A Lyric-Based Approach for Brazilian Music Knowledge Discovery: Brazilian Country Music as a Case Study

Jorge L. F. Silva Junior[1], **Rafael G. Rossi**[2], **Fábio M. F. Lobato**[1]

[1]Instituto de Engenharia e Geociências - Universidade Federal do Oeste do Pará
Santarém, PA, Brasil

[2]Universidade Federal de Mato Grosso do Sul - Três Lagoas, MS, Brasil

jorgeluizfigueira@gmail.com, rafael.g.rossi@ufms.br

fabio.lobato@ufopa.edu.br

Abstract. *Computational techniques can be used to identify musical trends and patterns, helping people filtering and selecting music according to their preferences. In this scenario, researches claim that the future of music permeates artificial intelligence, which will play the role of composing music that best fits the tastes of consumers. So, extracting patterns from this data is critical and can contribute to the music industry ecosystem. These techniques are well known in the field of Musical Information Retrieval. They consist of the audio characteristics extraction (content) or lyrics (context), being the latter preferable because it demands lower computational cost and presenting better results. However, when observing state of the art, it was found that there is a lack of antecedents that investigate the extraction of Brazilian music patterns through lyrics. In this sense, the main goal of this work is to fill this gap through text mining techniques, analyzing the songs classification in the subgenres of Brazilian country music. This analysis is based on lyrics and knowledge extraction to explain how subgenres differ.*

Resumo. *Técnicas computacionais podem ser usadas para identificar tendências e padrões musicais, ajudando as pessoas a filtrar e selecionar músicas de acordo com suas preferências. Nesse cenário, pesquisas afirmam que o futuro da música permeia a inteligência artificial, que desempenhará o papel de compor músicas que melhor atendam aos gostos dos consumidores. Portanto, extrair padrões desses dados é fundamental e pode contribuir para o ecossistema da indústria da música. Essas técnicas são bem conhecidas no campo da recuperação de informações musicais. Eles consistem na extração das características de áudio (conteúdo) ou letra (contexto), sendo o último preferível por exigir menor custo computacional e apresentar melhores resultados. No entanto, ao observar o estado da arte, verificou-se a falta de antecedentes que investiguem a extração de padrões musicais brasileiros por meio de letras. Nesse sentido, o objetivo principal deste trabalho é preencher essa lacuna por meio de técnicas de mineração de texto, analisando a classificação das músicas nos subgêneros do sertanejo. Essa análise é baseada em letras e extração de conhecimento para explicar como os subgêneros diferem.*

1. Introdução

No último ano, o mercado fonográfico brasileiro faturou US$ 298,8 milhões, apresentando um crescimento de 5,7%[1] acima da média do mercado global [Sobota 2019]. Dentro desse cenário, o gênero com maior destaque é o sertanejo [Darlington 2019]. A partir de uma pesquisa conduzida pelo Datafolha, o sertanejo foi apontado como o estilo com maior preferência pelos ouvintes, obtendo 37% dos votos (3.933) [Barcinski 2018]. O mercado nesse estilo musical mantém-se aquecido com a produção de *shows* que variam de R$ 200 mil à R$ 500 mil nos últimos anos [Sales 2018].

Graças às plataformas de *streaming*, a forma de consumo da música mudou drasticamente, obrigando as gravadoras musicais repensarem suas estratégias de vendas [Schedl et al. 2018]. Com uma ampla diversidade de artistas e músicas no mercado, tornou-se difícil para as pessoas selecionarem suas músicas [Katarya and Verma 2018]. Para contornar esses problemas, técnicas computacionais podem ser empregadas para identificar tendências e padrões nas músicas, assim ajudar pessoas filtrarem e selecionarem músicas de acordo com seus gostos. Essas técnicas são bastante conhecidas no campo da Recuperação de Informação Musical (*Music Information Retrieval - MIR*) [Deshmukh and Kale 2018].

Ainda nesse panorama, pesquisas afirmam que o futuro da música permeia a inteligência artificial e que esta desempenhará o papel de compor músicas que mais se adéquem aos gostos musicais ou humor dos ouvintes [Sturm et al. 2019]. Assim, extrair padrões e informações desses dados é fundamental e pode contribuir positivamente para todos os atores desse rico e complexo ecossistema da indústria musical (ouvintes, gravadoras, produtores musicais, publicitários etc.) [Schedl et al. 2017].

Os sistemas baseados em *MIR* consistem na extração de características de áudio (conteúdo) ou de letras de música (contexto) [Knees and Schedl 2015]. Abordagens baseadas no contexto são preferíveis em virtude do menor custo computacional e melhores resultados se comparado às abordagens baseadas em conteúdo [Bello et al. 2018]. Para extrair padrões considerando as letras das música, pode-se fazer uso das representações *Bag-of-Words* (*BoW*), fazer uso de *Part-of-Speech* (*PoS*) *tagging*, ou ainda utilizar um conjunto de Características Estatísticas Textuais (CET) extraídas de métricas de ocorrência das palavras, como apresentados respectivamente em [Yang 2018], [Barman et al. 2019] e [Fell and Sporleder 2014]. Essas representações são submetidas à algoritmos de aprendizado de máquina [Knees and Schedl 2013] os quais são responsáveis por identificar padrões nas músicas de diferentes gêneros ou subgêneros. A modelagem de tópicos e extração de regras também têm sido exploradas como observado no trabalho de [Choi 2018] e [Barkwell et al. 2018], respectivamente.

Por meio da pesquisa no estado da arte, percebeu-se a escassez de antecedentes que exploram a extração de conhecimento da música brasileira com a abordagem baseada em letras de música. Além disso, os trabalhos relacionados comparam somente os gêneros e não os subgêneros, dado que a extração de padrões tão parecidos (subgêneros) por vezes não é uma tarefa trivial [Armentano et al. 2017]. Nesse contexto, o presente estudo visa analisar a extração de padrões das letras de música do gênero sertanejo como um estudo de caso. Como observado em [Flynn et al. 2016] e [Rasmussen and Densley 2017], o

[1]Dados dos Produtores Fonográficos Associados (Pro-Música Brasil)

sertanejo pode ser compreendido ainda em dois subgêneros que se diferem. De um lado o sertanejo masculino, subgênero formado de músicas interpretadas por artistas masculinos e o "feminejo", artistas femininas. Para isso, foram empregadas técnicas de Mineração de Textos que visam responder às seguintes Perguntas de Pesquisa (PP):

1. Como classificar os subgêneros sertanejo masculino e o feminejo?
2. Quais são as marcas estilísticas que os diferenciam?
3. Qual a temática dominante de cada um?

O restante deste artigo está organizado da seguinte forma. Nas Seções 2 e 3 serão explanados, respectivamente, os trabalhos relacionados e a extração de características da música. Na Seção 4, é descrita a abordagem proposta. Por fim, nas Seções 5 e 6 serão apresentados os resultados bem como a conclusão e trabalhos futuros.

2. Trabalhos Relacionados

Inicialmente, a extração de conhecimento musical era amplamente feita por entradas de áudio [Li et al. 2011]. Essa forma de extração obteve muito sucesso no passado. Tarefas que utilizam de letras de música obtinham resultados insatisfatórios. No entanto, com o surgimento de técnicas de Processamento de Linguagem Natural (PLN), a abordagem baseada em letras recebeu destaque devido a melhoria nos resultados e ao fato de ser computacionalmente menos custosa que a abordagem baseada em áudio. [Bello et al. 2018].

Na literatura, observa-se que muitos trabalhos que investigam letras de música fazem a tarefa de classificação de gênero. Essa tarefa é realizada com técnicas clássicas de mineração de textos, no qual um conjunto de algoritmos de Aprendizado de Máquina (AM), por meio de uma representação *BoW*, podem extrair padrões dos gêneros musicais com base nos termos das músicas e suas respectivas categorias. Posteriormente é possível generalizar esses dados e fazer predições de classes [Aggarwal 2018], que nesse contexto equivale a classificação de gênero.

A exemplo disso, [Barman et al. 2019] defende que a *BoW*, como característica descritiva das músicas, por si só consegue prover resultados de classificação de diferentes gêneros de maneira acurada. Os autores realizaram a classificação de gênero para 8 tipos distintos, utilizando de algoritmos tradicionais de AM como *k-Nearest Neighbor* (*kNN*), *Support Vector Machine* (*SVM*), *Random Forest*, obtendo como resultado uma performance de classificação de 60% de acurácia na distinção dos gêneros. Em [Yang 2018], os resultados de [Barman et al. 2019] são corroborados, com experimentos de classificação para 7 gêneros com *BoW* e *PoS*. A representação *BoW* teve um importante papel na classificação, promovendo a melhor performance de classificação de acurácia (66%). Embora as características fornecidas pelo *PoS* resultaram em uma acurácia menor (63%), notou-se ser um forte descritor de estilo, o que pode fazer significantes distinções entre os dados e obter um melhor desempenho em um gênero em particular do que em outro.

No que diz respeito à modelagem de tópicos, o *Latent Dirichlet Allocation* (*LDA*) é utilizado em diversas aplicações como a redução de dimensionalidade, agrupamento de termos semanticamente relacionados em um único atributo, bem como separar textos que tratam de um mesmo tema [Aggarwal 2018]. Em [Choi 2018] e [Laoh et al. 2018], os autores utilizaram a modelagem de tópicos com o intuito de identificar os temas nas letras

de música. Como resultado, notaram que essa técnica permitiu uma categorização das músicas de acordo com seus temas.

Em uma linha semelhante, o trabalho de [Sasaki et al. 2014] propôs o desenvolvimento de um sistema de busca interativa de música, que por meio de *LDA*, indicava as músicas baseadas nos tópicos de interesse. Há diferentes paradigmas de AM, e consequentemente diferentes formas de exibir os padrões extraídos. Por exemplo, as árvores de decisão são facilmente interpretadas por um humano. Sendo assim, pertencem ao paradigma simbólico [Tan 2018]. Cada caminho de uma árvore de decisão corresponde à uma regra [Aggarwal 2018]. As regras cobrem exemplos majoritariamente ou unicamente pertencentes à uma classe do conjunto de dados. Cada regra é composta por testes lógicos considerando termos e frequência dos termos [Freitas 2014].

3. Extração de Características da Música

Para realizar a mineração de textos a partir de letras de músicas, estas precisam estar em uma representação estruturada que seja interpretável pelos algoritmos de aprendizado de máquina [Knees and Schedl 2013, Rossi 2016]. A representação no modelo espaço-vetorial é considerada por grande parte desses algoritmos [Aggarwal 2018]. Na Tabela 1 é ilustrado esse modelo de representação no formato matricial, matriz atributo-valor, para uma coleção com m letras de música e n atributos [Tan 2018]. Neste estudo, o conjunto de letras de música é denotado por $L = \{l_1, l_2, ..., l_m\}$ e o conjunto de atributos por $A = \{a_1, a_2, a_3, ..., a_n\}$.

Tabela 1. Representação no espaço vetorial para *m* letras de música e *n* atributos.

	$\mathbf{a}_1$	$\mathbf{a}_2$	$\mathbf{a}_3$	...	$\mathbf{a}_n$
$\mathbf{l}_1$	w_{l_1,t_1}	w_{l_1,t_2}	w_{l_1,t_3}	...	w_{l_1,t_n}
$\mathbf{l}_2$	w_{l_2,t_1}	w_{l_2,t_2}	w_{l_2,t_3}	...	w_{l_2,t_n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathbf{l}_m$	w_{l_m,t_1}	w_{l_m,t_2}	w_{l_m,t_3}	...	w_{l_m,t_n}

As células da matriz quantificam a ocorrência de uma característica da música. Palavras presentes nas letras de música podem ser utilizadas como características. Neste caso, a representação é denominada *bag-of-words* [Harris 1954]. Entretanto, outras características podem ser extraídas e utilizadas. As características consideradas neste trabalho são apresentadas nas subseções seguintes.

3.1. *Bag-of-Words*

As letras de músicas podem ser descritas por uma representação *Bag-of-Words*. Nela o conjunto de palavras da coleção musical é representado em um espaço multidimensional, onde cada música torna-se um vetor e cada palavra passa a ser uma característica, não importando sua ordem ou semântica. Diferentes esquemas de peso podem ser definidos para representar os respectivos valores dessas palavras. Podem ser utilizados os pesos: (i) *tf* a frequência de uma palavra no documento; (ii) *tf-idf*, no qual computa a frequência de uma palavra no documento pelo inverso da frequência de documentos; e a (iii) *binary*, que apenas indica a ocorrência de uma palavra no documento [Harris 1954]. Para tarefa de

classificação de subgêneros, foi escolhido o peso *tf-idf* em razão de sua eficácia em tarefas de mineração de texto [Vijayarani et al. 2015]. Vale ressaltar que os termos das letras podem passar por um processo de limpeza, padronização e simplificação, para diminuir o número de termos e melhorar a qualidade da representação.

3.2. *Part-of-Speech*

Part-of-Speech tagging ou marcação de partes da fala, é uma rotulação gramatical das palavras de acordo com sua definição e o contexto em que elas aparecem (*e.g.* substantivos, verbos, artigos). Em [Yang 2018], o autor sugere *PoS* como forte descritor de estilo. Diante disso, formulou-se a premissa de que diferentes gêneros podem se diferir nas classes gramaticais. Portanto, foram extraídos vários descritores de fala nas letras e feito a contagem de: substantivos, verbos, pronomes, adjetivos e conectivos. Computou-se também métricas derivadas como a incidência de conteúdo, dada pela somas das classes de substantivos, verbos e adjetivos, e a diversidade de conteúdo, uma taxa da incidência de conteúdo sobre o total de palavras do documento. Nessa tarefa foi utilizado um *PoS tagger* para o português brasileiro, validado em [Fonseca et al. 2015] com desempenho de 97,33% de acurácia. Como a base de dados lida com documentos de diferentes tamanhos, os valores dos atributos foram normalizados em uma escala de razão de 1000.

3.3. Estatísticas Textuais

Documentos de texto também podem ser descritos por medidas estatísticas simples [Choi 2018]. Características textuais como tamanho médio das palavras, proporção de palavras únicas no vocabulário e quantidade de sílabas foram elencadas. Uma das hipóteses desse trabalho é que estas características variam de acordo com os diferentes gêneros e podem dar uma indicação da complexidade, bem como o padrão de estruturação das letras. A lista de características consideradas neste trabalho é apresentada na Tabela 2.

Tabela 2. Características Estatísticas Textuais (CET).

Nome da característica	Descrição
Caracteres	Simples contagem
Tamanho Médio da Palavra	Razão de caracteres sobre quantidade de palavras
Palavras	Contagem das palavras
Palavras Únicas	Quantidade de palavras únicas
Sentenças	Quantidade de linhas
Média de Palavras por Sentença	Razão do total de palavras sobre total de sentenças
Sílabas	Contagem das sílabas
Média de Sílabas por palavra	Razão do total de sílabas sobre total de palavras
Taxa de palavras raras	Razão das palavras que ocorrem uma única vez sobre total de palavras
Diversidade Lexical	Razão de palavras únicas sobre total de palavras

4. Proposta

Nas próximas subseções são apresentados os detalhes da proposta do artigo referentes às coleções de letras de música e técnicas de pré-processamento utilizadas e tarefas comuns de *MIR* como: classificação de gênero, extração de regras e modelagem de tópicos. Ambas utilizadas para se realizar a descoberta de conhecimento musical baseado em letras de música.

4.1. Aquisição de dados

Neste estudo, os dados foram obtidos por meio da plataforma *online* Letras[2] que fornece letras de músicas de forma colaborativa, ou seja, seus próprios usuários enriquecem o repositório ao submeterem as letras de música que ainda não estão cadastradas. Existem diversos *websites* que provêm esse mesmo serviço. No entanto, a escolha dessa plataforma foi em razão de, na maioria dos casos, a música escrita corresponder com fidedignidade à música ouvida. Com o objetivo de prevenir resultados enviesados pela ocorrência de muitas letras de um mesmo artista, foi definido arbitrariamente a aquisição de 12 letras de músicas para cada um dos 5 principais artistas de cada subgênero, o que totalizou uma amostra de 120 letras.

4.2. Pré-processamento

Ao realizar aplicações com mineração de textos, uma tarefa determinante na qualidade dos resultados é a preparação dos dados. Isso exige um tratamento na coleção por meio de etapas que visam reduzir uma grande quantidade de ruído como: (i) conversão de caracteres para caixa baixa; (ii) remoção de acentuação; (iii) remoção de pontuação e caracteres especiais; (iv) remoção de números; (v) remoção de *stopwords*; (vi) radicalização das palavras [Aggarwal 2018]. Para conduzir essa preparação no conjunto de letras musicais, foi utilizado métodos de pré-processamento voltados para textos na língua portuguesa, semelhantes ao conduzido em [Cirqueira et al. 2018].

4.3. Classificação de Subgêneros

Para a classificação de subgêneros, foi utilizada a biblioteca *Scikit-Learn* [Pedregosa et al. 2011]. Foi executado um conjunto de experimentos que abrangem diversos algoritmos de classificação com diferentes parâmetros e combinações das características *BoW*, *PoS* e CET descritas na Seção 3. Os algoritmos de classificação bem como a parametrização utilizada é apresentada na Tabela 3. Vale ressaltar que o algoritmo *Naïve Bayes* não possui parâmetros, sendo que as descrições mencionadas correspondem às variações deste algoritmo (modelos de distribuição de probabilidade).

Tabela 3. Algoritmos de classificação.

Algoritmos	Parâmetros utilizados
kNN	k = [1-21], distance = {euclidean, cosine, manhattan}
Naïve Bayes	Bernoulli, Complement, Gaussian, Multinomial
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 = {linear, poly, rbf}
Decision Tree	criterion={gini, entropy}
Random Forest	n_estimators=[1-10], min_samples_split=[2-10]

4.3.1. Avaliação da Performance de Classificação

Com a finalidade de encontrar o melhor modelo de classificação, bem como descobrir quais características melhor contribuem para essa tarefa, os modelos foram avaliados pela

[2]https://www.letras.mus.br/

acurácia, uma métrica que pontua o percentual de acerto nas predições. Para validação dos resultados, cada base foi testada 10 vezes para cada algoritmo com o uso de *cross-validation* com 10 *folds*.

4.4. Extração de Regras

Algoritmos de árvores de decisão podem ser usados tanto como modelos descritivos quanto preditivos [Tan 2018]. Descritivos porque permitem entender como os documentos são classificados uma vez que pertencem ao paradigma de aprendizado simbólico, a partir das regras lógicas criadas para cada nó da árvore com base nos pesos dos atributos do conjunto de dados. Recuperar o conjunto dessas regras desempenha um papel à engenharia clássica de descoberta de conhecimento [Aggarwal 2018].

4.5. Modelagem de Tópicos

Para modelagem de tópicos, foi conduzida a mesma técnica utilizada em [Sasaki et al. 2014], [Choi 2018], [Laoh et al. 2018] e [Lobato et al. 2018], que optaram pela escolha do *LDA* devido sua simplicidade e rapidez. Este modelo considera como dado de entrada, uma representação *BoW* descrita na Seção 3, produzindo uma matriz termo-tópico. Deste modo, o *LDA* admite cada tópico como uma distribuição de probabilidade sobre um conjunto de palavras de um conjunto de documentos. Vale ressaltar que a extração de tópicos é um tipo de aprendizado não supervisionado, ou seja, não considera os rótulos, nesse caso o gênero das músicas no aprendizado. A análise dos resultados da modelagem de tópicos é por muitas vezes subjetiva [Chen et al. 2013].

Foram verificadas as seguintes opções de construção de tópicos e palavras (T-P), respectivamente: 3-5; 3-10; 5-3; 5-5; 5-10; 10-5; e 10-10. Ao alternar entre esses esquemas notou-se uma grande dispersão, como a ocorrência repetitiva de tópicos ou palavras; ou então as composições eram insuficientes para a rotulação pelos anotadores. Dado que a escolha desses esquemas é arbitrária [Brookes and McEnery 2019, Rodrigues et al. 2019], procurou-se a composição que melhor permitiria a facilidade de anotação dos tópicos. Nesse estudo a análise da modelagem de tópicos consiste em uma análise manual das palavras mais representativas de cada tópico. Em busca de encontrar as diferenças temáticas entre os subgêneros, primeiramente o conjunto de letras de música foi dividido entre as composições masculinas e femininas e posteriormente feita a modelagem sobre esses dois conjuntos.

5. Resultados

Na Tabela 4 são reportadas as maiores performances de classificação dos algoritmos considerados na avaliação, i.e., foi realizada a análise de melhor cenário. Estão assinaladas em negritos as maiores performances, e em caso de empate, está assinalada em sublinhado a representação com menor número de atributos. Na Tabela 5 é apresentada a parametrização que obteve melhor performance de classificação para cada experimento. As representações *BoW* e *PoS* obtiveram performance de acurácia semelhante à relatada em [Barman et al. 2019] e [Yang 2018]. A combinação das representações apresentou impacto significativo na performance apenas quando utilizado PoS/CET no algoritmo *Decision Tree*. No entanto, destaca-se que a representação CET atingiu a melhor performance na maioria dos casos. Diante da 1ª PP, conclui-se que CET melhor discrimina os subgêneros sertanejo masculino e feminejo.

Tabela 4. Melhores performances de classificação.

Exp.	Representação	Dim.	*kNN*	*Naïve Bayes*	*SVM*	*Decision Tree*	*Random Forest*	Média	Desvio Padrão
1	BoW	1130	0.57	0.58	0.57	0.50	0.58	0.56	±0.03
2	PoS	7	0.64	0.48	0.52	0.54	0.61	0.55	±0.06
3	CET	10	0.69	**0.58**	**0.62**	0.55	**0.70**	0.63	±0.06
4	BoW / PoS / CET	1147	0.69	0.57	0.61	0.57	0.68	0.62	±0.05
5	BoW / PoS	1137	0.59	0.57	0.53	0.59	0.55	0.56	±0.02
6	BoW / CET	1140	**0.70**	**0.58**	0.61	0.61	0.55	0.61	±0.05
7	PoS / CET	17	0.67	0.53	**0.62**	**0.64**	0.60	0.61	±0.05
	Média		0.65	0.55	0.58	0.57	0.61		
	Desvio Padrão		±0.05	±0.03	±0.04	±0.04	±0.06		

Tabela 5. Melhores parâmetros obtidos para cada experimento.

Exp.	*kNN*	*Naïve Bayes*	*SVM*	*Decision Tree*	*Random Forest*
1	euclidean, k: 3	Bernoulli	C: 0.0001, kernel: poly	entropy	min_samples_split: 4, n_estimators: 9
2	cosine, k: 9	Gaussian	C: 10000, kernel: linear	entropy	min_samples_split: 6, n_estimators: 1
3	manhattan, k: 17	Multinomial	C: 100000, kernel: rbf	gini	min_samples_split: 6, n_estimators: 8
4	manhattan, k: 17	Multinomial	C: 10, kernel: linear	gini	min_samples_split: 5, n_estimators: 9
5	cosine, k: 5	Bernoulli	C: 1, kernel: linear	gini	min_samples_split: 2, n_estimators: 5
6	manhattan, k: 17	Bernoulli	C: 10, kernel: linear	gini	min_samples_split: 2, n_estimators: 1
7	euclidean, k: 11	Bernoulli	C: 100000, kernel: rbf	gini	min_samples_split: 3, n_estimators: 5

Nas Tabelas 6, 7 e 8 são apresentadas um conjunto de regras extraídas por meio do algoritmo *Decision Tree* para os três tipos de representações utilizadas: *BoW*, *PoS* e CET. Foram extraídas as regras que possuem maior percentual de cobertura para cada coleção de letras. Para melhor compreensão das regras extraídas da *BoW* foi adotado para esta representação o esquema de peso termo-frequência. Ao comparar as regras da *BoW*, pode-se perceber que o feminejo não faz uso dos termos "*cachaça*", "*doida*" ou "*casa*", enquanto que o sertanejo masculino faz mais uso dos termos "*difícil*" e "*esquecer*".

Ao analisar as regras extraídas da representação *PoS*, foi observado que a incidência de verbos no feminejo é maior que no sertanejo masculino. Por fim, analisando os resultados das regras considerando a representação CET, pode-se observar que o número de palavras únicas é maior no feminejo bem como o número de caracteres. Portanto, as letras das músicas tendem a ser maiores no feminejo e responde a 2ª PP.

Tabela 6. Conjunto de regras extraídas da *BoW*

Subgênero	Regras	Abrangência
Sertanejo masculino	cachaça>=2, desespero >=1, difícil >=1, espalhando >=1, acabou >=3, casa >=1, aconteceu >=2, esquecer >=2	45%
Feminejo	cachaça <=2, encontro <=2, esquecer <= 2, sorriso <=3, doida <=1, casa<=1, somos <=1, paixão <=1, explico <=1, odiar <=1 e pode<=4	81%

Tabela 7. Conjunto de regras extraídas do *PoS*.

Subgênero	Regras	Abrangência
Sertanejo masculino	Incidência de Substantivos <= 0.079, Incidência de Conteúdo <= 0.137, Incidência de Adjetivos <= 0.008, Incidência de Verbos <=0.06, Diversidade de Conteúdo <= 0.456	48%
Feminejo	Incidência de Substantivos <= 0.079, Incidência de Conteúdo <= 0.137, Incidência de Adjetivos <= 0.008, Diversidade de conteúdo >= 0.064, Incidência de Verbos >= 0.06	27%

Tabela 8. Conjunto de regras extraídas da CET.

Subgênero	Regras	Abrangência
Sertanejo masculino	Palavras Únicas<=59.5, Caracteres>= 495, Palavras <=211	27%
Feminejo	Palavras Únicas>=59.5, Sílabas<= 495, Tamanho médio da palavra >=3.859, Caracteres>= 1077	38%

Mediante a análise do melhor cenário para modelagem de tópicos descrita na Seção 4, optou-se pela escolha da extração de 5 tópicos e seus 10 principais termos. O esquema de rotulação dos tópicos procedeu-se da seguinte forma: três avaliadores manualmente anotaram o tema que julgaram predominante analisando os principais termos de cada tópico, após, essas rotulações foram cruzadas a fim de gerar um consenso. Os resultados da modelagem de tópicos são apresentados nas Tabelas 9 e 10.

Tabela 9. Tópicos do subgênero sertanejo masculino.

Tópico	Principais termos
Romantismo	amor vida foi quero coracao tem boca falar sorriso ficar
Separação	amar medo hora sinceros revolta olhar sentimento calejado acabou passageiro
Desconfiança	sonha vida ficar faz medo vez amor mudou gente sou
Saudade	louca trocar tem ser povo coracao problema saudade deixa mao
Embriaguez/Sofrimento	esquecer amo coracao beber volta beijo amor sabe quero manda

Tabela 10. Tópicos do subgênero feminejo.

Tópico	Principais termos
Relacionamento	respeita chama corpo cama gostoso quiser desgrama prontas ser viu
Indignação	raiva volta pior leva ver aconteceu procurei ligacao sou rapariga
Embriaguez/Sofrimento	passa amor sofrer sorte quero deixa beber vida ter ver
Traição	mulher casa separada supera passa dando culpa beijo deixa sequencia
Sofrimento	ficar gente amor chorar preocupa fique negocio mordendo fiz amo

Ao analisar os tópicos encontrados pode-se observar que ambos subgêneros possuem temas em comum como "Embriaguez e Sofrimento". No entanto, ao considerar os temas que os diferem pode-se chegar à resposta da 3ª PP, na qual as letras de música no sertanejo masculino tendem à temáticas que retratam mais sobre romance e saudade enquanto a temática "decepção no relacionamento" é destacada no feminejo.

6. Conclusões e Trabalhos Futuros

De acordo com a literatura, características extraídas de letras podem ser utilizadas para desenvolver sistemas capazes de identificar e extrair padrões musicais. Ademais, a importância dessa atividade está nas contribuições para indústria musical obtidas ao analisar esses padrões. Contudo, foi observado a escassez de trabalhos que contemplem o análise de subgêneros, e possuam um escopo voltado para música brasileira, principalmente do sertanejo, fato este que motivou a conduta desta pesquisa.

Neste artigo, foram realizadas três principais tarefas pertencentes ao campo de Recuperação de Informação Musical: classificação de gêneros, extração de regras e modelagem de tópicos. Optou-se a escolha do gênero sertanejo como estudo de caso devido

ser o estilo musical de maior preferência e maior rentabilidade. Foram compreendidos como subgêneros do sertanejo as composições unicamente masculinas aqui denotadas por sertanejo masculino, e as composições femininas, aqui denominadas feminejo.

Os resultados apresentados neste trabalho demonstraram que é possível automatizar a classificação das músicas nos gêneros trabalhados, além de apresentar as distinções entre os subgêneros. Esse fato pode contribuir com sistemas de recomendação e recuperação, além de futuramente poder ser útil em sistemas de inteligência artificial para composição de músicas. Como trabalhos futuros, pretende-se estender esse trabalho considerando mais subgêneros tanto do sertanejo quanto de outros gêneros da música brasileira.

Agradecimentos

Os autores gostariam de agradecer a Universidade Federal do Oeste do Pará (UFOPA), Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação Tecnológica (Proppit) pelo financiamento parcial deste trabalho e a Pró-Reitoria de Ensino de Graduação (Proen) pela concessão de uma bolsa de intercâmbio por meio do Edital Nº 082/2018 - Programa de Mobilidade Acadêmica Externa Temporária Nacional, o qual viabilizou a parceria com o Grupo de Estudo e Pesquisa em Inteligência Computacional (UFMS / CPTL) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Processo nº 433082/2018-6.

Referências

Aggarwal, C. C. (2018). *Machine learning for text*. Springer.

Armentano, M. G., De Noni, W. A., and Cardoso, H. F. (2017). Genre classification of symbolic pieces of music. *Journal of Intelligent Information Systems*, 48(3):579–599.

Barcinski, A. (2018). Pesquisa comprova: no Brasil o sertanejo lidera, mas o futuro é do funk. Disponível em: `https://blogdobarcinski.blogosfera.uol.com.br/2018/07/24/pesquisa-comprova-no-brasil-o-sertanejo-lidera-mas-o-futuro-e-do-funk/`. Último acesso em: 28 de agosto de 2019.

Barkwell, K. E., Cuzzocrea, A., Leung, C. K., Ocran, A. A., Sanderson, J. M., Stewart, J. A., and Wodi, B. H. (2018). Big data visualisation and visual analytics for music data mining. In *Int Conf. Information Visualisation*, pages 235–240. IEEE.

Barman, M. P., Dahekar, K., Anshuman, A., and Awekar, A. (2019). It's only words and words are all I have. In *Eur. Conf. on Information Retrieval*, pages 30–36. Springer.

Bello, J. P., Grosche, P., Müller, M., and Weiss, R. (2018). Content-based methods for knowledge discovery in music. In *Springer Handbook of Systematic Musicology*, pages 823–840. Springer.

Brookes, G. and McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1):3–21.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., and Ghosh, R. (2013). Leveraging multi-domain prior knowledge in topic models. In *Twenty-Third Int. Joint Conf. on Artificial Intelligence*.

Choi, K. (2018). *Computational lyricology: quantitative approaches to understanding song lyrics and their interpretations*. PhD thesis, University of Illinois at Urbana-Champaign.

Cirqueira, D., Pinheiro, M. F., Jacob, A., Lobato, F., and Santana, Á. (2018). A Literature Review in Preprocessing for Sentiment Analysis for Brazilian Portuguese Social Media. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 746–749. IEEE.

Darlington, S. (2019). Why Brazil's music industry is booming again. Disponível em: `https://www.billboard.com/articles/columns/latin/8507548/brazil-music-industry-top-markets-streaming`. Acesso em: 4 de junho de 2019.

Deshmukh, P. and Kale, G. (2018). A survey of music recommendation system. *Int. Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(3):1721–1729.

Fell, M. and Sporleder, C. (2014). Lyrics-based analysis and classification of music. In *Proc. 25th Int. Conf. on Computational Linguistics: Technical Papers*, pages 620–631.

Flynn, M. A., Craig, C. M., Anderson, C. N., and Holody, K. J. (2016). Objectification in popular music lyrics: An examination of gender and genre differences. *Sex roles*, 75(3-4):164–176.

Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2.

Freitas, A. A. (2014). Comprehensible Classification Models: A Position Paper. *SIGKDD Explor. Newsl.*, 15(1):1–10.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Katarya, R. and Verma, O. P. (2018). Efficient music recommender system using context graph and particle swarm. *Multimedia Tools and Applications*, 77(2):2673–2687.

Knees, P. and Schedl, M. (2013). A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 10(1):2.

Knees, P. and Schedl, M. (2015). Music retrieval and recommendation: A tutorial overview. In *Proc. Int. Conf. on Research and Development in Information Retrieval*, pages 1133–1136. ACM.

Laoh, E., Surjandari, I., and Febirautami, L. R. (2018). Indonesians' song lyrics topic modelling using latent dirichlet allocation. In *Int. Conf. on Information Science and Control Engineering*, pages 270–274. IEEE.

Li, T., Ogihara, M., and Tzanetakis, G. (2011). *Music data mining*. CRC Press.

Lobato, F. M. F., Silva, M., Coelho, K., Silva, C., and Pontes, F. (2018). Vamos falar sobre deficiência? Uma análise dos Tweets sobre este tema no Brasil. In *Brazilian Workshop on Social Network Analysis and Mining*.

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.

Rasmussen, E. E. and Densley, R. L. (2017). Girl in a country song: Gender roles and objectification of women in popular country music across 1990 to 2014. *Sex Roles*, 76(3-4):188–201.

Rodrigues, L., Junior, J., and Lobato, F. (2019). A culpa é dela! É isso o que dizem nos comentários das notıcias sobre a tentativa de feminicıdio de Elaine Caparroz. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 47–58, Porto Alegre, RS, Brasil. SBC.

Rossi, R. G. (2016). *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. PhD thesis, Universidade de São Paulo.

Sales, B. (2018). Quem são os cantores sertanejos mais bem pagos da atualidade? Disponível em: `https://segredosdomundo.r7.com/quem-sao-os-cantores-sertanejos-mais-bem-pagos-da-atualidade/`. Último acesso em 28 de agosto de 2019.

Sasaki, S., Yoshii, K., Nakano, T., Goto, M., and Morishima, S. (2014). Lyricsradar: A lyrics retrieval system based on latent topics of lyrics. In *Int. Society for Music Information Retrieval Conf.*, pages 585–590.

Schedl, M., Knees, P., and Gouyon, F. (2017). New paths in music recommender systems research. In *Proc. Conf. on Recommender Systems*, pages 392–393. ACM.

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., and Elahi, M. (2018). Current challenges and visions in music recommender systems research. *Int. Journal of Multimedia Information Retrieval*, 7(2):95–116.

Sobota, G. (2019). Indústria da música no Brasil tem crescimento acima da média internacional em 2018. Disponível em: `https://cultura.estadao.com.br/noticias/musica,industria-da-musica-no-brasil-tem-crescimento-acima-da-media-internacional-em-2018,70002777115`. Último acesso em 28 de agosto de 2019.

Sturm, B. L., Ben-Tal, O., Monaghan, U., Collins, N., Herremans, D., Chew, E., Hadjeres, G., Deruty, E., and Pachet, F. (2019). Machine learning research that matters for music creation: A case study. *Journal of New Music Research*, 48(1):36–55.

Tan, P.-N. (2018). *Introduction to data mining*. Pearson Education India.

Vijayarani, S., Ilamathi, M. J., and Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *Int. Journal of Computer Science & Communication Networks*, 5(1):7–16.

Yang, J. (2018). Lyric-based music genre classifcation. Master's thesis, University of Victoria.