	quero	0,684		quer	0,709
	coração	0,623		amor	0,623
Gospel	Deus	1,808	**MPB**	amor	0,851
	senhor	1,021		tudo	0,466
	Jesus	0,948		vida	0,455
Pagode	amor	1,475	**Pop-Rock**	tudo	0,872
	quero	0,701		amor	0,767
	gente	0,583		quero	0,728
Rock	tudo	0,809	**Romântico**	amor	1,696
	quero	0,693		quero	0,628
	amor	0,601		coração	0,607
Samba	amor	1,094	**Sertanejo**	amor	1,270
	samba	0,489		vida	0,595
	vida	0,431		coração	0,591

Por fim, a Tabela 3 indica a proporção percentual de palavras do vocabulário de cada gênero (disposto nas linhas) que também pertencem ao vocabulário de um outro estilo (disposto nas colunas). Os gêneros estão sinalizados conforme se segue: forró (0), funk (1), gospel (2), MPB (3), pagode (4), pop-rock (5), rock (6), romântico (7), samba (8) e sertanejo (9). Para facilitar a análise, o menor e o maior valor observados para cada gênero (linha) são destacados com as cores vermelha e azul, respectivamente.

Tabela 3. Matriz de coocorrência de vocabulários

	(0)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
(0)	100	50,4	39,1	54,7	46,7	51,6	54,5	45,6	52,7	54,6
(1)	36,5	100	29,7	39,4	36,8	41,6	43,1	32,8	38,0	39,2
(2)	43,4	45,6	100	51,1	41,7	50,1	51,9	45,5	48,4	50,1
(3)	41,7	41,4	35,0	100	37,8	48,4	50,8	38,6	48,2	45,2
(4)	51,3	55,9	41,3	54,6	100	54,7	56,4	46,3	55,3	53,8
(5)	42,1	46,9	36,8	51,8	40,6	100	56,2	40,0	46,1	46,3
(6)	41,3	45,2	35,5	50,6	38,9	52,3	100	38,6	44,9	45,3
(7)	55,1	54,9	49,5	61,3	50,9	59,3	61,5	100	59,2	61,0
(8)	43,9	43,8	36,3	52,7	41,8	47,1	49,3	40,8	100	48,6
(9)	47,4	47,0	39,2	51,5	42,4	49,2	51,8	43,8	50,7	100

Nessa tabela verifica-se que a coluna (2) é a que contém os menores valores para todas as linhas, exceto a própria (2), sinalizando que o gospel é o estilo de menor coincidência com os demais vocabulários. Por sua vez, a coluna (6) mostra que o rock é o gênero de maior similaridade de vocabulário com os demais, exceto forró e samba. Cabe destacar que nem sempre as relações são recíprocas, ou seja, tal matriz não é simétrica. Por exemplo, há muitas palavras do gênero rock nas letras do gênero romântico (61,5%), porém há poucas do gênero romântico nas letras de rock (38,6%).

4. Metodologia

Nesta seção é apresentado breve resumo dos modelos avaliados. O LSTM é de uma arquitetura capaz de armazenar padrões de interesse ao longo de intervalos de iteração arbitrários. Por sua vez, o *Dropout* [Srivastava et al. 2014]: é uma estratégia de regularização que consiste em excluir aleatoriamente um percentual ("taxa de *Dropout*") das conexões de entrada e das conexões recorrentes a cada iteração de treinamento. O LSTM Bidirecional (BiLSTM) consiste em uma variante LSTM com duas camadas, com fluxos de informações em direções contrárias. Os *Bidirecional Encoder Representations from Transformers* [Devlin et al. 2019] (BERT) representam modelos pré-treinados, formados por camadas empilhadas de *Transformers*. Operam realizando o mapeamento das palavras do texto em *embeddings* que são submetidos a um cabeçote interno de classificação [HuggingFace 2018]. O modelo utilizado neste trabalho foi o BERTimbau-Base [Souza et al. 2020]. A estratégia BERT + Regressão Logística explora o modelo BERT pré-treinado e sem ajuste fino, que é utilizado apenas para a geração dos *embeddings*, submetidos a um classificador baseado em Regressão Logística. Por fim, o esquema *Term Frequency - Inverse Document Frequency* [Salton and Buckley 1988] (TF-IDF) + Regressão Logística se refere aos experimentos em que os *embeddings* são gerados através da técnica TF-IDF e classificados em sequência por Regressão Logística.

4.1. Geração de *Embeddings*

A geração de *embeddings* para os modelos LSTM e BERT envolveu os seguintes passos:

1. Cada palavra do *corpus* foi indexada por um número inteiro único (o tamanho do dicionário considerando todos os gêneros foi de 78307 palavras);
2. Cada letra de música foi associada a um vetor de inteiros, cujas componentes foram definidas pelos índices (em ordem) de suas palavras;
3. Cada vetor foi ampliado com zeros até que se alcançasse a dimensão daquele correspondente à letra com a maior quantidade (1996) de palavras, no caso do LSTM, e truncado ou ampliado com zeros até 512 palavras, no caso do BERT;
4. Cada vetor resultante foi mapeado num vetor real denso de 100 e 768 dimensões no caso dos modelos LSTM e BERT, respectivamente. No primeiro, por restrições computacionais. No segundo, por ser um valor comumente utilizado na literatura.

Cabe destacar que tais procedimentos foram realizados por códigos próprios para os modelos LSTM, e pela versão pré-treinada *BertTokenizerFast*, para o modelo BERT. Com relação às classificações por Regressão Logística, os *embeddings* gerados pelo BERT foram diretamente obtidos através do modelo pré-treinado através de sua versão *SentenceTransformer*. A abordagem TF-IDF, para uma melhor comparação dos resultados, considerou as 768 palavras mais frequentes das músicas, para todos os gêneros musicais.

5. Resultados

Este trabalho foi realizado utilizando a linguagem de programação *Python* e as bibliotecas de código aberto *Pandas* [Wes McKinney 2010], *Scikit-learn* [Pedregosa et al. 2011], *TensorFlow* [Abadi et al. 2016] e *Keras* [Chollet et al. 2015]. Os modelos foram treinados e avaliados através da plataforma "Google Colab", considerando uma GPU "Tesla

T4" de 12GB. O desenvolvimento e a avaliação dos modelos consideraram a estratégia de *hold-out* [Japkowicz and Shah 2011], conforme as divisões explicitadas na Subseção 3.2.

Os hiperparâmetros dos modelos foram sintonizados de maneira gulosa e consideraram os valores presentes na Tabela 4. Tal sintonia começou pela LSTM padrão, assumindo $BS = 32$, por ser o valor intermediário. A busca pelo melhor número de épocas resultou em $NE = 5$, por *early stop*. Em seguida, fixando-se este valor, buscou-se o melhor tamanho do lote, dado por $BS = 8$. Logo após, passou-se à LSTM com *Dropout* e, mantidos os demais hiperparâmetros, os melhores resultados foram obtidos com $TD = 0,5$. Por fim, a BiLSTM com *Dropout* considerou os valores sintonizados anteriormente. Com relação ao BERT, a sintonia começou pelo número de épocas, fixando-se a taxa de aprendizado em $LR = 2e - 5$. A partir deste ponto, encontrou-se $NE = 3$ por *early stop*. Em seguida, de posse deste valor, buscou-se a melhor taxa de aprendizado, resultando em $LR = 5e - 5$. Para os classificadores por Regressão Logística, todos os hiperparâmetros adotados consideraram os valores padrões da biblioteca [Pedregosa et al. 2011].

Tabela 4. Valores avaliados durante a sintonia de hiperparâmetros para as redes do tipo LSTM e para o modelo BERT

Hiperparâmetro	Valores para LSTM	Valores para BERT
Número de épocas (NE)	2, 5, 10	2, 3, 4
Tamanho do lote (BS)	8, 16, 32, 64, 128	16
Taxa de *Dropout* (TD)	0,1, 0,25, 0,5	Não se aplica
Taxa de aprendizado (LR)	Não se aplica	2e-5, 5e-5, 1,25e-4

Para todos os modelos, a função-custo utilizada foi a "Entropia Cruzada Categórica" [Goodfellow et al. 2016] e o método de otimização empregado foi o "Adam" [Kingma and Ba 2015]. Ademais, como a base é balanceada e visando uma comparação mais direta com outros trabalhos, a figura de mérito considerada foi a acurácia.

A Tabela 5 sintetiza os resultados obtidos. Cabe observar que as duas abordagens baseadas em BERT apresentaram um desempenho melhor do que as baseadas em LSTM. Tal fato era em parte esperado, pois o BERT é um modelo pré-treinado e mais complexo, portanto mais hábil para a captura de nuances semânticas no conteúdo das letras. Ademais, em relação aos modelos LSTM, a técnica de *Dropout* combinada com a abordagem bidirecional resultou no modelo de maior acurácia dentre os desta categoria: 52,4%. Numericamente, os ganhos da estratégia de *Dropout* ($\sim$1,6%) se mostraram bem superiores aos associados ao uso de camadas bidirecionais ($\sim$0,6%). Cabe ainda ressaltar que o tempo de treinamento do modelo bidirecional foi praticamente o dobro do unidirecional. A abordagem BERT com ajuste fino atingiu a maior acurácia: 61,6%, valor similar ao percentual de acertos obtido por pessoas ao ouvir amostras de áudio com uma duração de 3 segundos (70%) [Gjerdingen and Perrott 2008] e, com relação à classificação a partir do conteúdo de letras em português, superior aos 50% de acertos reportados em [Guimarães et al. 2020], mesmo que esse último tenha considerado uma menor quantidade (7) de gêneros.

As acurácias por gênero foram 58,4% (forró), 78,4% (funk), 89,7% (gospel), 47,2% (MPB), 63,7% (pagode), 40,5% (pop-rock), 58,3% (rock), 59,1% (romântico), 55,3% (samba) e 64,4% (sertanejo). Assim, o modelo se mostrou mais acurado para o

Tabela 5. Resultados obtidos pelos modelos

Abordagem	Melhor configuração	Tempo de treino (minutos)	Acurácia no teste (%)
BERT (com ajuste fino)	$NE = 3$ $LR = 5e-5$	$147,4$	$61,6$
BERT + Regressão Logística	(Sem ajuste fino)	$5,5$	$54,5$
BiLSTM com *Dropout*	$NE = 5, BS = 8$ $TD = 0,5$	$53,3$	$52,4$
LSTM com *Dropout*	$NE = 5, BS = 8$ $TD = 0,5$	$28,7$	$51,8$
LSTM	$NE = 5$ $BS = 8$	$25,9$	$50,2$
TF-IDF + Regressão Logística	(Sem ajuste fino)	$0,2$	$47,3$

gênero gospel, face ao seu vocabulário particular. Por sua vez, os maiores erros, 37,9% e 17,3%, se referem às letras de pop-rock classificadas como rock e vice-versa, respectivamente, o que está provavelmente associado à similaridade dos vocabulários, conforme as Tabelas 2 e 3. Outra medida relevante é a razão entre o total de canções corretamente classificadas em cada gênero pelo total de canções classificadas como pertences a este gênero, que foram 58,1% (forró), 87,4% (funk), 84,8% (gospel), 58,2% (MPB), 60,4% (pagode), 54,3% (pop-rock), 43,8% (rock), 46,8% (romântico), 63,4% (samba) e 66,6% (sertanejo). Convém notar que o gospel e o rock obtiveram o segundo maior e o menor entre todos os valores, respectivamente, o que está em acordo com a discussão ao final da Seção 3.

6. Conclusão

Este trabalho propôs a criação de uma base de dados balanceada, considerando letras de canções em português distribuídas em dez estilos musicais, e avaliou métodos de Aprendizado Profundo para automatizar sua classificação. O modelo BERT superou os demais, tendo sido atingida uma acurácia de 61,6%, valor próximo ao desempenho humano no reconhecimento de amostras de áudio com uma duração de 3 segundos.

Como trabalhos futuros, espera-se refinar os modelos, explorando a sintonia de outros hiperparâmetros e a pluralidade de estilos por artista, visto que a rotulagem pelo seu gênero mais frequente é uma das limitações do estudo presente. Ademais, almeja-se utilizar técnicas de validação cruzada para uma avaliação estatística mais rigorosa dos resultados. Considera-se também avaliar o modelo BERTimbau *large*, bem como modelos baseados na fusão de atributos derivados das letras e dos sinais de áudio das músicas.

7. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: A system for Large-Scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA. USENIX Association.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly.

Bonds, M. E. (2018). *Listen to This*. Pearson, 4th edition.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

da Silva Muniz, V. H. and de Oliveira e Souza Filho, J. B. (2023). Robust handcrafted features for music genre classification. *Neural Computing and Applications*, 35(13):9335–9348.

de Araújo Lima, R., de Sousa, R. C. C., Lopes, H., and Barbosa, S. D. J. (2020). Brazilian lyrics-based music genre classification using a blstm network. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part I*, page 525–534. Springer-Verlag.

de Oliveira, M. B. (2023). Conjunto de dados para classificação de gêneros a partir de letras de músicas em português. GitHub. `https://github.com/oliveiraamaatheus/Conjunto-de-dados-para-classificacao-de-generos-a-partir-de-letras-de-musicas-em-portugues` [Accessed: 2023-08-11].

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gjerdingen, R. and Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, 37(2):93–100.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Guimarães, P., Froes, J., Costa, D., and Freitas, L. (2020). A comparison of identification methods of Brazilian music styles by lyrics. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 61–63, Seattle, USA. Association for Computational Linguistics.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

HuggingFace (2018). Bert for sequence classification. Hugging Face Transformers Documentation. `https://huggingface.co/docs/transformers/mod`

el_doc/bert#transformers.BertForSequenceClassification [Accessed: 2023-06-19].

Japkowicz, N. and Shah, M. (2011). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, USA.

Jeong, I. and Lee, K. (2016). Learning temporal features using a deep neural network and its application to music genre classification. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States*, pages 434–440.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Neisse, A. (2022). Song lyrics from 79 musical genres. Kaggle. https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres [Accessed: 2023-06-19].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pimenta, M. F. and Pugliesi, J. B. (2022). Reconhecimento de gêneros musicais com técnicas de aprendizagem de máquina supervisionada. *Revista Eletrônica de Computação Aplicada*, 3(1):23–46.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*, page 403–417.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Vicente, D. (2022). Nem pagode, nem sertanejo: pisadinha faz o brasil dançar na pandemia. https://www.exame.com.br/bussola/nem-pagode-nem-sertanejo-pisadinha-faz-o-brasil-dancar-na-pandemia [Accessed: 2023-06-19].

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.