UNIVERSIDADE FEEVALE

GIOVANI POZZO JUNIOR

CONSTRUÇÃO DE UM *DATASET* DE INFORMAÇÕES MUSICAIS EXTRAÍDAS DE *POSTS* EM REDES SOCIAIS *ONLINE*

Novo Hamburgo

2019

GIOVANI POZZO JUNIOR

CONSTRUÇÃO DE UM *DATASET* DE INFORMAÇÕES MUSICAIS EXTRAÍDAS DE *POSTS* EM REDES SOCIAIS *ONLINE*

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação pela Universidade Feevale.

Orientador: Me. Roberto Scheid

Novo Hamburgo

2019

Giovani Pozzo Junior

Construção de um *Dataset* de Informações Musicais Extraídas de *Posts* em Redes Sociais *Online*

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação pela Universidade Feevale, sendo submetido à Banca Examinadora e considerado aprovado em __/__/2019.

_____

Prof. Me. Roberto Scheid
Professor Orientador

_____

Membro da Banca Examinadora

_____

Membro da Banca Examinadora

AGRADECIMENTOS

À minha avó, Delvair, por ter presenciado e se emocionado com todas as conquistas que tive na vida até aqui; por ter acompanhado de perto estes meus 5 anos de graduação; e pela certeza de que, mesmo tendo partido poucos meses antes deste término, está assistindo à minha gradução do camarote mais alto que existe.

À minha mãe, Silviah, pelo incondicional suporte e companheirismo, desde o dia em que eu nasci; à minha mãe, por absolutamente tudo.

Ao meu pai, Giovani, pelos valores e pelos tantos e sempre disponíveis abraços, festivos ou de consolo.

À minha irmã Danielle, por ser uma referência intelectual para mim, e por todo o suporte ao longo do curso e deste trabalho.

À minha irmã Rafaella, pela parceria irrestrita dos momentos difíceis aos mais prazerosos e divertidos.

Aos amigos que apoiaram e torceram por mim, todos; em especial aos de mais longa data, Ângelo, Dienifer e Hugo.

Aos amigos que fiz na FEEVALE, sem os quais essa trajetória acâdemica faria muito menos sentido, em especial aos sempre presentes Kevin, Andriel, Carlos e Fidelix.

Ao meu orientador, Prof. Me. Roberto Scheid, por todo o incentivo, compreensão e ensinamentos ao longo do curso e, especialmente, durante a construção deste trabalho.

Aos meus ídolos na música, por manterem vivo para mim o colorido do mundo; à Amy Winehouse, Billie Holiday, Caetano Veloso, Elza Soares, Maria Bethânia e tantos, tantos outros.

RESUMO

O crescente uso de redes sociais *online* gerou, na última década, um fenômeno capaz de alterar a maneira como as pessoas se comunicam e compartilham suas opiniões, gostos ou interesses, gerando diariamente uma vasta massa de dados. O *Global Digital Report* divulgou que a marca de usuários ativos na Internet chegou a 4 bilhões no ano de 2018, sendo, destes, 75% ativos nas redes sociais *online*. Ao mesmo passo, de acordo com a Federação Internacional da Indústria Fonográfica, no ano de 2016, 50% da receita total da indústria da música proveio de serviços digitais de distribuição de música. À medida em que a Internet é cada vez mais priorizada enquanto meio de distribuição de música, maior é a quantidade de faixas (comerciais e não comerciais) disponíveis aos consumidores – com um catálogo que atualmente ultrapassa 20 milhões de faixas. Por este motivo, ao longo dos últimos anos, pesquisas e empresas têm dedicado esforços em desenvolver soluções para garantir aos usuários recomendações musicais cada vez mais precisas. Dentro desta perspectiva, o presente trabalho busca responder à seguinte pergunta: é possível extrair informações de *posts* em redes sociais *online* que possam servir de insumo para geração de recomendações musicais relevantes? Para isto, através de métodos de processamento de linguagem natural e análise de sentimentos, é construído um *dataset* de contexto musical baseado em informações extraídas de *posts* de redes sociais *online* (*tweets*) que podem ser potencialmente utilizadas como insumo para geração de recomendações musicais. Além da geração deste *dataset*, este estudo apresenta também o algoritmo desenvolvido e utilizado para a coleta, análise e geração do *dataset* e introduz uma revisão bibliográfica acerca de temas relativos a redes sociais e sistemas de recomendação musical.

Palavras-chave: Sistemas de Recomendação. Música. Redes Sociais. *Twitter*.

ABSTRACT

The increasing use of *online* social networks has generated, in the last decade, a phenomenon capable of changing the way people communicate and share their opinions, tastes or interests, generating a vast mass of data daily. The Global Digital Report reported that the brand of active Internet users reached 4 billion in 2018, of which 75% are active in *online* social networks. At the same time, according the International Federation of the Phonographic Industry, in 2016, 50% of the total revenue of the music industry came from digital music distribution services. As the Internet is increasingly prioritized as a means of distributing music, the greater the number of (commercial and non-commercial) tracks available to consumers - with a catalog that currently exceeds 20 million tracks. For this reason, over the last few years, research and companies have dedicated efforts to develop solutions to ensure users with increasingly accurate music recommendations. From this perspective, this work seeks to answer the following question: is it possible to extract information from *posts* in *online* social networks that can serve as input for generating relevant musical recommendations? To answer this, through natural language processing methods and feelings analysis, a musical context *dataset* is constructed based on information extracted from *online* social networking *posts* (*tweets*) that can potentially be used as input for generating musical recommendations. In addition to the generation of this *dataset*, this study also presents the algorithm developed and used for the collection, analysis and generation of the *dataset* and introduces a bibliographic review about themes related to social networks and music recommendation systems.
Keywords: Recommender System. Music. Social Networks. *Twitter*.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Ao longo dos últimos anos, as tecnologias digitais têm assumido um papel essencial em mudanças drásticas observadas em diversos aspectos da vida em sociedade (CASTELLS, 2013). A consultoria global *We Are Social*, através do *Global Digital Report* (2018), divulgou que a marca de usuários ativos na Internet chegou a 4 bilhões no ano de 2018, sendo, destes, 75% usuários ativos nas redes sociais *online*.

Nas palavras de Mira e Bodoni (2011, p. 106),

> [...] a definição do termo rede social é um processo, por si só, interdisciplinar. [...] podemos constatar que estamos diante de um conceito absolutamente contemporâneo, que integra os seres humanos não mais como [...] uma multidão, e sim, um grupo organizado, que tem em suas conexões um padrão matematicamente avaliável (MIRA; BODONI, 2011, p. 106).

As redes sociais, para Tomaél, Alcará e Di Chiara (2005, p. 1), "constituem uma das estratégias subjacentes utilizadas pela sociedade para o compartilhamento da informação e do conhecimento“. Tais ferramentas proporcionaram, ainda, de acordo com Recuero (2009), que as pessoas pudessem construir suas identidades na rede de computadores, interagir e comunicar com outras pessoas e imprimir “rastros” na rede, que possibilitam o reconhecimento de padrões.

Ao mesmo passo em que o fenômeno das redes sociais *online* expande-se exponencialmente, a utilização em larga escala de serviços de *streaming* e consumo de música digital tem feito da Internet um dos principais meios de distribuição de música, permitindo-nos observar mudanças significativas no universo fonográfico; de acordo com dados divulgados pela Federação Internacional da Indústria Fonográfica (2017), estima-se que, no ano de 2016, 50% da receita total da indústria da música proveio de serviços digitais de distribuição de música, resultando em um faturamento de 7.85 bilhões de dólares americanos.

Em consequência a estes avanços, a vasta quantidade de faixas musicais disponíveis na rede pode dificultar aos usuários encontrarem músicas que apreciam, problema que remete ao fenômeno conhecido como “Paradoxo da Escolha” (SCHWARTZ, 2009). Desta maneira, observam Hu, Koren e Volinksy (2008), os sistemas de recomendação têm sido cada vez mais explorados pela sua capacidade de selecionar nas imensas quantidades de dados disponíveis itens relevantes a cada indivíduo. Para os autores, a importância e popularização do uso de sistemas de

recomendação se dá, ainda, pela sua habilidade de filtrar dados desnecessários ou irrelevantes da vasta e crescente massa de dados acessíveis.

De Sá (2009 apud LEVITIN, 2007) analisa que, com a democratização dos processos de produção, distribuição e armazenamento de música na nova realidade do mundo fonográfico, a seleção de faixas e a maneira ou facilidade com que o usuário as recupera assume papel mercadológico de potencial proeminência, argumento que endossa a relevância de se estudar recomendação musical automatizada no cenário aqui apresentado.

Diante do exposto, dá-se a seguinte pergunta de pesquisa: é possível extrair informações de *posts* em redes sociais *online* que possam servir de insumo para geração de recomendações musicais relevantes?

A partir desta pergunta, são descritos abaixo os objetivos que orientam a presente pesquisa.

1.1. OBJETIVOS

A seguir, estão descritos os objetivos da presente pesquisa.

1.1.1. Objetivo Geral

Através de métodos de processamento de linguagem natural e análise de sentimentos, construir um *dataset*[1] de dados extraídos de *posts* de redes sociais *online* que possam ser potencialmente utilizados como insumo para geração de recomendações musicais.

1.1.2. Objetivos Específicos

O escopo do trabalho delimita-se pela resolução dos seguintes objetivos específicos:

[1] Um *dataset* é um conjunto de dados estruturados. Estes dados devem ser recuperáveis no formato de uma entidade única, ou seja, estarem formatados em uma estrutura que permite a sua leitura e/ou processamento na íntegra (MELODA, 2019).

a) Analisar a bibliografia relacionada a redes sociais, sistemas de recomendação e ao estado-da-arte de algoritmos de recomendação musical baseada em dados em larga escala;

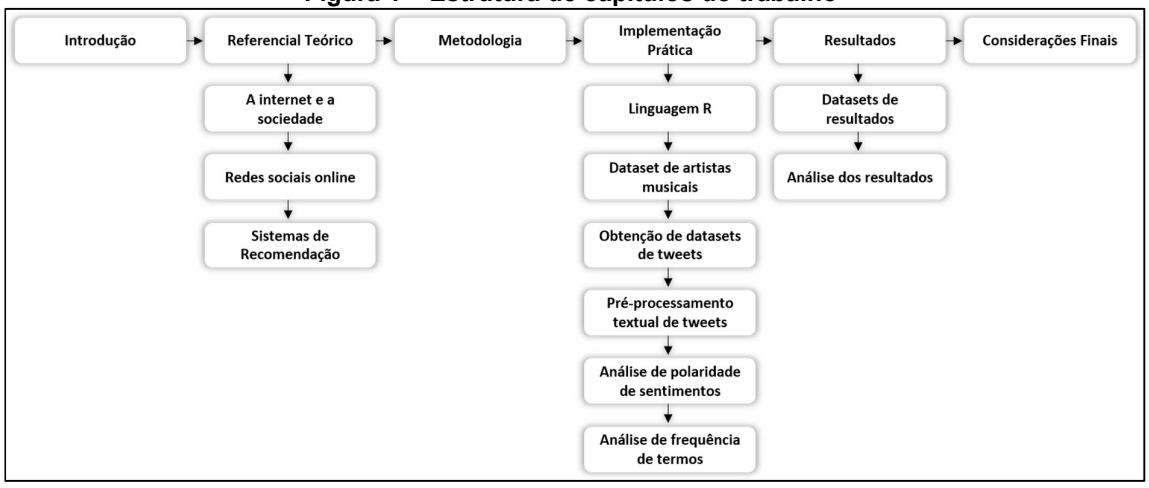
b) Realizar análise de *benchmarking* e analisar trabalhos relacionados, prospectando projetos equivalentes ou similares;

c) Desenvolver um algoritmo que possibilite a coleta, análise e geração de resultados da análise de maneira automatizada;

d) Realizar uma análise estatística simples dos dados contidos no *dataset* gerado.

Com relação aos objetivos específicos, optou-se por uma alteração na maneira de validar o modelo em comparação ao anteprojeto e versão inicial deste trabalho (TC I), em virtude de limitações técnicas a serem detalhadas no capítulo 4. Desta maneira, ao invés da aplicação de questionários a usuários finais (inicialmente planejado), optou-se pela construção do *dataset* como demonstração de viabilidade deste modelo.

Figura 1 – Estrutura de capítulos do trabalho



Fonte: O autor

A fim de facilitar o entendimento no que se refere ao encadeamento dos conteúdos apresentados neste trabalho, a Figura 1 apresenta a estrutura de capítulos deste trabalho.

2 REFERENCIAL TEÓRICO

Neste capítulo, pretende-se apresentar o embasamento científico do presente trabalho, discorrendo desde a definição de Internet e seu impacto na sociedade contemporânea à explanação dos conceitos de – e/ou relacionados a – sistemas de recomendação, redes sociais *online* e meios de distribuição e consumo de música digital.

2.1 A INTERNET E A SOCIEDADE CONTEMPORÂNEA

Para Castells (2003, p. 8), a Internet constitui o primeiro meio de comunicação na sociedade humana a possibilitar a comunicação de “muitos com muitos” em escala global, acarretando em mudanças radicais nas relações sociais.

A Internet oferece praticidade na realização de tarefas de diversos tipos por desobrigar a necessidade de deslocamento e/ou por tornar mais simples o acesso à informação. São exemplos desta comodidade a compra de produtos, a realização de transações bancárias e até mesmo o empreendimento do trabalho, uma vez que o trabalho remoto (ou *home office*) vem se tornando uma realidade cada vez mais presente ao longo dos últimos anos em empresas de todos os portes (CASTELLS, 2003).

A redução significativa no tempo investido para realização destas atividades é um dos fatores que leva as pessoas a ficarem cada vez mais conectadas e, por conseguinte, mais presentes em interações virtuais (GRAEML et al., 2004).

Tais vantagens contribuíram para o crescimento substancial no número de usuários ativos na Internet, evidenciado quando observados dados estatísticos como os da pesquisa do IBGE (2016), que estimou que, em 2014, mais da metade dos domicílios particulares permanentes no Brasil passaram a ter acesso à Internet, o equivalente a 36,8 milhões, representando 54,9% do total do país.

A facilidade de se obter acesso à Internet, por sua vez, criou espaço para o surgimento de novas mídias, diferentes das tradicionais como jornal, rádio e televisão. Saad (2003) aponta alguns exemplos de meios de comunicação que surgiram na *Web* e tornaram o compartilhamento de conhecimento mais democrático, como as páginas

de notícias, os *blogs*[2] e as *wikis*[3]. O surgimento da *Web* 2.0, a partir do início dos anos 2000, reforçou ainda mais esse processo de descentralização.

> A *Web* 2.0 potencializa a ação do usuário na rede por meio da oferta, quase sempre gratuita, de ferramentas de que permitem a expressão e o compartilhamento com outros usuários de opiniões, criações, desejos, reclamações, enfim, qualquer forma de comunicação interpessoal (SAAD, 2003, p. 149).

Para Saad (2003), o surgimento da *Web* 2.0 fomentou a criação de páginas mais interativas, que incentivam a participação dos usuários. Esse tipo de página possibilitou aos usuários interagirem com os veículos de comunicação em vez de apenas receber informações passivamente. O aumento de interatividade faz com que usuários da Internet tenham a oportunidade de atuar como criadores de conteúdo e não apenas como consumidores. Este advento tecnológico trata-se, para Bertaglia (2017), de uma significativa "mudança de paradigma na interação entre usuário e Internet".

Neste sentido, após o cenário explicitado até aqui, é importante destacar que nesta sociedade permeada e influenciada pela rede mundial de computadores e demais tecnologias em constante desenvolvimento, as formas de manter, cultivar ou estabelecer novas relações pessoais, assim como o avanço das mesmas, também sofrem transformações significativas (SAAD, 2003). No tocante a este ponto, um fenômeno que tem contribuído de forma relevante para as mudanças no relacionamento humano são as redes sociais *online* (SAAD, 2003).

Bertaglia (2017, p. 10) acrescenta que a "explosão de conteúdo gerado por usuários" que se deu em decorrência do surgimento da *Web* 2.0, somado à proliferação do acesso à Internet, já transpassou suas mídias de origem, especialmente as redes sociais *online*, e atualmente atinge também os meios tradicionais de comunicação. "Cada vez mais jornais e programas de televisão abrem espaço para que os consumidores também sejam produtores de conteúdo. Essa interação [...] pode ser considerada como pós-*Web* 2.0" (BERTAGLIA, 2017, p. 10). O

---

[2] Contração do termo *Weblog*. Consiste em um conjunto de páginas eletrônicas, estruturadas em postagens (que podem ser de textos, imagens, áudio, *links*, vídeos, etc.) organizadas cronologicamente, e que comumente permitem aos usuários/leitores escreverem comentários (AMARAL; RECUERO; MONTARDO, 2009).

[3] Plataformas web colaborativas usadas para compartilhamento de conteúdo e que possibilitam a edição coletiva de documentos, usualmente sem necessidade de identificação (registro) do usuário (VALENTE; MATTAR, 2007).

autor aponta este fato como uma justificativa para o aumento de pesquisas científicas envolvendo o conceito de *conteúdo gerado por usuário*, que se refere ao conteúdo criado pelo público comum, em vez profissionais pagos, e distribuído (principalmente) pela Internet (DAUGHERTY; EASTIN; BRIGHT, 2008).

Dentre os avanços e ferramentas emergidas na Internet ao longo das últimas duas décadas, podem-se destacar as redes sociais *online*, também conhecidas como redes sociais virtuais ou sites de redes sociais, como um dos principais agentes transformadores das relações sociais (COSTA et al., 2016). O autor observa que as redes sociais *online* têm estado no centro de diversas pesquisas devido à sua importância no que se refere, por exemplo, à difusão da informação e possibilidade de comunicação

2.2 REDES SOCIAIS *ONLINE*

Tomaél e Marteleto (2006, p. 75) definem rede social, de uma perspectiva sociológica, como uma congregação de atores e suas conexões ou relacionamentos. Para os autores, as

> [...] redes sociais referem-se a um conjunto de pessoas (ou organizações ou outras entidades sociais) conectadas por relacionamentos sociais, motivados pela amizade e por relações de trabalho ou compartilhamento de informações e, por meio dessas ligações, vão construindo e reconstruindo a estrutura social.

Ainda sob o espectro social, Recuero (2009, p. 24, grifo do autor apud WASSERMAN, 1994; DEGENNE, 1994) elucida: “uma rede social é definida como um conjunto de dois elementos: *atores* (pessoas, instituições ou grupos; os nós da rede) e suas *conexões* (interações ou laços sociais).”.

Já sob o espectro digital, ou na perspectiva de estudo de sistemas de informação, Recuero (2009) esclarece que as *redes sociais online* (ou sites de redes sociais) são, por sua vez, os espaços utilizados na Internet para a expressão das *redes sociais*, reforçando aqui a diferença entre os dois conceitos.

Sobre os elementos que formam a rede social (atores e conexões, conforme citado), a autora também esclarece como se diferencia a compreensão e estudo destes no mundo *offline* e no mundo *online*, respectivamente, a começar pelos atores:

> Os atores são o primeiro elemento da rede social, representados pelos nós (ou nodos). Trata-se das pessoas envolvidas na rede que se analisa. Como partes do sistema, os atores atuam de forma a moldar as estruturas sociais, através da interação e da constituição de laços sociais. Quando se trabalha com redes sociais na Internet, no entanto, os atores são constituídos de maneira um pouco diferenciada. Por causa do distanciamento entre os envolvidos na interação social, principal característica da comunicação mediada por computador, os atores não são imediatamente discerníveis. Assim, neste caso, trabalha-se com representações dos atores sociais, ou com construções identitárias do ciberespaço (RECUERO, 2009, p. 25).

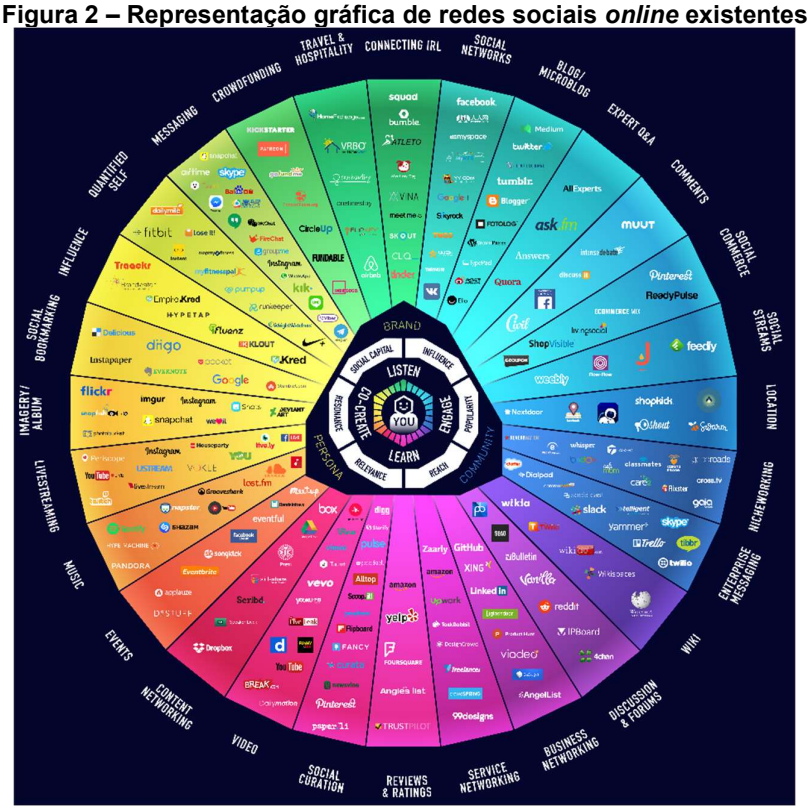
Ao usar a expressão “representações de atores sociais”, Recuero (2009) se refere ao fato de que, no âmbito de sites de redes sociais *online*, um ator social pode ser representado, por exemplo, por uma conta no *Twitter*, um perfil no *Facebook*, ou até mesmo por um *blog*. Cabe ressaltar o destaque da autora para o fato de que, no exemplo do *blog*, apesar deste representar um único nó na rede enquanto ferramenta, o mesmo pode ser mantido por atores múltiplos (no caso de um *blog* coletivo com diferentes autores), logo, consiste em uma representação de um ator social e não pode ser considerado o ator social em si.

Constante às conexões em redes sociais, Recuero (2009) as apresenta como resultado dos laços sociais formados através da interação entre atores. Estas, por sua vez, também possuem diversas particularidades e fatores diferenciais quando estudadas no âmbito de redes sociais *online*.

> O primeiro deles é que os atores não se dão imediatamente a conhecer. Não há pistas da linguagem não verbal e da interpretação do contexto da interação. É tudo construído pela mediação do computador. O segundo fator relevante é a influência das possibilidades de comunicação das ferramentas utilizadas pelos atores. Há multiplicidade de ferramentas que suportam essa interação e o fato de permitirem que a interação permaneça mesmo depois do ator estar desconectado do ciberespaço (RECUERO, 2009, p. 31).

Quanto a diversidades de sites de redes sociais existentes e suas diferentes naturezas, no que se refere à maneira como as pessoas utilizam cada rede, a Figura 2 ilustra o *Conversation Prism 5.0* (em tradução livre, “Prisma de Conversação”), uma representação gráfica do universo das redes sociais *online*, mapeando e categorizando diversas plataformas atualmente existentes na *Web* (SOLIS, 2018).

Figura 2 – Representação gráfica de redes sociais *online* existentes



Fonte: Adaptado de Solis (2018)

Conforme será abordado nos capítulos a seguir, a rede social *online Twitter* foi utilizada como base da implementação prática realizados neste trabalho. Assim, para garantir melhor compreensão dos capítulos posteriores, abaixo será introduzida uma visão geral desta plataforma.

2.2.1 *Twitter*

O *Twitter*, fundado em 2006, é uma das redes sociais *online* mais populares, atualmente com 126 milhões de usuários ativos diariamente (SHABAN, 2019).

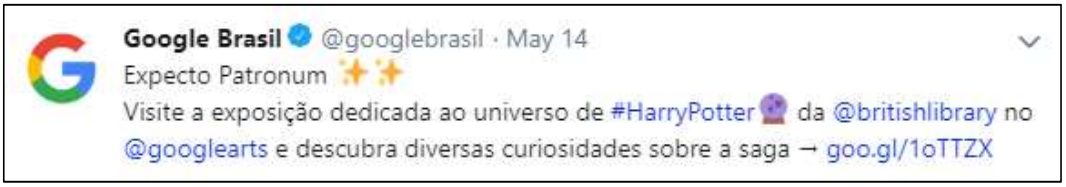
Esta rede social *online* disponibiliza um ambiente onde os usuários podem compartilhar informações ou gerar conteúdo de maneira instantânea. O compartilhamento destes conteúdos se dá através de pequenas mensagens denominadas *tweets*. Um *tweet* pode conter até 280 caracteres, principal característica desta plataforma, que, em virtude da limitação de caracteres, é comumente classificada como um serviço de *microblogging* (PINTO, 2018).

Conforme explana Pinto (2018), existem símbolos e padrões frequentemente utilizados por usuários do *Twitter*. Os principais, a serem abordados posteriormente neste trabalho, são:

a) ***Hashtag* (#):** qualquer palavra, nome ou frase precedida por um símbolo de sustenido (#) em um *tweet* são interpretadas como *hashtags*. Estas são utilizadas como *tags* de conteúdo, propriamente ditas, e utilizadas para facilitar a filtragem e identificações destes *tweets*;
b) **Menção (@):** para referenciar um outro usuário no *Twitter*, adiciona-se ao *tweet* o nome do usuário (*username*) precedido pelo símbolo "@". Neste caso, o usuário referenciado recebe uma notificação exibindo o *tweet* em que foi mencionado – é importante ressaltar que também é permitido aos usuários responderem outros *tweets*. Nestes *tweets* de resposta, o *Twitter* automaticamente já adiciona uma menção ao usuário autor do *tweet* original; e
c) **Retweet (RT):** quando a mensagem de um *tweet* é precedida pelas letras RT, significa que esta mensagem foi escrita por um usuário do *Twitter* e está sendo replicada por outro usuário, na forma de um *retweet*.

Os *tweets* podem, ainda, conter *emojis*[4], fotos, vídeos, geolocalização e outros elementos. Pinto (2018) ressalta, ainda, que devido à limitação de caracteres por *tweet*, é comum a abreviação de palavras, a utilização de marcadores e de URLs encurtadas.

Figura 3 – Exemplo de um *tweet*



Fonte: Twitter (2019)

[4] *Emojis* são "caracteres de imagem", ou pictogramas, populares na comunicação baseada em texto. Estes são popularmente utilizados em mensagens enviadas através de smartphones e no compartilhamento de conteúdo em redes sociais (MILLER et al., 2016).

Na Figura 3, é possível observar alguns dos códigos e padrões explanados acima. A figura mostra um *tweet* compartilhado pela empresa Google Brasil (*username* @googlebrasil), contendo uma menção aos usernames @britishlibrary e @googlearts, a *hashtag* #HarryPotter, e uma URL encurtada.

2.3 SISTEMAS DE RECOMENDAÇÃO

O aumento exponencial na quantidade de informações disponíveis na *Web* e a crescente emergência de novos modelos de negócios digitais expandiu a oferta de opções acessíveis na rede. Esse fenômeno, por sua vez, passou também a dificultar os processos de escolha dos usuários, problema que, conforme citado anteriormente, remete ao fenômeno conhecido como “Paradoxo da Escolha” (RICCI; ROKACH; SHAPIRA, 2010; SCHWARTZ, 2009).

No que tange ao processo de escolhas no mundo real, conforme Shardanand e Maes (1995), as pessoas frequentemente baseiam-se em recomendações de terceiros (“*word of mouth*”) para tomarem decisões referentes a tarefas cotidianas, como indicações de filmes, músicas, conteúdos e produtos. Para Ricci, Rokach e Shapira (2010), é através deste prisma, e como uma alternativa para minimizar a sobrecarga de informações potencialmente recebidas por um usuário, que surge a demanda por sistemas de recomendação.

O conceito de sistema de recomendação contempla técnicas e aplicações que objetivam prover sugestões de itens a usuários, sendo “item”, a denominação genérica atribuída ao elemento recomendado ao usuário (RESNICK; VARIAN, 1997). No âmbito da Tecnologia da Informação, conforme explana De Sá (2009), tratam-se de aplicações de software que buscam antecipar os interesses dos usuários na rede a fim de recomendar novos produtos. Para a autora, quando “incluídos na categoria de filtros de informação, eles complementam a atuação de buscadores, como o popular *Google*, localizando a informação que você *não* sabe que procura”.

Em termos práticos, os sistemas de recomendação podem gerar recomendações baseando-se em diversos tipos de *inputs*, ou entradas. A entrada mais conveniente é a de *feedbacks explícitos*, isto é, quando os usuários expressam voluntariamente seu nível de interesse em determinado item (por exemplo, dar nota a um filme, classificar uma música como “Gostei” ou “Não Gostei, etc.). No entanto, por estarem comumente disponíveis em escalas maiores que os feedbacks explícitos,

muitos sistemas inferem as preferências do usuário através dos *feedbacks implícitos*, que incluem observações de comportamento do usuário como histórico de consumo, histórico de navegação e padrões de busca (HU; KOREN; VOLINSKY, 2008).

Adomavicius e Tuzhilin (2005) indicam que os sistemas de recomendação podem ser definidos em 3 (três) categorias de acordo com o método utilizado para realizar as recomendações, denominados métodos de filtragem:

a) Sistemas de filtragem colaborativa;
b) Sistemas de filtragem baseada em conteúdo; e
c) Sistemas de filtragem híbrida.

O método de ***filtragem baseada em conteúdo***, segundo Pasquale, Gemmis e Semeraro (2010), parte do princípio de que os usuários tendem a interessar-se por itens similares aos que demonstraram interesse no passado. Sistemas que implementam este modelo de filtragem, indicam os autores, analisam um conjunto de itens ou descrições de itens previamente avaliados pelo usuário, e constroem um modelo ou perfil de interesses do usuário que se baseia nas características dos itens avaliados por aquele usuário, gerando "uma representação estruturada dos interesses do usuário" (PASQUALE; GEMMIS; SEMERARO, 2010, p. 75). Ainda na perspectiva dos autores, o processo de recomendação neste modelo consiste em confrontar os atributos do perfil do usuário com os atributos do item a ser recomendado. O resultado deste processo representa o nível de interesse do usuário no item analisado.

No âmbito específico de recomendação musical, De Sá (2009, p. 7) explana que:

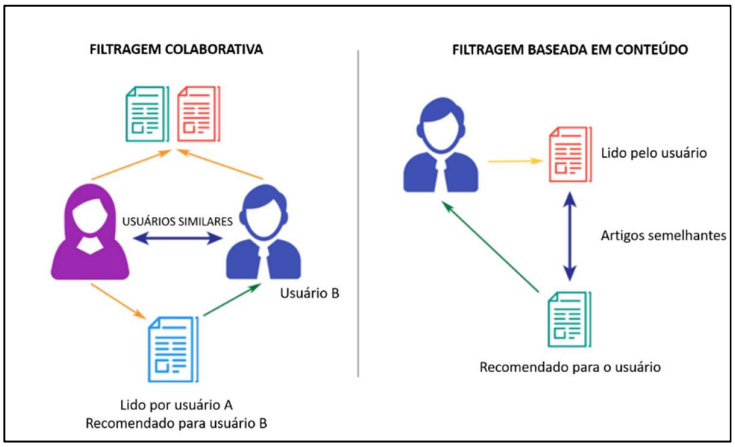
> [...] a ideia é a de que algoritmos podem ser desenvolvidos com base no registro da reação de consumidores, [...] buscando construir padrões de gosto musical. Padrões que podem acompanhar as categorias tradicionais – tais como gênero musical –, mas também podem ir além. Assim, seja através de recursos que solicitam ao usuário classificar se gosta ou não da música que está ouvindo, seja através do rastreamento do perfil do usuário a partir de comportamentos prévios ou da análise das *tags* utilizadas, o sistema coleta, armazena e cruza informações que serão usadas para futuras indicações. Estas, por sua vez, tornam-se cada vez mais precisas e confiáveis, uma vez que se baseiam nos próprios gostos do consumidor.

Já o método de ***filtragem colaborativa***, ao contrário do método baseado em conteúdo, tenta mensurar o nível do interesse do usuário no item sendo analisado com base na avaliação fornecida por outros usuários, usuários estes que tenham um

perfil de interesses similar ao do usuário em questão (ADOMAVICIUS; TUZHILIN, 2005).

A Figura 4 representa graficamente o funcionamento das filtragens colaborativa e baseada em conteúdo, respectivamente.

Figura 4 – Representação dos métodos de filtragem



Fonte: Adaptado de Calderon (2018)

Por sua vez, o método de ***filtragem híbrida*** busca contornar determinadas limitações dos métodos de filtragem colaborativa e baseada em conteúdo ao combinar características de ambos. Sistemas de recomendação podem se valer da abordagem híbrida ao implementar ambos os métodos separadamente e combinar seus resultados (“predições”), ou implementar apenas um dos métodos, porém combinando determinadas características do outro (ADOMAVICIUS; TUZHILIN, 2005).

A utilização de sistemas de recomendação tem sido cada vez mais necessária para garantir melhor aproveitamento na realização de diversas atividades na *Web*, incluindo o consumo de música, como explica De Sá (2009). Sob esta perspectiva, Pasick (2015) elucida que, dado o volume massivo de faixas musicais disponíveis na *Web*, torna-se cada vez mais complexo para os consumidores escolherem ou encontrarem algo já conhecido para ouvir ou, até mesmo, descobrir novas faixas ou artistas musicais.

2.3.1 Sistemas de recomendação musical

Conforme aponta Schedl (2018), o campo de pesquisa de métodos de recomendação musical enfrenta desafios específicos, justificados por algumas particularidades de faixas musicais enquanto itens de recomendação.

Para Schedl (2018), a fim de satisfazer as especificidades e necessidades de entretenimento musical de cada usuário, pesquisadores e designers de sistemas de recomendação musical devem considerar seus usuários de maneira “holística”, contemplando aspectos intrínsecos, extrínsecos e contextuais dos ouvintes. Por exemplo, a personalidade e o estado emocional dos ouvintes (aspectos intrínsecos), bem como sua atividade (extrínseco), são conhecidos por influenciar gostos e necessidades musicais. O mesmo ocorre com os fatores contextuais dos usuários, como condições climáticas, grupo social ou lugares de interesse.

A seguir, estão listados os principais aspectos que, segundo o autor, diferenciam os sistemas de recomendação musical de demais sistemas de recomendação, realizando uma comparação entre a recomendação de faixas musicais com a de livros e/ou filmes.

a) **Duração dos itens:** na recomendação tradicional de filmes, por exemplo, os itens de interesse têm uma duração típica de 90 minutos ou mais. Na recomendação de livros, o tempo de consumo é geralmente ainda mais longo. Em contraste, a duração de faixas de música popular geralmente varia entre 3 e 5 minutos. Por este motivo, os itens musicais podem ser considerados mais “descartáveis”;
b) **Extensão do catálogo:** o tamanho dos catálogos de faixas musicais está na faixa de dezenas de milhões de músicas, conforme mencionado anteriormente. Os serviços de distribuição de filmes trabalham com catálogo muito menores, normalmente com milhares ou dezenas de milhares de filmes e séries. A escalabilidade é, portanto, uma questão muito mais importante na recomendação musical do que na recomendação do filme;
c) **Sequência de consumo**: ao contrário dos filmes e livros, as músicas são mais frequentemente consumidas de maneira sequencial, ou seja, em uma lista de reprodução ou *playlist*. Isso gera desafios adicionais para identificação da sequência mais apropriada de itens em uma lista de recomendações;

d) **Repetição de itens já recomendados:** recomendar o mesmo item mais de uma vez não é tipicamente bem recebido por usuários de sistemas de recomendação de filmes ou de produtos, por exemplo. No âmbito musical, no entanto, a repetição de uma mesma música em um momento posterior pode ser apreciada pelo usuário e gerar uma experiência ainda mais positiva;

e) **Comportamento de consumo**: a música é muito frequentemente consumida passivamente, em segundo plano, o que influencia a maneira como o sistema deve identificar as preferências de um usuário. Ao basear-se no feedback implícito para inferir as preferências do ouvinte, o fato de um ouvinte não pular uma faixa pode ser interpretado erroneamente como um sinal positivo (falso-positivo), quando, na verdade, este não estava sequer ouvindo;

f) **Emoções:** é sabido que a música evoca emoções nas pessoas – entretanto, essa é uma relação mútua, uma vez que as emoções dos usuários também afetam suas preferências musicais. A forte relação entre música e emoções demanda uma área de pesquisa dedicada à identificação de sentimentos em música, comumente referida como *music emotion recognition* (em português, reconhecimento de emoção em música). O autor complementa que a regulação das emoções foi identificada como uma das principais razões pelas quais as pessoas ouvem música. Como exemplo, as pessoas podem ouvir gêneros musicais completamente diferentes quando estão tristes em comparação com quando estão alegre.

g) **Contexto:** contextos exercem forte influência na preferência musical momentânea de um usuário, na maneira como este consume e interage com a música. Por exemplo, ouvintes frequentemente criam *playlists* específicas para determinadas atividades ou ambientes (jantar romântico, caminhada, festa com os amigos, etc). Os tipos de contexto considerados com maior frequência incluem o local onde o usuário está no momento (no local de trabalho, em casa, na academia, etc) e horário (normalmente classificado por turno). O contexto pode, além disso, também se relacionar com a atividade do ouvinte, clima, ou o uso de diferentes dispositivos, como fones de ouvido. Como ouvir música também é uma atividade muito ligada à interação social, investigar o contexto social dos ouvintes é crucial para identificar suas preferências e comportamentos de consumo. A importância de considerar tais fatores contextuais para recomendações musicais relevantes é uma tendência de

pesquisa conhecida como *situation-aware music recommendation* (em tradução livre, recomendação musical baseada em situações).

Schedl (2018) defende que, de maneira geral, a preferência musical de um usuário depende da situação em que se encontra no momento da recomendação. A localização é um exemplo de situação que influencia fortemente uma recomendação musical e que, quando considerada, pode otimizar substancialmente a precisão do sistema – por exemplo, a preferência musical de um usuário tende a ser diferente em uma biblioteca e em uma academia.

A hora do dia, ou turno, é outro fator que pode ser levado em conta, uma vez que, frequentemente, a música que um usuário gostaria de ouvir pela manhã pode ser diferente do gosto musical noturno.

Outro aspecto situacional de substancial importância para recomendações musicais é o contexto social, uma vez que os gostos musicais e os comportamentos de consumo estão diretamente ligados às interações sociais dos ouvintes (por exemplo, é provável que um usuário prefira músicas diferentes quando está sozinho do que quando está em conjunto com amigos).

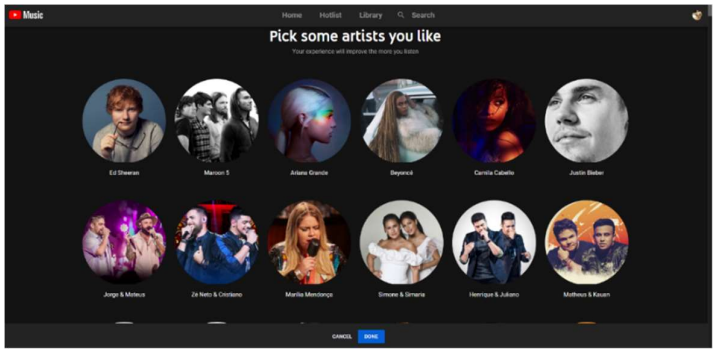
2.3.2 O problema de *cold-start* em sistemas de recomendação

> Um pré-requisito para qualquer sistema de recomendação é a disponibilidade dos dados que possam indicar as necessidades e preferências dos usuários. Embora o desempenho do algoritmo [de recomendação] tenha um papel importante [no processo de recomendação], a qualidade das recomendações com base em qualquer classe de sistemas de recomendação pode ser ruim se não houverem dados de qualidade suficientes fornecidos pelos usuários. Essa é uma situação conhecida como *cold-start*, que normalmente ocorre quando um novo usuário se registra no sistema e nenhum dado de preferência está disponível para esse usuário. Este é um grande problema em sistemas de recomendação, especialmente os com grande número de usuários (MOGHADDAM, 2018, p. 2, tradução nossa, grifo nosso).

Sobre o problema de *cold-start* no âmbito de sistemas de recomendação musical, Schedl (2018) complementa que este é um dos principais obstáculos, especialmente quando um novo usuário é registrado ao sistema e o sistema não tem dados suficientes associados a esse usuário. Neste caso, o sistema não consegue recomendar itens existentes apropriadamente por desconhecer o histórico e os gostos musicais do usuário ouvinte.

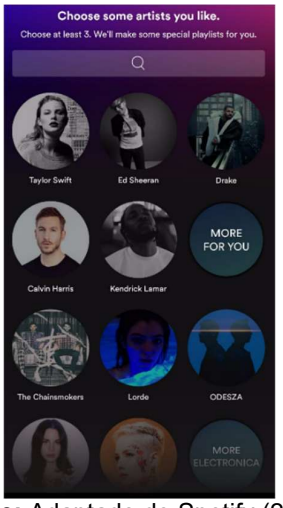
Uma das soluções mais comuns aplicadas às plataformas de streaming de música, para resolução do problema de *cold-start*, é solicitar ao novo usuário que ativamente informe ao sistema gêneros ou artistas musicais de sua preferência, conforme demonstrado nas Figuras 5, 6 e 7, que se referem a capturas de tela de 3 plataformas de streaming distintas que aplicam a mesma solução aqui mencionada.

Figura 5 – Tela de inicialização de novos usuários da plataforma *YouTube Music*



Fonte: Adaptado de YouTube Music (2019)

Figura 6 – Tela de inicialização de novos usuários da plataforma *Spotify*



Fonte: Adaptado de Spotify (2019)

Figura 7 – Tela de inicialização de novos usuários da plataforma *Deezer*

Fonte: Adaptado de Deezer (2019)

Para melhor compreensão dos conceitos de cunho musical abordados neste capítulo de embasamento teórico, o próximo subcapítulo apresenta uma revisão do atual cenário de distribuição e consumo de música digital.

2.4 DISTRIBUIÇÃO E CONSUMO DE MÚSICA DIGITAL

As práticas de produção e consumo de música vem se beneficiando de avanços tecnológicos desde o final século XIX, quando, com surgimento dos primeiros fonógrafos, finalmente deixava de ser necessária a presença física de um artista para a execução de obras musicais (RANDLE, 2001). O mesmo autor salienta que nas revoluções tecnológicas subsequentes, observou-se a invenção e consolidação do rádio e da televisão, por exemplo, que reconfiguraram os conceitos de comunicação de massa da época, tornando-se os principais meios de divulgação de produtos culturais de massa.

Durante muito tempo, e até poucos anos atrás, para ouvir uma faixa de música específica, por exemplo, era necessário ter acesso a ou possuir um disco de vinil, uma fita cassete, ou, mais recentemente, um *compact disc* (CD), outra invenção revolucionária para a indústria fonográfica (RANDLE, 2001).

Em tempo, de acordo com De Marchi (2011, p. 24),

> [...] a indústria fonográfica define-se como a produção de tecnologias de reprodução sonora, cujos conteúdos são constituídos predominantemente por repertório musical, sujeitos à regulação legal da propriedade através de uma variedade de direitos de propriedade intelectual.

A revolução das tecnologias digitais alterou permanentemente a indústria fonográfica. A ruptura neste mercado deu-se principalmente após a digitalização de gravações sonoras e a possibilidade de transformação em arquivos de áudio no sistema *Motion Picture Experts Group-1, Layer-3* (MP3), popularizado em meados dos anos 2000 (SÁ, 2006; PIRES; 2018).

Para De Sá (2006, p. 15, grifo nosso), a música sofreu, então, um processo de "desmaterialização [...] a partir de sua digitalização, transformando-se em bits que podem ser acessados, lidos e traduzidos em suportes variáveis", além da possibilidade de "ser reprocessada, *sampleada*[5] e reconectada com outros sons através de softwares específicos, num processo aberto e potencialmente infindável".

Após a ruptura que o surgimento do formato de música digital representou para o mercado fonográfico, diferentes modelos de negócio surgiram para atender à nova realidade. O modelo que vem conquistado preeminência nos anos 2010 é o de música sob demanda, disponível através das plataformas de *streaming*. De Marchi (2011) as conceitua como plataformas de comércio de fonogramas digitais, onde usualmente os assinantes pagam uma mensalidade para terem acesso ao catálogo completo de gravações. Atualmente, existem diversos serviços populares que atuam no modelo de *streaming*, como *Spotify*, *Deezer* e *YouTube Music* (SPOTIFY; DEEZER; YOUTUBE MUSIC, 2019).

Com grande aceitação também no Brasil, o jornal *O Globo* noticiou a popularização das plataformas de *streaming*, em 2016:

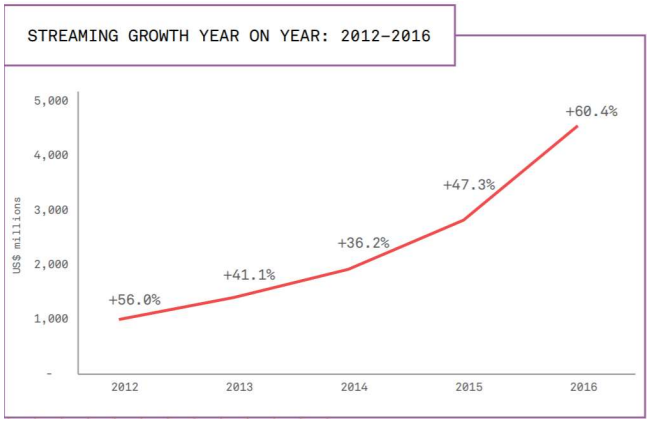
> [...] a comodidade de ouvir as músicas e os álbuns que se quer, na hora desejada, sem ter que ocupar estantes da casa ou megabytes de memória do computador era um futuro pelo qual muitos ansiavam — e que em 2016 virou, enfim, a mais corriqueira das realidades. Os sistemas de *streaming* [...] venceram de vez as resistências da indústria fonográfica e dos ouvintes mais tradicionais. E se tornaram o novo e desejável *mainstream* da música [...] (ESSINGER, 2016).

A modalidade de *streaming* vem expandindo de maneira significativa ao longo dos últimos anos. O *Spotify*, a plataforma de serviço de *streaming* que atualmente detém a liderança do mercado, possui 170 milhões de usuários ativos e mais de 35 milhões de músicas em seu catálogo (SPOTIFY, 2018).

[5] Anglicismo derivado de *sampling*. Em música, *sampling* se refere ao ato de usar amostras sonoras (gravações ou fragmentos) para criar novas sequências musicais (SILVEIRA, 2012).

A Figura 8, extraída do *Global Digital Report* de 2016 da Federação Internacional da Indústria Fonográfica, apresenta a comparação de receita obtida através destes serviços entre 2012 e 2016.

Figura 8 – Comparação de receita relativa a serviços de *streaming*



Fonte: Federação Internacional da Indústria Fonográfica (2017)

Este processo de digitalização da música alterou radicalmente as práticas relativas ao mercado fonográfico. Possibilitou, por exemplo, a bandas ou artistas divulgarem seu trabalho musical sem a intermediação de gravadoras, além de suprimir a preeminência da mídia física e, por conseguinte, das grandes gravadoras. Como resultado, a digitalização oportunizou, por fim, que tanto a produção quanto o consumo de música fossem democratizados, obrigando a reconfiguração do mercado no que tange o seu "circuito de produção-circulação-consumo" (SÁ, 2006).

Neste contexto, cabe salientar que os sistemas de recomendação musical se tornam artefatos quase indispensáveis na experiência do consumo de música, inclusive estando muitas vezes presentes e embutidos nos próprios *players* e serviços de *streaming* de música, a exemplo do *Spotify*, que disponibiliza semanalmente listas personalizadas de músicas para cada um de seus usuários com base em seu histórico de consumo nas semanas anteriores (PASICK, 2015).

2.5 TRABALHOS RELACIONADOS

Os subcapítulos anteriores desta seção de referencial teórico apresentaram os conceitos científicos necessários para embasamento e compreensão do modelo de

sistema que esta pesquisa objetiva propor, conforme apontam os objetivos específicos apresentados no capítulo de introdução.

Dando seguimento ao objetivo de delineamento teórico da pesquisa, e a fim de identificar contribuições científicas similares já comprovadas, foram analisadas algumas pesquisas acadêmicas relacionadas à proposta do presente trabalho, isto é, envolvendo sistemas de recomendação musical e análise textual ou de sentimentos utilizando informações provenientes de mídias sociais.

Não foram encontradas pesquisas que utilizem *posts* de redes sociais como insumo para inferir o gosto musical de usuários e, assim, gerar recomendações musicais, o que justifica o ineditismo desta proposta, especialmente em língua portuguesa.

Desta maneira, abaixo são apresentadas breves descrições de projetos acadêmicos de natureza equivalente ou similar à estudada nesta pesquisa.

Especificamente sobre a utilização de análise textual como fonte de recomendação musical, Ziwon, Lee e Lee (2013) analisaram uma base em larga escala de documentos enviados a uma estação de rádio. Cada documento analisado se refere a uma solicitação de música enviada por algum ouvinte, contendo a música solicitada e uma breve história pessoal. A pesquisa assume que tais estórias possam estar relacionadas ao contexto e sentimento das músicas sugeridas e apresenta, então, um sistema que realiza análise textual para recomendar músicas com base em histórias pessoais semelhantes. Os resultados das avaliações realizadas por usuários indicaram haver forte correlação entre a semelhança das músicas e dos documentos analisados.

Schedl e Schnitzer (2014), por sua vez, ressaltaram a importância de abordagens centradas no usuário, e que combinam diferentes métodos, para evolução dos algoritmos de recomendação musical. Nesta pesquisa, são propostos diversos algoritmos de recomendação musical que combinam informações de conteúdo e contexto das músicas, bem como contexto do usuário, especialmente informações de geolocalização. Um dos principais resultados da pesquisa é demonstrar que o componente “contexto do usuário” (como a geolocalização, neste caso), quando agregado a abordagens tradicionais (como o método de filtragem colaborativa) pode melhorar consideravelmente a qualidade das recomendações geradas.

Yamashita (2013) buscou melhorar a maneira como os usuários interagem com recomendações musicais através da utilização de *tags* explicativas, conhecidamente

uma forma de aumentar a credibilidade de sistemas de recomendação, garantindo maior transparência e, por conseguinte, satisfação ao usuário final. Através do desenvolvimento de um processo de coleta e cálculo de relevância das sugestões musicais, bem como de uma interface gráfica para exibição destas juntamente às *tags* escolhidas, a autora observou aumento considerável nos índices de inspeção, eficiência, eficácia e satisfação dos usuários com relação às recomendações recebidas.

Skowron et al. (2016) defendem que a presença e atividades *online* dos usuários são distribuídas por diferentes plataformas, devido às diferentes características destas (interativas e orientadas a conteúdo, por exemplo). Por isso, o conteúdo gerado por usuários demonstra fornecer informações importantes sobre os interesses, preferências e sentimentos dos usuários em relação a vários tópicos. Como os usuários deixam muitos traços de suas atividades digitais através das redes sociais *online*, estudos anteriores, analisados por Skowron et al. (2016), demonstraram que os recursos visuais, linguísticos e meta-linguísticos extraídos do conteúdo *online* gerado pelos usuários podem ser instrumentais para inferir traços de personalidade.

Com base nisso, Skowron et al. (2016) propuseram um método integrado de análise de texto e imagem de usuários baseado nas plataformas de redes sociais *Twitter* e *Instagram*. Os resultados preliminares indicaram que a análise conjunta e simultânea das atividades dos usuários nestas duas redes sociais populares levam a uma diminuição consistente nos erros de predição de traços de personalidade.

Para validação da proposta, Skowron et al. (2016) recrutaram usuários populares de redes sociais (Instagram e *Twitter*) que fossem nativos de língua inglesa e localizados nos Estados Unidos. Um questionário de personalidade foi administrado com o consentimento destes usuários. As respostas agregadas foram usadas para inferir os cinco traços básicos de personalidade dos participantes (abertura a experiências, *conscientiousness*, extroversão, amabilidade e neuroticismo); cada característica recebia um valor de 1 a 5. Em seguida, foram extraídos dados das contas de redes sociais destes mesmos participantes, excluindo aqueles com menos de 30 imagens no Instagram ou menos de 30 *tweets*: o conjunto final (com quantidade suficiente de dados em ambas as plataformas) foi composto por 62 usuários. Os dados finais foram confrontados para identificar padrões de personalidade convergentes.

Por fim, Rosa, Rodriguez e Bressan (2015), apresentaram um sistema de recomendação fundamentado por uma métrica avançada de análise de intensidade de sentimentos, que consiste na associação de uma métrica de sentimentos baseada em análise léxica com um fator de correção (que, por sua vez, é calculado a partir do perfil do usuário nas redes sociais *online*). Os sentimentos dos usuários são extraídos de frases postadas nas redes sociais *online* e a recomendação musical é realizada através de uma aplicação de baixa complexidade (além de baixo consumo de rede, memória e energia elétrica) para dispositivos móveis, que sugere músicas convergentes à intensidade de sentimento do usuário no momento presente. A validação da pesquisa se deu pela avaliação de usuários remotos que julgaram seu nível de satisfação com as recomendações geradas pelo sistema, alcançando 91% de nível de satisfação, em detrimento de recomendações geradas aleatoriamente, com nível de satisfação reportado em 65% pela amostra de usuários analisada.

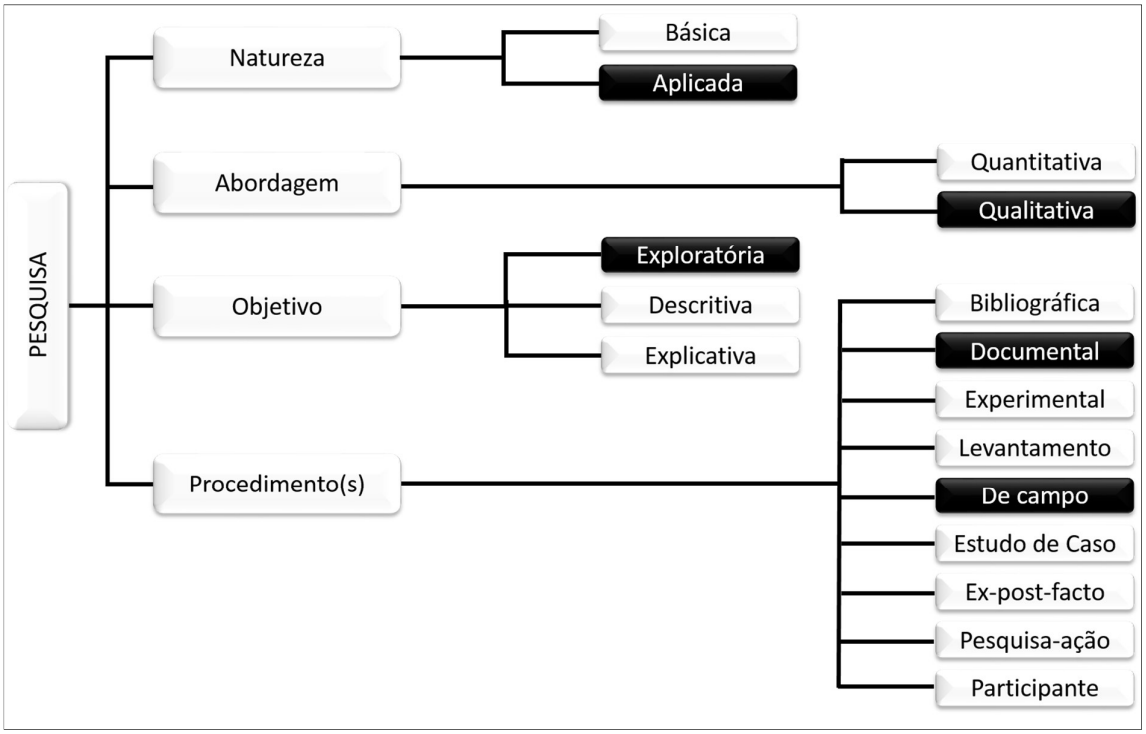
Esta breve revisão de trabalhos relacionados ao presente estudo encerra o embasamento teórico da pesquisa. O capítulo a seguir clarifica o enquadramento deste trabalho perante o método científico.

3 METODOLOGIA

Conforme Prodanov e Freitas (2013), uma pesquisa científica trata-se de um estudo planejado cuja finalidade é descobrir respostas para um problema o qual o repertório de conhecimento disponível não gera respostas adequadas, sendo realizado através da aplicação do método científico.

A Figura 9 apresenta de maneira resumida a qualificação metodológica desta pesquisa, tendo as caixas pretas representando as abordagens aqui utilizadas.

Figura 9 – Qualificação metodológica da pesquisa



Fonte: Adaptado de Bez (2011)

Este estudo tem sua natureza caracterizada como aplicada, que, definem Prodanov e Freitas (2013, p. 51), “objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos”.

Ainda de acordo com Prodanov e Freitas (2013, p. 43), a busca por conhecimento através de pesquisa científica é “a realização de um estudo planejado, sendo o método de abordagem do problema o que caracteriza o aspecto científico da

investigação". Uma vez que o resultado desejado para esta pesquisa não se dará unicamente por averiguação estatística, mas sim por uma análise da viabilidade do modelo proposto, a abordagem caracteriza-se como qualitativa.

Sob a ótica dos objetivos definidos, permite-se classificar a pesquisa como exploratória, que, segundo Prodanov e Freitas (2013), pretende proporcionar mais informações sobre o assunto estudado e possibilitar sua definição e seu delineamento.

Os procedimentos técnicos são caracterizados como de campo e documental. Prodanov e Freitas (2013) consideram que a pesquisa de campo tem por objetivo obter informações sobre uma hipótese que se deseja comprovar. Para os autores, a pesquisa documental, por sua vez, tem como base a utilização de materiais sem nenhum tratamento analítico prévio, ou que podem ser adaptados para utilização em conformidade com os objetivos da pesquisa. A utilização deste procedimento, para Prodanov e Freitas (2013, p. 56), "é destacada no momento em que podemos organizar informações que se encontram dispersas, conferindo-lhe uma nova importância como fonte de consulta".

Como é costumeiro em pesquisas de objetivo exploratório, o presente trabalho envolveu levantamento biográfico, em sua primeira fase de elaboração, buscando conhecimentos necessários para adquirir-se o estado-da-arte em conceitos envolvendo sistemas de recomendação, algoritmos de recomendação musical, análise textual e de sentimentos. A pesquisa se deu em artigos científicos, livros e especificações técnicas, como indicam Prodanov e Freitas (2013).

A fim de demonstrar a viabilidade de extrair informações de contexto musical contidas em *posts* de redes sociais, será detalhado, no capítulo 4, a construção de um algoritmo responsável pela coleta, análise e geração de resultados das análises realizadas em uma base de *tweets*. Como resultado, objetiva-se obter um *dataset* de padrões geralmente associados a artistas musicais, resultados estes que podem ser utilizados, em aplicações futuras, como insumo para geração de recomendações musicais.

4 IMPLEMENTAÇÃO PRÁTICA

A implementação prática proposta neste estudo consiste em analisar uma base de *posts* de contexto musical em redes sociais (que possuem menções a nomes de artistas musicais, como cantores e bandas), classificar o seu conteúdo textual em relação à sua polaridade de sentimento (positivo, negativo ou neutro), e realizar uma análise simples de frequência de termos, a fim de identificar quais palavras são mais frequentemente associadas a cada artista analisado.

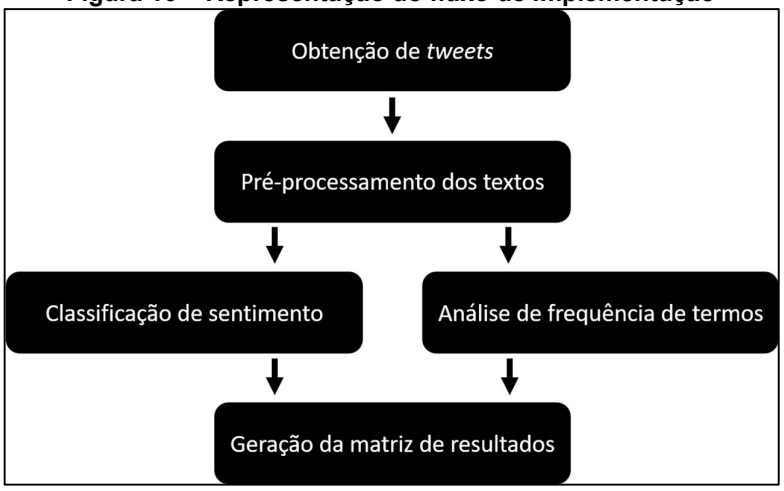
O resultado esperado é um algoritmo de coleta e análise de *tweets* e um *dataset* completo contendo os padrões comumente relacionados aos artistas analisados. Tais resultados podem servir de insumo para geração de recomendações musicais em aplicações futuras, conforme mencionado no capítulo anterior.

Para realizar esta implementação, o *Twitter* foi a rede social escolhida por oferecer uma *API*[6] que permite facilmente a extração de informações da plataforma.

Neste capítulo, é descrito o fluxo prático da implementação. O subcapítulo 4.1. introduz a linguagem R, na qual foi desenvolvido o algoritmo deste estudo. O subcapítulo 4.2. detalha o *dataset* de artistas musicais adotado como referência para a construção do *dataset*. Os subcapítulos seguintes apresentam o comportamento do algoritmo desenvolvido e aplicado neste estudo para execução das seguintes atividades: obtenção de *tweets* (subcapítulo 4.3), pré-processamento do conteúdo textual (subcapítulo 4.4), análise de sentimentos (subcapítulo 4.5) e análise de frequência de termos (subcapítulo 4.6).

[6] *Application Programming Interface* (em português, Interface para Programação de Aplicações) são conjuntos de padrões que permitem que programas de computador se comuniquem entre si e possam solicitar e receber informações. No caso do *Twitter*, a empresa disponibiliza acesso programático aos dados da plataforma através de uma *API* própria (*TWITTER*, 2019).

Figura 10 – Representação do fluxo de implementação



Fonte: O autor

O fluxo desta aplicação prática e do funcionamento do algoritmo utilizado neste estudo, detalhado nos subcapítulos 4.3 a 4.6, é também ilustrado na Figura 10.

4.1 LINGUAGEM R

R é uma linguagem e um ambiente de programação desenvolvidos para computação estatística. De acordo com o *website* oficial do projeto, consiste em um "conjunto integrado de recursos de software para cálculo, manipulação e visualização gráfica de dados" (R FOUNDATION, 2019, tradução nossa). Este ambiente oferece diversos métodos estatísticos, incluindo recursos para modelagem linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, *clustering*, dentre outros (R FOUNDATION, 2019).

Uma importante característica do R é a facilidade de expansão de funcionalidades através da instalação de *packages*, ou "pacotes", que se referem a funcionalidades adicionais desenvolvidas pela comunidade de desenvolvedores de R. Atualmente, existem mais de 14.000 pacotes disponibilizados na plataforma *Comprehensive R Archive Network*, popularmente denominada CRAN, o repositório oficial para pacotes de R (R FOUNDATION, 2019; COMPREHENSIVE R ARCHIVE NETWORK, 2019).

O ambiente R é, hoje, um dos mais popularmente utilizados para análises estatísticas de dados, e um dos recursos com maior projeção de crescimento, especialmente no âmbito acadêmico (ROBINSON, 2017).

De acordo com a introdução do presente capítulo, a linguagem e o ambiente R foram utilizados como base para o desenvolvimento do algoritmo implementado neste estudo. Para compreender sua aplicação neste estudoa, faz-se necessário, primeiramente, detalhar o *dataset* adotado como referência desta prática.

4.2 *DATASET* DE ARTISTAS MUSICAIS

Para o desenvolvimento desta prática, optou-se por utilizar um *dataset* de artistas que contivesse somente cantores e bandas brasileiros. Para isto, foi utilizado um *dataset* chamado “Os 500 brasileiros mais bombados hoje no Spotify”, que reúne os principais artistas brasileiros (de diferentes épocas e genêros musicais) classificados pela quantidade de ouvintes mensais registrados no Spotify (FELIX, 2017).

Conforme detalhado por Felix (2017), o *dataset* possui 550 registros e apresenta o nome do artista (cantor, dupla ou banda), o genêro musical e o nome da playlist do Spotify onde o artista é mais popular. A Figura 11 mostra o formato deste *dataset*.

Figura 11 – Formato do *dataset* de artistas musicais

	A	B	C
1	1Kilo	Rap	Top Brasil
2	A Banca 021	Rap	
3	A Banca 021	Rap	
4	A Banda Mais Bonita da Cidade	Pop	MPB
5	Adoniran Barbosa	Samba/Pagode	
6	Adriana Calcanhoto	MPB	Amor I Love You
7	Adriana Partimpim	Infantil	
8	Alceu Valença	MPB	Brasil Anos 80
9	Alcione	Samba/Pagode	
10	Alice Caymmi	Pop	Rock in Rio
11	Aline Barros	Gospel/Religioso	Sucessos Gospel
12	Aline Calixto	samba/Pagode	
13	Almir Guineto	Samba/Pagode	
14	Almir Sater	Sertanejo	
15	Alok	Eletrônico	Top Brasil
16	Ana Cañas	MPB	Rock in Rio
17	Ana Carolina	MPB	O Melhor da MPB
18	Ana Gabriela	Pop	Desplugado
19	Ana Muller	MPB	
20	Ana Nóbrega	Gospel/Religioso	

Fonte: O autor

Neste estudo, somente as primeiras duas colunas (constando o nome do artista, cantor, dupla ou banda, e o respectivo gênero musical) foram consideradas na análise, uma vez que o conteúdo da coluna referente à playlists do *Spotify* não possui relevância para o presente estudo.

O *dataset* completo é apresentado no Anexo A deste trabalho, no seguinte formato: Nome do artista (gênero musical). Para os artistas para os quais não há gênero musical registrado no *dataset*, o Anexo A apresenta apenas o nome.

4.3 OBTENÇÃO DE *DATASETS* DE *TWEETS*

A plataforma *Twitter* for Developers (2019a) oferece a versão básica da *API* ("*Standard API*") gratuitamente para possibilitar testes de integração, validação de conceitos e extração limitada de dados. Para projetos corporativos e/ou comerciais, são oferecidos planos pagos com níveis elevados de acesso aos dados e recursos para análise em larga escala, como as *Premium* e *Enterprise* APIs. A aplicação deste trabalho foi realizada na versão *Standard* da *API*.

Existe um limite de requisições que podem ser processadas pela *API*, documentadas na seção de *Rate Limits* da documentação oficial da *API* (*TWITTER* FOR DEVELOPERS, 2019b). Os limites são específicos para cada tipo de requisição e válidos por um período de tempo atualmente fixado em 15 minutos.

Para pesquisa e extração de *tweets*, podem ser realizadas até 450 requisições a cada 15 minutos, porém a quantidade total máxima de *tweets* retornados neste período é 18.000. Em termos de período compreendido, são disponibilizados somente os *tweets* dos 7 dias anteriores à requisição. Já para recuperar listas de seguidores de um usuário específico, o limite é de 15 requisições por período, podendo retornar uma quantidade total máxima de 75.000 seguidores.

A fim de viabilizar a integração entre o ambiente R e a *API* do *Twitter*, foi utilizado o pacote *rtweet* que, segundo a documentação oficial do pacote desenvolvido por Kearney (2018), consiste em um conjunto de chamadas para coleta de dados originados da *API*. Através deste pacote, é possível escrever em linguagem R comandos específicos da *API* (como, por exemplo, busca de *tweets*) e obter o retorno diretamente no ambiente R.

Para este trabalho, foi utilizada somente a função *search_tweets*, que retorna um conjunto de *tweets* que contenham um determinado termo de busca, conforme especificado por Kearney (2018).

Conforme constante na documentação oficial de Kearney (2018), a Figura 12 mostra o formato desta função e seus parâmetros.

Figura 12 – Formato da função *search_tweets*

```
search_tweets(q, n = 100, type = "recent", include_rts = TRUE,
  geocode = NULL, max_id = NULL, parse = TRUE, token = NULL,
  retryonratelimit = FALSE, verbose = TRUE, ...)
```

Fonte: Kearney (2018, p. 53).

Abaixo, estão especificados os parâmetros da função que foram utilizados para a aplicação neste trabalho.

- **q =** termo de busca. Utilizado para filtrar e selecionar os *tweets* a serem retornados pela *API*. Deve ser valor do tipo *string* e não exceder o limite máximo de 500 caracteres;
- **n =** quantidade de *tweets* desejados. Se não informado, o parâmetro considera 100 como padrão. Conforme mencionando anteriormente, o número máximo de resultados a cada 15 minutos é 18.000;
- **retryonratelimit =** valor do tipo *boolean* (“TRUE” ou “FALSE”) para indicar se a chamada deve permanecer ativa e esperar quando o limite da *API* for atingido. Isto é, se o máximo de 18.000 *tweets* permitidos a cada 15 minutos for atingido durante a execução da função, a chamada permanecerá ativa e será re-iniciada dentro de 15 minutos. Se não informado, o parâmetro considera “FALSE” como padrão. Neste trabalho, foi utilizado “TRUE”, para garantir a execução contínua do script;
- **include_rts =** valor do tipo *boolean* (“TRUE” ou “FALSE”) para indicar se os resultados devem conter *retweets* ou não. Se não informado, o parâmetro considera “TRUE” como padrão. Neste trabalho, foi utilizado “FALSE”, para que os retweets fossem desconsiderados;

- **lang =** indicador para restringir o idioma dos *tweets* retornados. Para este trabalho, foi utilizado o valor “pt”, estabelecido pela *API* do *Twitter* como relativo ao idioma Português.

Através da função *search_tweets*, o algoritmo desenvolvido para este estudo faz a busca massiva por *tweets* com menção ao nome cada um dos artistas contidos no *dataset* de artistas (descrito no subcapítulo 4.2). A busca ocorre pelo nome do artista e pela respectiva *hashtag* contendo o nome do artista. Desta forma, para cada artista, são realizadas 2 buscas: uma por *tweets* mencionando “nome do artista” e outra por *tweets* que contenham a *hashtag* “#nomedoartista”.

Para garantir um número relevante de *tweets* para cada artista e, ao mesmo tempo, limitar o tempo de processamento do script, o parâmetro n da função search_*tweets* foi fixado em 2.000. Desta maneira, a quantidade máxima de *tweets* a serem recuperados para cada busca foi 2.000 – totalizando um máximo de 4.000 tweets por artista, considerando as 2 requisições realizadas para cada.

Dos 550 artistas pesquisados, não foram identificados *tweets* contendo menções a 8 destes. Assim, ao final desta etapa, haviam sido gerados 542 *datasets* contendo a base de *tweets* obtida para cada um dos artistas analisados e com ao menos 1 *tweet* encontrado.

Para possibilitar a análise do conteúdo dos *tweets* obtidos, somente a coluna *text* (que armazena o conteúdo textual do *tweet* na íntegra) do *dataset* retornado pela *API* do *Twitter* foi considerada, uma vez que as demais informações dos *tweets* (como o nome do usuário autor, a data e hora do *tweet*, as coordenadas de geolocalização, etc) não foram utilizadas para a construção do *dataset* abordado neste estudo.

O Anexo B deste trabalho apresenta o trecho de código R implementado para a obtenção do *dataset* de *tweets*, conforme descrito neste subcapítulo.

Após obtenção dos *tweets*, foram utilizados alguns métodos de tratamento textual previamente à realização das análises propostas. O subcapítulo a seguir explana o pré-processamento a que estes pequenos textos foram submetidos.

4.4 PRÉ-PROCESSAMENTO TEXTUAL DE *TWEETS*

Para otimizar o processamento dos *tweets* analisados e potencializar a acurácia da análise, foram aplicados alguns métodos de pré-processamento no *dataset*, de modo a tratar e normalizar os textos.

Estes métodos de tratamento textual estão comumente presentes em algoritmos de Processamento de Linguagem Natural Natural (do inglês, *Natural Processing Language*, ou NLP), área do conhecimento que aborda métodos computacionais de interpretação e processamento da linguagem humana em formato textual ou falada (ALLEN, 2006).

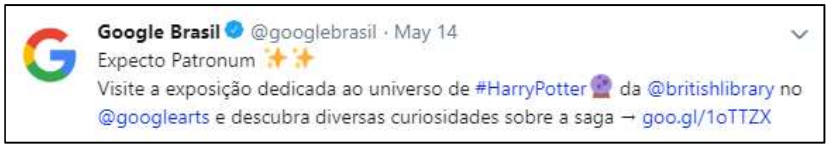
A seguir, são detalhados os tratamentos aplicados na devida sequência de aplicação.

a. **Conversão de caixa**: todo o conteúdo textual do *tweet* é convertido para letras minúsculas;
b. **Remoção das strings referentes ao nome do artista e *hashtag* do artista**: o nome do artista e a respectiva *hashtag* são retirados do texto dos *tweets*, a fim de garantir maior acurácia no cálculo de frequência de palavras. Uma vez que estes são os termos de busca utilizados na filtragem dos *tweets*, são potencialmente os termos de maior frequência, o que causaria ruído no cálculo de frequência;
c. **Remoção de pontuação:** são removidos os caracteres de pontuação através da expressão [:punct:] do ambiente R, que contempla os seguintes símbolos: ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [\ ] ^ _ ` { | } ~;
d. **Remoção de URLs:** todas as strings iniciadas com “http” são removidas do texto dos *tweets*, a fim de desconsiderar links compartilhados pelos usuários;
e. **Remoção de acentuação**: com o uso do pacote *stringi*, todos os caracteres que possuem acentuação são substituídos pelo seu equivalente sem acentuação
f. **Remoção de *emojis*:** os caracteres gráficos referentes a *emojis* são removidos com o auxílio da expressão [:alnum] do ambiente R
g. **Remoção de caracteres numéricos**: para a aplicação deste trabalho, optou-se pela remoção de caracteres numéricos por serem irrelevantes para as análises propostas (análise de sentimentos e de frequência de palavras);

h. **Remoção de *stopwords***: as *stopwords* (ou “palavras de parada”, em português) são palavras para as quais há grande ou a mesma probabilidade de aparecerem tanto em documentos relevantes quanto irrelevantes para determinada pesquisa (WILBUR; SIROTKIN, 1992). Deste modo, por não apresentarem relevância em termos de Processamento de Linguagem Natural, estas palavras são removidas dos textos analisados. As *stopwords* utilizadas como referência neste estudo estão disponíveis no Anexo C deste trabalho;
i. **Remoção de espaçamento adicional:** por fim, são removidos os espaços em branco deixados pela dedução dos elementos mencionados acima, de modo a manter o texto dos *tweets* o mais “fluído” possível.

A seguir, na Figura 13, é apresentado um exemplo de conteúdo textual de um *tweet* (original), seguido pelo texto pré-processado, conforme utilizado na análise deste estudo.

Figura 13 – Exemplo de texto original em *tweet*



Fonte: Adaptado de Twitter (2019)

O texto do *tweet*, após pré-processamento, é convertido para: “expecto patronum visite exposicao dedicada universo harrypotter britishlibrary googlearts descubra diversas curiosidades sobre saga”.

A Figura 14 apresenta o trecho de código R implementado e utilizado nesta prática para efetuar o pré-processamento textual descrito neste subcapítulo.

Figura 14 – Trecho de código para pré-processamento textual

```
tweets_clean <- select(tweets, id, text)
tweets_clean$clean_text <- tolower(tweets_clean$text) # Converte para letras minusculas
tweets_clean$clean_text <- gsub(nome_artista, '', tweets_clean$clean_text) # Retira nome do artista
tweets_clean$clean_text <- gsub(hashtag_artista, '', tweets_clean$clean_text) # Retira hashtag do artista
tweets_clean$clean_text <- removeWords(tweets_clean$clean_text,stopwords("pt")) # Retira stopwords
tweets_clean$clean_text <- gsub("[[:punct:]]", "", tweets_clean$clean_text) # Retira pontuação
tweets_clean$clean_text <- gsub("http\\w+","", tweets_clean$clean_text) # Retira links
tweets_clean$clean_text <- stri_trans_general(tweets_clean$clean_text, "Latin-ASCII") #Retira acentos
tweets_clean$clean_text <- gsub(nome_artista_sem_acento, '', tweets_clean$clean_text) # Retira nome do artista (sem acento)
tweets_clean$clean_text <- gsub(hashtag_artista, '', tweets_clean$clean_text) # Retira hashtag do artista (sem acento)
tweets_clean$clean_text <- gsub("[^[:alnum:][:blank:]?&/\\-]", "", tweets_clean$clean_text) #Retira emojis
tweets_clean$clean_text <- gsub('[0-9]', '', tweets_clean$clean_text) # Retira numeros
tweets_clean$clean_text <- gsub("[ |\t]{2,}", " ", tweets_clean$clean_text) # Retirar tabs (espaçamento)
tweets_clean$clean_text <- gsub("^ ", "", tweets_clean$clean_text) #Retira espaço em branco no inicio
tweets_clean$clean_text <- gsub(" $", "", tweets_clean$clean_text) #Retira espaço em branco no final
```

Fonte: O autor

Conforme fluxo apresentado na Figura 10, o passo seguinte à obtenção e tratamento do conteúdo dos *tweets* é a realização das análises propostas. O subcapítulo 5.5 aborda a análise de polaridade de sentimentos aplicada ao *dataset* após o pré-processamento textual.

4.5 ANÁLISE DE POLARIDADE DE SENTIMENTOS

Análise de sentimentos, também comumemente denominada análise de opinião, é a área de estudo de Processamento de Linguagem Natural que analisa as opiniões, emoções e sentimentos das pessoas em relação a entidades (indivíduos, organizações, produtos, etc) e seus atributos (LIU, 2012).

Segundo Liu (2012), os indicadores mais importantes em análises de sentimentos são os termos que indicam sentimentos ou opiniões, ou seja, as palavras que são frequentemente utilizadas para expressar sentimentos positivos (como “bom”, “maravilhoso”, “incrível”) ou negativos (como “mau”, “horrível”, “pavoroso”). Uma lista contendo estas palavras e as respectivas indicações de polaridade de sentimento (positivo, neutro ou negativo) é chamada léxico de sentimentos. Nas abordagens de análise de sentimentos baseadas em léxicos, a polaridade de um texto é identificada através da orientação semântica das palavras (e frases) que o compõem.

Partindo desta perspectiva, para possibilitar a classificação de sentimentos dos textos obtidos dos *tweets*, foi utilizado o pacote *lexiconPT* do ambiente R. Este pacote fornece acesso a dois léxicos para análise textual em língua portuguesa, respectivamente, **OpLexicon V3.0** e **SentiLexPT02**.

O pacote *lexiconPT* possui uma função específica para retornar o *data frame*[7] completo de cada um dos dois léxicos mencionados acima. Assim, para aplicação neste trabalho, ambos os *data frames* foram armazenados localmente no ambiente R para processamento.

Nas Figuras 15 e 16 são apresentadas as estruturas dos léxicos retornados pelas funções sentiLex_lem_PT02 e oplexicon_v3.0 do pacote lexiconPT. Em ambos, a coluna “term” armazena a palavra ou emoji a que se refere o registro. As colunas “grammar_category” (sentiLex_lem_PT02) e “type” (oplexicon_v3.0) apresentam a classificação do item, como “vb” (verbo) ou “adj” (adjetivo). A coluna “polarity” se refere

[7] Elemento do ambiente R utilizado para armazenamento de dados tabelados; uma lista de vetores em igual quantidade (R TUTORIAL, 2019).

à classificação numérica de polaridade de sentimento do item, onde valores menores ou iguais à -1 referem a um sentimento negativo, maiores ou iguais a 1, a um sentimento positivo, e o valor 0 indica um sentimento neutro. Por fim, as colunas “polarity_classification” (sentiLex_lem_PT02) e “polarity_revision” (oplexicon_v3.0) indicam se a polaridade de sentimento do item foi obtida manualmente ou automaticamente.

Neste trabalho, somente as colunas “term” e “polarity” de ambos os léxicos foram consideradas.

Figura 15 – Estrutura do léxico OptLexicon V3.0

	term	type	polarity	polarity_revision
521	¬¬	emot	-1	A
522	¬¬''	emot	-1	A
523	A¬	emot	-1	A
524	ab-rogar	vb	-1	A
525	ababadar	vb	0	A
526	ababelar	vb	-1	A
527	ababelar-se	vb	1	A
528	abacanar	vb	1	A
529	abacinar	vb	1	A
530	abafada	adj	-1	A
531	abafadas	adj	-1	A
532	abafado	adj	-1	A
533	abafados	adj	-1	A
534	abafante	adj	-1	A
535	abafantes	adj	-1	A

Fonte: O autor

Figura 16 – Estrutura do léxico SentiLexPT02

	term	grammar_category	polarity	polarity_classification
1	a-vontade	N	1	MAN
2	abafado	Adj	-1	JALC
3	abafante	Adj	-1	MAN
4	abaixado	Adj	-1	JALC
5	abalado	Adj	-1	JALC
6	abalizado	Adj	1	JALC
7	abalroado	Adj	-1	MAN
8	abalroar	V	1	MAN
9	abanar	V	1	MAN
10	abandalhado	Adj	-1	MAN
11	abandalhamento	N	-1	MAN
12	abandonado	Adj	-1	JALC
13	abandonar	V	-1	MAN
14	abarcante	Adj	0	MAN

Fonte: O autor

Com relação à quantidade de itens/palavras em cada léxico, o OpLexicon V3.0 possui 32.191 linhas, sendo, destas, 146 repetidas. O SentiLexPT02 apresenta 7.014 linhas, sendo 6 repetidas. Após a remoção dos itens repetidos, foi identificado que existe, ainda, um total de 5.122 palavras constantes nos dois léxicos.

Para contornar o problema de polaridades conflitantes nas palavras repetidas entre os léxicos, o algoritmo criado e utilizado neste trabalho priorizou a polaridade constante no léxico SentiLexPT02, por ter sido desenvolvido com base em análise de *datasets* de *tweets*, conforme documentação oficial do léxico. Para as palavras que não constam neste léxico, o algoritmo considera, então, a polaridade constante no léxico OpLexicon V3.0.

De modo a garantir a padronização da análise, os seguintes métodos de pré-processamento (descritos no subcapítulo 4.4) foram aplicados também nos termos dos léxicos de sentimentos: conversão de caixa, remoção de acentuação, remoção de pontuação e remoção de espaçamento adicional.

O algoritmo utilizado neste estudo realizou a classificação de sentimentos da seguinte maneira: cada *tweet* dos 542 *datasets* existentes foi fragmentado em palavras, de modo que foram criados 542 *datasets* temporários onde cada linha/registro continha uma palavra e o respectivo id do *tweet* original. O algoritmo faz a busca de cada palavra nos léxicos do pacote *lexiconPT* e determina o valor de polaridade de sentimento para cada palavra (-1, 0 ou 1). Após isto, o algoritmo realiza

a soma algébrica dos valores de polaridades obtidos para cada *tweet* e determina o valor final para o *tweet*, que passa a conter uma legenda com os labels “Positivo”, “Neutro” ou “Negativo”.

A Figura 17 apresenta parcialmente o trecho de código R responsável pela identificação de polaridade de sentimentos dos *tweets* analisados, conforme detalhado no presente subcapítulo.

Figura 17 – Trecho de código para análise de polaridade de sentimentos

```
# Criar uma linha para cada palavra
tweets_clean_final <- tweets_clean %>% unnest_tokens(term, clean_text)

# Remover palavras duplicadas dentro de um mesmo tweet
tweets_clean_final <- unique(tweets_clean_final)

# Verifica cada palavra nos lexicos
tweets_sentiment_analysis  <- tweets_clean_final %>%
                              left_join(op30, by = "term") %>%
                              left_join(sent %>% select(term, lex_polarity = polarity), by = "term") %>%
                              select(id, term, polarity, lex_polarity)

tweets_sentiment_analysis <- sqldf("select  id,
                                            term,
                                            sum(polarity) as 'polarity',
                                            sum(lex_polarity) as 'lex_polarity'
                                    from    tweets_sentiment_analysis
                                    group by
                                            id, term")

count2 = 1
tweets_sentiment_analysis$sentiment = NA

for(m in 1:nrow(tweets_sentiment_analysis)){ ## Analise de sentimento
    if ((stri_cmp(toString(tweets_sentiment_analysis$polarity[count2]), "NA")==-1)) {
        tweets_sentiment_analysis$sentiment[count2] = tweets_sentiment_analysis$polarity[count2]
    }

    else {
        tweets_sentiment_analysis$sentiment[count2] = tweets_sentiment_analysis$lex_polarity[count2]
    }
    count2 <- count2 + 1
}
```

Fonte: O autor

A fim de otimizar a acurácia da análise de sentimentos (e, posteriormente, da geração de recomendações), foi realizada uma análise de frequência de termos em todos os *tweets* obtidos. O subcapítulo a seguir detalha a análise realizada.

4.6 ANÁLISE DE FREQUÊNCIA DE TERMOS

De modo a identificar os termos mais comumentes utilizados quando determinado artista é referenciado nos *tweets* analisados, foram aplicados métodos

para obter os *n-gramas* mais frequentes nos *datasets* específicos de cada artista musical.

Vilela (2011) define um n-grama como “uma subsequência de n elementos de uma dada sequência”. A autora esclarece que um n-grama de tamanho 1 (isto é, de um único elemento) é comumente denominado unigrama. Os n-gramas de tamanho 2 são bigramas, e, os de tamanho 3, trigramas. Dos Santos (2016) complementa que, a partir de 4 elementos, estes são denominados n-gramas. Dos Santos (2016) endossa a utilização do atributo n-grama no contexto de análise textual, que pode ser bastante informativo e um recurso relevante para capturar estilos de linguagem e apoiar algoritmos de análise de sentimentos.

Neste trabalho, para cada artista analisado, foi identificado o unigrama, o bigrama e o trigrama mais frequente. Isto é, a palavra, o conjunto de 2 palavras (juntas), e o conjunto de 3 palavras (juntas) mais frequentemente utilizados ao mencionar o artista.

Para implementação prática no ambiente R, foi utilizado o pacote tidytext, que objetiva suportar tarefas de mineração de dados textuais (SILGE; ROBINSON, 2018)

Para a obtenção dos n-gramas, foi utilizada a função *unnest_tokens* deste pacote, que realiza a separação de palavras dos textos contidas no *dataset*. Para cada *dataset*, a função foi utilizada 3 vezes: geração de unigramas, bigramas e trigramas, retornando a respectiva listagem de n-gramas. Posteriormente, valendo-se de funções do pacote *sqldf* do ambiente R, as 3 listagem retornadas para cada *dataset* foram agrupadas e quantificadas, de modo a identificar o n-grama presente no maior número de *tweets*.

A Figura 18 mostra um trecho do código R responsável pela obtenção de unigramas. A mesma lógica foi replicada para bigramas, trigramas e para identificação dos n-gramas por tipo de sentimento, conforme descrito no subcapítulo 4.5 deste trabalho.

Figura 18 – Trecho de código para obtenção de unigramas

```
df <- tweets_clean %>% unnest_tokens( unigram, clean_text, token = "ngrams", n = 1)
unigram <- sqldf("  select  unigram
                    from    df
                    where   unigram <> 'NA'
                    group by
                            unigram
                    order by
                            count(*) DESC
                    LIMIT 1")
```

Fonte: O autor

Em conformidade com o fluxo de implementação apresentado na introdução deste capítulo, o passo seguinte à realização das análises de sentimentos e de frequência de termos é a geração da matriz de resultados. O formato e análise dos resultados obtidos com a execução do algoritmo e o *dataset* de referência adotado são abordados no capítulo a seguir.

5 RESULTADOS

O algoritmo implementado neste estudo gera, como resultado, 2 *datasets* distintos: um por artista, contendo a listagem completa de *tweets* analisados, e um *dataset* de resultados gerais, contemplando as métricas referentes a todos os artistas analisados.

5.1 *DATASETS* DE RESULTADOS

O Quadro 1 apresenta o nome e a descrição de conteúdo das colunas presentes nos *datasets* gerados individualmente por artista analisado.

Quadro 1 – Estrutura dos *datasets* individuais por artista

Nome da coluna	Descrição
Id	Identificador do *tweet* no *dataset*
Texto	Texto original do *tweet* analisado
texto_final	Texto do *tweet* após aplicação dos métodos de pré-processamento
sentimento_valor	Soma dos valores de polaridade de sentimento de cada palavra analisada
sentimento_legenda	Classificação de polaridade de sentimento do *tweet*, baseado na coluna sentimento_valor, podendo ser “Negativo”, “Neutro”, “Positivo” ou “Não identificado”. A legenda “Não identificado” significa que não constam nenhum dos termos do texto do *tweet* em nenhum dos léxicos analisados.

Fonte: O autor

Por sua vez, através da análise automatizada dos *datasets* individuais (descritos acima), o algoritmo gera um *dataset* de resultados finais, apresentando os dados de cada artista de maneira resumida.

O Quadro 2 apresenta o nome e a descrição de conteúdo das colunas presentes no *dataset* de resultados das análises.

Quadro 2 – Estrutura do *dataset* de resultados gerais

Nome da coluna	Descrição
nome_artista	nome do artista analisado, em letra minúscula e sem acentos
hashtag_artista	*hashtag* do artista analisado
genero_musical	classificação de gênero musical conforme *dataset* de referência
total_tweets	quantidade total de *tweets* analisados
total_tweets_positivos	quantidade total de *tweets* classificados com polaridade positiva de sentimento
total_tweets_negativos	quantidade total de *tweets* classificados com polaridade negativa de sentimento
total_tweets_neutros	quantidade total de *tweets* classificados com polaridade neutra de sentimento
total_tweets_nao_encontrados	quantidade total de *tweets* para os quais não foi encontrada polaridade de sentimento
perc_tweets_positivos	porcentagem de *tweets* (em relação à quantidade total de *tweets* analisados) classificados com polaridade positiva de sentimento
perc_tweets_negativos	porcentagem de *tweets* (em relação à quantidade total de *tweets* analisados) classificados com polaridade negativa de sentimento
perc_tweets_neutros	porcentagem de *tweets* (em relação à quantidade total de *tweets* analisados) classificados com polaridade neutra de sentimento
perc_tweets_nao_encontrados	porcentagem de *tweets* (em relação à quantidade total de *tweets* analisados) para os quais não foi encontrada polaridade de sentimento
Unigrama	unigrama (palavra) mais frequente nos *tweets* analisados
Bigrama	bigrama (2 palavras juntas) mais frequente nos *tweets* analisados
Trigrama	trigrama (3 palavra juntas) mais frequente nos *tweets* analisados
unigrama_positivos	unigrama (palavra) mais frequente nos *tweets* analisados e classificados com polaridade positiva de sentimento

bigrama_positivos	bigrama (2 palavras juntas) mais frequente nos *tweets* analisados e classificados com polaridade positiva de sentimento
trigrama_positivos	trigrama (3 palavra juntas) mais frequente nos *tweets* analisados e classificados com polaridade positiva de sentimento
unigrama_negativos	unigrama (palavra) mais frequente nos *tweets* analisados e classificados com polaridade negativa de sentimento
bigrama_negativos	bigrama (2 palavras juntas) mais frequente nos *tweets* analisados e classificados com polaridade negativa de sentimento
trigrama_negativos	trigrama (3 palavra juntas) mais frequente nos *tweets* analisados e classificados com polaridade negativa de sentimento
unigrama_neutros	unigrama (palavra) mais frequente nos *tweets* analisados e classificados com polaridade neutra de sentimento
bigrama_neutros	bigrama (2 palavras juntas) mais frequente nos *tweets* analisados e classificados com polaridade neutra de sentimento
trigrama_neutros	trigrama (3 palavra juntas) mais frequente nos *tweets* analisados e classificados com polaridade neutra de sentimento

Fonte: O autor

A prática descrita no capítulo anterior, por sua vez, originou os 2 tipos de *datasets* de resultados descritos acima; isto é, foram gerados 542 *datasets* de resultados individuais (referentes a cada um dos artistas analisados), e 1 *dataset* de resultados finais.

O subcapítulo a seguir discute os resultados obtidos no *dataset* de resultados gerais resultante desta prática.

5.2 ANÁLISE DE RESULTADOS

Com relação às quantidades, foi analisado um total de **298.124** *tweets*, sendo, em média, **550** *tweets* para cada um dos 542 artistas analisados. Destes, o artista com maior número de *tweets* analisados foi **Sandy e Júnior (3161 *tweets*)**. Já com relação a menor quantidade de *tweets* analisados, foi obtido apenas **1** tweet para 6 dos artistas analisados.

O Quadro 3 apresenta a relação dos 10 artistas para os quais foram obtidas as maiores quantidades de *tweets,* podendo ser considerado como um índice de popularidade destes dentro da amostra analisado.

Quadro 3 – Artistas com maior quantidade de *tweets* analisados

Nome do artista	Hashtag do artista	Total de Tweets	Tweets Positivos	Tweets Negativos	Tweets Neutros
sandy e junior	#sandyejunior	3161	1316	618	710
luan santana	#luansantana	2522	856	784	487
Cartola	#cartola	2516	450	1118	383
paula fernandes	#paulafernandes	2227	772	746	432
los hermanos	#loshermanos	2191	867	501	439
wesley safadão	#wesleysafadão	2063	633	485	418
Sandy	#sandy	2021	730	479	420
bruno & marrone	#bruno&marrone	1993	222	1344	399
marília mendonça	#maríliamendonça	1984	816	346	432
gusttavo lima	#gusttavolima	1974	696	389	347

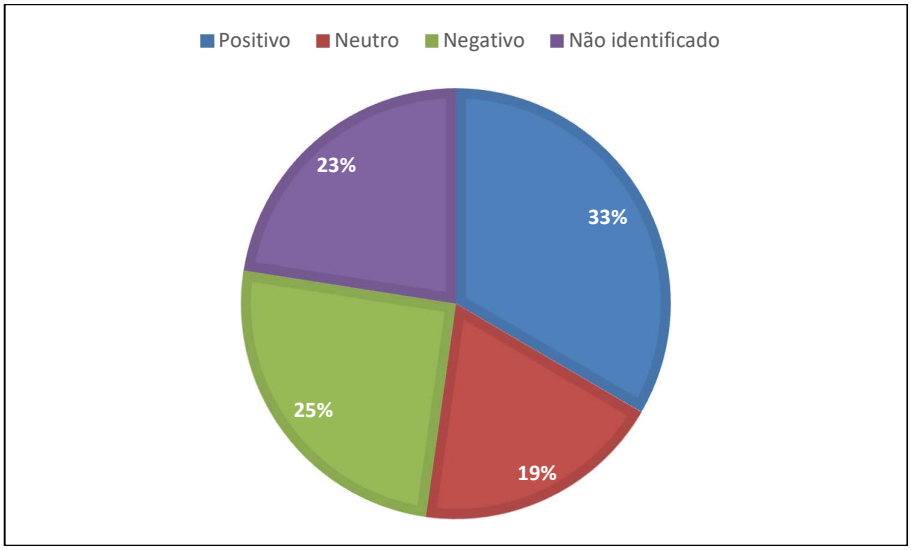
Fonte: O autor

No que tange às classificações de polaridade de sentimento, na sequência constam as porcentagens médias identificadas para cada uma das classificações praticadas neste estudo (Positivo, Neutro, Negativo e Não identificado):

a) Porcentagem média de *tweets* positivos: **33.41%**
b) Porcentagem média de *tweets* neutros: **18.83%**
c) Porcentagem média de *tweets* negativos: **25.23%**
d) Porcentagem média de *tweets* com polaridade não identificada: **22.53%**

O Gráfico 1 apresenta a distribuição das porcentagem médias referente às polaridades de sentimento identificadas.

Gráfico 1 – Distribuição de porcentagem de classificação por polaridade de sentimento



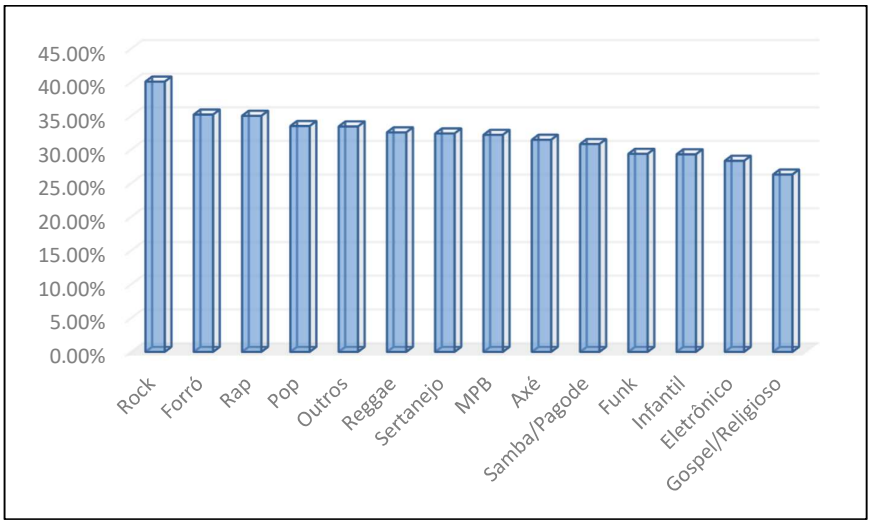
Fonte: O autor

Ainda no tocante à polaridade de sentimentos, na sequência estão listados os artistas para os quais foram observadas as maiores porcentagens de *tweets* classificados com polaridade positiva, neutra, negativa e não identificada.

a) Maior porcentagem de *tweets* positivos: **79.03%** (Ponto de Equilíbrio)
b) Maior porcentagem de *tweets* neutros: **58.02%** (Tropkillaz)
c) Maior porcentagem de *tweets* negativos: **67.44%** (Bruno & Marrone)
d) Maior porcentagem de *tweets* com polaridade não identificada: **66.83%** (Humberto & Ronaldo)

No que tange a relação entre a classificação de polaridade de sentimentos e os diferentes gêneros musicais analisados, o Gráfico 2 apresenta a distribuição entre gêneros musicais de porcentagem média de *tweets* classificados com polaridade de sentimento positiva; o Gráfico 2, apresenta a mesma distribuição para os classificados com polaridade negativa.

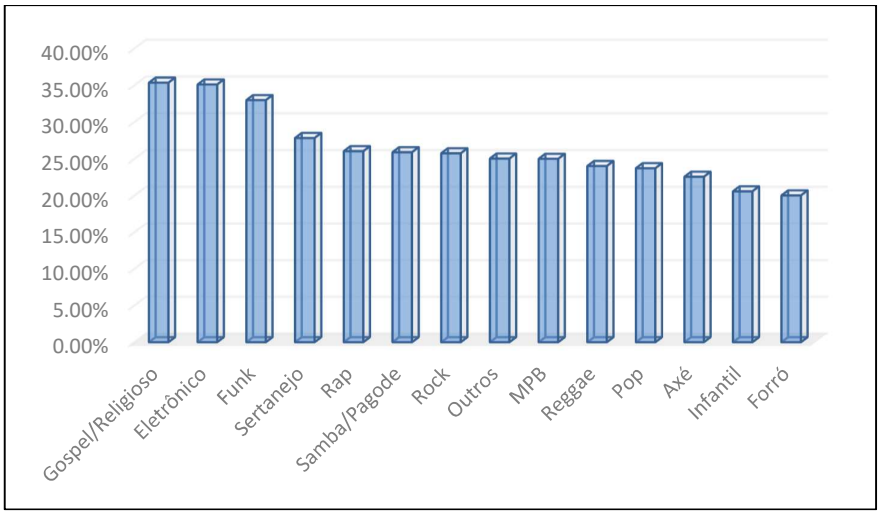
Gráfico 2 – Distribuição de porcentagem média de *tweets* classificados com polaridade de sentimento positiva por gênero musical



Fonte: O autor

Conforme pode ser visto nos Gráficos 2 e 3, o gênero Rock foi o que obteve o maior índice de comentários classificados positivamente, enquanto Gospel/Religioso ficou em último lugar na classificação de *tweets* positivos e também no primeiro lugar na classificação de negativos.

Gráfico 3 – Distribuição de porcentagem média de *tweets* classificados com polaridade de sentimento negativa por gênero musical



Fonte: O autor

O dataset de resultados aqui descrito, bem como os 542 *datasets* de resultados individuais obtidos, estão disponíveis para consulta e podem ser solicitados através do e-mail giovani.pozzo@gmail.com.

6 CONSIDERAÇÕES FINAIS

Como visto no subcapítulo 2.1, o surgimento da *Web* 2.0 incentivou a criação de páginas mais interativas, e, pela primeira vez, possibilitou aos usuários interagirem com os veículos de comunicação de maneira ativa (SAAD, 2003). Um exemplo de contribuição desta mudança de paradigma é a emergência das redes sociais *online*, que, afirmam Costa et al. (2016), têm estado no centro de diversas pesquisas devido à sua importância no que se refere, por exemplo, à difusão da informação e possibilidade de comunicação.

Ao mesmo passo, a democratização do acesso à Internet ao longo dos últimos anos também contribuiu para o aumento significativo de usuários ativos. No Brasil, por exemplo, a marca de domicílios particulares permanentes com acesso à Internet passou de 50% em 2014, indica pesquisa do IBGE (2016).

O consequente aumento exponencial na quantidade de informações disponíveis na rede passou, por conseguinte, a dificultar aos usuários encontrarem com facilidade informações relevantes para si, fenômeno estudado por Schwartz (2009) e denominado “Paradoxo da Escolha”. Como uma alternativa a esta sobrecarga de informações, surge a demanda por sistemas de recomendação (RICCI; ROKACH; SHAPIRA, 2010).

A revolução digital também afetou drasticamente as práticas do universo fonográfico, aponta De Sá (2009). No que tange às plataformas de *streaming*, com um catálogo de mais de 35 milhões de faixas musicais atualmente disponíveis no *Spotify* (SPOTIFY, 2018), esta sobrecarga de informações também pode ser facilmente observada no âmbito musical.

Como apresentado no capítulo 3, no tocante aos trabalhos relacionados, embora os temas de recomendação musical e análise textual e de sentimentos de postagens de redes sociais *online* tenham sido largamente explorados ao longo dos últimos anos, não foi identificada nenhuma pesquisa que tenha como foco utilizar os *posts* compartilhados por um usuário nas redes sociais *online* para inferir seus gostos musicais e servir como fonte para recomendações musicais; há exceção da pesquisa apresentada por Rosa, Rodriguez e Bressan (2015), que utiliza-se da coleta de tais *posts* somente para inferir o estado emocional do usuário, porém sem levar em conta preferências musicais.

Partindo desta perspectiva, o presente estudou buscou detalhar um método de construção de um *dataset* de informações musicais obtidas através de análise textual e análise de sentimentos aplicadas a uma base de *tweets*.

Como resultado deste estudo, foram obtidos 542 *datasets* contendo *tweets* com menções a artistas musicais brasileiros, a respectiva classificação de polaridade de sentimento (por *tweet*) e seu respectivo texto normalizado. Além disso, foi gerado um *dataset* de resultados gerais contendo uma análise de frequência de termos (n-gramas) para cada artista musical, de modo a indicar os termos mais frequentemente mencionados quando determinado artista é referenciado.

A base de *tweets* utilizada neste estudo foi obtida em 19 de Maio de 2019 e, em conformidade com as limitações estabelecidas pela *API* do *Twitter*, o período de obtenção de histórico de *tweets* compreende os 7 dias anteriores ao momento do envio da requisição. Neste sentido, vale ressaltar que este dataset possui como limitação a alta sensibilidade a eventos específicos – isto é, os resultados das análises podem sofrer forte influência de eventuais acontecimentos populares da semana anterior (lançamentos musicais, acontecimentos midiáticos, declarações públicas, etc)

Ainda quanto às limitações dos datasets gerados neste estudo, deve-se levar em conta que a busca por tweets baseada no nome dos artistas (e suas respectivas hashtags) assume que a maior quantidade de *tweets* retornados se refere ao artista pesquisado. Esta hipótese torna-se inválida em diversos casos, como quando um nome artístico é também largamente utilizado em outros contextos linguísticos que não possuem relação com o artista pesquisado (por exemplo, os cantores “Silva” e “Cartola”), ou nos casos de múltiplas celebridades que compartilham o mesmo nome (por exemplo, os cantores “Roberto Carlos” e “Falcão”).

Para pesquisas futuras, sugere-se aplicar o método proposto em bases de dados maiores, contemplando maior quantidade de artistas musicais, por um período de tempo mais abrangente, de modo a aumentar a relevância do *dataset.*

É possível, ainda, aplicar o algoritmo desenvolvido neste estudo em um dataset de tweets de usuários específicos, de maneira a gerar recomendações personalizadas.

Por fim, sugere-se, para trabalhos futuros, o estudo de vaiabilidade da aplicação do modelo aqui proposto como uma potencial solução à problemática de *cold-start* apresentada no embasamento teórico deste estudo.

REFERÊNCIAS

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 6, 2005, p. 734-749.

ALLEN, J. F. **Natural Language Processing:** The Cambridge Handbook of Second Language Acquisition. Rochester, New York, USA: University of Rochester, 2006. Disponível em: <http://ebooks.cambridge.org/ref/id/CBO9781139051729A033>. Acesso em 2 mai. 2019.

AMARAL, A.; RECUERO, R.; MONTARDO, S. P. Blogs: mapeando um objeto. *In*: _____. (Org.). **Blogs.com**: estudos sobre blogs e comunicação. São Paulo: Momento Editorial, 2009, p. 27-131.

BERTAGLIA, T. F. **Normalização textual de conteúdo gerado por usuário**. 2017. 160p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2017.

BEZ, M. R. **O uso de tecnologia para apoiar a implantação de métodos ativos nos currículos de medicina.** 2011. 117p. Proposta de Tese (Doutorado em Informática na Educação) – Programa de Pós-graduação em Informática na Educação, Centro Interdisciplinar de Novas Tecnologias na Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011.

CALDERON, Pio. **An overview of recommendation systems**. DATA meets MEDIA. 23 mai. 2018. Disponível em: <http://datameetsmedia.com/an-overview-of-recommendation-systems/> Acesso em 02 jun. 2018.

CASTELLS, M. **A galáxia da Internet**: reflexões sobre a internet, os negócios e a sociedade. Rio de Janeiro: Editora Paz e Terra, 2003.

COMPREHENSIVE R ARCHIVE NETWORK. **Contributed Packages**. 2019. Disponível em: <https://cran.r-project.org/web/packages/index.html> Acesso em 2 mai. 2019.

COSTA, L. F. et al. O uso de mídias sociais por revistas científicas da área da ciência da informação para ações de marketing digital, **Revista ACB: Biblioteconomia em Santa Catarina**, v. 21, n. 2, p. 338-358, abr./jul. 2016.

DAUGHERTY, T; EASTIN, M. S.; BRIGHT, L. Exploring consumer motivations for creating user-generated content, **Journal of Interactive Advertising**, v. 8, n. 2, 2008, p.16-25.

DE MARCHI, L. G. **Transformações estruturais da indústria fonográfica no Brasil 1999-2009:** desestruturação do mercado de discos, novas mediações do comércio de fonogramas digitais e consequências para a diversidade cultural no

mercado de música. 2011. 289p. Tese (Doutorado em Comunicação e Cultura) – Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2011.

DE SÁ, S. P. A música na era de suas tecnologias de reprodução, **Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação**, v. 6, ago. 2006. 19 p.

DE SÁ, S. P. Se vc gosta de Madonna também vai gostar de Britney! Ou não? Gêneros, gostos e disputa simbólica nos Sistemas de Recomendação Musical, **Revista da Associação Nacional dos Programas de Pós-Graduação em Comunicação**, v. 12, n. 2, maio/ago. 2009. 20 p.

DEEZER. **Deezer**. 2019. Disponível em: <https://deezer.com>. Acesso em 16 mai. 2019.

DOS SANTOS, J. C. **Avaliação Automática de Questões Discursivas Usando LSA.** 2016. 118p. Tese (Doutorado em Computação Aplicada) – Programa de Pós-Graduação em Engenharia Elétrica, Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2016.

ESSINGER, S. O ano quem que o streaming conquistou milhões. **O Globo**, Rio de Janeiro, 27 dez. 2016.

FEDERAÇÃO INTERNACIONAL DA INDÚSTRIA FONOGRÁFICA. **Global Music Report 2017**. 2017. Disponível em: <http://www.ifpi.org/news/IFPI-GLOBAL-MUSIC-REPORT-2017> Acesso em: 20 mar. 2018.

FELIX, V. **Os 500 brasileiros mais bombados hoje no Spotify**. Red Bull. 2017. Disponível em: <https://www.redbull.com/br-pt/os-500-brasileiros-mais-bombados-hoje-no-spotify> Acesso em 3 abr. 2019.

GONZAGA, S. **Package 'lexiconPT'**. 2017. Disponível em: <https://cran.r-project.org/web/packages/lexiconPT/lexiconPT.pdf> Acesso em 3 abr. 2019.

GRAEML, K. S. et al. O impacto do uso (excessivo) da Internet no comportamento social das pessoas, **Revista Psicologia Corporal**, v. 5, 2004. p. 1-6.

HU, Y.; KOREN, Y.; VOLINSKY, C. Collaborative filtering for implicit feedback datasets, **IEEE International Conference on Data Mining 2008**, 2008, p. 263–272.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal: 2014.** Rio de Janeiro: 2016. 84p. Disponível em: <https://biblioteca.ibge.gov.br/index.php/ biblioteca-catalogo?view=detalhes&id=295871>. Acesso em 25 abr. 2018.

KEARNEY, M. W. **Package 'rtweet'.** 2018. Disponível em: <https://cran.r-project.org/Web/packages/rtweet/rtweet.pdf> Acesso em 3 abr. 2019.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, 2009. Disponível em: < https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>. Acesso em 10 mai. 2019.

MELODA. **Dataset definition**. 2019. Disponível em: <http://www.meloda.org/dataset-definition/> Acesso em 10 mai. 2019.

MILLER, H. et al. “Blissfully Happy” or “Ready to Fight”: Varying Interpretations of Emoji. **Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)**, 2016, p.259-268.

MIRA, J.; BODONI, P. Os impactos das redes sociais virtuais nas relações de jovens e adultos no ambiente acadêmico nacional, **Revista de Educação**, v. 14, n. 17, 2011, p. 103-115.

MOGHADDAM, F.; ELAHI, M. Cold Start Solutions For Recommendation Systems. In: **Big Data Recommender Systems**: Recent Trends and Advances. 2018.

PASICK, A. **The magic that makes Spotify’s ‘Discover Weekly’ playlists so damn good**. Quartz Media. 21 dez. 2015. Disponível em: <https://qz.com/571007/the-magic-that-makes-spotifys-discover-weekly-playlists-so-damn-good/> Acesso em 04 jun. 2018.

PASQUALE, L.; GEMMIS, M.; SEMERARO, G. Content-based Recommender Systems: State of the Art and Trends. In: RICCI, F. et al. (Org.). **Recommender Systems Handbook**. New York: Springer, 2010. cap. 3, p. 73-105.

PINTO, G. F. **Comportamento Informacional e Mineração Textual no Twitter**. 2018. 63p. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos, São Carlos, 2018.

PIRES, M. M. **Distribuição de música na cultura digital:** a relação de artistas autônomos com as plataformas de Streaming. 2017. 115p. Dissertação (Mestrado em Tecnologias da Inteligência e Design Digital) – Programa de Pós-graduação em Tecnologias da Inteligência e Design Digital, Pontifícia Universidade Católica de São Paulo, São Paulo, 2017.

PRODANOV, C. C.; FREITAS, E. C. de. **Metodologia do trabalho científico**. 2. ed. Novo Hamburgo: Feevale, 2013.

R FOUNDATION. **What is R?**. 2019. Disponível em: <https://www.r-project.org/about.html> Acesso em 2 mai. 2019.

R TUTORIAL. **Data Frame**. 2019. Disponível em: <http://www.r-tutor.com/r-introduction/data-frame>. Acesso em 5 mai. 2019.

RANDLE, Q. A historical overview of the effects of new mass media introductions on magazine publishing during the 20th century, **Peer-reviewed Journal on the Internet**, v. 6, n. 9, set. 2001.

RECUERO, R. **Redes sociais na Internet**. Porto Alegre: Meridional, 2009.

RESNICK, P.; VARIAN, H. R. Recommender systems, **Communications of the ACM**, v. 40, n. 3, 1997, p. 55-58.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to Recommender Systems Handbook. In: RICCI, F. et al. (Org.). **Recommender Systems Handbook**. New York: Springer, 2010. cap. 1, p. 1-37.

ROBINSON, David. **The Impressive Growth of R**. StackOverflow Blog. 2017. Disponível em: <https://stackoverflow.blog/2017/10/10/impressive-growth-r/> Acesso em 2 mai. 2019.

ROSA, R. L.; RODRIGUEZ, D. Z.; BRESSAN, G. Music recommendation system based on user's sentiments extracted from social networks, **IEEE Transactions on Consumer Electronics**, v. 61, n. 3, p. 359-367, ago. 2015.

SAAD, B. **Estratégias para a mídia digital:** internet, informação e comunicação. São Paulo: Senac São Paulo, 2003.

SCHEDL, M. et al. Current challenges and visions in music recommender systems research, **International Journal of Multimedia Information Retrieval**, v. 7, n. 2, 2018, p. 95-116.

SCHEDL, M.; SCHNITZER, D. Location-Aware Music Artist Recommendation, In: **MultiMedia Modeling (MMM 2014)**, v. 8326, 2014, p. 205-213.

SCHWARTZ, B. **The paradox of choice**: why less is more. New York: Ecco, 2004.

SHABAN, Hamza. **Twitter reveals its daily active user numbers for the first time**. The Washington Post, 2019. Disponível em: <https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/?noredirect=on&utm_term=.e46c9c8abc7f>. Acesso em 1 mai. 2019.

SHARDANAND, U.; MAES, P. Social information filtering: algorithms for automating "Word of Mouth", **Human Factors in Computing Systems**, 1995, p. 210-217.

SILGE, J.; ROBINSON, D. **Introduction to tidytext**. 2018. Disponível em: <https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html>. Acesso em 10 mai. 2019.

SILVEIRA, H. J. **Colagem musical na música eletrônica experimental**. 2012. 202p. Dissertação (Mestrado em Música) – Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo, 2012.

SKOWRON, M. et al. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram, **WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web**, 2016, p.107-108.

SOLIS, Brian. **Conversation Prism 5.0**. 2018. Disponível em: <https://conversationprism.com/> Acesso em 02 mai. 2018.

SPOTIFY. **Company Info:** Fast Facts. Disponível em: <https://newsroom.spotify.com/companyinfo> Acesso em 08 jun. 2018.

SPOTIFY. **Spotify**. Disponível em: <https://music.youtube.com. Acesso em 16 mai. 2019.

TOMAÉL, M. I.; MARTELETO, R. M. Redes sociais: posição dos atores no fluxo da informação, **Revista Eletrônica de Biblioteconomia e Ciência da Informação**, 2006, p. 75-91.

TOMAÉL, M.; ALCARA, A.; DI CHIARA, I. Das redes sociais à inovação, **Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 34, n. 2, p. 93-104, maio/ago. 2005.

TWITTER FOR DEVELOPERS. **Products overview**. 2019. Disponível em: <https://developer.twitter.com/en/products/products-overview>. Acesso em 10 mai. 2019.

TWITTER FOR DEVELOPERS. **Rate Limits**. 2019. Disponível em: <https://developer.twitter.com/en/docs/basics/rate-limits>. Acesso em 10 mai. 2019.

TWITTER. **Help Center:** About Twitter's APIs. Disponível em: <https://help.twitter.com/en/rules-and-policies/twitter-API> Acesso em 10 mai. 2019.

TWITTER. **Twitter**. Disponível em: <https://twitter.com>. Acesso em 16 mai. 2019.

VALENTE, C.; MATTAR, J. **Second Life e Web 2.0 na educação**: o potencial revolucionário das novas tecnologias. São Paulo: Novate, 2007.

VILELA, P. C. **Classificação de Sentimento para Notícias sobre a Petrobrás no Mercado Financeiro.** 2011. 44p. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Informática, Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2011.

WE ARE SOCIAL. **Global Digital Report 2018.** Londres: 2018. Disponível em: <https://wearesocial.com/uk/blog/2018/01/global-digital-report-2018> Acesso em: 20 mar. 2018.

WILBUR, W. J.; SIROTKIN, K. **The automatic identification of stop words**. Journal of Information Science, v. 18, n. 1, p. 45–55, 1992.

YAMASHITA, J. S. **Visualização de tags para explicar e filtrar recomendações de músicas**. 2013. 185p. Dissertação (Mestrado em Ciências) – Programa de Pós-graduação em Ciência da Computação, Universidade de São Paulo, São Paulo, 2013.

YOUTUBE MUSIC. **YouTube Music**. Disponível em: <https://spotify.com>. Acesso em 16 mai. 2019.

ZIWON, H.; LEE, K.; LEE, K. Music recommendation using text analysis on song requests to radio stations, **Expert Systems with Applications**, v. 14, n. 5, abr. 2013, p. 2608-2618.

ANEXO A – *DATASET* DE ARTISTAS MUSICAIS

1Kilo (Rap), A Banca 021 (Rap), A Banda Mais Bonita da Cidade (Pop), Adoniran Barbosa (Samba/Pagode), Adriana Calcanhoto (MPB), Adriana Partimpim (Infantil), Alceu Valença (MPB), Alcione (Samba/Pagode), Alice Caymmi (Pop), Aline Barros (Gospel/Religioso), Aline Calixto (Samba/Pagode), Almir Guineto (Samba/Pagode), Almir Sater (Sertanejo), Alok (Eletrônico), Ana Cañas (MPB), Ana Carolina (MPB), Ana Gabriela (Pop), Ana Muller (MPB), Ana Nóbrega (Gospel/Religioso), Ana Vilela (Pop), Anavitória (Pop), Anderson Freire (Gospel/Religioso), André Valadão (Gospel/Religioso), Andy Bianchini (Eletrônico), Angra (Metal), Anitta (Pop), Apanhador Só (Rock), Ara Ketu (Axé), Arlindo Cruz (Samba/Pagode), Armandinho (Reggae), Arnaldo Antunes (MPB), Art Popular (Samba/Pagode), As Bahias e a Cozinha Mineira (MPB), Astrud Gilberto (MPB), Avine Vinny (Forró), Aviões do Forró (Forró), Babado Novo (Axé), Baco Exu do Blues (Rap), Banda A Favorita (Brega), Banda Calypso (Brega), Banda do Mar (MPB), Banda Eva (Axé), Banda Uó (Pop), Barão Vermelho (Rock), Baviera (Rap), Bebel Gilberto (MPB), Belchior (MPB), Belo (Samba/Pagode), Benito di Paula (MPB), Beth Carvalho (Samba/Pagode), Beto Guedes (MPB), Bezerra da Silva (Samba/Pagode), Bhaskar (Eletrônico), Biel (Pop), Biquini Cavadão (Rock), BK (Rap), Black Alien (Rap), Blitz (Rock), Bokaloka (Samba/Pagode), Bom Gosto (Samba/Pagode), Bonde da Stronda (Rap), Bonde do Tigrão (Funk), Bonde R300 (Funk), Boogarins (Rock), BRAZA (Reggae), Breno & Caio Cesar (Sertanejo), Bruna Karla (Gospel/Religioso), Bruninho e Davi (Sertanejo), Bruno & Barretto (Sertanejo), Bruno & Marrone (Sertanejo), Bruno Be (Eletrônico), Bruno Martini (Eletrônico), Cacife Clandestino (Rap), Caetano Veloso (MPB), Calcinha Preta (Forró), Capital Inicial (Rock), Carlinhos Brown (MPB), Cartola (Samba/Pagode), Cássia Eller (Rock), Cat Dealers (Eletrônico), Cazuza (Rock), César Menotti e Fabiano (Sertanejo), Céu (MPB), Charlie Brown Jr. (Rock), Chiclete com Banana (Axé), Chico Buarque (MPB), Chico César (MPB), Chico Rey e Paraná (Sertanejo), Chico Science (Rock), Chimarruts (Reggae), Chitãozinho & Xororó (Sertanejo), Chrystian e Ralf (Sertanejo), Cidade Negra (Reggae), Clara Nunes (Samba/Pagode), Clarice Falcão (Pop), Class A (Rap), Claudia Leitte (Pop), Claudinho e Buchecha (Funk), Cleber & Cauan (Sertanejo), ConeCrewDiretoria (Rap), Conrado & Aleksandro (Sertanejo), Costa Gold (Rap), CPM 22 (Rock), Criolo (Rap), Cristiano Araújo (Sertanejo), Cristina Mel (Gospel/Religioso), Dadá Boladão (Funk), Dalsin (Rap), Dani Russo (Funk),

Daniel (Sertanejo), Daniela Araújo (Gospel/Religioso), Daniela Mercury (Axé), Davi Sacer (Gospel/Religioso), David Quinlan (Gospel/Religioso), Day e Lara (Sertanejo), Daya Luz (Pop), Delano (Funk), Dennis DJ (Funk), Detonautas Roque Clube (Rock), Diante do Trono (Gospel/Religioso), Diego e Victor Hugo (Sertanejo), Dilsinho (Samba/Pagode), Diogo Nogueira (Samba/Pagode), Dj Batata (Funk), DJ Caique (Rap), DJ Marlboro (Funk), DJ PV (Gospel/Religioso), Djavan (MPB), Djonga (Rap), Don L (Rap), Dona Ivone Lara (Samba/Pagode), Dorgival Dantas (Forró), Dorival Caymmi (MPB), Dream Team do Passinho (Funk), Dudu Nobre (Samba/Pagode), É o Tchan (Axé), Edi Rock (Rap), Edson e Hudson (Sertanejo), Edu Chociay (Sertanejo), Eduardo Costa (Sertanejo), Ego Kill Talent (Rock), Elba Ramalho (MPB), Elekfantz (Eletrônico), Eli Soares (Gospel/Religioso), Eliane Elias (Jazz), Elis Regina (MPB), Elza Soares (MPB), Emicida (Rap), Emilio Santiago (MPB), Engenheiros do Hawaii (Rock), Erasmo Carlos (Rock), Esteban Tavares (Pop), Evokings (Eletrônico), Exaltasamba (Samba/Pagode), Eyshila (Gospel/Religioso), Fábio Jr. (MPB), Fafa de Belem (MPB), Fagner (MPB), Falamansa (Forró), Far From Alaska (Rock), Felguk (Eletrônico), Felipe Araújo (Sertanejo), Felipe Ret (Rap), Felten (Eletrônico), Fernanda Brum (Gospel/Religioso), Fernanda Takai (MPB), Fernandinho (Gospel/Religioso), Fernando & Sorocaba (Sertanejo), Ferrugem (Samba/Pagode), Fiduma & Jeca (Sertanejo), Filipe Labre (Sertanejo), Flora Matos (Rap), Fly (Pop), Forfun (Rock), Fred e Gustavo (Sertanejo), Frejat (Rock), Fresno (Rock), Froid (Rap), Ftampa (Eletrônico), Gabily (Pop), Gabriel Camcam (Rap), Gabriel Diniz (Sertanejo), Gabriel Elias (Reggae), Gabriel Gava (Sertanejo), Gabriel Guedes de Almeida (Gospel/Religioso), Gabriel O Pensador (Rap), Gabriela Rocha (Gospel/Religioso), Gal Costa (MPB), Galinha Pintadinha (Infantil), Geo (Pop), George Henrique & Rodrigo (Sertanejo), Gian e Giovani (Sertanejo), Gilberto Gil (MPB), Gloria Groove (Pop), Gonzaguinha (MPB), Gregory e Gabriel (Sertanejo), Grupo Fundo de Quintal (Samba/Pagode), Grupo Revelação (Samba/Pagode), Guilherme Arantes (MPB), Guilherme de Sá (Gospel/Religioso), Guilherme e Santiago (Sertanejo), Gustavo Mioto (Sertanejo), Gusttavo Lima (Sertanejo), Haikaiss (Rap), Harmonia do Samba (Axé), Heloisa Rosa (Gospel/Religioso), Henrique e Diego (Sertanejo), Henrique e Juliano (Sertanejo), Hugo Del Vecchio (Sertanejo), Humberto & Ronaldo (Sertanejo), Hungria Hip Hop (Rap), Igor Guimarães, Imaginasamba (Samba/Pagode), Inimigos da HP (Samba/Pagode), Ira (Rock), Isadora Pompeo (Gospel/Religioso), Israel e Rodolffo (Sertanejo), Israel Novaes (Sertanejo), Ivan Lins (MPB), Ivete Sangalo (Pop), IZA

(Pop), Jade Baraldo (Pop), Jads & Jadson (Sertanejo), Jammil (Axé), Jão (Pop), Jefferson Moraes (Sertanejo), Jeito Moleque (Samba/Pagode), Jerry Smith (Funk), João Bosco & Vinicius (Sertanejo), João Bosco (MPB), João Brasil (Eletrônico), João Donato (MPB), João Gilberto (MPB), João Neto & Frederico (Sertanejo), João Paulo e Daniel (Sertanejo), Johnny Hooker (MPB), Jonas Esticado (Sertanejo), Jord (Eletrônico), Jorge Aragão (Samba/Pagode), Jorge Ben Jor (MPB), Jorge e Mateus (Sertanejo), Jorge Vercillo (MPB), Jose Augusto (Brega), Jota Quest (Pop), Joyce (MPB), Júlia & Rafaela (Sertanejo), Kamau, Karol Conká (Rap), Katinguelê (Samba/Pagode), Kell Smith (Pop), Kelly Key (Pop), Kid Abelha (Rock), Kleber Lucas (Gospel/Religioso), Kleo Dibah e Rafael (Sertanejo), Knust (Rap), Koringa (Funk), Kyber Krystals (Pop), Laura Souguellis (Gospel/Religioso), Leandro e Leonardo (Sertanejo), Legião Urbana (Rock), Lenine (MPB), Léo Santana (Axé), Leonardo Gonçalves (Gospel/Religioso), Lexa (Pop), Lia Clark (Pop), Liniker (MPB), Little Joy (Rock), liu (Eletrônico), LIVIT (Eletrônico), Los Hermanos (Rock), Loubet (Sertanejo), Luan Estilizado (Sertanejo), Luan Santana (Sertanejo), Lucas Lucco (Sertanejo), Luccas Carlos (Rap), Ludmilla (Pop), Luisa Sonza (Pop), Luiz Bonfá (MPB), Luiz Lins (Rap), Luiz Melodia (MPB), Luiza Possi (Pop), Lulu Santos (Rock), Luma Elpidio (Gospel/Religioso), Maglore (Rock), Mahmundi (Pop), Maiara e Maraisa (Sertanejo), Mallu Magalhães (Pop), Mamonas Assassinas (Rock), Maneva (Reggae), Mano Brown (Rap), Manu Gavassi (Pop), Mar Aberto (Pop), Marcello Gugu (Rap), Marcelo Camelo (MPB), Marcelo CIC (Eletrônico), Marcelo D2 (Rap), Marcelo Jeneci (MPB), Márcia Fellipe (Sertanejo), Marcos & Belutti (Sertanejo), Marcos & Fernando (Sertanejo), Marcos Valle (MPB), Maria Bethânia (MPB), Maria Cecília e Rodolfo (Sertanejo), Maria Gadú (MPB), Maria Rita (MPB), Mariana Aydar (MPB), Mariana Fagundes (Sertanejo), Mariana Nolasco (Pop), Marília Mendonça (Sertanejo), Marina Lima (Rock), Marisa Monte (MPB), Martinho da Vila (Samba/Pagode), Mart'nália (Samba/Pagode), Matanza (Rock), Matheus e Kauan (Sertanejo), Matogrosso e Mathias (Sertanejo), Maykow e Bruno (Sertanejo), Maysa (MPB), MC 2k (Funk), MC Bin Laden (Funk), MC Brinquedo (Funk), MC Brisola (Funk), MC CL (Funk), Mc Davi (Funk), MC Dede (Funk), MC DG (Funk), Mc Don Juan (Funk), MC Dudu (Funk), MC Fioti (Funk), MC G15 (Funk), Mc Gui (Funk), MC Guimê (Funk), MC Gustta (Funk), Mc Gw (Funk), Mc Hariel (Funk), MC Hollywood (Funk), MC Jhey (Funk), Mc Jhowzinho e MC Kadinho (Funk), MC João (Funk), MC Kekel (Funk), MC Kevin (Funk), MC Kevinho (Funk), MC Lan (Funk), MC Léléto (Funk), MC Leozinho (Funk), MC Livinho

(Funk), MC Lustosa (Funk), MC Marcinho (Funk), MC Menor da VG (Funk), MC Mirella (Funk), MC MM (Funk), MC Pedrinho (Funk), MC Phe Cachorrera (Funk), MC Pikachu (Funk), MC Pocahontas (Funk), MC Rahell (Funk), MC Rodolfinho (Funk), MC Th (Funk), MC WM (Funk), Mc Zaac (Funk), Melanina Carioca (MPB), Melim (Pop), Menor (Funk), Michel Teló (Sertanejo), Milionário e José Rico (Sertanejo), Milton Nascimento (MPB), Ministério Zoe (Gospel/Religioso), MOB79 (Rap), Molejo (Samba/Pagode), Mumuzinho (Samba/Pagode), Mundo Livre (Rock), Munhoz e Mariano (Sertanejo), MV Bill (Rap), Nação Zumbi (Rock), Naiara Azevedo (Sertanejo), Naldo Benny (Pop), Nana Caymmi (MPB), Nando Reis (Pop), Nara Leão (MPB), Natiruts (Reggae), Nego do Borel (Pop), Negra Li (Rap), Nelson Freire (Clássico), Nenhum de Nós (Rock), Netinho (Axé), Ney Matogrosso (MPB), Nossa Toca (Pop), Novos Baianos (MPB), Nx Zero (Rock), O Rappa (Rock), O Teatro Mágico (MPB), O Terno (Rock), Oba Oba Samba House (Samba/Pagode), Oficina G3 (Gospel/Religioso), Onze:20 (Reggae), Oriente (Rap), Os Arrais (Gospel/Religioso), Os Cretinos (Funk), Os Mutantes (Rock), Os Paralamas do Sucesso (Rock), Os Travessos (Samba/Pagode), Oswaldo Montenegro (MPB), Otto (MPB), OutroEu (Pop), Pabllo Vittar (Pop), Pablo Martins (Rap), Padre Fábio de Melo (Gospel/Religioso), Padre Marcelo Rossi (Gospel/Religioso), Parangolé (Axé), Pato Fu (Rock), Paula Fernandes (Sertanejo), Paula Mattos (Sertanejo), Paulinho da Viola (Samba/Pagode), Paulinho Moska (MPB), Paulo César Baruk (Gospel/Religioso), Paulo Miklos (MPB), Paulo Ricardo (Rock), Pedra Letícia, Pedro e Benicio (Sertanejo), Pedro Paulo e Alex (Sertanejo), Pentagono (Rap), Perera DJ (Funk), Perícles (Samba/Pagode), Perlla (Funk), Phill Veras (MPB), Pikeno e Menor (Funk), Pitty (Rock), Pixote (Samba/Pagode), Planet Hemp (Rock), Planta e Raíz (Reggae), Plutão Já Foi Planeta (Rock), POLLO (Rap), Ponto de Equílibrio (Reggae), Preta Gil (Pop), Preto no Branco (Gospel/Religioso), Primeiramente (Rap), Priscilla Alcântara (Gospel/Religioso), Projota (Rap), Psirico (Axé), Raça Negra (Samba/Pagode), Racionais (Rap), Rael (Rap), Raimundos (Rock), Rashid (Rap), Raul Seixas (Rock), Ravena (Pop), Reginaldo Rossi (Brega), Reinaldo (Gospel/Religioso), Renato Russo (Rock), RICCI (Eletrônico), Rick e Renner (Sertanejo), Rincon Sapiência (Rap), Rita Lee (Rock), Ritchie (Rock), Roberta Campos (Pop), Roberta Miranda (Sertanejo), Roberta Sá (Samba/Pagode), Roberto Carlos (MPB), Rodolfo Abrantes (Gospel/Religioso), Rodrigo Amarante (MPB), Rodrigo Marim (Sertanejo), Rodriguinho (Samba/Pagode), Rosa de Saron (Gospel/Religioso), Rouge (Pop), Roupa Nova (Pop), RPM (Rock), Rubel (Pop), RZO

(Rap), Sabotage (Rap), Sain (Rap), Sandra de Sá (MPB), Sandy (Pop), Sandy e Junior (Pop), Santti (Eletrônico), Scalene (Rock), Secos e Molhados (MPB), Selva (Eletrônico), Sepultura (Metal), Sérgio Mendes & Brasil'66 (MPB), Sérgio Mendes (MPB), Sergio Reis (Sertanejo), Seu Jorge (MPB), Silva (Pop), Simone e Simaria (Sertanejo), Sinara (Reggae), Skank (Pop), Só Pra Contrariar (Samba/Pagode), Sofia Oliveira (Pop), Solange Almeida (Sertanejo), Sorriso Maroto (Samba/Pagode), Soweto (Samba/Pagode), Strike (Rock), Supercombo (Rock), Tá Na Mente (Samba/Pagode), Tati Zaqui (Funk), Tchakabum (Axé), Teresa Cristina (Samba/Pagode), Thaeme & Thiago (Sertanejo), Thalles, Thalles Roberto (Gospel/Religioso), Thiago Brava (Sertanejo), Thiago Martins (Samba/Pagode), Thiaguinho (Samba/Pagode), Tiago Iorc (Pop), Tiê (Pop), Tihuana (Rock), Tim Maia (MPB), Titãs (Rock), Tom Jobim (MPB), Tom Zé (MPB), Toque no Altar (Gospel/Religioso), Toquinho (MPB), Tribalistas (MPB), Tribo da Periferia (Rap), Tribo de Jah (Reggae), TrintaTrinta (Rap), Trio Parada Dura (Sertanejo), Tropkillaz (Eletrônico), Tulipa Ruiz (MPB), Turma do Pagode (Samba/Pagode), Ultraje a Rigor (Rock), UM44K (Pop), Valesca Popozuda (Funk), Vanessa da Mata (MPB), Vanguart (Rock), Victor e Léo (Sertanejo), Villa Baggage (Sertanejo), Vingadora (Axé), Vinicius de Moraes (MPB), VINNE (Eletrônico), Vintage Culture (Eletrônico), Vitor Kley (Pop), Wanessa (Pop), Wesley Safadão (Forró), Wilson Simonal (MPB), Xamã (Rap), Xande de Pilares (samba/Pagode), Xuxa (Infantil), Zé Felipe (Sertanejo), Zé Neto & Cristiano (Sertanejo), Zé Ramalho (MPB), Zeca Baleiro (MPB), Zeca Pagodinho (Samba/Pagode), Zeeba (Pop), Zélia Duncan (MPB), Zerky (Eletrônico), Zezé Di Camargo e Luciano (Sertanejo), Zimbra (Rock), Zizi Possi (MPB)

ANEXO B – TRECHO DE CÓDIGO PARA OBTENÇÃO DE TWEETS

```
# Limpa o ambiente
rm(list=ls())

# Carrega o pacote utilizado
library("rtweet")
library("stringr")
library("dplyr")
library("tidytext")
library("tm")
library("stringi")

# Autentica no Twitter conforme dados fornecidos na plataforma Twitter for Developers
twitter_token <- create_token(
  app = "XXXXXXXXXXXXXXXXX",
  consumer_key = "XXXXXXXXXXXXXXXXX",
  consumer_secret = "XXXXXXXXXXXXXXXXXX",
  access_token = "XXXXXXXXXXXXXXXXX",
  access_secret = "XXXXXXXXXXXXXXXXX")

# Carrega dataset inicial de artistas (o arquivo deve conter 1 coluna e a ultima linha em branco)
artistas <- read.csv("C:/Users/I860982/Desktop/artistas.csv", encoding="UTF-8")
names(artistas)[1] = "nome_artista"

count = 1
count_tweets = 0

for(i in 1:nrow(artistas)){ # Percorre o dataset de artistas
    nome_artista <- tolower(toString(artistas$nome_artista[count])) # Obtem o nome do artista
    hashtag_artista = gsub(" ", "", paste('#', nome_artista), fixed = TRUE) # Monta a varivel hashtag do artista

    # Pausa de 15 minutos a cada vez que o número de requisições disponíveis é menor que 20
    if(select(rate_limit(twitter_token, "search_tweets"), remaining) < 20) { Sys.sleep(930) }

    # Busca tweets que contenham o nome do artista
    tweets_name     <- search_tweets(q = nome_artista, n=2000, lang="pt", include_rts = FALSE, retryonratelimit = FALSE)
    # Busca tweets que contenham a hashtag do artista
    tweets_hashtag  <- search_tweets(q = hashtag_artista, n=2000, lang="pt", include_rts = FALSE, retryonratelimit = FALSE)

    tweets <- rbind(tweets_name, tweets_hashtag) # Unifica ambos os objetos
    count_tweets <- count_tweets + nrow(tweets)
    rm("tweets_name", "tweets_hashtag") # Remove os objetos originais
    tweets <- unique(tweets) # Remove tweets duplicados

    # Remove virgulas para evitar conflitos na geracao do CSV
    tweets$text <- gsub(",", "", tweets$text)

    # Salva o arquivo final
    write.csv(tweets$text, gsub(" ", "", paste('C:/Users/I860982/Desktop/tweets/tweets_',nome_artista,'.csv')), fileEncoding="UTF-8")

    rm("tweets")
    count <- count+1
}
```

ANEXO C – LISTA DE *STOPWORDS*

de, a, o, que, e, do, da, em, um, para, com, não, uma, os, no, se, na, por, mais, as, dos, como, mas, ao, ele, das, à, seu, sua, ou, quando, muito, nos, já, eu, também, só, pelo, pela, até, isso, ela, entre, depois, sem, mesmo, aos, seus, quem, nas, me, esse, eles, você, essa, num, nem, suas, meu, às, minha, numa, pelos, elas, qual, nós, lhe, deles, essas, esses, pelas, este, dele, tu, te, vocês, vos, lhes, meus, minhas, teu, tua, teus, tuas, nosso, nossa, nossos, nossas, dela, delas, esta, estes, estas, aquele, aquela, aqueles, aquelas, isto, aquilo, estou, está, estamos, estão, estive, esteve, estivemos, estiveram, estava, estávamos, estavam, estivera, estivéramos, esteja, estejamos, estejam, estivesse, estivéssemos, estivessem, estiver, estivermos, estiverem, hei, há, havemos, hão, houve, houvemos, houveram, houvera, houvéramos, haja, hajamos, hajam, houvesse, houvéssemos, houvessem, houver, houvermos, houverem, houverei, houverá, houveremos, houverão, houveria, houveríamos, houveriam, sou, somos, são, era, éramos, eram, fui, foi, fomos, foram, fora, fôramos, seja, sejamos, sejam, fosse, fôssemos, fossem, for, formos, forem, serei, será, seremos, serão, seria, seríamos, seriam, tenho, tem, temos, tém, tinha, tínhamos, tinham, tive, teve, tivemos, tiveram, tivera, tivéramos, tenha, tenhamos, tenham, tivesse, tivéssemos, tivessem, tiver, tivermos, tiverem, terei, terá, teremos, terão, teria, teríamos, teriam