Análise das Letras das Músicas Brasileiras mais Tocadas nas Rádios das Últimas Seis Décadas

Italo Lourenço Trindade[1], **Leandro Colombi Resendo**[2], **Jefferson Oliveira Andrade**[2], **Karin S. Komati**[2]

[1]Coordenação de Informática

[2]Programa de Pós-graduação em Computação Aplicada (PPComp)
Campus Serra do Instituto Federal do Espírito Santo (IFES)

italolt10@gmail.com, {leandro,jefferson.andrade,kkomati}@ifes.edu.br

Abstract. *This work carried out an analysis of the most played songs on Brazilian radios over 6 decades, from 1960 to 2019. The goal was to assess whether there was any difference in the characteristics of the songs' lyrics of the songs over the decades in relation to the level of word repetition and word quantities. Software components were developed for finding the lyrics of the most played songs for the desired period, as well as for cleaning and processing the data. The results obtained indicate a great variation in musical styles over the decades, as well as a significant increase in the number of words and the average number of words repeated in the lyrics. Thus, it is possible to see that Brazilian songs have become more repetitive in the last two decades and that they were much shorter in the 1960s.*

Resumo. *Este trabalho realizou uma análise das músicas mais tocadas nas rádios brasileiras em 6 décadas, de 1960 à 2019. O objetivo foi avaliar se houve diferença nas características das letras das músicas ao longo das décadas em relação ao nível de repetições de palavras e quantidades de palavras. Foram desenvolvidos componentes capazes de coletar as letras das músicas mais tocadas para o período determinado, realizar a tarefa de limpeza e processamento dos dados. Os resultados obtidos indicam uma grande variação nos estilos musicais ao longo das décadas, bem como um aumento significativo no número de palavras e no número médio de palavras repetidas nas letras. Foi possível perceber que as músicas brasileiras ficaram mais repetitivas nas últimas duas décadas, e que eram bem mais curtas na década de 60.*

1. Introdução

A recuperação de informações musicais (MIR, do inglês *Music Information Retrieval*) é considerada um tema de estudo de caráter interdisciplinar que consiste em recolher e analisar informações em músicas, como a sua representação simbólica (partitura), texto (letra), imagem (fotografia de um artista ou capa de álbum), gênero, artista, intérprete e outras características, através de diferentes técnicas e softwares [Schedl et al. 2014]. As informações recuperadas por MIR podem ser utilizadas em aplicações de [de Melo Faria et al. 2015], [de Araújo Lima et al. 2020],

[Araujo et al. 2019]: classificação de gênero de música, identificação do "mood" (sentimento que uma música provoca em uma pessoa), reconhecimento de uma música através de um pequeno trecho dela, similaridade de música, identificação do cantor, *ranking* de artistas, predição de sucesso de música, geração automática de *playlists* de músicas, sistemas de recomendação e transcrição automática.

Uma das formas do MIR é uso da letra da música. A letra da música é o texto contido nas composições para ser cantado ou, às vezes, recitado. As letras das músicas podem ser utilizadas como recurso didático para fomentar discussões acerca de assuntos polêmicos, conteúdos sociolinguísticos e sócio-históricos [Pereira 2015]. A letra pode servir como entrada de dados para diversos tipos de pesquisa e aplicações na área, tais como classificação automática de gêneros musicais, indexação de músicas, identificação de emoções, dentre outros [Ribeiro and Silla 2014].

Neste trabalho foi realizada a análise comparativa das letras das músicas mais tocadas nas rádios nas últimas seis décadas, de 1960 a 2019. Quando cita-se "a década de 60", considera-se que é o intervalo de 1960 a 1969, e assim por diante para as décadas de 70, 80, 90, 2000 e 2010. Como não existe uma base de dados disponível com as informações das letras, foi necessário criar um processo de extração (*web scraping*) e limpeza dos dados, que estão descritos na Seção 3. Análise dos resultados na Seção 4, concluindo o trabalho na Seção 5.

2. Trabalhos Correlatos

Nesta seção será descrito o estudo apresentado por [Powell-Morse 2015], o qual analisa as letras de 225 músicas inglesas, que ficaram pelo menos três semanas como #1 nas paradas da Billboard, entre os anos de 2005 e 2014. O objetivo era responder às perguntas: "Qual gênero é o mais sofisticado?", "Quais artistas são os mais idiotas?" e "Alguma canção de sucesso pode ser lida confortavelmente por um aluno da 1ª série?". Como o estudo não é acadêmico/científico, há o uso de adjetivos coloquiais como "idiota".

Foi calculada a média anual da quantidade de palavras das músicas de cada gênero musical, e com isso a resposta da primeira pergunta foi que o gênero "R&B/Hip-Hop" é mais sofisticado por ter uma média maior que os outros gêneros em quase todos os anos, exceto 2010. Para avaliar as outras duas perguntas, foi usada a métrica de análise de escrita da língua inglesa, o índice de notas Flesch-Kincaid [Solnyshkina et al. 2017], que calcula a dificuldade de compreensão de um texto em inglês, indicando à qual série escolar está associado. A resposta da segunda pergunta foi o grupo "Three Days Grace" (não muito conhecido no Brasil), e quanto à terceira pergunta, a resposta é "sim, todas as letras de todas as músicas". Também foi gerado um gráfico que indicava o nível de complexidade médio das letras das músicas dos últimos 10 anos, e a curva foi descendente, indicando que a complexidade do nível de leitura diminuiu. O trabalho de [Powell-Morse 2015] motivou este presente estudo, ao nos questionarmos: "será que as letras das músicas nacionais também teriam um comportamento de estarem ficando mais simples com o decorrer dos anos?" e "será que as letras das músicas nacionais também teriam um comportamento de estarem ficando mais repetitivas?". Na língua portuguesa, infelizmente, não há um índice tal qual o Flesch-Kincaid. Assim, este estudo se limitou a analisar a quantidade de palavras das letras musicais.

O trabalho de [Ribeiro and Silla 2014] apresenta um sistema para a recuperação

automática de letras de músicas na web, denominado Ethnic Lyrics Fetcher (ELF). O objetivo do trabalho é propor um novo mecanismo para a detecção e extração automática de letras de músicas. O presente trabalho se diferencia da proposta ELF, pois o objetivo maior não é o sistema em si, mas a análise sobre a evolução das letras das músicas no decorrer de seis décadas.

3. Materiais e Métodos

Nesta seção são detalhados os processos para a criação da base de dados das letras das músicas mais tocadas das últimas seis décadas. Para se obter a base, o processo foi separado em 3 componentes principais: *Search Component* (Componente de Pesquisa), *Cleaning Component* (Componente de Limpeza) e *Processing Component* (Componente de Processamento). O fluxo geral do sistema é uma sequência dos processos, como ilustrado na Figura 1.

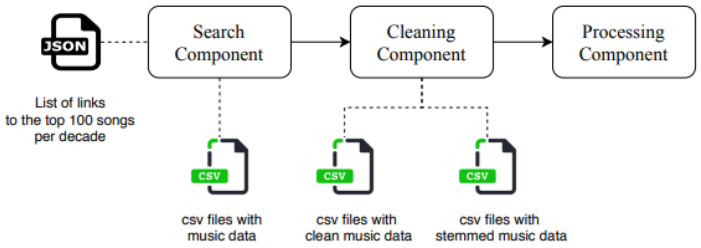

Figura 1. Fluxo geral do sistema.

O *Search Component* recebe como entrada um arquivo JSON (do inglês *JavaScript Object Notation*) contendo uma lista de *links* para a página que disponibiliza as 100 músicas mais tocadas nas rádios brasileiras separadas por década[1]. De acordo com esses dados, o componente é capaz de realizar a busca da música pelo seu título e nome do artist[2]. Assim, para cada década é gerado um arquivo **csv** contendo os dados brutos (BASE BRUTA) de cada letra de música (*csv files with music data*). O *Cleaning Component* tem a função de receber cada um desses arquivos **csv** gerados pelo *Search Component* e aplicar os métodos de limpeza de dados do texto, como retirada de pontuações e cifras. Como resultado final, o *Cleaning Component* também gera arquivos *csv* separados por décadas, a BASE LIMPA (*csv files with clean music data*). Ainda sobre a BASE LIMPA é realizado o processo de *stemming* (que transforma as palavras em sua forma canônica), gerando a BASE STEM (*csv files with stemmed music data*). Por último, o *Processing Component* responsável por realizar a leitura de cada arquivo *csv* gerado pelo *Cleaning Component* e processar esses dados para geração de gráficos e relatórios.

Como não foi encontrada base pública *online* com as letras das músicas brasileiras mais populares ao longo dos anos, foi desenvolvido módulo para coleta de dados. O trabalho de [Silva et al. 2019] de tornar uma base de dados de música ampla pública se concentra nos resultados da Billboard, já este trabalho se concentra nas músicas nacionais. O *Search Component* coleta as letras das 100 músicas mais tocadas nas rádios brasileiras separadas por década. O componente foi implementado em Node.js[3] e utiliza bibliotecas

[1]https://maistocadas.mus.br/
[2]https://github.com/italolourenco/music-web-scraping
[3]https://nodejs.org/en/

como Puppeteer[4] e Jsdom[5] para realizar os processos de *web scraping*. O componente faz a leitura desse arquivo JSON e cria objetos JSON para acessar e manipular as informações da página. Para cada objeto JSON processado, obtém-se a lista de músicas para o *link* informado. Para cada música da lista, busca-se a letra da música no site Vagalume, tal como em [da Silva et al. 2020], e armazena-se: o título da música, o nome do artista e a letra da música. Em caso de erro na busca do Vagalume, faz-se a pesquisa via Google. Caso a pesquisa seja um sucesso, armazenam-se as mesmas informações, caso contrário, descarta-se a música da base de dados. Para coletar apenas as letras em língua portuguesa foi usado o campo **lang** do resultado de busca na API do Vagalume.

O componente Cleaning Component inicia com a toquenização[6], depois a remoção de *stop words*, cifras (símbolos criados para representar o acorde), numerais, caracteres especiais e a preparação das letras para a BASE LIMPA e de *stemming* para a BASE STEM. Foi desenvolvido em Python 3, utilizando a biblioteca NLTK (*Natural Language Toolkit*). A tarefa de remoção de *stop words* retira do texto as palavras que ocorrem frequentemente (tais como 'a', 'de', 'o', 'da', 'que', 'e', 'do', portanto correspondem aos artigos, preposições e numerais, além de conjunções e pronomes), mas que na maioria das vezes não são informações relevantes para a construção da compreensão do texto. O *stemming* é a técnica que reduz cada palavra ao seu radical, a raiz de uma palavra pode ser encontrada eliminando os prefixos e sufixos que indicam variação na forma da palavra, como plural e tempos verbais. As palavras "meninas", "meninos" e "menininhos" se reduziriam a "menin". A Figura 2 apresenta um exemplo da letra após cada um dos passos.

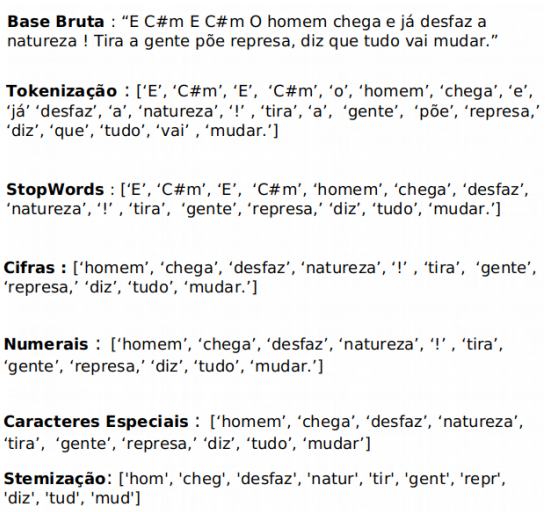

Figura 2. Exemplo do resultado da limpeza de dados do Cleaning Component, passo a passo.

4. Resultados e Discussão

O resultado da coleta de dados gerou três bases de dados: a base de dados inicial (BASE BRUTA), a base de dados inicial após o processamento de limpeza, denominada de BASE LIMPA e a base de dados após o *stemming*, a BASE STEM. A Tabela 1 traz os números consolidados de músicas coletadas, as que foram descartadas por serem de outro idioma e as que não foram encontradas.

Esta análise geral é dada pelo estilo mais marcante de cada década, de acordo com os artistas com mais músicas nas mais ouvidas nas rádios. Essa análise é simplista e foi feita baseada no principal gênero do artista. Não é o mais preciso, pois o mesmo

[4]https://pptr.dev/

[5]https://github.com/jsdom/jsdom

[6]*Tokens* são as unidades de textos e geralmente corresponde a uma palavra ou pontuação ou caractere especial.

artista pode versar sobre diferentes gêneros, e a análise mais específica deveria ser pela música em si. Mas foi o suficiente para analisar a evolução da popularização dos gêneros. A Figura 3 apresenta os gráficos das mais tocadas por gênero e por década, resumindo graficamente a descrição a seguir.

Tabela 1. Quantidade de música por status de cada década

	1960	1970	1980	1990	2000	2010	Total
Salvo	59	52	62	44	55	72	344
Outro Idioma	27	34	36	47	31	17	192
Não Encontrado	14	14	2	9	14	11	64

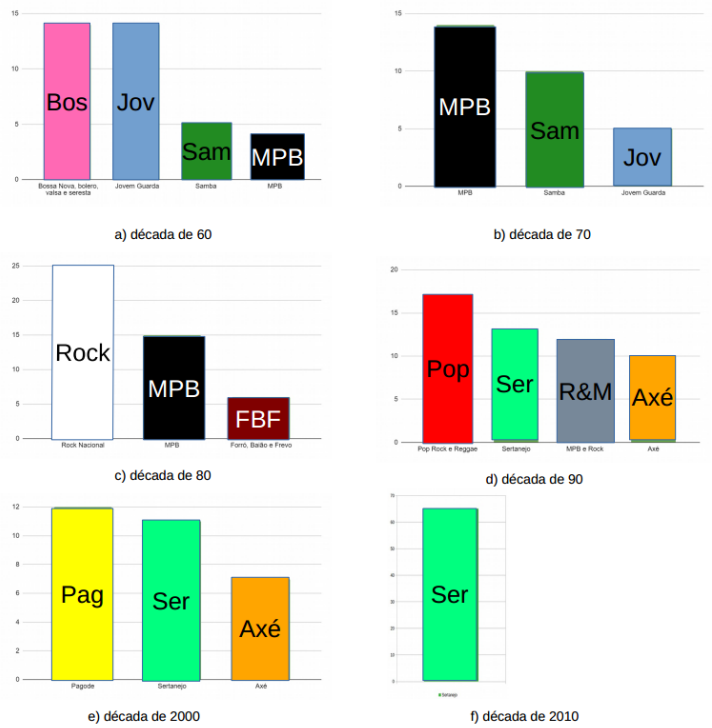

Figura 3. Gráficos das mais tocadas por gênero e por década.

No início da década de 60, há vários artistas de músicas românticas, no estilo de bossa nova, bolero, valsa e seresta contando com 14 músicas (barra rosa) que mais tocaram nas rádios, além do samba (barra verde escura). Começa a onda Jovem Guarda (barra azul) e o início do MPB (barra preta). A década de 70 foi muito diversa, mas o gênero de músicas românticas (bossa nova, bolero, valsa e seresta) que fizeram muito sucesso nos anos 50-60 perdem espaço e não aparecem mais no gráfico. A Jovem Guarda se mantém apenas com o Roberto Carlos, e o samba e o MPB são dominantes. A década de 80 foi marcada pelo rock nacional (barra branca), seguido pela MPB e com forte representação de músicas típicas brasileiras, tais como o forró, baião, e frevo (barra vinho, com o texto FBF). A década de 90 foi muito marcada pelo crescimento do sertanejo (barra verde clara) e o aparecimento do axé (barra laranja) e do pagode nas rádios. O pop (barra vermelha) se consagra e há uma mistura de rock com MPB e às vezes *reggae* (barra cinza com o texto 'R&M'). Na década de 2000, o axé continua, mas com menos representantes. O sertanejo e o pagode se mantém com muitos sucessos nas rádios. Já na década de 2010, o sertanejo

dominou a lista com 67 músicas das 72 músicas nacionais encontradas[7].

O gráfico *boxplot* da Figura 4 exibe a quantidade de palavras por música encontradas na BASE BRUTA, o eixo x a década e no eixo y a quantidade de palavras por música. Pelo gráfico, as músicas mais antigas (década de 60) apresentam a menor mediana. E a mediana apresenta uma tendência de crescimento com o passar das décadas. Na década de 1990, há muitas exceções como os pontos acima do limite superior da caixa, este fato se deve principalmente pelo sucesso das músicas de "Gabriel, o Pensador" que possuem letra extensa. O gráfico *boxplot* da Figura 5 exibe a quantidade de palavras repetidas por música encontradas na base STEM. É notável que a caixa da década de 2010 é maior que as demais, indicando que há muitas músicas que repetem palavras que possuem o mesmo radical em suas letras.

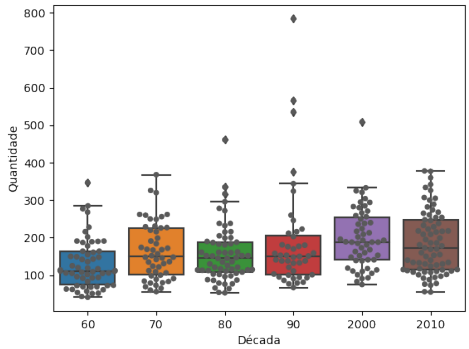

Figura 4. Gráfico da quantidade de palavras por música para a BASE BRUTA.

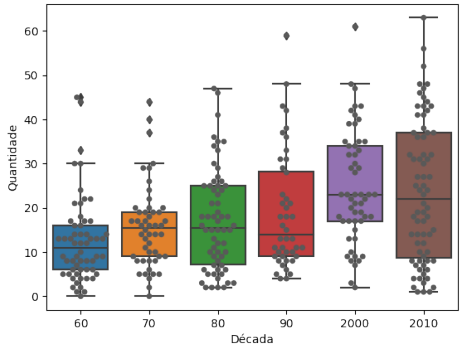

Figura 5. Gráfico da quantidade de palavras repetidas por música para a BASE STEM.

5. Considerações Finais

Neste trabalho foi realizada a coleta e análise das músicas mais tocadas nas rádios brasileiras no período de 1960 a 2019. Para realizar essa tarefa foram coletadas ao todo 344 letras de músicas. Nota-se a preferência por artistas românticos durante a década de 1960; a dominação do MPB e do rock nacional durante as décadas de 1970 e 80; o início da popularização do sertanejo, axé e pagode em 1990; a consolidação do pagode e sertanejo na década 2000 e por fim o sertanejo dominando as músicas mais tocadas nas rádios brasileiras na década 2010. Esse trabalho também avaliou a letra das músicas a fim de entender a variação da quantidade de palavras e repetições de palavras. A resposta da pergunta "será que as letras das músicas nacionais também teriam um comportamento de estarem ficando mais simples com o decorrer dos anos?" é de que pela quantidade de palavras por letra musical, a resposta é não, as letras ficaram cada vez mais longas pelo aumento do valor da mediana. E "será que as letras das músicas nacionais também teriam um comportamento de estarem ficando mais repetitivas?", pela análise de quantidade de palavras repetidas com o mesmo radical, a resposta é que sim, as letras de músicas ficaram mais repetitivas nas últimas duas décadas (2000 e 2010).

Como trabalhos futuros, pretende-se realizar a análise de mais décadas anteriores à década de 60, implementar técnicas para extrair o gênero musical de cada música e não

[7]https://veja.abril.com.br/cultura/sertanejos-ofuscam-musicos-internacionais-nas-radios-brasileiras/

pelo artista, avaliar as mais tocadas no YouTube BR, Spotify BR, e outros serviços de *streaming*, para avaliar como a preferência de cada canal é diferente das rádios, realizar análise de sentimentos nas letras obtidas e analisar o comportamento dos artistas ao longo das décadas, quanto à sua popularidade, comprimento das letras das músicas e mudança de estilo.

Referências

Araujo, C., Cristo, M., and Giusti, R. (2019). Predicting music popularity on streaming platforms. In *Anais do XVII Simpósio Brasileiro de Computação Musical*, pages 141–148, Porto Alegre, RS, Brasil. SBC.

da Silva, A. C. M., Silva, D. F., and Marcacini, R. M. (2020). 4mula - a multitask, multimodal, and multilingual dataset of music lyrics and audio features. In *Anais do XXVI Simpósio Brasileiro de Multimídia e Web*, pages 305–308, Porto Alegre, RS, Brasil. SBC.

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 *International Conference on Artificial Intelligence and Soft Computing*, pages 525–534. Springer.

de Melo Faria, F. L., Pereira Jr, Á. R., and Merschmann, L. H. (2015). Prediction of artists' rankings by regression. In *SBSI*, pages 95–102.

Pereira, P. G. (2015). As relações entre língua, cultura, música e o processo de ensino-aprendizagem de língua estrangeira. *Revista Estudos Anglo-Americanos*, (43):62–83.

Powell-Morse, A. (2015). Lyric intelligence in popular music: A ten year analysis. https://www.seatsmart.com/blog/lyric-intelligence/.

Ribeiro, R. and Silla, C. (2014). Recuperação inteligente de letras de músicas na web. In *Anais do XXXIII Concurso de Trabalhos de Iniciação Científica da SBC*, pages 41–50. SBC.

Schedl, M., Gómez, E., and Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8:127–261.

Silva, M. O., de Alencar Rocha, L. M., and Moro, M. M. (2019). MusicOSet: An enhanced open dataset for music data mining. In *XXXII Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2019 Companion*, Fortaleza, CE, Brazil. SBC.

Solnyshkina, M., Zamaletdinov, R., Gorodetskaya, L., and Gabitov, A. (2017). Evaluating text complexity and flesch-kincaid grade level. *Journal of Social Studies Education Research*, 8(3):238–248.