Text Representations for Lyric-Based Identification of Musical Subgenres

Fabrício Almeida do Carmo, Jorge Luiz F da Silva Junior, Rafael G Rossi and Fábio M F Lobato

***Abstract*—The advancement of techniques and computational tools for data mining has been boosting the music market with applications focused on user experience. These techniques explore musical data looking for patterns and trends that can guide business strategies. One of the key steps in these applications is the vector representation of the original text. This work approaches textual representation techniques applied to the problem of classifying musical sub-genres, a gap in the literature in musical information retrieval, whose complexity lies in the difficult identification of the separation boundary between the sub-classes of the same genre since both carry several features in common. For this, exhaustive experiments were carried out aiming to find the best combination between classifier and textual representation models. The results showed enriched Bag-of-Words (BoW) with the SVM and Logistic Regression algorithms obtained better results than embeddings models and deep neural networks. The conclusions obtained could guide future studies for classifying texts whose separability surfaces are subtle and challenging.**

***Index Terms*—Music Classification, Text Representations, Bag-of-words, Word Embeddings, Neural Networks, Deep Learning**

I. Introdução

O grande volume de dados produzido com o avanço da internet carrega oportunidades de negócios nos mais diversos setores da sociedade [1]. Em contraponto, essa sobrecarga de informações provoca desafios para a construção de sistemas de recomendação [2], [3]. No setor musical, as plataformas de *streaming* buscam, em ferramentas robustas de processamento de áudio e texto, soluções que possam oferecer experiências mais segmentadas e satisfatórias aos seus clientes. No entanto, realizar tais procedimentos de filtragens e extração de padrões ainda é desafiador considerando a diversidade musical e o alto custo computacional inerente ao processo de Recuperação de Informação Musical, do inglês, *Music Information Retrieval* (MIR) [4].

No campo da MIR, diversas técnicas computacionais são utilizadas na busca por padrões contidos nos dados musicais. Nessa direção, estudos como identificação do gênero musical [5] e classificação de humor [6] vêm se destacando principalmente quando impulsionados por técnicas robustas do Processamento de Linguagem Natural (PLN), como as representações de palavras que incorporam elementos contextuais e com os recursos do Aprendizado Profundo, do inglês, *Deep Learning* (DL) [7]. Vários trabalhos já apontam essas tecnologias como promissoras por possuírem menor custo de processamento se comparadas com soluções baseadas em áudio [3], [4].

As representações de dados textuais visam modelagens vetoriais representativas de um determinado texto. Elas são elementos fundamentais em PLN dado sua capacidade de transformar os documentos de entrada em vetores numéricos que preservam informações originais, fornecendo representações interpretáveis por modelos de *Machine Learning* (ML) e DL. Por exemplo, a frequência dos termos podem ser computadas e analisadas - e.g. *Bag-of-Words* (BoW). Outra alternativa explorada na literatura é a extração das estatísticas da estrutura do texto (e. g., verbos, pronomes, tamanho de sentenças). Ambas podem fornecer vetores significantes para algoritmos de ML. No entanto, esses modelos não consideram as relações entre as palavras (contexto), dificultando sua aplicação em problemas com maior complexidade, como a classificação de subgêneros musicais - onde a fronteira de separação entre classes não é trivial.

Nessa linha, as representações baseadas em *word embeddings* vêm fornecendo resultados promissores em diferentes tarefas envolvendo PLN. Estas representações conseguem adicionar informações semânticas no processo representacional [8]. Em um alinhamento com os modelos *word embeddings*, algumas arquiteturas de redes neurais multicamadas, como *Convolutional Neural Network* (CNN) e *Recurrent Neural Network* (RNN), já são utilizadas para tarefas de classificação textual em diversos segmentos de mercado, dado sua acuracidade e capacidade de lidar com maior volume de dados [7].

Imerso no campo da MIR, este trabalho atua no problema da classificação de subgêneros musicais utilizando o conteúdo textual, um tópico ainda pouco explorado na literatura [9]. Em tal problema, há um aumento de complexidade se comparado a classificação do gênero musical tradicional devido às subclasses de um mesmo gênero carregarem diversas características em comum [10]. Considerando essa lacuna na literatura, o objetivo deste trabalho é encontrar combinações de representações textuais e modelos preditivos que melhor se adequem ao problema em questão. Para isso, é realizado uma análise experimental com diversas técnicas de representações de palavras e algoritmos de ML e DL dispostos no estado da arte da MIR e da PLN. Foram utilizados nos experimentos um conjunto de dados com letras de músicas dos gêneros *rock* = *{heavy metal, punk* e *soft}* e *pop* = *{tecnopop, rock* e *power}*, construído para o presente estudo e disponibilizado em repositório público, caracterizando uma contribuição técnica do trabalho.

Fabrício Almeida do Carmo, Programa de Engenharia da Computação, Universidade Estadual do Maranhão, Brasil e-mail:fabrycio30@gmail.com.

Jorge Luiz Figueira da Silva Junior, Fábio Manoel França Lobato, Instituto de Engenharia e Geociências, Universidade Federal do Oeste do Pará, Brasil e-mails:(jorgeluizfigueira@gmail.com, fabio.lobato@ufopa.edu.br)

Rafael Geraldeli Rossi, Specialist Data Scientist at iFood, Brasil. e-mail:rafael.geraldeli@ifood.com.br

Os experimentos computacionais apontam que representações mais simples como a BoW quando combinadas com algoritmos tradicionais de ML, como *Support Vector Machine* (SVM) e *Logistic Regression*, oferecem resultados competitivos na classificação de subgênero musical. As análises realizadas mostraram que estas combinações produziram resultados superiores em comparação aos cenários envolvendo *word embeddings* e redes neurais profundas.

O restante deste artigo está organizado como segue. Na Seção 2 são dispostos os trabalhos relacionados, na seção 3 é apresentada a definição formal do problema. Na Seção 4 são descritos os experimentos. Os resultados e análises são discutidos na Seção 5. Por fim, na Seção 6 são apresentadas as conclusões e os trabalhos futuros.

II. Trabalhos Relacionados

A literatura mostra que técnicas de PLN têm impulsionado análises textuais nos mais variados campos de pesquisa [7], [11]. Em MIR, tópicos de estudos têm apontado resultados promissores, principalmente com a utilização de representações de dados que incorporam informações contextuais (e.g. *word embeddings*) e com arquiteturas de DL voltadas para o processamento de texto [5], [6], [12].

Trabalhando com classificação de gêneros musicais, em [5], os autores realizaram experimentos para a identificação de gêneros musicais brasileiros, utilizando diferentes técnicas de representação de palavras, a saber: *Word2Vec* [8], *Wang2Vec* [13], *FastText* [14] e *Glove* [15], para extração de informações a partir das letras. Os resultados apontaram que a rede *Bidirectional Long Short-Term Memory* (BLSTM) obteve maior desempenho que algoritmos tradicionais como SVM e *Random Forest*. Em uma combinação com o *Wang2vec*, o BLSTM obteve F1-score médio de 0,481 considerando 14 gêneros musicais.

Em [7], os autores também aplicaram arquiteturas multicamadas no processo de identificação de temas/classes de pedidos de acesso à informação dirigidos ao Poder Executivo Federal brasileiro. Experimentos com CNN e RNN, do tipo *Long Short-Term Memory* (LSTM), e uma combinação de ambas foram realizados, visando encontrar a arquitetura mais adequada para o problema. Os resultados mostraram que a CNN traz melhores resultados para textos curtos. Corroborando o uso destas arquiteturas, [12] realizam experimentos que apontam ganho no desempenho quando alinhadas com representações vetoriais sensíveis ao contexto, indicando que RNN supera as limitações das janelas de tamanhos fixos dos modelos n-gramas e a *Gated Recurrent Unit* (GRU) e LSTM capturam relacionamentos mais distantes no espaço de treinamento.

Um problema análogo ao tratado no presente estudo é a classificação de humor, dada a suas nuances e dificuldades em estabelecer uma superfície de separabilidade clara no processo de descoberta de informação musical. Neste sentido, [6] utilizaram técnicas de representações mais simples como *Term-Frequency/Inverse-Document Frequency* (TF-IDF) e modelos *word embedding* para a tarefa de classificação. As análises mostraram que as redes CNNs atingiram 71% de acurácia, associadas ao *Glove*. No estudo, também foram avaliadas redes BLSTM e uma combinação das duas, tal como mostrado em [7] e [16]. No tratamento de subgêneros musicais, em [10], os autores utilizaram Características Estatísticas Textuais (CET), BoW e *Part-of-Speech* (PoS) *tagging* como estratégias de representação dos textos musicais. Aplicados em tarefas de classificação, os resultados apontaram que estas estatísticas simples, como: contagem de palavras, caracteres, sílabas e sentenças, também oferecem vetores representativos de determinada música.

Os trabalhos mencionados acima destacam que diferentes arranjos de representações *word embeddings* com arquiteturas de redes neurais profundas produzem resultados relevantes na descoberta de conhecimento baseados em contexto. No entanto, considerando nosso melhor esforço na busca de trabalhos relacionados, não foram encontrados estudos que utilizassem tais representações no problema da classificação de subgêneros musicais. Em uma linha alternativa, [10] mostra que abordagens mais simples, como CET, podem produzir resultados promissores para o problema em questão. Buscando preencher esta lacuna, o presente trabalho realiza uma análise mais ampla, estendendo o *framework* experimental de [10], adicionando técnicas mais robustas apresentadas nos outros trabalhos relacionados. Tal cobertura experimental é uma importante contribuição desta pesquisa.

III. Definição do Problema

A literatura da MIR fornece um volume de trabalhos significativos para o problema de classificação de gênero musicais. No entanto, há poucos trabalhos direcionados para indentificação dos subgêneros. Esse problema pode ser definido como uma tarefa de classificação multiclasse, visando encontrar a superfície de separação entre as classes dos subgêneros. A exemplo, uma música do gênero *rock* pode ser considerada *Heavy Metal*, *Punk* ou *Soft*, mostrando que, por pertencer ao mesmo gênero, o aprendizado do modelo não é trivial.

Definindo o problema como aprendizado multiclasse, de acordo com [7], diz-se que: dado um conjunto de músicas $M = \{m_1, m_2, m_3, .., m_k\}$ de um determinado gênero musical e um conjunto de classes com seus subgêneros $S = \{s_1, s_2, s_3, ..., s_n\}$, busca-se uma função de relação entre os dois conjuntos. Tal que: $F(m_i, s_j) \rightarrow \{0, 1\}$, onde: 1 representa a existência da relação da música m_i com o subgênero s_j e 0 caso contrário.

Dada a definição do problema, os pontos positivos e as lacunas apresentadas na Seção II, este trabalho apresenta um *framework* experimental para a classificação de subgêneros musicais com base na análise textuais das letras das músicas. Os detalhes de cada etapa do *framework* são apresentados na próxima seção.

IV. *Framework* Experimental

Esta seção discorre sobre as principais atividades realizadas no desenho do *framework* experimental. Destacam-se aqui: os passos da obtenção e preparação dos dados; modelagem dos vetores de representação musical; técnicas de classificação utilizadas, e, por fim, as métricas e procedimentos para a

TABELA I
CONFIGURAÇÕES DO CONJUNTO DE DADOS.

Gênero	Subgênero	Amostra
Rock	*Heavy Metal*	6.371
	Punk	4.697
	Soft	1.955
Pop	*Tecnopop*	2.169
	Power	1.162
	Rock	5.505
	Total de amostras	21.859

validação dos resultados. Visando prover reprodutibilidade, os códigos-fonte e dados utilizados nos experimentos encontram-se integral e publicamente disponíveis no repositório https://github.com/fabiolobato/music-representation.

A. Base de Dados

Neste estudo, foram adotados dois gêneros musicais popularmente conhecidos: *rock* e *pop*. Para isso, foram extraídos os dados textuais, em inglês, disponíveis na plataforma Letras [17], localizado em https://www.letras.mus.br, um repositório aberto e alimentado pelos próprios usuários. Tal abordagem foi adotada porque não se encontrou dados com tais características disponíveis na literatura. Análises prévias mostraram que a plataforma tem uma maior precisão em relação à fidelidade da letra original, evitando assim, a adição de maiores ruídos no experimento. A Tabela I, apresenta as principais características do conjunto de dados adquirido.

B. Vetores de Representação Musical

As representações de palavras são essenciais para que modelos de ML e DL realizem suas tarefas. Tais modelos, são dependentes de representações vetoriais numéricas como parâmetro de entrada. A literatura dispõe de diversas abordagens que realizam esse procedimento, algumas formando vetores esparsos (BoW), outras adicionando dados contextuais *(word embeddings)*, criando vetores densos e que incorporam informações semânticas. Procedimentos representacionais mais simples também foram considerados nesta pesquisa, como: CET e as análises das classes gramaticais por meio da PoS.

Em vetorização do tipo BoW, cada texto de entrada é representado por um vetor unidimensional do tamanho de N, onde N é o conjunto de todas características extraídas (palavras) do texto. Dessa forma, o vetor resultante de uma música é a contagem das frequências de cada palavra *(token)* extraídas de seu texto musical. Dado que pode haver um determinado termo que pode se repetir entre os diferentes gêneros, nesse trabalhos foi utilizado o esquema de pesos TF-IDF, que visa ponderar a frequência de cada termo pelo inverso do número de documentos que o termo ocorre. Com isso, termos que ocorrerem em vários documentos e que podem não serem úteis para discriminar os subgêneros, terão o seu peso diminuído.

A representação *BoW*, dado suas características, não considera fatores importantes em sua estrutura, como o tratamento da esparsividade do vetor, a relação entre as palavras de um determinado texto e palavras com significados parecidos são tratadas como atributos diferentes. À vista disso, modelos *word embedding* buscam compreender as conexões entre um *token* e seus vizinhos, fornecendo uma projeção de cada palavra no espaço vetorial, de forma que os relacionamentos entre vetores expressem as relações semânticas entre as palavras. Uma das formas bastante utilizadas para obter esta projeção, é treinando redes neurais rasas para prever palavras e seus contextos. Destacam-se aqui as arquiteturas *Continuous Bag Of Words* (CBoW) e *Skip-gram* [8]. Outro modelo utilizado na geração dos vetores é o *FastText*, com essa técnica obtém-se uma representação vetorial para uma sequência de caracteres e não para um *token* completo como é realizado em CBoW e *Skip-gram*. Com isso, a representação da palavra é obtida através das representações das sequências de caracteres, possibilitando a obtenção de representações para palavra não vistas no conjunto de treinamento [14].

TABELA II
CONJUNTO DE PARÂMETROS DOS MODELOS *embedding* ADOTADOS.

Modelos	Parâmetros
CBoW/Skip-Gram/FastText	Janela de contexto = [5,10]
	Dimensões = [25, 50, 100, 300]
	Épocas de treinamentos = [3, 5, 50]
	Mínimo n-gramas = 3
	Máximo n-gramas = 6
Glove	Dimensões = [100, 300]

TABELA III
CONJUNTO DE ESTATÍSTICAS TEXTUAIS.

Cód.	Característica	Descrição
C1	Caracteres	Quantidade de caracteres da música
C2	Tamanho médio da palavra	Razão entre caracteres e palavras
C3	Palavras	Quantidade de palavras
C4	Palavras Únicas	Quantidade de palavras únicas
C5	Sentenças	Quantidade de linhas
C6	Média de palavras por sentença	Razão entre palavras e sentenças
C7	Sílabas	Quantidade de sílabas
C8	Média de sílabas por palavras	Razão entre sílabas e palavras
C9	Taxa de palavras raras	Razão entre palavras raras e total
C10	Diversidade lexical	Razão entre palavras únicas e total

Modelos *Glove* pré-treinados e validados em [15] também foram adotados nos experimentos. Essa técnica realiza o aprendizado por meio da construção de uma matriz de co-ocorrência (palavras-contexto) que computa a frequência com que uma palavra aparece em um determinado contexto. A Tabela II, apresenta os principais parâmetros das representações geradas, tal como em [10].

Além das representações cujos atributos são derivados das representações das palavras dos textos, pode-se também utilizar representações derivadas das estatísticas (e.g. tamanho médio das palavras, proporção de palavras únicas no vocabulário e quantidade de sílabas) contidas no texto de entrada. [10] indica que essas características estruturais, vide tabela III, fornecem elementos de distinções entre subgêneros. Em PoS, é realizado um mapeamento das sentenças de acordo sua estrutura gramatical, aplicando rótulos lexicais na palavras. Dessa forma, é possível obter estatísticas representativas dos textos musicais, analisando as ocorrências das palavras segundo

TABELA IV
PARAMETRIZAÇÕES DOS ALGORITMOS UTILIZADOS NO CENÁRIO 1.

Algoritmo	Parâmetros Utilizados
kNN	k = [3-21], weights= uniform,distance, metric = {euclidean,manhattan,cosine}
SVM	$C = \{ 10^{-5}\ 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}\}$ kernel={sigmoid,rbf,linear,poly}, max-iter=[25000]
Naïve Bayes	types = {Bernoulli, Complement, Gaussian, Multinomial}
Decision Tree	criterion = {gini, entropy}
Logistic Regression	solver = lbfgs, penalty = {l1,l2}

suas respectivas classes gramaticais. Para os experimentos foram computados: substantivos, verbos, pronomes, adjetivos e conectivos, além de duas novas medidas: (i) a incidência de conteúdo, dada pela soma das classes de substantivos, verbos e adjetivos, e (ii) a diversidade de conteúdo, dada pela razão da taxa de incidência de conteúdo e o total de palavras do documento musical.

C. Configuração Experimental

Para os experimentos computacionais, foram configurados dois cenários de classificação que se distinguem pelos vetores de representação textual e modelos preditivos adotados. No primeiro, foram utilizados algoritmos tradicionais de ML e modelos de representações textuais mais simples (CET, PoS e BoW), expandindo [10]. No segundo, são contempladas as arquiteturas de redes neurais profundas e as representações baseadas em *word embeddings*. Nos dois cenários foram adotadas F1-macro e acurácia como métricas de avaliação. A F1-macro é utilizada devido ao desbalanceamento do conjunto de dados, ela oferece uma análise de desempenho por classe, apresentando um resultado mais equilibrado do modelo. Para treinar e testar as diferentes abordagens, dividimos os conjuntos conjunto de dados em conjuntos de treinamento (70%) e teste (30%) de forma estratificada, visando preservar as disposições das classes.

No primeiro cenário, foi utilizado os recursos da biblioteca *scikit-learn* [18] para a implementação dos algoritmos: SVM, *k-Nearest Neighbors* (kNN), *Decision Tree*, *Random Forest*, *Naïve Bayes* e *Logistic Regression*. A Tabela IV apresenta os algoritmos e suas respectivas parametrizações. Para a construção das representações textuais foi utilizado a biblioteca *Natural Language Toolkit* (NLTK) [19].

No segundo cenário, foram adotadas três arquiteturas de redes neurais profundas que têm oferecido resultados significativos em problemas de classificação textual: CNN, LSTM e GRU. Ambas recebendo representações textuais baseadas em *word embedding* como entrada, tal como descrito na Tabela II. Para a construção dos vetores *embeddings* foi utilizado os recursos da biblioteca python *Gensim* [20] e para o desenvolvimento das arquiteturas multicamadas foi adotado a biblioteca python *TensorFlow* [21].

Nos experimentos com DL, foram aplicados testes com várias configurações de parametrização nas camadas das arquiteturas buscando sempre a melhor combinação. Na CNN, foram avaliados os filtros convolucionais $f = \{50, 100\}$, tamanhos de *kernel* $k = \{2, 3, 4, 5\}$ e utilizando a função ativação ReLu. Na camada de *pooling* foi adotado a estratégia do *maxpooling* e, visando a prevenção de sobre-ajustes, foram aplicadas taxas de 50% de *dropout* durante os treinamentos. Na camada densa foi adotada a função de ativação *softmax*. A Fig. 1 mostra a estrutura sequencial em camadas das redes utilizadas.

Nos experimentos com a LSTM, foram adotadas quantidades diferentes de neurônios/unidades $units = \{10, 20, 50, 70, 100\}$ na camada LSTM, com função de ativação ReLu. Para a camada de *pooling*, foram observadas três configurações: (i) sem estratégia de *pooling*, com o parâmetro $return\_sequence$ desativado, ou seja, sem o repasse de informações entre os neurônios; (ii) com $return\_sequence$ ativado e utilizando *GlobalMaxPooling* e (iii) também com comunicação entre os neurônios e utilizando *GlobalAveragePooling*. Nesta arquitetura também foram adotadas taxas de 50% de *dropout* nos treinamentos e uma camada densa com função de ativação *softmax* para obtenção das distribuições de probabilidades de classes dos subgêneros. Outra rede neural recorrente adotada nesta pesquisa é a GRU. Dado que sua estrutura é similar a LSTM, as configurações experimentais para esta rede são as mesmas.

É importante ressaltar que os dados utilizados nos experimentos passaram por filtros de tratamentos típicos da fase de pré-processamento de PLN, visando identificar e eliminar elementos ruidosos nos dados. Nesta etapa, foram realizadas: *(i)* remoção de *stopwords*, números, pontuação e caracteres especiais, *(ii)* conversão de caracteres para caixa baixa e *(iii)* radicalização das palavras. Tais procedimentos têm impacto direto na qualidade dos resultados do modelo final [22], [23].

V. RESULTADOS E ANÁLISES

As Tabelas V, VII, VI e VIII apresentam os resultados com os melhores desempenhos dos algoritmos em uma análise de melhor caso. Estão assinaladas em negritos as maiores acurácias, em caso de empate é sublinhado a representação com a menor dimensão, assumindo que tal representação tem maior eficiência em discriminar os dados, tal como realizado em [10]. A coluna "Dimensão" detalha os comprimentos dos vetores gerados.

Nos experimentos com o gênero *rock*, as representações mais simples e os algoritmos de ML obtiveram as maiores avaliações. A representação BoW produziu melhor resultado geral considerando a acurácia média para todos os classificadores ML, com 61%, em F1-macro BoW também é superior aos demais, com média de 53%. Do ponto de vista do algoritmo de classificação, o SVM foi o que obteve melhor resultado médio, com acurácia de 62% e F1-macro de 53%.

Os resultados também apontaram que combinações do BoW com as representações CET e PoS produzem ganho de desempenho no processo final de classificação, mostrando que tais características estruturais e gramaticais do conteúdo musical fornecem elementos relevantes para o aprendizado do algoritmo, como afirmado em [10]. Esse desempenho é visto no arranjo BoW/PoS, alcançando o melhor caso geral para

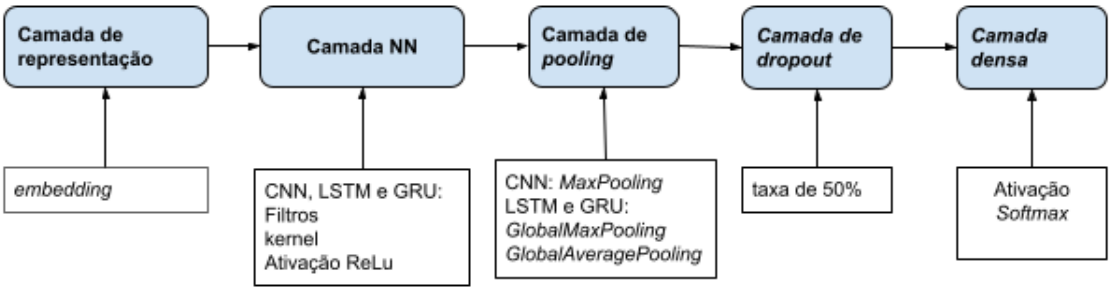

Fig. 1. Configuração das camadas da redes neurais utilizadas.

TABELA V
RESULTADOS DA ACURÁCIA PARA O GÊNERO *rock*.

Representação	Dimensão	KNN	SVM	*Random Forest*	*Naïve Bayes*	*Decision Tree*	*Logistic Regression*	CNN	LSTM	GRU	Média	Desvio Padrão
BoW	4.980	0.44	0.70	**0.62**	0.65	0.54	**0.68**	-	-	-	0.61	±0.09
PoS	7	0.46	0.48	0.47	0.49	0.41	0.49	-	-	-	0.46	±0.03
CET	10	0.50	0.51	0.49	0.48	0.46	0.51	-	-	-	0.49	±0.01
BoW|PoS	4.987	0.44	**0.71**	0.59	0.65	**0.54**	**0.68**	-	-	-	0.60	±0.10
BoW|CET	4.990	0.47	0.70	0.61	**0.66**	0.53	**0.68**	-	-	-	0.61	±0.09
PoS|CET	17	**0.51**	0.51	0.51	0.49	0.45	0.51	-	-	-	0.50	±0.02
BoW|PoS|CET	4.997	0.46	0.70	0.60	0.65	0.53	**0.68**	-	-	-	0.60	±0.09
CBoW	100	-	-	-	-	-	-	0.38	0.46	0.45	0.43	±0.04
CBoW	300	-	-	-	-	-	-	0.40	**0.50**	0.40	0.43	±0.06
Skip-gram	100	-	-	-	-	-	-	0.44	0.41	0.41	0.42	±0.02
Skip-gram	300	-	-	-	-	-	-	0.42	0.44	0.46	0.44	±0.02
FastText	100	-	-	-	-	-	-	**0.45**	0.47	0.45	0.46	±0.01
FastText	300	-	-	-	-	-	-	0.42	0.47	0.46	0.45	±0.03
Glove	100	-	-	-	-	-	-	0.44	0.48	**0.48**	0.47	±0.02
Glove	300	-	-	-	-	-	-	0.47	0.49	**0.48**	0.48	±0.01
Média	-	0.47	0.62	0.56	0.58	0.49	0.60	0.43	0.470	0.45	-	-
Desvio Padrão	-	±0.03	±0.11	±0.06	±0.09	±0.05	±0.09	±0.03	±0.03	±0.03	-	-

TABELA VI
RESULTADOS DA F1-MACRO PARA O GÊNERO *rock*.

Representação	Dimensão	KNN	SVM	*Random Forest*	*Naïve Bayes*	*Decision Tree*	Logistic Regression	CNN	LSTM	GRU	Média	Desvio Padrão
BoW	4.980	0.31	0.64	**0.51**	**0.63**	**0.48**	**0.63**	-	-	-	0.53	±0.13
PoS	7	0.31	0.33	0.36	0.36	0.35	0.30	-	-	-	0.33	±0.03
CET	10	0.37	0.38	0.41	0.41	0.40	0.35	-	-	-	0.39	±0.02
BoW|PoS	4.987	0.31	**0.65**	0.47	**0.63**	**0.48**	**0.63**	-	-	-	0.53	± 0.13
BoW|CET	4.990	0.35	0.64	0.50	**0.63**	**0.48**	**0.63**	-	-	-	0.54	±0.12
PoS|CET	17	**0.38**	0.42	0.41	0.40	0.39	0.36	-	-	-	0.39	±0.02
BoW|PoS|CET	4.997	0.34	0.64	0.49	**0.63**	0.47	**0.63**	-	-	-	0.53	±0.12
CBoW	100	-	-	-	-	-	-	0.29	0.32	**0.35**	0.32	±0.03
CBoW	300	-	-	-	-	-	-	**0.35**	0.33	0.28	0.32	±0.04
Skip-gram	100	-	-	-	-	-	-	0.31	**0.35**	0.33	0.33	±0.02
Skip-gram	300	-	-	-	-	-	-	0.32	0.32	**0.35**	0.33	±0.02
FastText	100	-	-	-	-	-	-	0.30	0.32	0.33	0.32	±0.02
FastText	300	-	-	-	-	-	-	0.32	0.32	0.33	0.32	±0.01
Glove	100	-	-	-	-	-	-	0.32	0.26	0.30	0.29	±0.03
Glove	300	-	-	-	-	-	-	0.31	0.29	0.33	0.31	±0.02
Média	-	0.34	0.53	0.45	0.53	0.44	0.50	0.32	0.31	0.33	-	-
Desvio Padrão	-	±0.03	±0.14	±0.06	±0.13	±0.05	±0.16	±0.08	±0.03	±0.02	-	-

as avaliações com *rock* com 71% de acurácia e 65% de F1-macro, no experimento com SVM. Resultados também corroborados nos experimentos com *pop* para SVM, destacados nas combinações BoW/CET e BoW/PoS/CET e em algumas combinações nas análises com KNN.

Os resultados dos classificadores para o gênero *pop*, apresentado nas Tabelas VII e VIII, mostram que a *Logistic Regression* obteve maior desempenho para a acurácia (com média de 65%) e *Naïve Bayes* para F1-macro (com média de 47%). Em análise de melhor caso, SVM e *Logistic Regression* alcançaram 67% de acurácia com BoW e algumas de suas combinações. Para F1-macro, o algoritmo *Naïve Bayes* obteve o melhor resultado também impulsionado por representações BoW.

Analisando os desempenhos das arquiteturas de redes neurais profundas e *word embedding*, os resultados mostraram uma vantagem da arquitetura de rede neural recorrente LSTM na análise da acurácia, com um pico de 50% no experimento com a representação CBoW de dimensão 300, e com 47% de acurácia média no cenário geral com *rock*. Para F1-macro, percebe-se um equilíbrio entre as três arquiteturas, ambas atingiram 35% quando utilizaram representações CBoW e *Skip-gram* como entradas.

Para as músicas *pop*, a arquitetura GRU obteve a melhor

TABELA VII
RESULTADOS DA ACURÁCIA PARA O GÊNERO *pop*

Representação	Dimensão	KNN	SVM	*Random Forest*	*Naïve Bayes*	*Decision Tree*	Logistic Regression	CNN	LSTM	GRU	Média	Desvio Padrão
BoW	4.980	0.59	0.61	0.64	**0.65**	**0.56**	**0.67**	-	-	-	0.62	±0.04
PoS	7	0.61	0.62	0.60	0.59	0.47	0.62	-	-	-	0.59	±0.06
CET	10	**0.62**	0.61	0.60	0.55	0.51	0.63	-	-	-	0.59	±0.05
BoW|PoS	4.987	0.60	0.59	**0.65**	**0.65**	0.53	**0.67**	-	-	-	0.62	±0.05
BoW|CET	4.990	0.61	**0.67**	0.63	**0.65**	0.54	**0.67**	-	-	-	0.63	±0.05
PoS|CET	17	**0.62**	0.60	0.61	0.49	0.49	0.63	-	-	-	0.57	±0.07
BoW|PoS|CET	4.997	**0.62**	**0.67**	0.64	0.62	0.54	**0.67**	-	-	-	0.63	±0.05
CBoW	100	-	-	-	-	-	-	0.61	**0.62**	0.62	0.62	±0.01
CBoW	300	-	-	-	-	-	-	**0.62**	0.61	0.57	0.60	±0.03
Skip-gram	100	-	-	-	-	-	-	**0.62**	0.61	**0.63**	0.62	±0.01
Skip-gram	300	-	-	-	-	-	-	0.60	**0.62**	0.62	0.61	±0.01
FastText	100	-	-	-	-	-	-	**0.62**	**0.62**	0.62	0.62	0
FastText	300	-	-	-	-	-	-	**0.62**	**0.62**	0.62	0.62	0
Glove	100	-	-	-	-	-	-	0.60	0.61	0.62	0.61	±0.01
Glove	300	-	-	-	-	-	-	0.60	0.61	0.61	0.61	±0.01
Média	-	0.61	0.62	0.62	0.60	0.52	0.65	0.61	0.62	0.61	-	-
Desvio Padrão	-	±0.01	±0.03	±0.02	±0.06	±0.03	±0.02	±0.01	±0.01	±0.02	-	-

TABELA VIII
RESULTADOS DA F1-MACRO PARA O GÊNERO *pop*.

Representação	Dimensão	KNN	SVM	*Random Forest*	*Naïve Bayes*	*Decision Tree*	Logistic Regression	CNN	LSTM	GRU	Média	Desvio Padrão
BoW	4.980	0.31	0.37	**0.41**	**0.53**	**0.42**	0.45	-	-	-	0.42	±0.07
PoS	7	0.30	0.30	0.34	0.36	0.35	0.26	-	-	-	0.32	±0.04
CET	10	**0.34**	0.32	0.38	0.42	0.38	0.28	-	-	-	0.35	±0.05
BoW|PoS	4.987	0.30	0.35	0.40	0.53	0.41	0.46	-	-	-	0.41	±0.08
BoW|CET	4.990	0.30	0.49	0.39	**0.53**	**0.42**	**0.47**	-	-	-	0.43	±0.08
PoS|CET	17	0.33	0.39	0.38	0.41	0.37	0.29	-	-	-	0.36	±0.04
BoW|PoS|CET	4.997	0.28	**0.50**	0.40	0.52	0.41	**0.47**	-	-	-	0.43	±0.09
CBoW	100	-	-	-	-	-	-	0.33	0.31	0.30	0.31	±0.02
CBoW	300	-	-	-	-	-	-	0.33	0.30	0.32	0.32	±0.02
Skip-gram	100	-	-	-	-	-	-	0.32	**0.33**	0.31	0.32	±0.01
Skip-gram	300	-	-	-	-	-	-	0.30	**0.33**	0.29	0.31	±0.02
FastText	100	-	-	-	-	-	-	**0.35**	0.30	0.30	0.32	±0.03
FastText	300	-	-	-	-	-	-	0.31	0.30	0.30	0.30	±0.01
Glove	100	-	-	-	-	-	-	0.33	0.30	0.33	0.32	±0.02
Glove	300	-	-	-	-	-	-	0.32	0.32	**0.34**	0.33	±0.02
Média	-	0.30	0.39	0.39	0.47	0.39	0.38	0.32	0.31	0.31	-	-
Desvio Padrão	-	±0.02	±0.07	±0.02	±0.07	±0.03	±0.10	±0.02	±0.01	±0.02	-	-

avaliação geral via acurácia (63%), utilizando representação *Skip-gram* de dimensão 100, e a CNN para F1-macro (com 35%) em uma combinação com *FastText* de dimensão 100. Para este gênero, os resultados médios apontam para uma competitividade entre as três arquiteturas adotadas. No entanto, aponta-se diferença significativa entre as duas métricas, indicando influência do desbalanceamento do conjunto de dados.

Em comparação com os algoritmos de ML, os resultados com as redes neurais mostraram-se inferiores, o que pode ser justificado pelo tamanho da amostra de dados utilizados nas representações de entrada. É importante ressaltar que os modelos *embeddings* requerem um volume maior de dados para o treinamento vetores mais representativos. Um outro indicativo, é que usualmente os modelos de DL oferecem resultados mais acurados quando treinadas com amostras maiores de dados.

Observando os modelos pré-treinados, bem como os treinamentos das redes neurais, os resultados do *Glove* apresentam baixa variação para ambos os gêneros e para todas as arquiteturas adotadas. A dimensão do vetor também apresenta diferença mínima para estes modelos *Glove*, fator mais evidente nos testes com músicas do gênero *rock*. Os resultados também não indicam vantagens significativas para os modelos treinados com os dados musicais.

Fazendo um comparativo entre gêneros, os resultados mostram que *pop* apresenta melhor desempenho em relação a acurácia na identificação de subgêneros nos dois cenários experimentais. Fatores que podem justificar este resultado são: o desbalanceamento maior das amostras de *pop* em relação a *rock* e as diferenças nas estruturas dos textos. Em *pop* notou-se que músicas são mais extensas do que no *rock*. Por exemplo, em *pop* a média de sentenças, palavras, sílabas e caracteres são: 37, 227, 257 e 1.068, respectivamente, já em *rock* os correspondentes médios são: 35, 214, 245 e 1.017. Em PoS, também verificou-se uma maior incidência de classes gramaticais (verbos e pronomes) nas músicas *pop*, no subgênero *Power* foi detectado a maior diversidade de conteúdo. Tais indicativos podem oferecer uma maior variedade de palavras e seus relacionamentos contextuais para os modelos *embedding* adotados.

VI. CONCLUSÃO

A literatura mostra que técnicas de Processamento de Linguagem Natural têm impulsionado análises de dados nos mais variados domínios de pesquisa [7], [11]. Em MIR, essas técnicas podem ser aplicadas para identificar tendências e padrões a partir dos dados musicais. Inserida nesse campo, esta pesquisa investiga representações de palavras e textos que melhor discriminam os dados musicais (subgêneros baseados em letras), visando encontrar combinações de representação de texto e algoritmos de classificação que ofereçam melhores resultados. O estudo considerou dois gêneros musicais bem conhecidos pelo público: o *rock* = {*Heavy Metal, Punk, Soft*} e o *pop* =

{*Tecnopop, Power, Rock*}. Os experimentos mostraram que a representação *BoW* obteve o melhor desempenho considerando todos os cenários de testes e os algoritmos tradicionais de *ML* mostraram-se competitivos frente às técnicas baseadas em redes neurais profundas, com destaque para *Logistic regression* e SVM.

Em suma, ressalta-se que a ampla cobertura experimental, envolvendo vários modelos de representação textual e diversas estratégias de classificação, e a aplicação em um problema em aberto no campo da MIR cuja fronteira de separabilidade não é trivial, são contribuições importantes deste trabalho. Tais contribuições podem guiar novos estudos envolvendo classificação de subgêneros musicais e problemas de classificação baseados em texto de forma geral. Como trabalhos futuros, pretende-se aplicar análises estatísticas, adotar estratégias de representação de dados sensíveis ao contexto, ampliar o número de amostras e de gêneros musicais, além de adoção de estratégias focadas na melhora de desempenho dos modelos preditivos. Destaca-se que os códigos-fonte e dados utilizados nos experimentos encontram-se integral e publicamente disponíveis no repositório https://github.com/fabiolobato/music-representation visando proporcionar a reprodutibilidade do estudo.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) -DT-308334/2020; pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/2021; e Acordo de Cooperacão Técnica N° 02/2021 (Processo N° 38328/2020 -TJ/MA). Agradecemos tambem aos revisores(as) pelas sugestões que muito auxiliaram na melhora do trabalho.

REFERÊNCIAS

[1] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *Journal of Big Data*, vol. 6, p. 44, Jun 2019.

[2] J. Sun and S. K. Gupta, "Variational fuzzy neural network algorithm for music intelligence marketing strategy optimization," *Intell. Neuroscience*, vol. 2022, p. 10, jan 2022.

[3] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, pp. 303–319, April 2011.

[4] J. P. Bello, P. Grosche, M. Müller, and R. Weiss, "Content-based methods for knowledge discovery in music," in *Springer Handbooks*, Springer Handbooks, pp. 823–840, Springer, 2018.

[5] R. de Araújo Lima, R. C. C. de Sousa, H. Lopes, and S. D. J. Barbosa, "Brazilian lyrics-based music genre classification using a blstm network," in *Artificial Intelligence and Soft Computing* (L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J. M. Zurada, eds.), (Cham), pp. 525–534, Springer International Publishing, 2020.

[6] R. Akella and T.-S. Moh, "Mood classification with lyrics and convnets," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 511–514, 2019.

[7] E. Paiva, A. Paim, and N. Ebecken, "Convolutional neural networks and long short-term memory networks for textual classification of information access requests," *IEEE Latin America Transactions*, vol. 19, p. 826–833, Jun. 2021.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[9] A. Caparrini, J. Arroyo, L. Pérez-Molina, and J. Sánchez-Hernández, "Automatic subgenre classification in an electronic dance music taxonomy," *Journal of New Music Research*, vol. 49, no. 3, pp. 269–284, 2020.

[10] J. S. Junior, R. Rossi, and F. Lobato, "Uma abordagem baseada em letras para a descoberta de conhecimento da música brasileira: o sertanejo como um estudo de caso," in *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, (Porto Alegre, RS, Brasil), pp. 949–960, SBC, 2019.

[11] A. Patel and A. Arasanipalai, *Applied Natural Language Processing in the Enterprise*. O'Reilly Media, 2021.

[12] R. Patil, N. Bowman, and J. Wood, "Analysis of different types of word representations and neural networks on sentiment classification tasks," in *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 0473–0478, Oct 2021.

[13] W. Ling, C. Dyer, A. Black, and I. Trancoso, "Two/too simple adaptations of word2vec for syntax problems," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 1299–1304, Association for Computational Linguistics, 2015.

[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[16] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2392–2396, 2017.

[17] Letras, "Letras platform," 2022 [Online]. Avaliable: https://www.letras.mus.br/mais-acessadas/.

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

[19] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

[20] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, (USA), p. 265–283, USENIX Association, 2016.

[22] W. Etaiwi and G. Naymat, "The impact of applying different preprocessing steps on review spam detection," *Procedia Computer Science*, vol. 113, pp. 273–279, 2017. The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017).

[23] M. Pita and G. L. Pappa, "Strategies for short text representation in the word vector space," in *7th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 266–271, IEEE, 2018.

Fabrício Almeida do Carmo Received the B.S. degree in Computer Science from Federal University of Western Pará (UFOPA), Pará, Brazil, in 2018. He is currently pursuing Master's degree in Computer and Systems Engineering from State University of Maranhão (UEMA), Maranhão, Brazil. His current research interests include machine learning, deep learning and natural language processing.

Jorge Luiz Figueira da Silva Junior is graduated in Computer Science from the Institute of Engineering and Geosciences (IEG) of the Federal University of Western Pará (UFOPA), in Santarém/PA (2021). Since 2018 he has been working on Text Mining projects. More specifically, he has worked in the areas of data collection, pre-processing, topic extraction and supervised learning for text classification.

Rafael Geraldeli Rossi received the B.S degree in Information Systems and Ms and PhD degrees in Computer Science and Computational Mathematics from University of Sao Paulo, Brazil. He is currently an Data Scientist at the biggest FoodTech Company in Brazil (iFood). His research interests include machine learning, text mining and graph-based methods.

Fábio Manoel França Lobato is a Lecturer of Computing at the Federal University of West Pará and leads the Applied Computing Research Group. He has a Productivity Grant in Technological Development and Innovative Extension from the National Council for Scientific and Technological Development (CNPq). His research interests are data science, social media analytics, and electronic markets.