Letícia Felix Batista

Classificação da Polaridade de Tweets Relacionados a Artistas da Indústria Musical Brasileira

Niterói - RJ, Brasil

21 de dezembro de 2022

Universidade Federal Fluminense

Letícia Felix Batista

Classificação da Polaridade de Tweets Relacionados a Artistas da Indústria Musical Brasileira

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Douglas Rodrigues Pinto

Niterói - RJ, Brasil

21 de dezembro de 2022

Universidade Federal Fluminense

Letícia Felix Batista

Classificação da Polaridade de Tweets Relacionados a Artistas da Indústria Musical Brasileira

Monografia de Projeto Final de Graduação sob o título *"Classificação da Polaridade de Tweets Relacionados a Artistas da Indústria Musical Brasileira"*, defendida por Letícia Felix Batista e aprovada em 21 de dezembro de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Douglas Rodrigues Pinto
Departamento de Estatística – UFF

Profa. Dra. Karina Yuriko Yaginuma
Departamento de Estatística – UFF

Profa. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Niterói, 21 de dezembro de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

B333c Batista, Letícia Felix
Classificação da Polaridade de Tweets Relacionados a Artistas da Indústria Musical Brasileira / Letícia Felix Batista. - 2022.
57 f.: il.

Orientador: Douglas Rodrigues Pinto.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.

1. Aprendizado de máquina. 2. Mineração de texto. 3. Indústria musical. 4. Twitter (Site de relacionamentos). 5. Produção intelectual. I. Pinto, Douglas Rodrigues, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

Neste trabalho foram exploradas técnicas de classificação de texto com o intuito de realizar o monitoramento da reputação de artistas da indústria musical brasileira através da classificação da polaridade de dados gerados por usuários do Twitter expressando opiniões em relação ao artista de interesse. Foram coletados tweets com menções à artista brasileira Anitta e rotulados manualmente dentro das classificações "negativo" e "positivo", refletindo a opinião do autor do tweet em relação à artista. Utilizando a base de tweets rotulados, foram construídos três tipos de classificadores: o de Gradient Boosting, Regressão Logística e um classificador não supervisionado baseado no Léxico de Polaridades Rotuladas na língua portuguesa"Oplexicon" (SOUZA et al., 2011). Os resultados mostraram que os classificadores baseados em Aprendizado de Máquinas obtiveram melhores desempenhos do que a classificação básica por Léxico Rotulado. As classificações por Léxico Rotulado se mostraram enviesadas a classificar as amostras como "positivas", ficando com especificidade baixa. Técnicas mais complexas de pré-processamento exploradas no trabalho trouxeram benefícios aos classificadores baseados em Léxico, mas não acarretaram melhorias nos classificadores baseados em Aprendizado de Máquinas.

Palavras-chave: Aprendizado de Máquinas. Mineração de Texto. Indústria Musical. Twitter.

Agradecimentos

Agradeço à minha família, que sempre me deu suporte em todos os sentidos e que compartilhou comigo todas as dores e vitórias desta jornada. Obrigada por todo amor, cuidado, conselhos, incentivos e apoio em todos os momentos da minha vida.

À Universidade Federal Fluminense, por ter sido a minha segunda casa durante esses anos de estudo, sendo um ambiente de respeito, cooperação e cultivo de conhecimento que me forneceu, do início ao fim, mesmo passando por uma pandemia, a estrutura necessária para o meu desenvolvimento.

Ao Professor Douglas Rodrigues que, além de ter me orientado neste trabalho com toda paciência e disponibilidade, foi uma inspiração ao fazer de suas aulas um ambiente de desenvolvimento do pensamento crítico e de amadurecimento. Que, com sua dedicação ao ensino de qualidade, nos mostrou como a Estatística poderia ser um assunto fascinante e sempre prezou por ir além de um sistema de ensino engessado, tratando seus alunos como indivíduos e valorizando o tipo de inteligência de cada um.

Às Professoras Karina Yaginuma e Jéssica Kubrusly, por terem aceitado o convite para compor a minha banca e por todas as contribuições feitas não só neste trabalho, mas também na minha formação, durante a qual tive o prazer de conhecer um pouco de suas mentes brilhantes. Ver figuras femininas tão fortes na minha área de formação sempre foi uma grande inspiração para mim.

A todos os professores que contribuíram com meu aprendizado e com a minha evolução como pessoa, sempre me incentivando a ultrapassar meus limites e provando que eu era capaz de ir mais longe. Se hoje eu pulo de peito aberto para novos desafios, é pelos aprendizados valiosos que vocês me proporcionaram.

Aos amigos que fiz na UFF, agradeço pelas trocas maravilhosas do dia a dia e por terem tornado essa jornada bem menos árdua.

E a mim mesma, por todo esforço que fiz para chegar até aqui.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A música é uma forma de arte que inspira e emociona, mas que também é o principal produto de um mercado global e muito rentável. Os artistas, parte mais exposta da indústria, estão constantemente caminhando em uma linha tênue entre a verdadeira expressão de si mesmos e o comportamento esperado pelo público.

Na era das plataformas digitais de música, os ouvintes passaram a ter mais controle sobre o conteúdo que consomem. Junto da voz fornecida à eles pelas redes sociais, este poder de escolha se potencializou em poder de influência sobre o sucesso e o fracasso da carreira de artistas da indústria, que estão a todo momento sujeitos a boicotes com o intuito de prejudicá-los, bem como a multirões organizados para gerar picos de consumo que os beneficiem nas principais paradas de música.

A opinião em massa do público sobre um artista muitas vezes guarda a explicação de tendências de consumo que não podem ser explicadas por outros fatores. Essas opiniões estão presente de maneira pública e acessível em redes sociais como o Twitter, e podem ser analisadas com o uso de recursos de mineração de textos, análise de sentimento e classificação textual.

Nesse contexto, o objetivo deste trabalho é explorar técnicas de classificação de texto e construir um algoritmo capaz de classificar automaticamente grandes quantidades de tweets sobre uma entidade de interesse. Para isso, foi criada uma base de dados à partir da coleta e rotulação manual de tweets relacionados à artista brasileira Anitta, escolhida como artista alvo neste estudo devido à vasta quantidade de conteúdo gerado por usuários da plataforma mencionando-a diariamente.

Todas as etapas de código e as bases de dados desenvolvidas neste trabalho estão disponíveis no Github[1], a fim de atender o modelo conceitual de reprodutibilidade (MONDELLI; PETERSON; GADELHA, 2019).

[1] https://github.com/leticiafelix/classificador-polaridade-tweets

1.1 Motivação

1.1.1 A Música Como Indústria

A música sempre teve um papel importante na história. Na posição de uma das mais apreciadas formas de arte, ela representa parte crucial da cultura e identidade de um povo, emocionando e embalando os mais importantes momentos da vida das pessoas.

Com os avanços tecnológicos e a chegada das plataformas de *streaming* de música, o acesso instantâneo à milhares de músicas chegou à palma das mãos de qualquer indivíduo que possua um dispositivo com acesso à internet, tornando a música um fator ainda mais presente em diversos momentos do dia a dia.

Entre os inúmeros papéis que assume na sociedade, hoje, a música também é protagonista de uma indústria que cresce globalmente e movimenta bilhões de dólares por ano.

Segundo o relatório global da indústria fonográfica publicado em 2022 pela IFPI (Federação Internacional da Indústria Fonográfica) (IFPI, 2022), associação que representa a indústria fonográfica internacionalmente, as receitas geradas pela indústria fonográfica no ano de 2021 alcançaram US$25,9 bilhões, sendo 65% deste valor (US$16,9 bilhões) proveniente do consumo de música em plataformas de *streaming* como o Spotify, Deezer, Apple Music, Amazon Music, entre outras.

A Figura 1 mostra a evolução da receita gerada pela indústria entre 1999 e 2021 segmentada por fonte de receita. No gráfico é possível notar a brusca mudança no mercado no decorrer dos anos, quando o *streaming* passou a substituir as mídias físicas no posto de maior gerador de receita.

No Brasil, 11º país em consumo no mercado global em 2021 (IFPI, 2022), a receita gerada pela indústria fonográfica no ano foi de R$2,1 bilhões, sendo R$1,8 bilhões (85,6%) vindos do consumo em plataformas de *streaming*.

A partir destes dados, é notório que, apesar da indústria fonográfica ser baseada em uma forma de arte que emociona, sensibiliza e é o cerne do sonho de artistas e fãs, esta, assim como os demais segmentos de negócios, também possui seu lado estratégico e mercadológico que lida com a característica de "produto" embutida nas carreiras e músicas, tratando-as como parte de um negócio rentável que precisa ser monitorado e gerenciado através de estratégias de *marketing*, vendas e comunicação, para assim galgar espaço em um mercado cada vez mais competitivo e se adaptar às mudanças que vêm

com os avanços tecnológicos.

"Na visão mercadológica, os artistas podem ser vistos como produto, os quais envolvem um mercado, no qual a venda depende em muito da imagem e da constante divulgação do mesmo, que, no caso é a música." (JORGETTO, 2009)

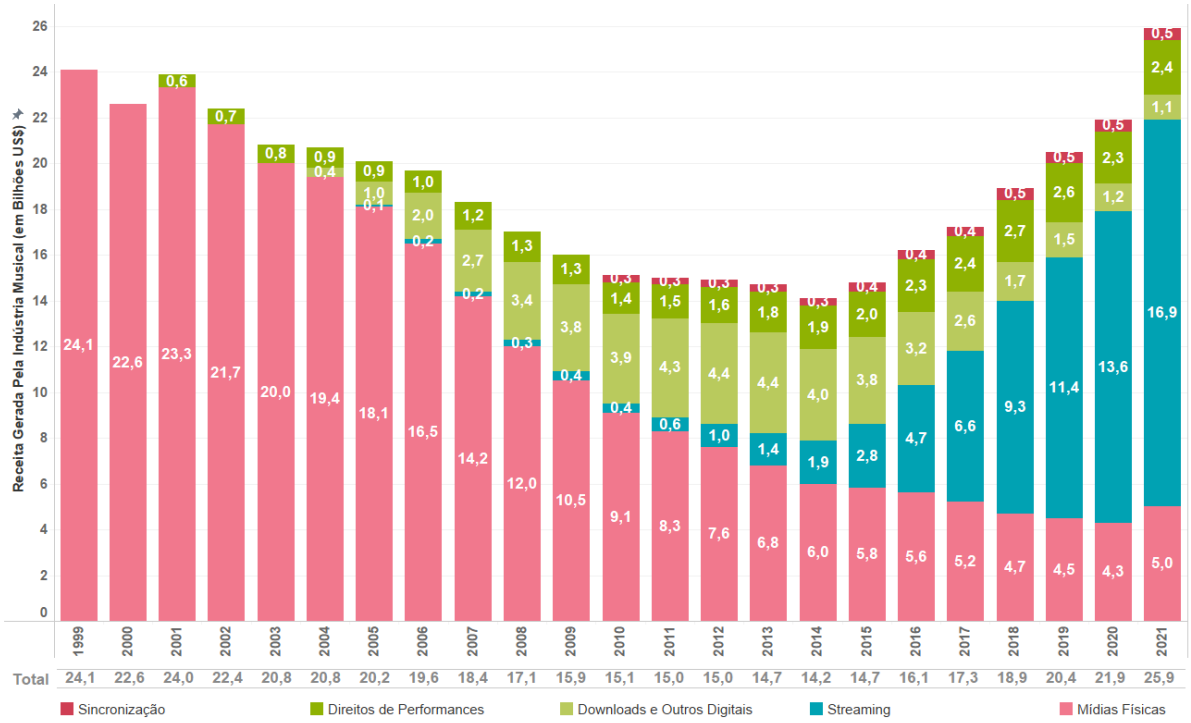

Figura 1: Evolução da receita gerada pela Indústria Fonográfica por fonte de consumo de 1999 a 2021. (IFPI, 2022)

1.1.2 A Dualidade: Dádiva e Produto

Neste sentido, a indústria musical vive uma dualidade. Os atributos necessários para lidar com seres humanos e suas características e valores individuais costuma ser diretamente oposto à maneira essencialmente fria e estratégica com que se lida com produtos constituídos por objetos. Lidar com ambos estes aspectos na indústria da música envolve desafios.

Sendo característica intrínseca do mercado da música, essa dualidade se estende para todos os níveis da indústria. As plataformas de *streaming*, gravadoras, artistas, produtores de eventos e todas as partes envolvidas em processos neste universo precisam lidar, em alguma proporção, com essas duas facetas tão opostas da indústria. Não há dúvidas,

entretanto, de que o artista, que é o próprio produto, é quem vive essa dualidade em seu nível mais profundo.

Nos dias atuais, um artista da indústria fonográfica precisa, mais do que nunca, ter a habilidade de equilibrar seu lado humano com seu lado produto, caminhando assim em uma linha tênue entre a mais verdadeira expressão de sua personalidade e o comportamento esperado pelo público.

> "O *marketing* pessoal artístico em muito se equivale ao político. Ambos devem preservar a boa imagem, o interesse e a preferência de seus públicos, garantindo para si uma atenção especial. Ter consciência do poder da opinião pública, saber agir de forma a convencê-la e, principalmente, considerar o poder do *marketing* na construção de imagens perante à opinião pública são fatores de grande importância para um artista."
> (JORGETTO, 2009)

Ser apreciado aos olhos do público é um fator fundamental para o sucesso da carreira de um artista. É o público quem dita o que é relevante no mercado através de sua aceitação e consumo orgânico. Ter uma imagem que provoca o sentimento de identificação e admiração do público e que fomenta nas pessoas o desejo de associação também é um aspecto importante por diversas razões. Dentre elas:

- O bom relacionamento com outros artistas da indústria, com quem um artista pode trocar apoio e realizar colaborações que rendem visibilidade e alcance a novos públicos;
- O apoio das plataformas de *streaming*, que, na posição de detentoras do principal meio pelo qual a música é distribuída ao público atualmente, possuem o poder de impulsionar ou suprimir a visibilidade do conteúdo disponibilizado nelas;
- A visibilidade em programas de televisão, rádios, e eventos que cultivam a relevância do artista e servem como meios de divulgação de seu trabalho;
- O apoio das grandes empresas, que se associam à imagem de artistas através do patrocínio de eventos e campanhas publicitárias.

Em resumo, ao mesmo tempo em que o artista é um indivíduo com emoções e opiniões próprias, existe uma relação de dependência entre seus interesses e os interesses dos demais segmentos relacionados à indústria em seu produto. Sendo assim, muito da carreira de

um artista depende do sucesso deste equilíbrio e da preservação da sua imagem como um bem ao qual exista o desejo de associação, seja por parte de fãs, artistas, plataformas ou empresas.

> "Mediante a adequação do discurso, a proximidade e a relação com seus públicos e também da manutenção da imagem artística, um profissional da carreira artista consegue prestígio não somente para si, como fornece também uma estrutura positiva no relacionamento geral entre públicos, mídias e financiadores." (JORGETTO, 2009)

1.1.3 A Forma de Consumir Música

Assim como todas as indústrias, a fonográfica também se adapta e se renova no ritmo dos avanços tecnológicos. Há alguns anos, as formas mais comuns de se ouvir música eram através das rádios, programas de TV e mídias físicas, como fita cassete, CDs e DVDs.

Atualmente, com a chegada das plataformas de *streaming*, além do acesso ao conteúdo musical ter sido extremamente simplificado, podendo ser feito através de, praticamente, qualquer dispositivo com acesso à internet, o usuário passou a ter mais liberdade para escolher o que vai ouvir, dado que, com uma assinatura paga em uma plataforma de *streaming*, ele tem acesso à milhares de músicas catalogadas nos aplicativos.

Desta forma, é descartada a necessidade de compra de mídias físicas ou de *downloads* para que se deixe de ouvir música de forma passiva, como nas rádios e outros meios em que o usuário não tem o poder de escolha, e se passe a consumir música à partir de uma busca ativa por conteúdo de seu interesse.

> "A popularização das plataformas de *streaming* mudou totalmente nossa forma de consumir música. O ouvinte não precisa mais escolher o disco que vai comprar e ouvir. Numa plataforma, ele tem todos (ou quase todos) à disposição, numa prateleira virtual que irmana o artista iniciante e o consagrado. Isso é inovador e democrático." - Mauro Ferreira, crítico de música do G1.

Outro ponto importante de se abordar é o fato de que, mesmo sem uma assinatura paga, as pessoas têm o hábito de ouvir música em plataformas de *streaming*, apesar de terem menos poder de escolha do que os usuários pagantes. Como pode ser observado na Figura 2, que mostra que 17,7% de toda a receita gerada pela indústria fonográfica em 2021 veio do consumo de usuários com contas gratuitas em plataformas de *streaming*. Diferente

dos usuários que pagam um valor mensal para usufruir de todas as funcionalidades das plataformas, os usuários com contas gratuitas geram a receita chamada *"Ad-Supported"* ao ouvirem anúncios.

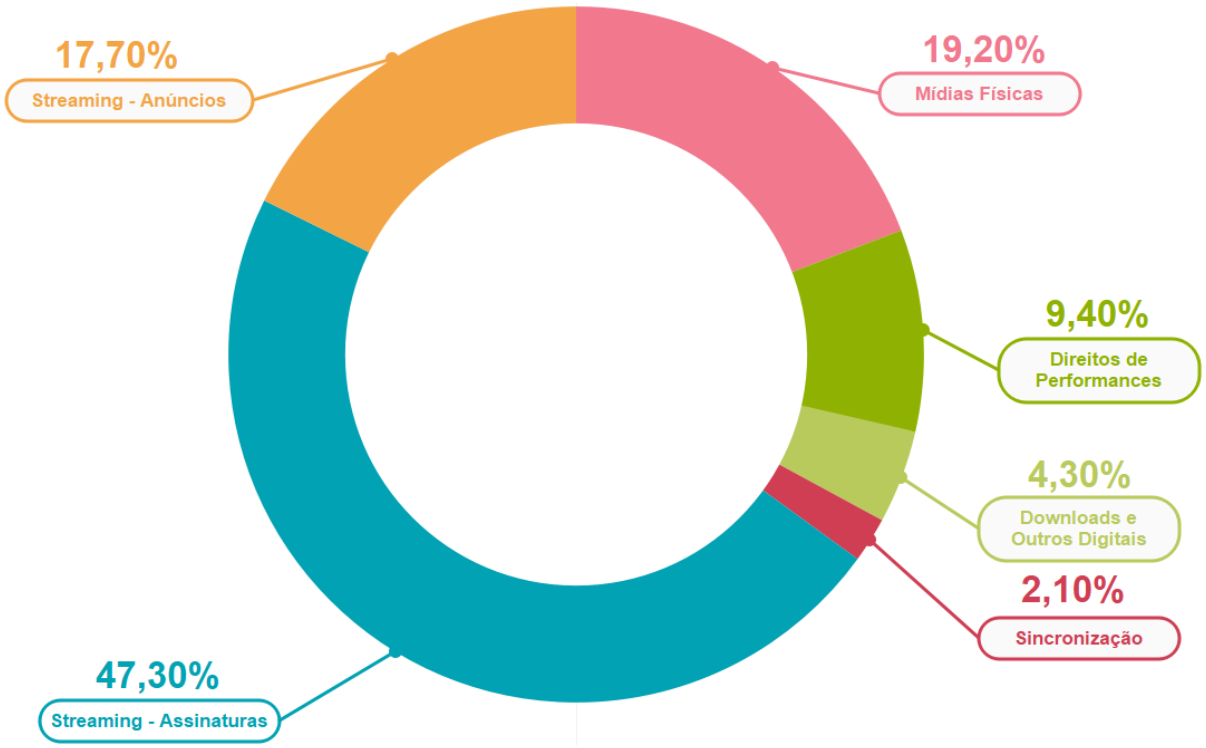

Figura 2: Proporção das fontes de consumo na receita gerada pela Indústria Fonográfica em 2021. (IFPI, 2022)

Com uma assinatura gratuita, o usuário não pode selecionar faixas específicas que deseja ouvir. Ele fica dependente de consumir *playlists* curadas pelo próprio aplicativo ou por outros usuários, o que concede poder de influência, também, às plataformas sobre o que é consumido dentro delas, bem como à alguns usuários sobre os outros.

A receita gerada pelas assinaturas pagas em plataformas de *streaming* é ainda maior que a receita *"Ad-Supported"*. Ela corresponde à 47,3% de toda receita gerada pela indústria fonográfica global em 2021 (Relatório Global IFPI 2022). Com uma assinatura paga nas plataformas, o usuário possui total liberdade de direcionamento do seu próprio consumo. Esta liberdade utilizada conjuntamente por milhões de pessoas confere ao público um grande poder de movimentar o mercado e ditar tendências de consumo.

Um exemplo da força deste movimento ocorreu em março de 2022, com a faixa *"Envolver"* da artista brasileira Anitta[2], que, ao viralizar nas redes sociais e atingir o Top 10 Global do Spotify, recebeu apoio de grande parte da população brasileira, que se mobili-

[2]https://g1.globo.com/pop-arte/musica/noticia/2022/04/14/envolver-de-anitta-e-hit-global-mas-1o-lugar-teve-impulso-brasileiro-veja-graficos.ghtml

zou para intensificar ainda mais o crescimento da faixa e impulsioná-la a posições ainda maiores nas paradas globais. Este esforço resultou na permanência da faixa na posição de mais escutada mundialmente no aplicativo durante 3 dias.

O mesmo poder que uma base de usuários unida tem de impulsionar uma faixa à posição #1 global, também pode impactar negativamente um artista em caso de boicote. Em 2021, o rapper brasileiro Projota viu sua carreira ser fortemente impactada após uma participação no reality show BBB (*Big Brother Brasil*) que desagradou o público. O artista foi eliminado do programa em março de 2021 com 91,89% dos votos, e, em janeiro de 2022, afirmou em entrevista ao Jornal O Globo[3]: "Perdi amizades, fãs, contratos e prestígio, perdi boa parte do alicerce que levei 20 anos para construir.".

Exemplos como estes evidenciam a horizontalização na relação consumidor x indústria fonográfica que vem se intensificando com o crescimento das plataformas de *streaming* de música nos últimos anos. Com o poder de escolha do público em relação ao seu próprio consumo, este deixa de ser receptor e se torna protagonista na criação de tendências de consumo no mercado.

Agora, o público que consome música através de assinaturas pagas nas plataformas de *streaming*, responsáveis por quase metade de toda a receita gerada mundialmente pelo mercado fonográfico, é o principal mobilizador do mercado, escolhendo o que deseja consumir e impulsionar na cauda longa do catálogo musical disponível nas plataformas de *streaming*. Isso exige da indústria a capacidade de observar e se adaptar às tendências criadas pelo público.

1.1.4 O Poder do Público na Web 2.0

Além de ter mudado a forma como se consome música, a evolução da *Internet* também aumentou o poder de influência dos indivíduos uns sobre os outros e a polarização de opiniões em grandes grupos a partir das redes sociais.

"Web 2.0" é o nome dado ao fenômeno da mudança na forma como os usuários navegam a *Internet* após o aumento da velocidade e da facilidade do uso de aplicativos *online*. Com recursos suficientes, a navegação deixou de ser estática e passiva e se tornou dinâmica, possibilitando a rápida troca de informações entre usuários e a geração de conteúdo por eles.

[3]https://oglobo.globo.com/cultura/musica/projota-ja-perdi-muito-na-vida-mas-ali-veio-tudo-deuma-vez-25351460

"Na década passada, se tornou comum olhar a *Internet* como mais que uma fonte de informação, mas um lugar onde pessoas comuns podem contribuir com conteúdo através de *blogs*, avaliações de usuários, ou outras postagens públicas." (BLANK; REISDORF, 2012)

Nos dias atuais, as pessoas compartilham conteúdo 24 horas por dia em diferentes redes sociais, emitindo suas opiniões, compartilhando fotos, vídeos, e registrando os acontecimentos de suas vidas em tempo real. Nunca antes se teve acesso a tanto conteúdo gerado por usuários, e isto é uma fonte preciosa de informações que abre espaço para acompanhar o que está se passando na cabeça do público em relação a qualquer tópico de interesse.

Diversos setores já têm se atentado à quantidade de informação útil presente no oceano de conteúdo gerado por usuários da *Internet* e gerado metodologias para extrair o melhor dela. Uma dessas metodologias é a Análise de Sentimentos, uma área de estudos desenvolvida com a finalidade de identificar o sentimento latente, ou a polaridade positiva ou negativa, de textos provenientes de conteúdo postado por usuários em *sites*, *blogs*, avaliações de produtos, entre outros.

Outra consequência da *Web 2.0* foi a possibilidade do público expressar suas indignações e ser ouvido pelas marcas, sabendo que pode afetá-las financeiramente ao organizar boicotes e causar um déficit de consumo em massa. Com uso das redes sociais, os internautas conseguem facilmente influenciar as marcas a romperem contratos com pessoas públicas envolvidas em polêmicas, sob ameaça de boicote.

Isso ocorreu com a rapper brasileira Karol Conká, que viu diversas marcas anunciarem publicamente quebras de contrato e cancelamento de sua participação em eventos devido à rejeição que enfrentou do público durante a sua participação no BBB (*Big Brother Brasil*) em 2021[4] . Apesar disso, poucos meses após o ocorrido, a rapper conseguiu reconquistar o apoio do público através da reconstrução de sua imagem nas redes sociais[5].

Do mesmo modo que o público exige posicionamento das marcas sobre questões importantes e, quando sente que a marca fere valores cruciais para eles, deixa de consumi-la e organiza boicotes, ele também faz isso com artistas e celebridades. O boicote do público acontece de forma ainda mais sensível na indústria musical devido à competitividade da indústria e ao poder de escolha do usuário sobre seu próprio consumo, citado anterior-

[4]https://g1.globo.com/pop-arte/musica/noticia/2021/02/02/festival-rec-beat-suspende-exibicao-departicipacao-de-karol-conka-atitudes-vao-contra-principios-basilares.ghtml

[5]https://oglobo.globo.com/ela/gente/de-volta-cena-karol-conka-conta-como-se-reconstruiu-comajuda-de-terapia-queriam-que-eu-desistisse-25399200

mente.

Quando um usuário escolhe deixar de consumir um artista nas plataformas de *streaming*, há milhares de opções similares para substituir o que ele costumava consumir. Enquanto para outros setores, como o de *delivery* (aplicativos de entrega de comida), por exemplo, há poucas opções de substituição, fazendo com que, mesmo em casos em que o público repudia a marca, precise continuar usufruindo de seus produtos e serviços.

1.1.5 Monitoramento de Reputação

Diante da sensibilidade do consumo no mercado musical atrelado à reputação dos artistas e no contexto da *Web 2.0*, com abundância de conteúdo gerado por usuários expressando seus pensamentos em tempo real, o cenário atual se configura ideal para realizar o monitoramento da reputação de artistas da indústria musical.

Realizar este monitoramento, no entanto, exige um tipo complexo de análise devido à grande quantidade de conteúdo existente nas redes e ao fato do sentimento do público, no geral, estar oculto neste vasto conjunto de conteúdo.

Assim, este trabalho tem como objetivo desenvolver métodos automáticos de realizar a classificação da polaridade negativa ou positiva das opiniões do público sobre um artista da indústria musical brasileira, explorando técnicas de análise de texto e de desenvolvimento de algoritmos de classificação, a fim de tornar possível a classificação em massa de tweets sobre um artista, o que permitirá o uso destas informações nos mais variados contextos de monitoramento de reputação.

1.2 Objetivos

O objetivo principal deste trabalho é construir um algoritmo capaz de classificar a polaridade da opinião expressa por usuários do Twitter em relação a um artista de interesse, no contexto da indústria musical brasileira. O objetivo secundário é estudar e desenvolver técnicas de classificação de polaridade de textos provenientes do Twitter, explorando e contribuindo com os materiais disponíveis para este fim na língua portuguesa.

1.3 Organização

O Capítulo 2 apresenta os Materiais e Métodos explorados no trabalho, ele é dividido entre as Seções 2.1, onde são descritos os Materiais utilizados para a criação do classificador, como a base de dados montada e o Léxico Rotulado. Na Seção 2.2 de Métodos, são apresentados os algoritmos de classificação e as etapas de pré-processamento exploradas. No Capítulo 3 está todo o desenvolvimento dos classificadores, bem como a avaliação de suas métricas de qualidade. O Capítulo 4 contém as conclusões obtidas com este trabalho.

2 Materiais e Métodos

2.1 Materiais

2.1.1 Twitter

O Twitter é uma das redes sociais mais utilizadas para se expressar opiniões atualmente. Diferente de outras redes sociais onde predominam as publicações de fotos e vídeos, o Twitter é focado na publicação de textos curtos, de até 280 caracteres, e tem toda a sua interface pensada para estimular os usuários a publicarem suas ideias a qualquer momento do dia.

Lançada em 2006, a plataforma segue mantendo sua relevância e crescendo em número de usuários, tendo acumulado, atualmente, cerca de 217 milhões de usuários diários.

> "No terceiro trimestre de 2021, o número do Twitter de usuários ativos diários monetizáveis (mDAUs) atingiram 211 milhões. O site de rede social e microblog teve 187 milhões de mDAUs no terceiro trimestre de 2020, o que representa um aumento de 12.8%. 500 milhões de tweets são publicados a cada dia. 350,000 tweets são postados a cada minuto."[1]

Sendo depósito de milhões de frases postadas por usuários expressando seus pensamentos diariamente, o Twitter é uma fonte preciosa de informações sobre qualquer assunto de interesse e, por isto, foi utilizado como fonte de dados neste trabalho.

A plataforma possui, ainda, uma API (*Application Programing Interface*) onde disponibiliza seus dados de maneira facilitada para usuários que possuem permissão de desenvolvedor. Esta permissão pode ser solicitada por qualquer pessoa com interesse em realizar estudos com dados da rede social.

[1] https://www.websiterating.com/pt/research/twitter-statistics

2.1.2 Banco de Dados

2.1.2.1 Escolha do Artista

A artista escolhida como alvo nesta análise foi a brasileira Anitta, dada sua relevância no cenário fonográfico brasileiro atual e forte presença nas redes sociais. A artista tem cerca de 26 milhões de ouvintes mensais na maior plataforma de *streaming* de música da atualidade, o *Spotify*.

No Twitter, a artista possui cerca de 18 milhões de seguidores e 53 mil tweets, o que evidencia sua constante atividade na rede social. Além disto, à partir de seus tweets é possível observar que sua atividade nesta rede não se limita à interações imparciais com fãs e à divulgação de seu trabalho, mas apresenta um histórico conhecido pela abordagem de assuntos polêmicos como política e causas sociais, o que a faz centro de constantes discussões e alvo de milhares de comentários dos usuários da plataforma.

Constantemente citada nos portais de notícias, a artista atrai um grande volume diário de tweets emitindo opiniões sobre os assuntos em que está inserida, o que fornece uma abundante base de exemplos para serem extraídos e analisados neste tipo de estudo.

A Figura 3 mostra exemplos do posicionamento da artista em assuntos polêmicos no Twitter.

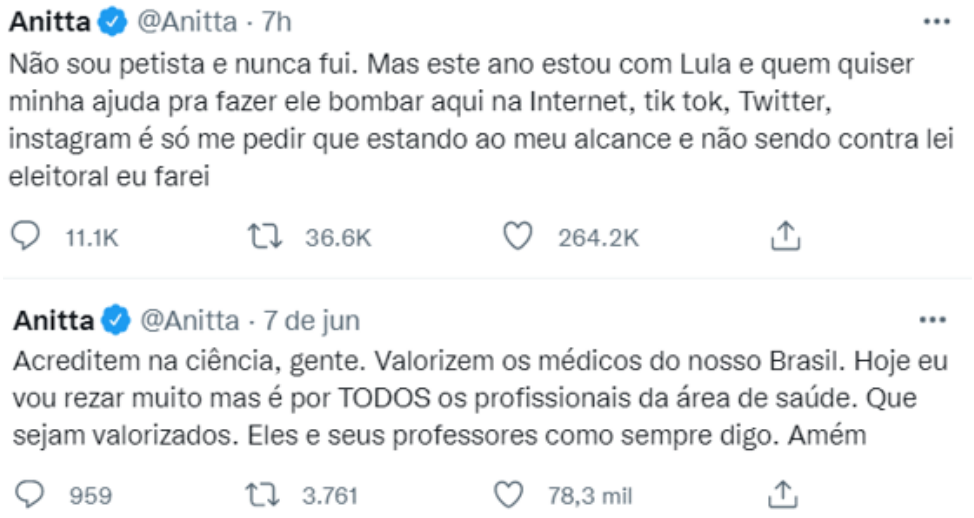

Figura 3: Exemplo de tweet da artista Anitta.

2.1.2.2 Criação do Banco de Dados

O banco de dados utilizado para a criação do classificador foi gerado à partir de tweets extraídos com o uso da API do Twitter. A interação com a API foi feita por um *script* na linguagem de programação R (R, 2009), com o uso do pacote "rtweet" (KEARNEY, 2022).

Foram extraídos tweets que continham a palavra "Anitta", descartando *retweets* (compartilhamento de um tweet de outro usuário), *replies* (tweet em resposta a outro tweet), tweets contendo links ou mídias e tweets feitos por contas verificadas, a fim de captar tweets que representassem a opinião primária do usuário, sem a interferência de contextos externos.

A extração dos tweets foi feita mensalmente no decorrer do ano de 2022 e, durante o período de coleta, foram captados tweets com a opinião dos usuários acerca de diversos tópicos nos quais a artista se envolveu, como: premiações internacionais, lançamento de álbum e produtos, eleições presidenciais, participações em eventos, conflitos com outros artistas da indústria, entre outros.

Os tweets extraídos foram rotulados manualmente dentro das categorias "positivo" ou "negativo", refletindo a polaridade da fala do usuário em relação à artista analisada.

Parte da base de tweets utilizados foi cedida pelos alunos da turma de Aprendizado de Máquinas do período 2022.2 da UFF, ministrada pelo Professor Douglas Rodrigues, orientador deste trabalho.

Alguns exemplos de tweets rotulados como positivos ou negativos podem ser vistos nas Figuras 4 e 5.

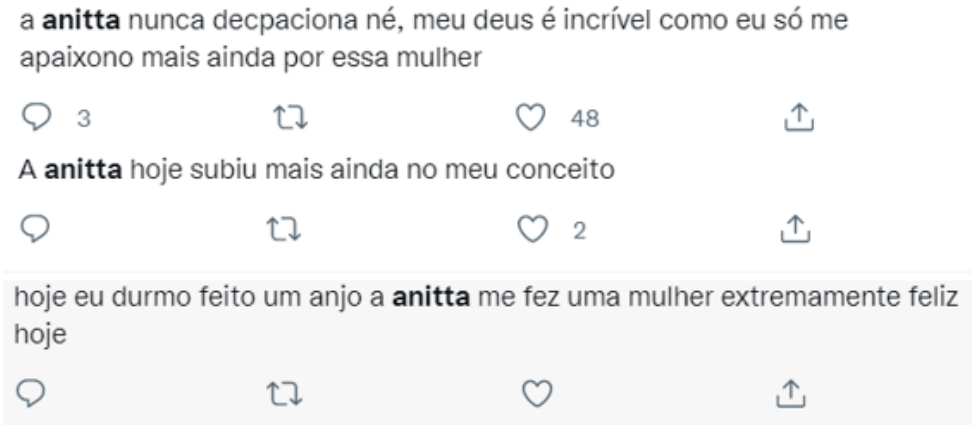

Figura 4: Exemplo de tweets rotulados como "positivos".

Figura 5: Exemplo de tweets rotulados como “negativos”.

Ao todo, foram rotulados 5320 tweets, dos quais 2180 receberam rótulo “negativo” e 3140 receberam rótulo “positivo”, como mostra a Figura 6.

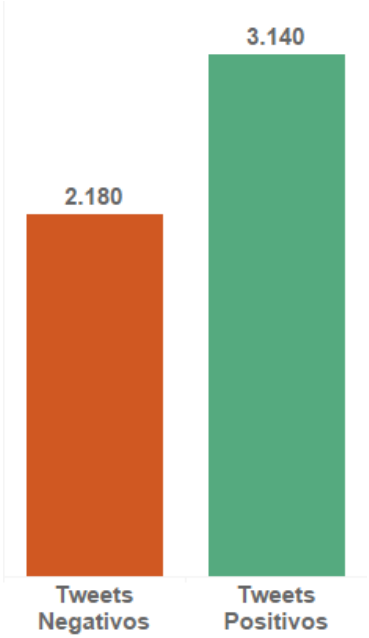

Figura 6: Quantidade de tweets positivos e negativos na amostra.

2.1.3 Matriz Termo-Documento

Uma matriz termo-documento consiste em uma estrutura que faz o agrupamento dos termos observados em todos os elementos que compõem o *corpus* de documentos, neste caso, o conjunto de *tweets* rotulados, deixando-os no formato necessário para serem utilizados em um algoritmo de treinamento de um modelo de Aprendizado de Máquinas.

Sua estrutura é a de uma matriz $[i, j]$, onde cada coluna j corresponde a um termo, cada linha i corresponde a um *tweet*, e o conteúdo de $[i, j]$ é preenchido com a frequência do termo j no documento i.

Neste trabalho, a frequência utilizada foi a proporção do termo no total de termos do tweet, seguindo a Fórmula 2.1.

$$\frac{f_{j,i}}{t_i} \tag{2.1}$$

Onde, $f_{j,i}$ = número de vezes que o termo j aparece no documento i e t_i = total de termos contidos no documento i.

Desta maneira, a matriz fica estruturada conforme a Tabela 1.

tweet	n	representar	vc	vivar	viver	diferente	fazer	tomar	amar
1	0,13	0,25	0,13	0,13	0,25	0,00	0,00	0,00	0,00
2	0,00	0,00	0,00	0,00	0,00	0,40	0,20	0,20	0,00
3	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,33
4	0,00	0,00	0,00	0,00	0,00	0,00	0,11	0,00	0,00
10	0,00	0,00	0,09	0,00	0,00	0,00	0,00	0,00	0,09
12	0,00	0,00	0,00	0,00	0,00	0,00	0,11	0,00	0,00
13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,33	0,00
24	0,00	0,00	0,25	0,00	0,00	0,00	0,00	0,00	0,00
27	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,25
29	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,25
30	0,00	0,00	0,00	0,00	0,00	0,00	0,25	0,00	0,00
32	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,17
33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,33
34	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,50
54	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,00	0,00
60	0,00	0,00	0,00	0,00	0,00	0,00	0,20	0,00	0,00
61	0,00	0,00	0,00	0,00	0,00	0,00	0,33	0,00	0,00
62	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,25
70	0,00	0,00	0,00	0,00	0,00	0,00	0,13	0,00	0,00

Tabela 1: Estrutura de uma matriz termo-documento.

Para gerar a matriz termo documento, é preciso selecionar os termos com relevância entre os tweets observados. Para isso, é necessário primeiro fazer a *tokenização* dos documentos, neste caso, a separação individual de cada termo que compõe o tweet. O caractere de espaço foi usado como separador.

Com os termos separados, foram realizadas as etapas de padronização descritas na Seção 2.2.1.

Desta maneira, os termos foram agrupados e tiveram sua frequência no *corpus* contabilizada.

Com o objetivo de converter estes termos em uma matriz termo-documento, onde cada termo se torna uma variável preditora no modelo de Aprendizado de Máquinas, foi necessário reduzir a quantidade de termos para preservar apenas os mais importantes no

conjunto de documentos.

Além das reduções realizadas pelas etapas de pré-processamento, outro critério usado para redução da quantidade de termos foi o da seleção dos termos com frequência de aparição de pelo menos 16 vezes entre os documentos. Segundo (FELDMAN; SANGER, 2007), "Evidencias experimentais sugerem que usar apenas o top 10% das palavras mais frequentes não reduz a performance dos classificadores.". Dado que o total de termos contidos no conjunto de tweets foi de 6047, foi escolhida a frequência mínima de 16, que manteve os 543 termos mais frequentes entre os tweets, uma quantidade bastante próxima do top 10%.

A Tabela 2 mostra algumas das palavras que mais apareceram no *corpus*.

Palavra	Frequência
ir	677
pra	657
fazer	567
ta	561
musicar	513
musico	512
falar	449
ver	427
q	384
amar	345
poder	343
to	288
saber	280
querer	269
todo	263
gente	262
tao	246
achar	245
ficar	237
mulher	235

Tabela 2: Palavras com maior frequência no corpus.

Com os termos padronizados, reduzidos e com frequência calculada, com algumas linhas de código foi feita a pivotização da tabela para obter o formato desejado, conforme o exemplo mostrado anteriormente na Tabela 1.

2.1.4 Léxico Rotulado

Um léxico rotulado consiste em um dicionário no qual as palavras possuem rótulos que refletem a sua polaridade, indicando se seu significado expressa um sentimento negativo, positivo ou, em alguns casos, neutro.

O léxico pode ser desenvolvido de diferentes maneiras. As palavras contidas nele podem ser associadas à uma classificação binária representando os rótulos "positivo" e "negativo". Ou podem, em vez disso, ser associadas a um *score* que represente seu nível de "positividade" ou "negatividade", de acordo com um peso equivalente à intensidade de seu significado.

Por exemplo, um dicionário com palavras classificadas com polaridades dentro do intervalo de -1 a 1, sendo -1 o mais negativo e 1 o mais positivo, pode atribuir um *score* -1 à palavra "péssimo", e -0,7 à palavra "ruim", dado que a palavra "péssimo" é "mais negativa" do que a palavra "ruim".

Outro critério que pode ser considerado ao atribuir *scores* às palavras de um dicionário é a classe gramatical da palavra. Um léxico pode dar mais peso às palavras cujas classes possuem maior importância na definição da polaridade de uma frase, como os adjetivos. Desta forma, as palavras pertencentes às demais classes gramaticais têm seu impacto reduzido na definição do rótulo do texto.

Existem diversos léxicos rotulados já concebidos na língua inglesa, devido à vasta gama de estudos nesta área no exterior. Neste estudo, foi explorado o léxico rotulado na língua portuguesa "Oplexicon" (SOUZA et al., 2011), que é apresentado com mais detalhes na Subseção 2.1.4.1.

2.1.4.1 Oplexicon

O léxico da língua portuguesa com polaridade rotulada escolhido para ser utilizado neste trabalho foi o "Oplexicon" (SOUZA et al., 2011). Este léxico é composto por 32.191 palavras com polaridades rotuladas em -1 (negativa), 0 (neutra) ou 1 (positiva). Além da polaridade, o léxico também traz a informação da classe gramatical das palavras.

O Oplexicon contém palavras em sua forma flexionada em gênero, número e em outras formas de escrita.

A Tabela 3 mostra exemplos de palavras relacionadas aos adjetivos "bonito" e "feio" no léxico.

bonita	adj	1
bonitas	adj	1
bonitinho	adj	1
bonito	adj	1
bonitos	adj	1

feia	adj	-1
feianchao	adj	-1
feiao	adj	-1
feiarrao	adj	-1
feias	adj	-1
feio	adj	-1
feios	adj	-1
feiosa	adj	-1
feiosas	adj	-1
feioso	adj	-1
feiosos	adj	-1

Tabela 3: Exemplo dos adjetivos “bonito” e “feio” no Oplexicon.

2.2 Métodos

2.2.1 Pré-Processamento

O pré-processamento é a etapa da mineração de textos onde se realiza uma série de tratamentos nos dados coletados, a fim de se obter uma amostra o mais padronizada e limpa possível para ser utilizada no treinamento do classificador.

Para isto, é necessário um tratamento pesado, levando em conta os potenciais causadores de erros que se pretendem evitar na análise.

Ao utilizar textos escritos por milhões de indivíduos diferentes como dados de entrada em uma análise quantitativa, é necessário lidar com inúmeros obstáculos, como:

- O uso da linguagem informal, com presença de gírias, erros gramaticais e milhares de diferentes padrões de digitação particulares de cada usuário.
- A utilização da linguagem enfática, com repetições propositais de letras e uso não convencional da acentuação nas frases.
- A presença de ironia e sarcasmo nos textos, que é um fator que dificulta a interpretação automática da polaridade das frases, dado que uma análise computacional costuma analisar as palavras em seu sentido literal e isolado de seu contexto.

Para amenizar a interferência de problemas como os citados acima, foram realizados os seguintes tratamentos nos dados:

- **Remoção de acentos e pontuação:** A remoção de acentos e pontuação foi realizada a fim de otimizar o cruzamento dos termos com o léxico rotulado, que é composto apenas por palavras sem acento, e também para melhorar a padronização e o agrupamento dos termos contidos nos tweets, dado que no contexto das redes sociais é comum dispensar a acentuação das palavras.

  Apesar dos benefícios, a remoção dos acentos acarreta alguns prejuízos, como a perda da diferenciação de palavras que se tornam iguais sem acento, como "más" e "mas".

  Entretanto, foi decidido manter esta etapa do tratamento, dado que os prejuízos causados por ela são pontuais e ganha-se mais com a remoção dos acentos.

  Este tratamento foi feito em poucas linhas de código, utilizando uma função que substitui os caracteres acentuados por sua versão não acentuada.

- **Padronização de palavras flexionadas:** A padronização de palavras flexionadas tem como objetivo reduzir o universo de palavras analisadas, agrupando palavras flexionadas por uma única palavra de origem. Por exemplo, as palavras "amada", "amado", "amando" e "amo" são todas ligadas à palavra "amor" e possuem a mesma polaridade, logo, podem ser reduzidas ao termo "amor".

  Reduzir à uma única palavra de origem todas as variações na qual uma palavra pode aparecer torna a análise mais concisa e eficaz, já que o universo de palavras tem sua esparsidade reduzida e as palavras presentes no banco de dados se centralizam em menos palavras com maiores frequências de aparição.

  Este tratamento também aumenta a abrangência do léxico rotulado, dado que uma palavra em sua forma de origem tem mais chances de ser encontrada no léxico do que quando esta está escrita em uma forma flexionada.

  A padronização foi realizada através do uso de um Léxico de Lematização (MECH, 2018), construído para este fim. O léxico é constituído de uma lista de mais de 850 mil palavras flexionadas relacionadas a cerca de 92 mil palavras de origem.

  A partir de um cruzamento simples do léxico com os termos extraídos, foi feita a substituição da forma flexionada das palavras pela palavra de origem.

- **Remoção de *stop words*:** As *stop words* são palavras utilizadas para auxiliar na estruturação de frases, mas que não carregam peso semântico e, portanto, costumam não contribuir para análises textuais. Geralmente, são as palavras que aparecem com

maior frequência no conjunto de documentos. Alguns exemplos de *stop words* são: as, e, os, de, para, com, sem, foi, etc.

A remoção de *stop words* foi feita com uso da lista de stop words disponibilizada em (LOPES, 2012). Foram cruzadas e removidas as palavras dos tweets coletados que estavam presentes na lista.

2.2.2 Analisador de Dependência Entre as Palavras de um Texto

Neste trabalho, foram exploradas as relações de dependência entre as palavras do texto a fim de verificar a possibilidade de ser obter uma melhor classificação dos tweets ao direcionar mais atenção às palavras diretamente relacionadas com a entidade analisada.

Para isto, foi utilizada a ferramenta VISL (*Visual Interactive Syntax Learner*) (DENMARK, 2022), que disponibiliza em um *site* diversos tipos de analisadores de dependência entre palavras desenvolvidos na língua portuguesa. O método utilizado pelo *site* é o do sistema de análise PALAVRAS (BICK, 2000).

O VISL foi um facilitador para o uso do PALAVRAS pois já fornece toda a aplicação da metodologia de análise do PALAVRAS em uma interface amigável, onde insere-se uma frase como *input* e recebe-se como retorno uma saída organizada das relações de dependência entre as palavras contidas no texto. Além disso, o *site* permite escolher o tipo de saída desejada. Um exemplo de funcionamento do analisador de dependência em estrutura de árvore aplicado a uma amostra real pode ser visto na Figura 7.

O processamento dos tweets no analisador de frases do *site* VISL foi feito por um *script* na linguagem R. As etapas do *script* consistiam em realizar uma requisição do *site* com cada frase, fazer o *download* do arquivo *html* com o conteúdo do *site* após carregar a análise da frase e realizar uma série de tratamentos de texto para obter a análise de dependências no formato necessário para ser utilizado.

Para interpretar a análise visual mostrada na Figura 7 em forma de texto, foram usados os argumentos “Visualization” = “Source” e “Notational convention” = “CG-style”, que retornam a análise no formato da Figura 8.

Após a exploração de diversos tratamentos na tentativa de extrair informação desta estrutura, foi obtido um resultado satisfatório realizando os passos na lista a seguir, que também estão ilustrados pela Tabela 4.

1 - Primeiro, separou-se o resultado obtido pelo analisador em duas colunas. A pri-

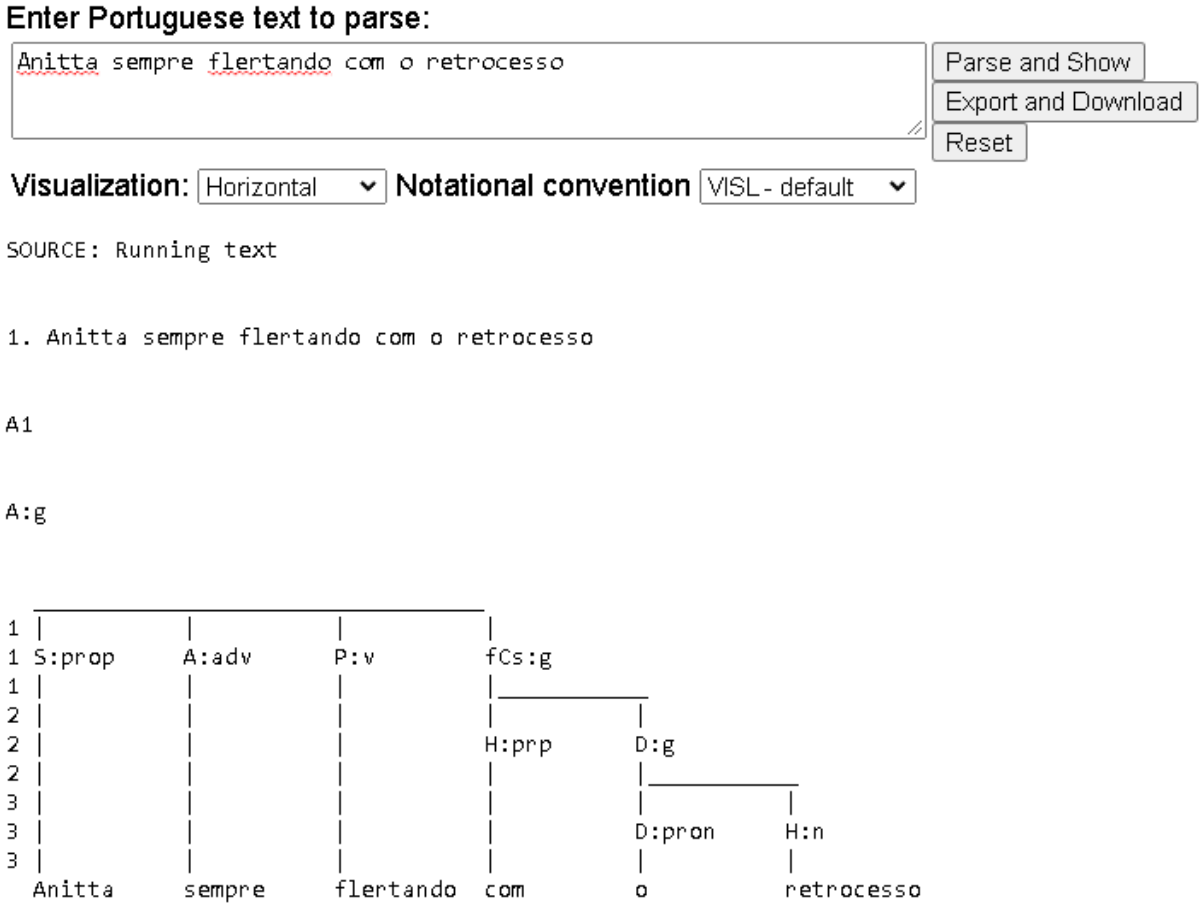

Figura 7: Exemplo de funcionamento da ferramenta VISL.

```
SOURCE: Running text
1. Anitta sempre flertando com o retrocesso
A1
UTT:ap
SUBJ:prop("Anitta" <*> M/F S £hum)      Anitta
ADVL:adv("sempre" <left>)       sempre
P:v-ger("flertar" §AG)  flertando
PRED:pp
=H:prp("com" <right>)   com
=P<:np
==>N:pron-det("o" <artd> DET M S)       o
==H:n("retrocesso" M S) retrocesso

SOURCE: live
1. running text
A1
```

Figura 8: Exemplo de extração de informação pela ferramenta VISL.

meira sendo tudo à esquerda do caractere ")" e a segunda sendo tudo à direita do caractere ")". Assim, foi possível identificar as linhas que continham os termos da frase analisada.

2 - Foi feita a contagem de símbolos "=" em cada linha. Este símbolo representa a

"altura" do galho em que a palavra está.

3 - As linhas que não continham palavras analisadas, mas estavam no meio de duas linhas com palavras analisadas, foram categorizadas como "links", ou seja, representavam conexões na frase.

4 - As linhas que não eram links e nem continham palavras foram excluídas.

Saída Parser	Palavra	Cont (=)	Link	Excluir
SOURCE: Running text		0	1	1
1. Anitta sempre flertando com o retrocesso		0	1	1
A1		0	1	1
A:g(advp		0	1	1
S:prop("Anitta" <*> M/F S £hum	Anitta	0	0	0
A:adv("sempre" <left>	sempre	0	0	0
P:v(ger "flertar" §AG	flertando	0	0	0
PRED:pp		0	1	0
=H:prp("com" <right>	com	1	0	0
=P<:np		1	1	0
==N:pron-det("o" <artd> DET M S	o	2	0	0
==H:n("retrocesso" M S	retrocesso	2	0	0

Tabela 4: Etapas de extração de informação do analisador de dependências.

Após os tratamentos anteriores, foi aplicada a seguinte lógica para definir se cada linha estava relacionada ou não com a anterior:

1 - Se "Cont (=)" da linha atual for igual ao "Cont (=)" da linha anterior, as palavras estão no mesmo nível e são seguidas uma da outra. Então a linha atual recebe o mesmo identificador de "Relação" da anterior.

2 - Se o teste anterior for falso, é testado se "Cont (=)" da linha atual é igual ao"Cont (=)" da linha anterior +1, o que significa que a linha atual é um nível mais fundo da linha anterior, então está em um galho dentro dela. Logo, "Relação" da linha atual = "Relação" da linha anterior.

3 - Se o teste 2 é falso, porém a linha atual representa um "link", então, "Relação" da linha atual = "Relação" da linha anterior.

4 - Se o teste 3 também for falso, então "Relação" da linha atual = "Relação" da linha anterior +1, para começar um novo conjunto de relações.

Após os passos descritos no parágrafo anterior, foi obtida uma base conforme a Tabela 5. Neste exemplo, todas as palavras ficaram com identificador de relação = 0, ou seja, estão todas no mesmo subconjunto.

Assim, passou-se a explorar as maneiras como as informações trazidas poderiam ser incorporadas no treinamento do modelo.

Saída Parser	Palavra	Cont (=)	Link	Excluir	Relação
S:prop("Anitta" <*> M/F S £hum	Anitta	0	0	0	0
A:adv("sempre" <left>	sempre	0	0	0	0
P:v(ger "flertar" §AG	flertando	0	0	0	0
PRED:pp		0	1	0	0
=H:prp("com" <right>	com	1	0	0	0
=P<:np		1	1	0	0
==N:pron-det("o" <artd> DET M S	o	2	0	0	0
==H:n("retrocesso" M S	retrocesso	2	0	0	0

Tabela 5: Etapas de extração de informação do analisador de dependências com coluna "Relação" identificando cada subconjunto de palavras.

Foram testadas duas possibilidades para o uso do analisador de dependências, apresentadas a seguir.

2.2.2.1 Redução do Texto por Relações de Dependência

A primeira tentativa de uso do analisador de dependência foi na seleção de palavras dos tweets coletados.

Com o identificador de cada subconjunto de palavras dentro das frases construído na Seção 2.2.2, foi selecionado o subconjunto que continha o termo "Anitta", descartando os demais.

Desta maneira, foi gerada uma segunda versão da base de tweets rotulados, contendo apenas as palavras diretamente ligadas e próximas à entidade de interesse na frase. Alguns tweets com este tratamento podem ser vistos na Tabela 6.

2.2.2.2 Ponderação do Texto por Relações de Dependência

Nesta maneira de utilizar o parser de dependência, o subconjunto de palavras próximas e diretamente ligadas ao termo "Anitta" foi duplicado na frase e os demais termos, indiretamente, ou não ligados à entidade de interesse, foram mantidos.

Assim, foi gerada uma terceira versão da base de tweets rotulados, com os termos diretamente ligados à "Anitta" com frequência 2x maior do que os demais. Alguns tweets com este tratamento podem ser vistos na Tabela 6.

Pelos exemplos da Tabela 6, é possível notar que este método de utilização do analisador de dependências é eficiente em alguns casos e ineficiente em outros. Na frase 1, o

Polaridade	Texto Original	Texto Parcela Anitta Duplicada	Texto Reduzido à Parcela Anitta
1	Agradecimentos de TCC : Agradeço a Anitta por todas as músicas que me deram energia pra escrever	Agradeço a Anitta Agradeço a Anitta Agradecimentos de TCC por todasas músicas que me deram energia pra escrever	Agradeço a Anitta
1	Amo as antigas da Anitta sem condições	as antigas de a Anitta as antigas de a Anitta Amo	as antigas de a Anitta
-1	Anitta fica se diminuindo e indo nessa coisas que envolvem G-Kay, pra que? Que mico.	Anitta fica se diminuindo e indo em Que mico Anitta fica se diminuindo e indo em Que mico essa coisas	Anitta fica se diminuindo e indo em Que mico
1	Anitta nunca errou quando disse que o latino era um falido	Anitta nunca errou quando disse Anitta nunca errou quando disse que o latino era um falido	Anitta nunca errou quando disse

Tabela 6: Exemplos de tweets rotulados: original, ponderado e reduzido.

analisador beneficia a parcela “Agradeço a Anitta”, captando pontualmente o sentimento principal expressado pelo usuário em relação à artista. Entretanto, na frase 2, o analisador penaliza a parcela da frase que contém a palavra “Amo”, que carrega o sentimento principal.

2.2.3 Variável Score por Léxico Rotulado

O Léxico rotulado foi utilizado para calcular o *score* dos tweets. De maneira simples, foram cruzadas as palavras do léxico rotulado com as dos tweets. As palavras encontradas no léxico tiveram polaridade -1 (negativa) ou 1 (positiva) atribuídas, de acordo com sua classificação no léxico, as que não foram encontradas no léxico receberam polaridade 0 (neutra).

O *score* calculado foi a soma das polaridades das palavras de cada tweet.

Em um teste inicial, o *score* entrou como uma variável na matriz termo-documento para ser considerada no modelo, mas essa prática precisou ser descartada por gerar um viés na classificação, que ficou quase 100% baseada na variável *score*.

2.2.4 O Problema da Classificação

A tarefa de gerar um algoritmo capaz de classificar a polaridade, ou o sentimento latente expressado em um texto, está dentro do campo de estudos da Mineração de Opiniões, ou Análise de Sentimentos. Segundo (PANG; LEE, 2008), "A Mineração de Opiniões e a Análise de Sentimentos é o estudo computacional de opiniões, sentimentos e subjetividade em textos.".

Essa é uma área de estudo que tem evoluído significativamente nos últimos anos, buscando utilizar a tecnologia como aliada no desenvolvimento de técnicas cada vez mais sofisticadas para lidar com os desafios da análise de textos, conseguindo assim extrair uma maior quantidade de informações, com mais precisão e agilidade, de todo o conteúdo gerado por usuários disponível na internet, que tende a crescer no decorrer dos anos.

A linguagem foi desenvolvida pelos seres humanos há milhares de anos como uma forma eficiente de comunicar seus pensamentos. O cérebro humano desenvolveu a capacidade de interpretar os padrões linguísticos e dar sentido ao que eles comunicam. No entanto, apesar de fazer isto com extrema precisão, ele enfrenta limitações quando se trata de realizar esta tarefa rapidamente para milhões de *inputs*.

A leitura deste grande volume de dados é facilmente realizada pela geração atual de computadores. Entretanto, estes não possuem a capacidade de interpretar o significado dos dados recebidos, a não ser que sejam ensinados.

Com o avanço da tecnologia e dos métodos de programação, se tornou possível unir a capacidade de uma máquina de receber milhões de *inputs* com a habilidade de entender padrões similares com os que o nosso cérebro utiliza para decodificar seus significados.

Neste contexto, surge a ideia de treinar uma máquina a fim de fornecer à ela esta habilidade. Isto geralmente é feito de duas principais maneiras, que serão apresentadas nas seções a seguir.

2.2.5 Classificação por Léxico Rotulado

Uma das maneiras de ensinar uma máquina a realizar a classificação de dados textuais é através do uso de um léxico rotulado, como o apresentado na Subseção 2.1.4.1. Esta abordagem consiste em fornecer à maquina a classificação prévia da polaridade de um grande número de palavras isoladas, para que as polaridades conhecidas das palavras contidas no texto sejam usadas para classificar a polaridade do texto como um todo.

A maneira mais simples de utilizar um léxico rotulado na classificação da polaridade de textos é calculando o *score* da frase como um todo, com base no *score* individual dos termos contidos nela, e aplicando uma regra de classificação sobre o *score* calculado para dizer se o texto tem polaridade negativa ou positiva.

Tanto o cálculo do *score* quanto a regra de classificação utilizada pelo algoritmo pode variar entre inúmeros métodos, dos mais simples aos mais complexos, à critério do desenvolvedor.

Um exemplo casual de regra seria:

"O *score* da frase é a soma das polaridades individuais dos termos contidos nela, e a regra de classificação da polaridade da frase é: se o *score* for menor que zero, atribua à frase a polaridade negativa, se for maior ou igual a zero, atribua à frase a polaridade positiva."

Assim, com um léxico rotulado, pode-se construir um algoritmo que realiza o cruzamento das palavras analisadas com as palavras contidas nele, calcula um *score* para o texto com base no *score* individual das palavras e utiliza a regra pré-definida pelo desenvolvedor para atribuir um rótulo ao texto de interesse. Este processo pode ser ilustrado pela Figura 9.

As principais vantagens de métodos como este são o baixo custo computacional e o fato de que não dependem de uma amostra rotulada para gerar um algoritmo de classificação. Entretanto, em análises em que o interesse está em identificar a polaridade da opinião do autor do texto em relação a um alvo específico mencionado nele, e não do texto como um todo, este método não possui a capacidade de analisar contexto ou identificar padrões, ficando limitado somente à polaridade individual das palavras que compõem o texto. Desta maneira, mostra-se interessante explorar outros métodos de classificação baseados na identificação de padrões nos dados, ou até mesmo as relações de dependência entre as palavras contidas no texto, e não somente sua polaridade individual.

O sentimento latente no Tweet da Figura 10 como um todo é a tristeza, que seria classificada com polaridade negativa no espectro dos sentimentos. No entanto, ao analisar o contexto, entende-se que o significado oculto que a frase manifesta é equivalente à frase "Gostaria de muito ir ao show da Anitta", e expressa, na verdade, um sentimento positivo em relação à artista.

No exemplo da Figura 10, a utilização do léxico poderia levar à conclusão equivocada de que a frase tem polaridade negativa em relação à artista devido ao peso negativo do

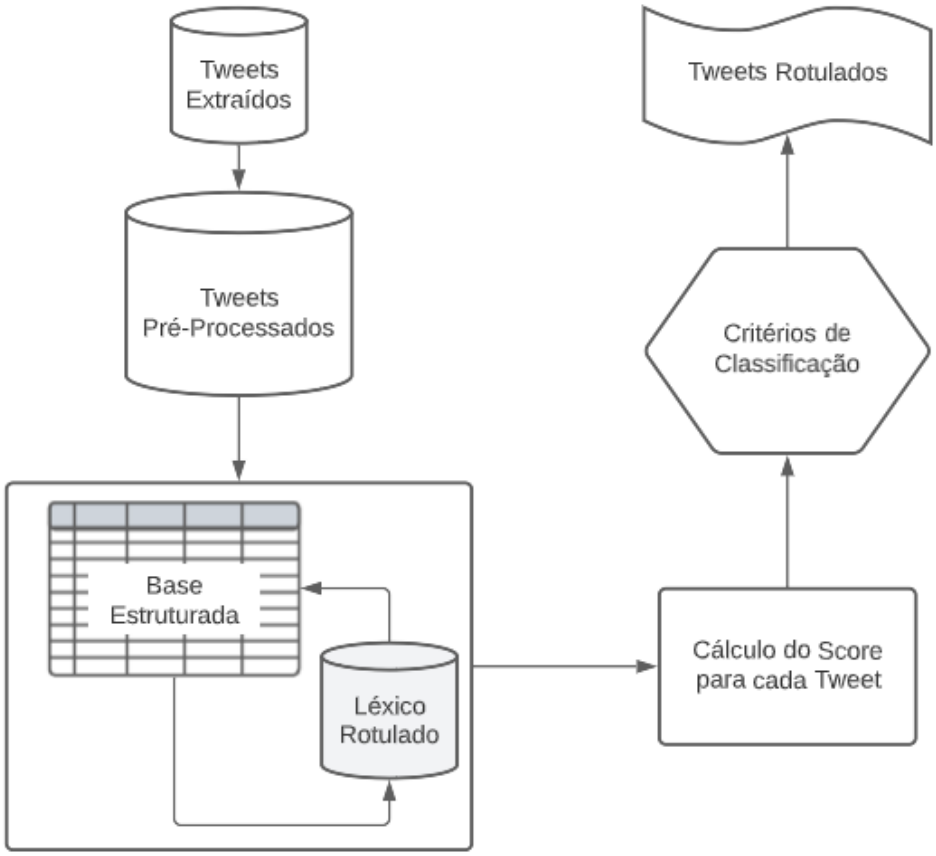

Figura 9: Processo de classificação com uso de Léxico Rotulado.

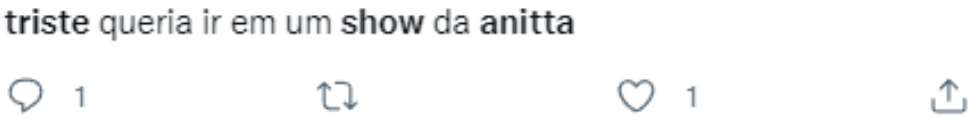

Figura 10: Exemplo de tweet com divergência entre classificação a nível de documento e a nível de entidade.

termo “triste”, enquanto utilizando-se um modelo treinado a partir da identificação de padrões em tweets rotulados onde este tipo de frase costuma ser classificado como positivo, este conjunto de palavras poderia ter mais chances de receber um rótulo positivo.

Uma análise empírica de Tweets extraídos inicialmente mostrou que são frequentes os casos de Tweets que se referem a mais de uma entidade em um mesmo texto, com o intuito de expressar uma opinião positiva em relação a uma das entidades, em detrimento da outra.

“Abordagens baseadas em léxico têm sido extensivamente aplicadas em textos convencionais como blogs, fóruns e avaliação de produtos. Entretanto, são menos explorados na Análise de Sentimento de Tweets. A principal razão é a unicidade de textos no Twitter

que, não só contêm um grande número de peculiaridades textuais e expressões coloquiais, mas também têm uma natureza dinâmica com novas expressões e hashtags surgindo de tempos em tempos." (GIACHANOU; CRESTANI, 2016)

2.2.5.1 Adaptação de Domínio do Léxico Rotulado

Um aspecto importante a ser considerado na classificação da polaridade de textos utilizando um léxico com polaridade rotulada é o universo em torno do assunto no qual está inserida a análise, comumente chamado de domínio.

A questão do domínio consiste em levar em conta o fato de que as palavras podem ter polaridades diferentes dependendo do contexto em que estão inseridas. Por exemplo, analisar textos relacionados à artistas da indústria musical é diferente de analisar textos sobre outros segmentos como o de roupas, comida ou eletrodomésticos.

Além da diferença entre domínios, existe ainda a diferença dos meios pelos quais os textos foram extraídos. As características observadas em um texto extraído de um microblog diferem das observadas em textos de revisões de usuários em sites de vendas, blogs ou artigos.

As palavras "destruição", "matar" e "nojo", são exemplos de expressões com conotação negativa no dicionário tradicional, mas que no contexto das redes sociais são utilizadas como elogios. Isto pode ser visto nos exemplos das Figuras 11, 12 e 13.

Figura 11: Exemplos de tweets utilizando a palavra "destruição" com conotação positiva.

Por este motivo, se mostra interessante adaptar o léxico de polaridades para lidar com as nuances dos diferentes significados que podem ser atribuídos a uma palavra no

Figura 12: Exemplos de tweets utilizando a palavra “mata” com conotação positiva.

Figura 13: Exemplos de tweets utilizando a palavra “nojo” com conotação positiva.

contexto de tweets relacionados a artistas da indústria musical.

Existem diferentes maneiras de realizar a adaptação de um Léxico Rotulado. Neste trabalho, o método escolhido para este fim foi o de Frequência de Termos, utilizado em (JIMENEZ-ZAFRA et al., 2016)

No artigo, são utilizadas duas abordagens para adaptar o léxico iSOL ao domínio de avaliaçao de filmes. O léxico é composto por palavras da língua espanhola com polaridades anotadas como “positiva” ou “negativa”. Ambas as abordagens melhoraram a performance do classificador baseado neste léxico.

O método da Frequência de Termos, escolhido para ser utilizado neste trabalho devido à sua simplicidade, é explicado de forma mais detalhada a seguir.

Adaptação do Léxico por Frequência de Termos: A adaptação do Léxico baseada na frequência de termos consiste em atribuir rótulos de polaridade “negativa” ou “positiva” nas palavras de acordo com sua frequência em registros rotulados como “negativos” ou “positivos”.

Por exemplo, se a palavra “ok” aparece em 500 registros rotulados, e 400 deles possuem rótulo “negativo”, então este método sugere que o rótulo individual da palavra “ok” deva ser “negativo”, por mais que no léxico rotulado original ela apareça com polaridade “positiva”.

Para realizar este tratamento, é preciso um léxico rotulado inicial e um grande conjunto de textos rotulados. Desta maneira, é possível quantificar a frequência de cada palavra em textos negativos e positivos, e, à partir de uma regra de decisão baseada

nessas frequências, optar entre manter ou modificar a polaridade da palavra no léxico convencional.

É importante que as amostras utilizadas na adaptação do léxico não sejam reutilizadas para avaliar o desempenho do classificador baseado no léxico adaptado, para que elas não causem viés na avaliação.

Realizando este tratamento, o domínio da análise, ou seja, o tema em torno dos textos extraídos, é levado em consideração na definição da polaridade dos termos no léxico.

2.2.6 Classificação por Aprendizado de Máquinas

Outro método muito utilizado na tarefa de treinar uma máquina para classificar a polaridade de textos é utilizando algoritmos de Aprendizado de Máquinas. "A maioria dos métodos propostos para lidar com a Análise de Sentimento no Twitter emprega um classificador do campo de Aprendizado de Máquinas que é treinado à partir de diversas características dos tweets." (GIACHANOU; CRESTANI, 2016)

Esta área de estudos tem o objetivo de treinar algoritmos à partir da observação e identificação de padrões em dados "do mundo real" no contexto analisado, de maneira que não seja necessário definir previamente um conjunto de todas as regras e condições que o algoritmo deve considerar para classificar dados. O próprio algoritmo define o processo de classificação que melhor se encaixa ao tipo de dado que está sendo analisado, baseado nas características que foram aprendidas através de uma amostra dos próprios dados. "A área de reconhecimento de padrões está comprometida com o descobrimento automático de regularidades em dados através do uso de algoritmos computacionais, e de utilizar essas regularidades para realizar ações como a classificação de dados em diferentes categorias." (BISHOP, 2006)

No exemplo da identificação da polaridade de textos, diferente da classificação com o uso do léxico rotulado, a classificação feita com um algoritmo de Aprendizado de Máquinas supervisionado partiria do princípio de "aprender" a melhor forma de atribuir rótulos aos textos à partir da observação de uma grande quantidade de textos previamente rotulados fornecida como *input*. Isto significa ter um banco de dados com diversos exemplos representativos dos textos que se pretende classificar futuramente, com suas respectivas polaridades já rotuladas. Essa rotulação costuma ser realizada de forma manual por alguém que conheça os dados e a forma correta na qual eles deveriam ser classificados, e deve ser feita com cautela, dado que o algoritmo de classificação será gerado com base

nela.

Desta maneira, esses dados rotulados são fornecidos como *input* e o algoritmo de aprendizado de máquinas os utiliza na busca de padrões os quais entende tratar-se das particularidades que levaram cada texto ao seu rótulo. Futuramente, o algoritmo modelado com base nesses padrões servirá para rotular *inputs* futuros, ainda não vistos e não rotulados. Este processo é ilustrado pelo diagrama da Figura 14.

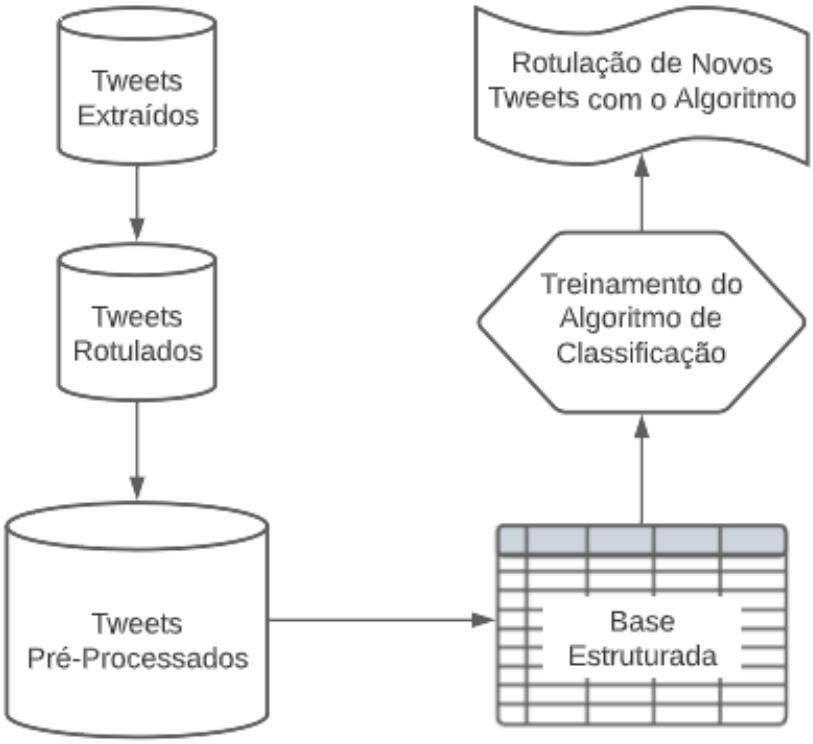

Figura 14: Processo de classificação por Aprendizado Supervisionado.

2.2.7 Algoritmos Baseados em Árvores

2.2.7.1 Árvores de Decisão

Uma árvore de decisão é um sistema de classificação binária que utiliza uma estrutura de árvore para gerar uma série de segmentações nos dados realizando testes em seus atributos, definindo assim o caminho até a sua classificação final. "Toda pergunta feita no processo de decisão é um teste em um atributo. Todo teste leva, ou à conclusão, ou a um teste adicional condicionado à pergunta atual." (ZHOU; LIU, 2016)

Árvores de decisão são compostas por uma raiz, que corresponde ao teste inicial, múltiplos nós, correspondentes aos testes adicionais, e múltiplas folhas, equivalentes às

classificações finais. Uma ilustração desta estrutura pode ser vista na Figura 15, que simula um processo de predição de atraso (ou não atraso) de uma pessoa testanto atributos do seu trajeto.

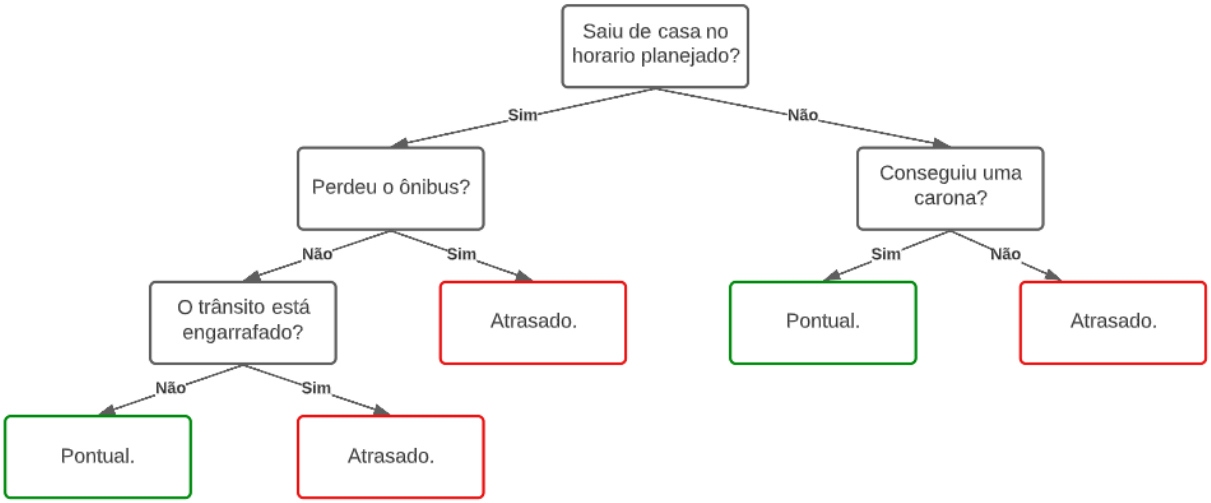

Figura 15: Estrutura de uma árvore de decisão.

"Cada nó divide a amostra em nós 'filhos', e cada caminho da raiz até a folha é uma sequência de decisões. O objetivo é gerar uma árvore capaz de prever a classificação de novas amostras." (ZHOU; LIU, 2016)

Para otimizar a seleção dos atributos que melhor segmentam as amostras, busca-se que o nó acumule mais amostras em uma classe, em outras palavras, o nó com maior "pureza".

A pureza de um nó pode ser medida de diferentes formas, dentre elas pela entropia que, conforme em (ZHOU; LIU, 2016), é calculada da seguinte maneira:

$$Ent(D) = -\sum_{k=1}^{|\gamma|} p_k \log_2 p_k. \tag{2.2}$$

Onde p_k é a proporção da classe k na base de dados atual D, onde $k = 1, 2, ..., |\gamma|$.

Quanto menor a $Ent(D)$, mais puro é o nó.

"Suponha que o atributo discreto a possui V possíveis valores $a_1, a_2, ..., a_V$. Então, dividindo a base de dados D pelo atributo a, produz-se V nós filhos, onde o $v$ésimo nó D^v inclui todas as amostras em D que possuem o valor a^v do atributo a. A entropia de D^v pode ser calculada usando 2.2. Dado que há diferentes números de amostras nos nós filhos, o peso $frac|D^v||D|$ é atribuído para refletir a importância de cada nó, ou seja, quanto

maior o número de amostras, maior o impacto do nó de ramificação. Então, o ganho de informação dividindo o conjunto de dados D com o atributo a é calculado conforme 2.3." (ZHOU; LIU, 2016)

$$Gain(D, a) = Ent(D) - \sum_{v=1}^{|V|} \frac{|D^v|}{|D|} Ent(D^v). \tag{2.3}$$

"Em geral, quanto maior o ganho de informação, mais melhora em pureza podemos esperar dividindo D pelo o atributo a. Assim, o ganho de informação pode ser utilizado para a seleção de atributos." (ZHOU; LIU, 2016)

2.2.7.2 Florestas Aleatórias

O método de classificação por floresta aleatória utiliza diversas árvores de decisão combinadas para classificar uma amostra. Primeiro, são treinadas F árvores individualmente, em seguida, as árvores treinadas são utilizadas para classificar novas amostras, de maneira que as predições das árvores para cada elemento da amostra são consideradas como votos para cada classe. A classe mais votada pelo conjunto de árvores se torna a classificação final daquele elemento da amostra, conforme ilustra a Figura 16.

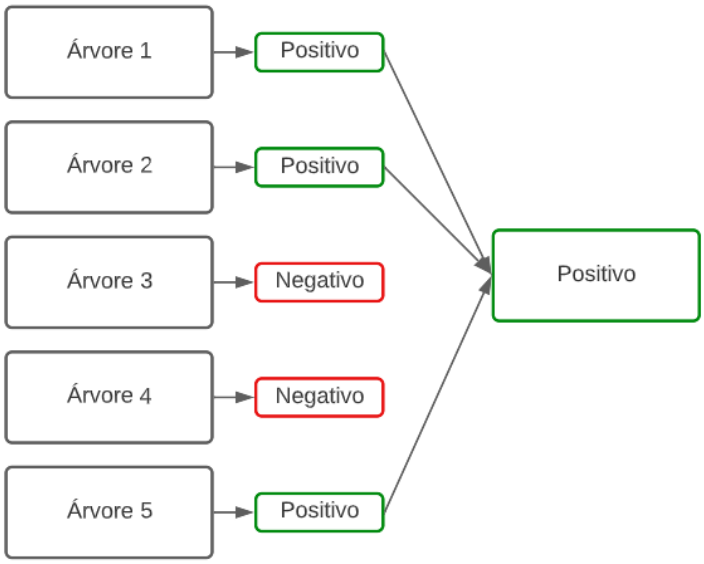

Figura 16: Processo de votação de uma floresta aleatória.

"Combinando múltiplos algoritmos de aprendizado, a habilidade de generalização de um *ensemble* é, geralmente, muito mais forte do que a de um algoritmo individual, especialmente para algoritmos fracos." (ZHOU; LIU, 2016)

Neste modelo de classificação, diversas árvores são treinadas paralelamente para que seja feita a votação da classe predita. Para garantir que as árvores não são correlacionadas entre si, os atributos considerados no treinamento de cada árvore são selecionados aleatoriamente e a seleção da amostra é feita utilizado o método de *bagging*, que consiste em realizar reamostragens com reposição na amostra treino para que cada árvore seja treinada com uma amostra diferente.

"Random Forest (RF) (Breiman 2001) é uma extensão do *bagging*, onde a aleatorização da seleção dos atributos é introduzida além do *bagging*. Especificamente, uma árvore de decisão tradicional seleciona um atributo de divisão ótimo entre o conjunto de atributos de cada nó, enquanto uma floresta aleatória seleciona de um subconjunto aleatório de k atributos do conjunto de atributos do nó. O parâmetro k controla a aleatoriedade, onde a divisão é a mesma que em árvores de decisão tradicionais se k é igual ao total de atributos, e um atributo de divisão é aleatoriamente selecionado se $k = 1$. Tipicamente, o valor recomendado de k é $k = log_2(d)$ (Breiman 2001), sendo d o total de atributos." (ZHOU; LIU, 2016)

2.2.7.3 Gradient Boosting Machine

O Gradient Boosting é um método que tem como objetivo melhorar um classificador focando em seus erros e tentando reduzi-los. Diferente do modelo de Floresta Aleatória, o Gradient Boosting não treina as árvores paralelamente. Este modelo é focado em gerar árvores melhores em prever os erros das árvores anteriores, o que faz com que o modelo seja treinado com uma melhora gradual do erro. "*Boosting* é uma família de algoritmos que converte classificadores fracos em classificadores fortes." (ZHOU; LIU, 2016)

No Gradient Boosting, um classificador base é treinado e, em seguida, um segundo classificador é treinado a partir de um ajuste na amostra de treino, que é feito com o intuito de direcionar mais atenção às amostras que forram classificadas incorretamente pelo primeiro classificador. Esse processo segue de maneira iterativa, até que sejam treinados um número pré definido de classificadores. Então, os classificadores gerados são combinados entre si, cada um com um peso.

O exemplo a seguir, retirado de (AYYADEVARA, 2018), mostra de maneira clara a lógica por tras do algoritmo de Gradient Boosting:

Suponha um modelo M, baseado em árvores de decisão. O modelo M pode ser escrito de acordo com :

$$Y = M(x) + \epsilon. \tag{2.4}$$

Onde Y é a variavel dependente e $M(x)$ é a árvore de decisão usando as X variáveis independentes.

O erro da árvore de decisão anterior pode ser predito por 2.5

$$\epsilon = G(x) + \epsilon_2. \tag{2.5}$$

$G(x)$ é outra árvore de decisão que tenta prever o erro utilizando as X variáveis independentes;

O próximo passo é similar ao anterior. É construído um modelo que tenta prever o $erro_2$ usando as X variáveis independentes 2.6:

$$\epsilon_2 = H(x) + \epsilon_3. \tag{2.6}$$

Combinando todos os modelos, tem-se 2.7.

$$Y = M(x) + G(x) + H(x) + \epsilon_3. \tag{2.7}$$

É esperado que a Equação 2.7 tenha acurácia maior que a do modelo inicial $M(x)$, dado que em 2.7 são consideradas três árvores de decisão.

O algoritmo de Gradient Boosting retorna um número entre 0 e 1 que expressa a probabilidade da amostra ser classificada como positiva. Para atribuir uma classificação binária negativa ou positiva à amostra, é necessário definir um ponto de corte a partir do qual a probabilidade será testada e, caso esteja acima ou seja igual ao ponto de corte, será classificada como positiva, caso contrário, será classificada como negativa. Este ponto de corte pode ser definido arbitrariamente, ou podem ser utilizados métodos como a Curva ROC, apresentada na Subseção 2.2.9.

No treinamento de um classificador baseado em Gradient Boosting pode-se ajustar os valores de seus parâmetros. Neste trabalho, foram explorados os ajustes nos seguintes parâmetros:

- Número de árvores: Corresponde ao número máximo de árvores treinadas pelo algoritmo.

- Profundidade: É o número máximo de nós que uma árvore pode ter.
- Taxa de aprendizado: É um valor entre 0 e um que determina o peso de cada árvore na concepção do modelo final e na minimização do resíduo gerado.

2.2.8 Regressão Logística

O modelo de Regressão Logística é utilizado para prever variáveis binárias, que descrevem se um evento ocorreu ou não. Este modelo faz uso de uma curva sigmoidal para classificar variáveis binárias. Uma curva sigmoidal varia de 0 a 1 e é descrita pela Fórmula 2.8:

$$S(t) = \frac{1}{1 + e^{-t}}. \tag{2.8}$$

Esta curva pode ser ilustrada pela Figura 17.

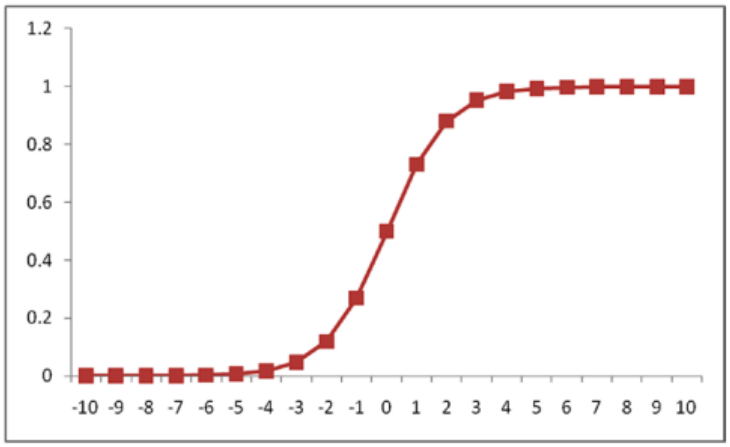

Figura 17: Gráfico de uma curva sigmoidal. Fonte: (AYYADEVARA, 2018)

"Uma Regressão Linear assume relação linear entre a variável dependente e as independentes. Ela é escrita como $Y = a+bX$. A Regressão Logística se afasta do pressuposto de que todas as relações são lineares aplicando uma curva sigmoidal." (AYYADEVARA, 2018)

A fórmula matemática do modelo de Regressão Logística é 2.9.

$$E(Y) = \frac{1}{1 + e^{-(a+b*X)}}. \tag{2.9}$$

Os valores entre 0 e 1 que a equação 2.9 retorna podem ser interpretados como probabilidades. Ou seja, quando $X = x$, a probabilidade da amostra ser classificada como

positiva é $E(Y)$.

Para converter essas probabilidades em classificações binárias, é necessário definir um ponto de corte c para o qual, se $Y >= c$, a classificação é positiva, se $Y < c$, a classificação é negativa.

Isto pode ser feito com o auxílio da Curva ROC, apresentada em 2.2.9.

2.2.9 Curva ROC

A Curva ROC (Curva Característica de Operação do Receptor) é um recurso que pode ser utilizado para auxiliar na escolha do ponto de corte em modelos que retornam uma probabilidade e não uma classificação binária.

A curva é montada seguindo as seguintes etapas: as amostras são ordenadas pela predição, então o ponto de corte é movido gradualmente de cima para baixo na lista ordenada equanto calcula-se, para cada ponto de corte, a taxa de Falsos Positivos (amostras negativas classificadas como positivas), representada no eixo x, e a taxa de Verdadeiros Positivos (amostras positivas classificadas corretamente), representada no eixo y. Um exemplo pode ser observado na figura 18.

A diagonal representa um modelo de palpites aleatórios, e o ponto (0,1) representa o modelo ideal, com 100% de acerto nas amostras positivas.

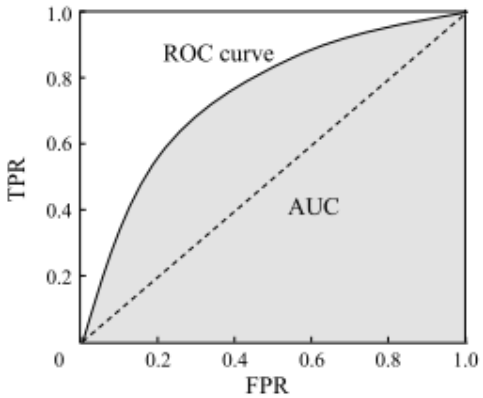

Figura 18: Gráfico de uma curva ROC. Fonte: (ZHOU; LIU, 2016)

Desta maneira, a curva ROC permite estudar as variações entre sensibilidade e especificidade para cada ponto de corte e escolher o que maximiza ambas as métricas.

Para testar este recurso neste trabalho, foi utilizado o pacote “ROCR” (SING, 2020) no R.

2.2.10 Métricas de Avaliação do Classificador

2.2.10.1 Matriz de Confusão

Uma matriz de confusão é um recurso que permite avaliar o desempenho de um algoritmo de classificação observando quantas amostras de cada classe foram classificadas corretamente e incorretamente por ele.

Segundo (ZHOU; LIU, 2016), "Em problemas de classificação binária, existem quatro combinações entre as verdadeiras classes e as classes preditas, chamadas verdadeiro positivo, falso positivo, verdadeiro negativo e falso negativo, e denotamos o número de amostras em cada caso como VP, FP, VN e FN, respectivamente."

A matriz de confusão resume o número de amostras classificadas pelo modelo dentro dos quatro tipos de resultado citados no parágrafo anterior, conforme mostra a Figura 19.

		Rótulos Verdadeiros	
		Positivo	**Negativo**
Predições	**Positivo**	VP	FP
	Negativo	FN	VN

Figura 19: Estrutura de uma matriz de confusão.

Vale notar que as posições [1,1] e [2,2] da matriz somam o número de acertos do classificador, enquanto as posições [1,2] e [2,1] somam o número de erros de classificação.

2.2.10.2 Acurácia

A acurácia é uma medida que avalia a capacidade do modelo de classificar os dados corretamente. Para isto, é calculada a proporção de acertos do modelo no total de amostras, como mostra a Fórmula 2.10.

$$\frac{VP + VN}{VP + VN + FP + FN} \tag{2.10}$$

2.2.10.3 Sensibilidade

A sensibilidade avalia os acertos do modelo em amostras da classe "positiva", calculando a proporção de verdadeiros positivos no total de amostras positivas, de acordo com a Fórmula 2.11.

$$\frac{VP}{VP + FN} \tag{2.11}$$

2.2.10.4 Especificidade

A especificidade é uma medida usada para avaliar os acertos da classe "negativa". É calculada a partir da razão entre os verdadeiros negativos e a quantidade total de amostras negativas, conforme a Fórmula 2.12.

$$\frac{VN}{FP + VN} \tag{2.12}$$

3 Resultados

3.1 Treinamento dos Modelos

Os modelos foram treinados a partir de um script na linguagem R (R, 2009), com uso dos pacotes "caret" (KUHN, 2022) e "gbm" (GREENWELL, 2022).

Para treinar os modelos, a amostra foi particionada em "treino" e "teste", com 75% usada para treino e 25% para teste.

Dado que a amostra possuía 44% tweets positivos a mais que tweets negativos, foi realizado o balanceamento por *undersampling* para equilibrar a quantidade de tweets das classes e evitar viéses no modelo. O balanceamento da amostra foi feito com uso do pacote "themis" (HVITFELDT, 2022).

Foram treinados, ao todo, 9 modelos de Gradient Boosting para cada uma das 3 bases de tweets rotulados (simples, ponderada e reduzida por dependência de palavras). Os 9 modelos foram treinados com *ntrees* (número de árvores) = 10000, e combinações dos parâmetros *shrink* (taxa de aprendizado) variando entre 0.1, 0.01 e 0.001 e *depth* (profundidade das árvores) variando entre 1, 3 e 5, conforme a Tabela 7.

Além disso, foi utilizado o comando *gbm.perf()* do pacote "gbm" (GREENWELL, 2022) para selecionar o número ótimo de árvores de cada modelo. Ou seja, o número de árvores a partir do qual não se tem mais ganho significativo adicionando mais árvores.

Também foi feita a comparação da qualidade dos modelos utilizando um *cutoff* fixo de 0.5 e o *cutoff* indicado pela curva ROC.

Os modelos de Regressão Logística (com os parâmetros padrão da função *train()* do *"caret"* (KUHN, 2022)) e um algoritmo simples baseado em classificação por Léxico Rotulado e Léxico Rotulado Adaptado ao Domínio foram treinados a fim de se obter opções de comparação com o modelo de Gradient Boosting, foco deste trabalho.

A classificação com uso do léxico foi feita calculando o *score* conforme em 2.2.3, e

ntrees	shrink	depth
10000	0.1	1
10000	0.1	3
10000	0.1	5
10000	0.01	1
10000	0.01	3
10000	0.01	5
10000	0.001	1
10000	0.001	3
10000	0.001	5

Tabela 7: Combinações de parâmetros para treinamento do modelo de Gradient Boosting

atribuindo polaridade "negativa" aos tweets com score menor que zero, e "positiva" aos tweets com score maior ou igual a zero. Foram testadas classificações com o léxico em sua versão original e com o léxico adaptado pelo método apresentado em **??**.

3.2 Avaliação dos Resultados

Dentre todos os modelos gerados, o que apresentou o melhor resultado foi o GBM com 459 árvores, taxa de aprendizado = 0.1, profundidade = 5 e cutoff = 0.5. O modelo alcançou acurácia de 74,7%, sensibilidade de 74% e especificidade de 75,9%.

A Tabela 8 mostra os resultados das métricas dos modelos organizados por algoritmos.

Na Tabela 9, os modelos estão ordenados da maior para a menor acurácia.

Nota-se que os algoritmos de aprendizado de máquina obtém melhores desempenhos do que a classificação básica por Léxico Rotulado.

As bases de tweets sem os tratamentos de redução ou ponderação das palavras ligadas à entidade de interesse geraram melhores classificadores com os algoritmos de aprendizado de máquinas do que as bases que consideraram este fator. Entretanto, a classificação com o léxico obteve maior acurácia com os tratamentos de redução, em seguida com o de ponderação, e as menores acurácias com as bases sem tratamento.

As classificações por Léxico Rotulado apresentaram acurácias mais baixas que os classificadores de aprendizado de máquinas. Além disso, se mostraram enviesadas a classificar as amostras como "positivas", ficando com especificidade baixa.

A adaptação do léxico impactou positivamente os resultados deste método, melhorando a acurácia e o equilibrio entre as medidas de sensibilidade e especificidade.

Algoritmo	Matriz Termo-Documento	Ntrees	Shrink	Depth	Cutoff	Acurácia	Sensibilidade	Especificidade	VP	VN	FP	FN
GBM	Simples	459	0,1	5	0,5	74,7%	74,0%	75,9%	600	393	125	211
GBM	Reduzidos	430	0,1	5	0,5	71,1%	68,4%	75,1%	536	406	135	248
GBM	Ponderados	7295	0,01	3	0,5	71,6%	68,8%	75,8%	539	410	131	245
Regresão Logistica	Simples					72,5%	71,6%	73,9%	581	383	135	230
Regresão Logistica	Reduzidos					70,3%	69,0%	72,1%	541	390	151	243
Regresão Logistica	Ponderados					70,6%	67,9%	74,5%	532	403	138	252
Léxico Rotulado	Simples					60,4%	87,1%	19,5%	2342	342	1416	346
Léxico Rotulado	Reduzidos					61,0%	87,2%	20,4%	2346	355	1385	345
Léxico Rotulado	Ponderados					60,9%	87,4%	19,7%	2358	341	1393	339
Léxico Rotulado Adptado	Simples					60,6%	66,6%	51,5%	1791	905	853	897
Léxico Rotulado Adptado	Reduzidos					61,5%	65,2%	55,7%	1754	970	770	937
Léxico Rotulado Adptado	Ponderados					61,0%	62,2%	59,0%	1678	1023	711	1019

Tabela 8: Comparação de resultado dos modelos.

Ranking	Algoritmo	Matriz Termo-Documento	Ntrees	Shrink	Depth	Cutoff	Acurácia	Sensibilidade	Especificidade	VP	VN	FP	FN
1	GBM	Simples	459	0,1	5	0,5	74,7%	74,0%	75,9%	600	393	125	211
2	Regresão Logistica	Simples					72,5%	71,6%	73,9%	581	383	135	230
3	GBM	Ponderados	7295	0,01	3	0,5	71,6%	68,8%	75,8%	539	410	131	245
4	GBM	Reduzidos	430	0,1	5	0,5	71,1%	68,4%	75,1%	536	406	135	248
5	Regresão Logistica	Ponderados					70,6%	67,9%	74,5%	532	403	138	252
6	Regresão Logistica	Reduzidos					70,3%	69,0%	72,1%	541	390	151	243
7	Léxico Rotulado Adptado	Reduzidos					61,5%	65,2%	55,7%	1754	970	770	937
8	Léxico Rotulado	Reduzidos					61,0%	87,2%	20,4%	2346	355	1385	345
9	Léxico Rotulado Adptado	Ponderados					61,0%	62,2%	59,0%	1678	1023	711	1019
10	Léxico Rotulado	Ponderados					60,9%	87,4%	19,7%	2358	341	1393	339
11	Léxico Rotulado Adptado	Simples					60,6%	66,6%	51,5%	1791	905	853	897
12	Léxico Rotulado	Simples					60,4%	87,1%	19,5%	2342	342	1416	346

Tabela 9: Ranking dos modelos com melhor acurácia.

Na Figura 20 é possível observar as variáveis que tiveram o maior peso no modelo de Gradient Boosting selecionado para as amostras simples, ponderada e reduzida. Nota-se algumas variáveis correspondentes à assuntos pontuais nos quais a artista alvo esteve envolvida no ano de 2022, como “vma” e “vmas”, que se referem à uma premiação da indústria da música, e “ep”, que diz respeito a um album lançado pela artista. Isto evidencia a necessidade de atualização periodica do classificador com novas amostras.

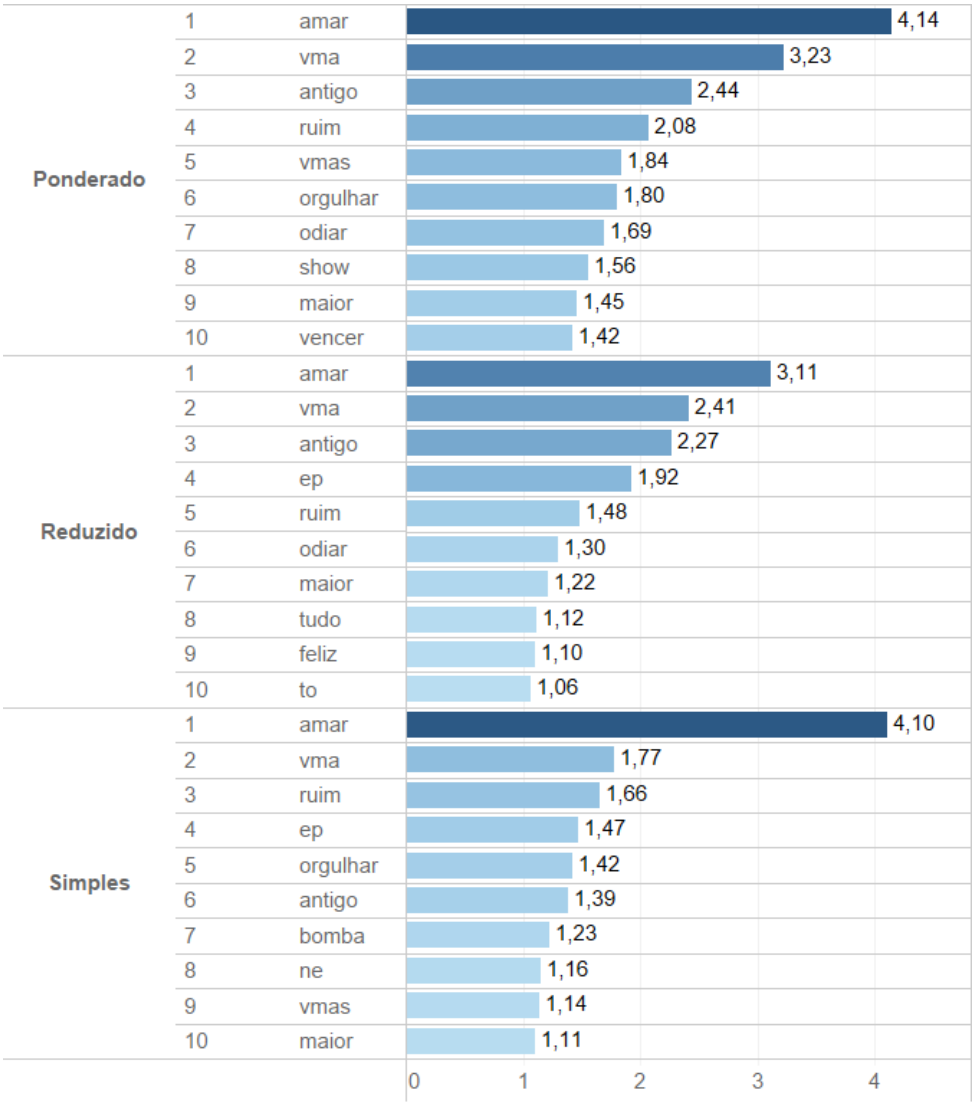

Figura 20: Variáveis com maior peso nos modelos de Gradient Boosting de cada amostra treino.

4 Conclusão

Levando em consideração o peso que a reputação de um artista da indústria musical desempenha em sua carreira, este trabalho teve como objetivo utilizar a vasta gama de conteúdo gerado por usuários no Twitter como uma fonte de informação sobre a polaridade, positiva ou negativa, da opinião das massas sobre ele.

Para isto, foram explorados algoritmos de classificação de texto baseados em aprendizado de máquinas e em léxico de polaridade rotulada. Os algoritmos baseados em aprendizado de máquinas foram os que apresentaram os melhores resultados, tendo melhor desempenho do que todos os modelos baseados em léxico. O algoritmo de Gradient Boosting com 459 árvores, taxa de aprendizado = 0.1, profundidade = 5 e cutoff = 0.5 alcançou acurácia de 74,7%, sensibilidade de 74% e especificidade de 75,9%.

Os algoritmos de Regressão Logística gerados com os parâmetros *default* da função *train()*, tiveram todos acurácia acima de 70%, enquanto os algoritmos baseados em léxico ficaram com acurácia abaixo de 62%.

Além disso, os algoritmos baseados em léxico apresentaram viés de classificação direcionado à classe "positiva", ficando com a especificidade (taxa de acertos da classe "negativa") muito mais baixa que a sensibilidade (taxa de acertos da classe "positiva").

Desta maneira, conclui-se que, apesar de exigir um maior custo de produção, tanto computacional, no treinamento do modelo, quando manual, na rotulação das amostras, os algoritmos de aprendizado de máquinas realizaram a tarefa da classificação de textos de maneira mais eficaz e confiável, alcançando uma boa taxa de acertos e um bom equilíbrio entre sensibilidade e especificidade.

Os métodos de análise da relação de dependência entre as palavras do texto utilizado neste trabalho não trouxeram melhora nas classificações. Ao contrário do que se esperava, a redução dos tweets à parcela diretamente ligada à entidade de interesse e a atribuição de peso à parcela diretamente ligada à entidade de interesse trouxeram prejuízos para o modelo, diminuindo as métricas de avaliação obtidas com o modelo simples sem este

tratamento.

Para trabalhos futuros, sugere-se que as relações de dependência sejam estudadas mais a fundo e que sejam explorados outros métodos para incorporar esta informação no treinamento dos modelos.

Apesar de não terem mostrado bom desempenho, os algoritmos baseados em léxico foram beneficiados pelos tratamentos de dependência entre as palavras. Vale ressaltar também que a adaptação de domínio do léxico acarretou melhora no equilibrio entre as métricas de sensibilidade e especificidade. Para trabalhos futuros, sugere-se explorar métodos mais sofisticados de cálculo de score por léxico, bem como diferentes regras de classificação a partir dos scores.

Referências

AYYADEVARA, V. K. *Pro Machine Learning Algorithms*. [S.l.]: APress, 2018.

BICK, E. *The Parsing System "Palavras"*. [S.l.]: Aarhus University Press, 2000.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006.

BLANK, G.; REISDORF, B. C. The participatory web, information, communication & society. 2012.

DENMARK, U. o. S. *Visual Interactive Syntax Learning*. 2022. Url:https://visl.sdu.dk/.

FELDMAN, R.; SANGER, J. *The Text Mining Handbook*. [S.l.]: Cambridge University Press, 2007.

GIACHANOU, A.; CRESTANI, F. Like it or not: A survey of twitter sentiment analysis methods. *Universita della Svizzera Italiana*, 2016.

GREENWELL, B. *Generalized Boosted Regression Models*. 2022.

HVITFELDT, E. *themis: Extra Recipes Steps for Dealing with Unbalanced Data*. 2022.

IFPI. Global music report 2022. 2022.

JIMENEZ-ZAFRA, S. et al. Domain adaptation of polarity lexicon combining term frequency and bootstrapping. *Department of Computer Science, Universidad de Jaén (Spain)*, 2016.

JORGETTO, J. H. Relações públicas e web 2.0: A manutenção do relacionamento entre artista e público pela internet. *UNESP*, 2009.

KEARNEY, M. W. *R: Collecting and Analyzing Twitter Data - featuring rtweet*. 2022.

KUHN, M. *caret: Classification and Regression Training*. 2022.

LOPES, A. *Stop Words*. 2012. Url:https://gist.githubusercontent.com/alopes/5358189.

MONDELLI, M. L. B.; PETERSON, A. T.; GADELHA, L. Exploring reproducibility and fair principles in data science using ecological niche modeling as a case study. 2019.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. 2008.

R, D. C. T. *R*. 2009.

SING, T. *ROCR: Visualizing the Performance of Scoring Classifiers*. 2020.

SOUZA, M. et al. Construction of a portuguese opinion lexicon from multiple resources. *8th Brazilian Symposium in Information and Human Language Technology*, 2011.

ZHOU, Z.-H.; LIU, S. *Machine Learning*. [S.l.]: Tsinghua University Press, 2016.