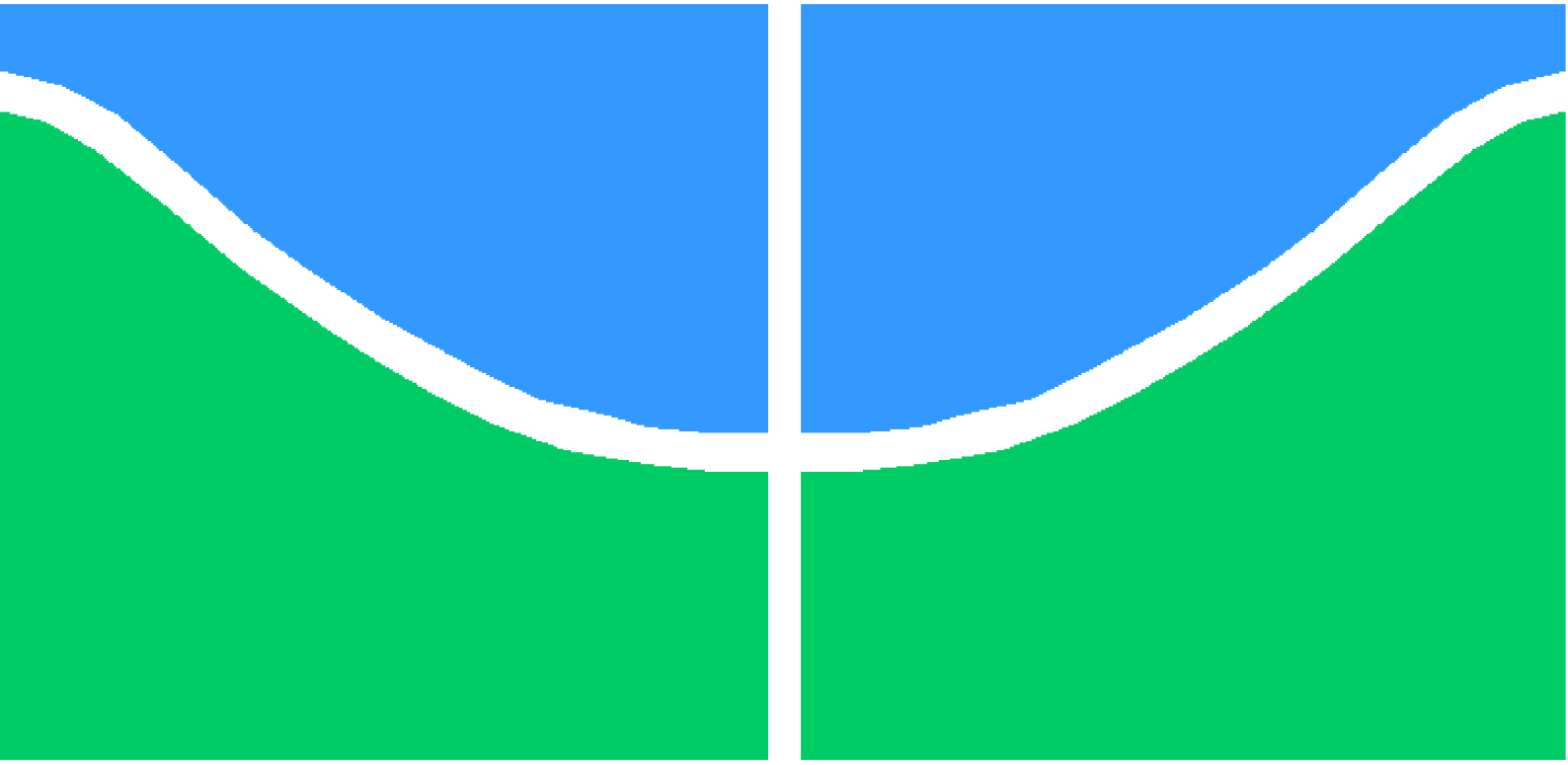

Universidade de Brasília – UnB
Faculdade UnB Gama – FGA
Engenharia de Software

Predição de comentários em mídias sociais sobre discursos racistas.

Autor: Marcelo Augusto Araújo dos Reis
Orientador: Doutora Carla Silva Rocha Aguiar

Brasília, DF
2021

Marcelo Augusto Araújo dos Reis

Predição de comentários em mídias sociais *sobre* discursos racistas.

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Orientador: Doutora Carla Silva Rocha Aguiar

Brasília, DF

2021

Marcelo Augusto Araújo dos Reis
Predição de comentários em mídias sociais *s*obre discursos racistas./ Marcelo Augusto Araújo dos Reis. – Brasília, DF, 2021-
62 p. : il. (algumas color.) ; 30 cm.

Orientador: Doutora Carla Silva Rocha Aguiar

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2021.

1. data science. 2. machine learning. I. Doutora Carla Silva Rocha Aguiar. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Predição de comentários em mídias sociais *s*obre discursos racistas.

CDU

Marcelo Augusto Araújo dos Reis

Predição de comentários em mídias sociais *s*obre discursos racistas.

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, :

Doutora Carla Silva Rocha Aguiar
Orientador

Brasília, DF
2021

Resumo

Com o aumento da utilização de redes sociais no Brasil, problemas existentes na sociedade começaram a ser bastante observados no meio *online*, como o discurso de ódio. O discurso de ódio tem se intensificado pela utilização da internet e suas redes sociais, sendo potencializados tanto pela capacidade de publicação instantânea de conteúdos quanto pela sensação de anônimato que os usuários tem nas redes sociais. Como não é possível para um ser humano analisar cada um dos *conteúdos* inseridos na internet, cada vez mais novas tecnologias são aprimoradas e surgem para esse propósito, como é o caso dos algoritmos de aprendizado de máquina, que ganharam bastante destaque nos ultimos anos com o avanço das capacidades computacionais. Com esses algoritmos é possível selecionar um conjunto de dados e treinar um modelo que consiga fazer a predição de um novo dado inserido. Desta forma, este trabalho busca entender como se dá as interações racistas nas redes sociais, a partir da busca de dados na *API* do *Twitter* e então construir um modelo de aprendizado de máquina que consiga identificar comentários com teor racista, assim como elencar as dificuldades existentes nesse processo.

Palavras-chaves: aprendizado de máquina, racismo, *fairness*

Abstract

With the increase in the use of social networks in Brazil, existing problems in society began to be widely observed in the online environment, such as hate speech. The hate speech has been intensified by the use of the internet and social networks, being made possible both by the ability to instantly publish content and by the anonymous feeling that users have on social networks. As it is not possible for a human being to analyze each of the content inserted in the internet, more and more new technologies are improved and appear for this purpose, as is the case of the machine learning algorithms that have gained a lot of prominence in the last years with the advancement of computational capacities. With these algorithms it is possible to select a set of data and train a model that can make the prediction of a new inserted data. In this way, this work seeks to understand how racist interactions take place in social networks, based on the search for data in the Twitter API and then build a machine learning model that can identify comments with racist content, as well as list the difficulties that exist in this process.

Key-words: Machine learning, racism, fairness

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

FGA	Faculdade do Gama
UnB	Universidade de Brasília
IBGE	Instituto Brasileiro de Geografia e Estatistica
IE	Extração da informação
NLP	Processamentode linguagem natural
KNN	K - Nearest Neighbor
API	Interface de Programação de Aplicativos
ML	Aprendizado de máquina
COMPAS	Perfil de Gerenciamento Corretivo de Infratorespara Sanções Alternativas
NLTK	Natural Language Toolkit
SVM	Support vector machine

Sumário

1 Introdução

Este capítulo tem como objetivo justificar a escolha o tema (seção 1.1), apresentar os problemas de pesquisa que tornam a escolha do tema válida (seção 1.2), apresentar a questão de pesquisa a ser investigada (seção 1.3) e descrever os objetivos gerais e específicos a serem alcançados (seções 1.4.1 e 1.4.2).

1.1 Justificativa

Em 2017, 74,9% dos domicílios brasileiros tinham acesso à internet. A principal finalidade do uso da internet para os brasileiros é a utilização de mensagens de texto, voz ou imagens por aplicativos diferentes de e-mail. Ou seja, o principal objetivo de quem se conecta à rede é o uso das redes sociais como Facebook, Whatsapp, Instagram e Twitter. (IBGE, 2017)

Com esse grande quantidade de pessoas utilizando as redes sociais para a comunicação, os sites de redes sociais tornaram-se cada vez mais presentes nas estratégias de comunicação e *markenting* das empresas e governos. Contudo, ainda que as redes sociais sejam uma grande aliada na divulgação das informações e opiniões das corporações aos seus *stakeholders*, a instantaneidade com que os comentários positivos e negativos chegam até as plataformas ainda é um grande desafio para quem administra as páginas de Facebook e outras redes sociais. (SALVAGNI, 2018)

A tarefa de detecção de discurso de ódio nas redes sociais é bem recente, assim como a regularização desses crimes online. No Brasil a lei que regulariza os crimes cibernéticos entrou em vigor apenas em 2012 (BRASIL, 2017). Com esta dificuldade na administração dos conteúdos expostos nas páginas, falta de regulamentação e ao abuso de um certo anonimato, as redes sociais começaram a servir como um palanque para que pessoas façam comentários preconceituosos e promovam discursos de ódio contra vários grupos da sociedade (MARTINS, 2014).

O discurso do ódio tem então se intensificado pela utilização da internet e das redes sociais que reduzem, por um lado, a interação social direta entre os atores que passam a ser produtores de mensagens e não apenas receptores, e por outro, potencializam o anonimato e permitem a publicação instantânea de conteúdos com uma velocidade gigantesca (STROPPA, 2015).

Em 2017, algumas redes sociais lançaram diretrizes sobre o discurso de ódio, o Twitter começou a promover ações que estão tentando livrar sua rede social de tal conteúdo. Então, a automatização do reconhecimento do discurso de ódio é um apoio funda-

mental para fazer as redes sociais livre desse tipo de violência (PEREIRA, 2018).

1.2 Problema de Pesquisa

A análise de sentimentos foi tratada como uma tarefa de Processamento de linguagem natural em vários níveis de granularidade. Começando por ser uma tarefa de classificação em nível de documento (TURNEY, 2002) e (AGARWAL B. XIE; PASSONNEAU, 2011), foi tratado no nível de sentença (HU; LIU, 2004b) e (HU; LIU, 2004a), mais recentemente, no nível da frase (AGARWAL; MCKEOWN, 2009) e (HU; LIU, 2005).

Conforme (AGARWAL B. XIE; PASSONNEAU, 2011), os sites de microblog evoluíram para se tornar uma fonte de tipos variados de informações. Isso se deve à natureza dos microblogs nos quais as pessoas postam mensagens em tempo real sobre suas opiniões sobre diversos tópicos, discutem questões atuais, reclamam e expressam sentimentos positivos e negativos pelos produtos que usam na vida cotidiana. Um desafio é criar tecnologia para detectar e resumir um sentimento geral.

O Twitter é uma rede social muito popular, com milhões de mensagens sendo emitidas diariamente, criando então um ótimo *dataset* para mineração de opniões (PAK; PAROUBEK, 2010), conforme (PEREIRA, 2018) o uso de tweets como fonte de dados para análise de sentimentos tem recebido uma crescente atenção.

Recentemente, o estudo das mensagens nas redes sociais não está apenas focado na detecção da polaridade das mensagens, mas também a deteção de discursos de ódio (PEREIRA, 2018).

Desta forma, este trabalho busca obter um conjunto de dados da rede social Twitter com a finalidade de analisar se é possível construir um modelo de aprendizado de máquina que possa identificar *tweets* que estejam promovendo discursos de ódio, assim como elencar as dificuldades e as melhores estratégias de classificação natural de linguagem e análise de sentimentos, verificando qual ou quais modelos tem o melhor resultado para a classificação destes dados.

1.3 Questão de Pesquisa

O objetivo deste trabalho é responder a seguinte questão: É possível construir um modelo de aprendizado de máquina que possa identificar comentários racistas no *Twitter*, quais são as dificuldades inerentes à criação deste modelo, e qual ou quais são os algorítmos de *machine learning* que tem o melhor resultado para a categorização de *tweets* que promovem discurso de ódio relacionados ao racismo.

1.4 Objetivos

Os seguintes objetivos gerais e específicos guiaram este trabalho:

1.4.1 Objetivo Geral

Implementar um modelo de categorização de tweets com a utilização de aprendizado de máquina, de forma a auxiliar a identificação de publicações que estejam promovendo discurso de ódio, para maior rapidez e melhoria na identificação desses conteúdos.

1.4.2 Objetivos Específicos

A fim de atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

- Obter e rotular um conjunto de dados de *tweets* em relação ao teor racista;
- Tratar o conjunto de dados de forma a extrair apenas as informações necessárias para o treinamento dos algoritmos;
- Aplicar diferentes estratégias de classificação ao conjunto de dados categorizados;
- Analisar os resultados das diferentes estratégias utilizadas;

2 Discurso de ódio

Segundo (MARTINS, 2011), o discurso de ódio é composto por dois elementos básicos: discriminação e externalidade. O discurso de ódio é caracterizado pela dicotonia da superioridade do emissor em relação ao atingido (a discriminação), e passa a existir quando é dada a conhecer por outrem que não é o próprio autor (a externalidade). Corroborando com esta ideia (BRUGGER, 2010) nos diz que "o discurso do ódio refere-se a palavras que tendem a insultar, intimidar ou assediar pessoas em virtude de sua raça, cor, etnicidade, nacionalidade, sexo ou religião, ou que têm a capacidade de instigar violência, ódio ou discriminação contra tais pessoas".

Ou seja, o discurso de ódio consiste na externalização, tanto escrita quanto verbal para outrem de palavras que tendem a difundir e estimular o ódio racial, a homofobia, a xenofobia e quaisquer outras formas de ódio baseada na intolerância.

2.1 Racismo

Segundo (SALLOUM1 et al., 2017), o racismo parte da ideia de que existem diferentes raças humanas, cada qual com suas respectivas diferenças e portanto, com algumas superiores às outras. Entretanto, do ponto de vista biológico e social, essas diferenças são um mito que se perpetua erroneamente de forma manter este tipo de pensamento na cabeça de algumas pessoas.

Corroborando com este ponto (MARTINS, 2014) nos diz o racismo é um conjunto de teorias e etnias que estabelecem uma hierarquia entre raças e etnias, onde uma raça se considera superior as outras, sendo assim um preconceito extremado contra indivíduos pertencentes a uma raça ou etnia diferente, considerada inferior.

No Brasil, o racismo se originou no sistema colonial escravista, época em que teve início da exploração do homem pelo homem, vez que a economia se baseava no trabalho forçado de negros e índios. E segundo (SALLOUM1 et al., 2017) este tipo de preconceito vem sendo perpetuado devido a grande diferença social existente no pais.

No âmbito jurídico, o primeiro marco de legislação brasileira em defesa dos descriminados foi a denominada “Lei Afonso Arinos” (Lei nº1390 de 1951) onde os “atos resultantes de preconceitos de raça ou de cor” começaram a ser tratados como infração penal (SANTOS, 2004). Entretanto, esta lei foi objeto de muitas críticas dado as suas penas brandas. Por fim, em 05 de janeiro de 1989 foi promulgada a denominada “Lei Caó” (Lei nº 7.716/89) que tornou crime os "atos resultantes de preconceito de raça ou cor".

A "Lei Caó"recebeu alguns aperfeiçoamentos desde então através das Leis nºs. 8.081/90, 8.882/94 e 9.459/97, esta última inclui uma norma penal incriminadora “Praticar, induzir ou incitar a discriminação ou preconceito de raça, cor, etnia, religião ou procedência nacional. Pena – reclusão, de um a três anos e multa” (SANTOS, 2004).

Segundo (NUNES, 2010), hoje em dia, as pessoas de modo geral se dizem contra o racismo e dizem que esse racismo claro e tradicional tem que ser combatido. Entretanto a condenação do racismo pela maior parte da sociedade não é sinônimo de sua inexistência, com o passar do do tempo pode-se notar uma mudança na manifestação do racismo, onde ele não se mostra mais de forma clara e aberta, mas sim de forma mais sutil e disfarçada, para então ser justificado e não adimitido como tal. Então este racismo disfarçado, ou sutil, é o mais praticado na sociedade atualmente. Entretanto, ainda existem inúmeros casos de uma manifestação explícita, principalmente em redes sociais. Esse tipo de racismo é o que será avaliado pelos modelos de aprendizado inteligente desenvolvidos neste trabalho.

2.1.1 O racismo nas redes sociais

As redes sociais vem se tornando um grande centro de divulgação de informações e propagação de ideias, devido a fatores como a sensação de anonimato, velocidade de difusão das informações e recente facilidade de se adquirir um plano de internet hoje em dia. Com esses fatores não é de se espantar que este meio também seria utilizado para pessoas propagarem discursos de ódio (MARTINS, 2011).

Como já citado na introdução deste trabalho, a tarefa de detecção e punição desses crimes online é muito recente, e com isso ainda é possível se ver muitos casos em que as pessoas não sofrem consequências em relação à esses crimes cometidos em chats, redes sociais e jogos online e utilizam esses meios sem medo de sofrerem sanções legais.

O enfoque dessa dissertação é a verificação desses discursos de ódio explícitos em redes sociais, como por exemplo os *tweets* das imagens 1, 2 e 3, onde um *youtuber* famoso faz diversos comentários racistas no *Twitter*.

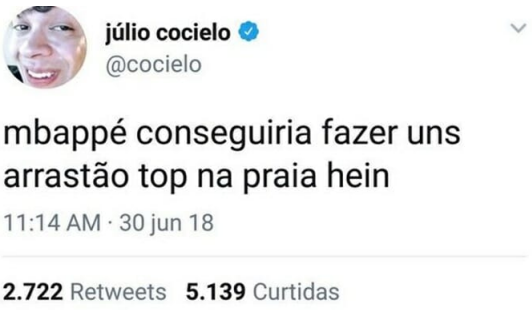

Figura 1 – Tweet com conteúdo racista feito por um youtuber famoso. Fonte: Twitter

Figura 2 – Tweet com conteúdo racista feito por um youtuber famoso. Fonte: Twitter

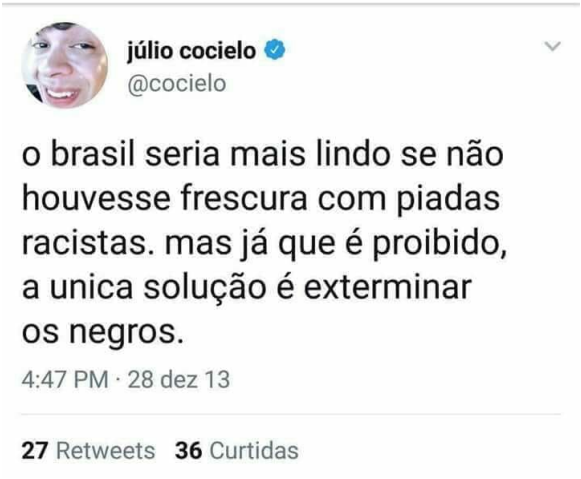

Figura 3 – Tweet com conteúdo racista feito por um youtuber famoso. Fonte: Twitter

Esse racismo explícito também pode ser visto atingindo pessoas famosas, como Maria Júlia Coutinho, que foi constantemente ofendida ao apresentar a previsão do tempo no Jornal Nacional (Figura 4).

No âmbito deste trabalho, é importante utilizar esses dados para formar um dicionário de palavras que são utilizadas pelas pessoas racistas para propagar seu discurso de ódio, este dicionário será importante na hora de se construir os modelos de aprendizado inteligente.

Figura 4 – Comentários com conteúdo racista feitos no facebook para atingir a aprensentadora Maju Coutinho. Fonte: Facebook

3 Categorização de texto

3.1 Mineração de textos

Segundo (SALLOUM et al., 2017), a mineração de textos é um domínio de conhecimento que foi incorporado em vários campos de pesquisa, como linguística computacional, recuperação de informação e mineração de dados. A mineração de texto tem a capacidade de ler uma forma não estruturada de dados para fornecer padrões de informações significativos.

E (TAN, 1999) nos diz que a Mineração de textos é um termo que se refere ao processo de extração de padrões ou conhecimento não-trivial, interessante e previamente desconhecido de documentos de texto.

Segundo (MINER et al., 2012), a mineração de textos pode ser dividida em sete áreas de atuação: Pesquisa e recuperação da informação, Clusterização de documentos, Classificação de documentos, Web Mining, Extração da informação (IE), Processamento de linguagem natural (NLP) e Extração de conceitos. A melhor alternativa de mineração de textos para este trabalho é a classificação de documentos, pois o presente trabalho pretende classificar *tweets* em categorias pré-definidas.

3.2 Categorização de texto

Segundo (PAWAR; GAWANDE MEMBER, 2012), a categorização ou classificação de texto é uma tarefa voltada para a classificação de padrões para mineraração de texto, sendo esta uma tarefa necessária para o controle das informações textuais. Ela refere-se ao processo de atribuir uma ou mais categorias a um texto dentre categorias predefinidas (PAWAR; GAWANDE MEMBER, 2012) e (CAVNAR W.B. E TRENKLE, 1994).

Dentro deste mesmo ponto de vista de definição de categorização de texto, (CAVNAR W.B. E TRENKLE, 1994) nos diz que a categorização é uma tarefa fundamental no processamento de documentos, facilizando o procesamento e manuseio destes documentos.

Segundo (SEBASTIANI, 2002), a categorização remonta ao início dos anos 1960, mas somente nos anos 1990 que se tornou um importante subcampo de sistemas de informação, devido à disponibilidade de *hardwares* mais potentes. Com o aumento da capacidade computacional e a grande quantidade de dados trafegamento nas redes sociais a categorização desses dados deve ser feita, para então se retirar informações úteis.

Embora a categorização de textos hoje em dia esteja quase sempre correlacionada à

aprendizagem de máquina, ela começou a ser feita no início dos anos 1960 através de uma abordagem baseada na Engenharia do Conhecimento, onde um especialista codificava um classificador através de regras que definiam cada categoria (SEBASTIANI, 2002). Com o avanço da tecnologia, a partir dos anos 1990 a abordagem da aprendizagem de máquina ganhou força e predomina ainda hoje.

Portanto, a categorização de textos atualmente é uma disciplina do campo da aprendizagem de máquina, compartilhando várias características com outras tarefas, como extração de informações e conhecimentos de textos e mineração de textos (SEBASTIANI, 2002).

Para categorização de textos é feito um classificador, que então dado o texto irá classificá-lo em uma das categorias predefenidas.

3.2.1 Tipos de classificadores

Segundo (SEBASTIANI, 2002), cada tipo específico de problema exige diferentes restrições, que devem então ser aplicadas à categorização de texto. O autor ainda cita dois tipos de classificadores de texto: monocategórico e multicategórico. O classificador monocategórico é quando uma categoria deve ser atribuida a cada documento do conjunto de documentos, já o classificador multicategório é capaz de atribuir mais de uma categoria a cada documento. O classificador utilizado neste trabalho deve ser monocategórico, pois uma postagem deve pertencer a exatamente uma categoria.

No ponto de vista da modelagem de um categorizador de textos, segundo (SEBASTIANI, 2002), existem duas maneiras diferentes de se fazer. A primeira é quando deseja encontrar todas as categorias pertinentes a um documento, denominada categorização oriantada a documento, a segunda alternativa é quando se deseja encontrar todos os documentos pertinentes a uma categoria, denominada categorização orientada a categoria. Para o presente trabalho, será feita uma categorização orientada a categoria, pois se deseja encontrar todos os *tweets* da categoria racista.

Segundo (SOARES, 2013), a abordagem mais comum para se fazer um classificador de textos é utilizando técnicas de aprendizagem de máquina, utilizando um processo indutivo para criar um modelo de classificação com base em dados de treinamento pré-rotulados, para classificar novos itens com classe desconhecida.

3.3 Aprendizado de máquina

Segundo (BOSE; MAHAPATRA, 2001, p. 212), o aprendizado de máquina é o estudo de métodos computacionais para tornar a aquisição de conhecimentos automática a partir de exemplos. Corroborando com essa ideia, (MITCHELL et al., 1997) nos diz

que o aprendizado de máquina pesquisa métodos computacionais relacionados à aquisição automática de novos conhecimentos, novas habilidades e novas formas de organizar o conhecimento já existente. A Figura 5 descreve o processo pelo qual a obtenção de um modelo de aprendizado de máquina passa para resolver um problema. Um sistema de aprendizado de máquina é um programa de computador que toma decisões com base em experiências ou exemplos previamentos aprendidos.

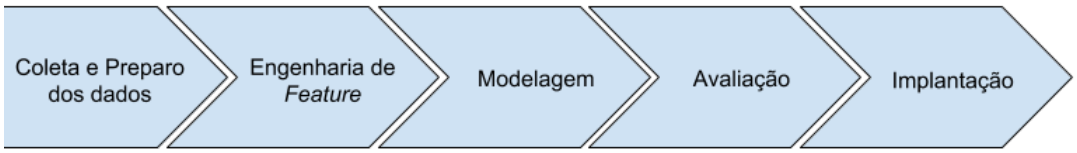

Figura 5 – Processo do aprendizado de máquina.

Para utilizar algoritmos de aprendizado de máquina, é necessário ter acesso aos dados, analisá-los e tratar os dados para que eles estejam limpos. Ou seja, é necessário um pré-processamento dos dados. A partir desses dados, é necessário entender esse conjunto de dados, fazendo uma engenharia de features, analisar a melhor maneira de apresentar esses dados para os algoritmos e então escolher o melhor modelo e algoritmo. Então, é necessário fazer uma avaliação deste modelo, com o objetivo de identificar sua acurácia ao testar novos dados. Por fim, com uma acurácia dentro do esperado, vem a implantação do modelo para a resolução do problema proposto (RICHERT, 2013).

3.3.1 Tipos de aprendizado

O aprendizado de máquina pode ser supervisionado ou não-supervisionado. No aprendizado supervisionado os dados já estão pré-rotulados com classes, então o objetivo é predizer a quais dessas classes novos dados pertencem. Já no aprendizado não-supervisionado, se busca padrões nos dados similares para agrupá-los e identificar possíveis classes. Na Figura 6, pode ser visto os tipos de aprendizado.

3.3.1.1 Aprendizado supervisionado

No aprendizado supervisionado, o algoritmo aprende com base em um conjunto de exemplos de treinamento, no qual todos os exemplos são constituidos de entradas e saídas corretas, esse algoritmo tem como objetivo classificar a qual classe pertence um novo exemplo (PILA, 2001).

No aprendizado supervisionado, todos os dados devem possuir um atributo, que é a classe ou rótulo, este atributo é a meta do que se deseja aprender e poder fazer previsões de novos dados a seu respeito (MONARD; BARANAUSKAS, 2003). Quando esses rótulos possuem valores discretos, é necessário utilizar o método da classificação, já quando estes rótulos possuem valores contínuos, o método a ser utilizado é a regressão.

Figura 6 – Tipos de aprendizado de máquina.

Dentre os algoritmos de aprendizado de máquina utilizados para classificação, temos as redes neurais, árvores de decisao, KNN (L-Nearest Neighbor), para a regressão temos as regressões linear simples e múltipla (LAROSE, 2014).

Um problema comum no aprendizado supervisionado, são os chamados *underfitting* e *overfitting*.

O *underfitting* ocorre quando o modelo não consegue achar relações entre as variáveis do conjunto de treinamento, ocasionando um desempenho muito ruim na acurácia do modelo e com o aumento desse conjunto a melhora no desempenho é muito pequena (MONARD; BARANAUSKAS, 2003).

No *overfitting*, o próprio conjunto de treinamento induz o modelo a ter um bom desempenho no conjunto de treinamento, entretanto, ao ser testado em outros conjuntos o seu desempenho piora (MONARD; BARANAUSKAS, 2003).

De modo a reduzir resultados tendenciosos, é possível serializar os dados em subconjuntos, sendo eles, dados de treinamento, de validação e de teste. Deste modo, o algoritmo vai treinar inicialmente nos dados de treinamento, e os demais dados servirão para uma avaliação posterior da acurácia do modelo (TAVARES; LOPES; LIMA, 2007).

Outro problema que é possível surgir na hora do treinamento dos modelos, é o desbalanceamento das classes onde um conjunto de dados onde as classes não estão igualmente representadas. Este aspecto pode influenciar o modelo criado, e quanto maior for a diferença entre a representação das classes, maior vai ser a valorização da classe dominante em detrimento da classe minoritária (SANTOS, 2016). Segundo (SANTOS, 2016), existem classes de métodos para balancear a distribuição das classes:

(i) *Under-sampling*: Que consiste em balancear as classes com a exclusão dos dados da classe majoritária.

(ii) *Over-sampling*: Que consiste em reproduzir os exemplos da classe minoritária de modo a obter uma distribuição mais igualitária.

3.3.1.2 Aprendizado não supervisionado

No aprendizado não-supervisionado, os exemplos não tem rótulos ou classes, ou seja, não há um rótulo ao qual deseja-se identificar. Os algoritmos recebem um conjunto de dados não rotulados e tenta agrupá-los conforme sua similaridade, esses agrupamentos são chamados de *clusters*. Após a clusterização dos dados, é necessário estudar os resultados e identificar o que cada um dos *clusters* significa (MONARD; BARANAUSKAS, 2003). Alguns dos principais algoritmos de clusterização são o *k-means*, o *clustering* hierárquico e as redes neurais.

3.3.2 Fases do aprendizado

3.3.2.1 Coleta e preparo dos dados

A primeira etapa do processo para se fazer um modelo de aprendizado de máquina é coletar os dados, nesta etapa é normal se ter previamente acesso a um banco de dados com essas informações. Entretanto, há casos em que você deverá acessar esses dados a partir de outro banco de dados, como por exemplo um API de uma rede social, deste modo você precisará coletar esses dados antes de começar a prepará-los.

Após obter esses dados é necessário prepará-los, geralmente em um formato tabular. Esse formato é como uma planilha na qual os dados são distribuídos em linhas e colunas, onde cada linha corresponde ao valor exemplo de interesse, ou seja, os valores do rótulos ou features que serão analisados. A maioria dos algoritmos de aprendizado de máquina precisam dos dados dessa forma, para tratá-los (BRINK JOSEPH RICHARDS, 2016).

Durante o preparo dos dados algumas técnicas podem ser utilizadas de acordo com o contexto e propósito do modelo.

(i) Remoção de linhas duplicadas na base de dados: Uma prática muito comum em redes sociais é a republicação do conteúdo de outra pessoa (*repost* ou *retweet*), caso os dados coletados forem acessados a partir de um banco de dados de uma rede social, é possível que venham dados duplicados por conta dessa republicação, então a técnica de remoção de linhas duplicadas pode ser utilizada.

(ii) *Stopwords*: São palavras e termos frequentes que não tem relevância nos dados. Exemplos: um, uma, com, de, da, as, os. É possível aplicar uma técnica para remoção dessas palavras.

(iii) *Stemming*: É a técnica de reduzir uma palavra o seu radical, removendo prefixo e sufixos de uma palavra.

(iv) Remoção de caracteres indesejados: Pontuações e links não adicionam informações extras ao se tratar dos textos, então é possível aplicar uma técnica para remoção dessas informações.

(v) *Lemmatization*: É a técnica que reduz as palavras flexionadas adequadamente, garantindo que a palavra raiz pertença ao idioma. Exemplo: tiver, tenho, tinha tem são do mesmo lema ter.

(vi) *Tokenização*: É o processo de dividir uma string ou texto em uma lista de *tokens*. Exemplo: É possível *tokenizar* a frase 'Este documento está incompleto' em quatro tokens ['Este ', 'documento', 'está', 'incompleto']

3.3.2.2 Engenharia de Feature

Após tratar os dados e escrevê-los de uma forma que os algoritmos de aprendizado de máquina o entendam, é necessário fazer uma engenharia de *feature*. A *feature* é a descrição de alguma característica do exemplo, ou seja, pode existir uma *feature* sexo, cujo os exemplos podem ter valores 'Masculino' ou 'Feminino'.

As *features* podem ter valores de dois tipos: Nominais ou contínuas. As Nominais descrevem valores que não possuem uma ordem entre si, como o exemplo do sexo acima e as contínuas, descrevem valores que possuem uma ordem linear, como, a idade de uma pessoa (MONARD; BARANAUSKAS, 2003).

Para fazer uma construção de um sistema de aprendizado de dados bem-sucedido é preciso analisar as *features* existentes e procurar uma pergunta que possa ser respondida pelos dados existentes. Dentre as features existentes, é possível que existam atributoso que não são diretamente relevantes, ou até mesmo irrelevantes. E existem atributos que tem mais relevância. Ou seja, é necessário escolher as *features* mais relevantes sobre o contexto analisado, para se ter um resultado bem-sucedido (MONARD; BARANAUSKAS, 2003) e (BRINK JOSEPH RICHARDS, 2016).

Todo esse processo de análise das *features* mais importantes e que serão utilizadas, exclusão ou combinação de *features* é a Engenharia de *Features*.

3.3.2.3 Modelagem

Nesta etapa do processo, será definido de acordo com as features existentes e o problema a ser solucionado quais serão os tipos de algoritmos que podem ser utilizados, assim como se esses algoritmos vão ser de aprendizado supervisionado ou algoritmos de

aprendizado não supervisionados. E todos eles tem o objetivo de: estimar a relação funcional entre os recursos de entrada e a variável de destino (BRINK JOSEPH RICHARDS, 2016).

3.3.2.4 Avaliação

Após ajustar os possíveis modelos de aprendizado de máquina para um determinado problema é necessário avaliar a precisão deste modelo, pois diferentes problemas exigem diferentes abordagens e modelos (BRINK JOSEPH RICHARDS, 2016). Não existe um algoritmo que irá apresentar o melhor resultado em todos os contextos (MONARD; BARANAUSKAS, 2003).

Para se fazer a validação desse modelo é possível utilizar uma técnica chamada *cross-validation* ou validação cruzada. Os dois métodos mais comuns de *cross-validation* são o *holdout method* e o *k-fold cross-validation*. No *holdout method*, um subgrupo dos dados, entre 20% a 40%, é setado como conjunto de teste que será utilizado para avaliar a precisão do modelo. Já no *k-fold cross-validation*, os dados são separados em k subgrupos, então todos os dados restantes são utilizados para o treinamento, por fim os dados do subgrupo são utilizados para avaliar a precisão do modelo, este processo se repete para cada um dos k subgrupos independentes (BRINK JOSEPH RICHARDS, 2016).

Se o resultado do modelo for dentro do esperado, é possível então implantar esse modelo para analisar novos dados em produção. Caso este resultado não esteja dentro do esperado e não seja bom o suficiente, será necessário tomar algumas medidas para tentar otimizar a precisão do modelo, como revisar os seus dados e *features* ou reconsiderar o algoritmo (BRINK JOSEPH RICHARDS, 2016).

3.3.2.5 Implantação

Após se ter um modelo que consiga predizer novos exemplos com uma boa precisão, é possível então implatá-lo no seu ambiente de produção.

3.4 *Fairness* em *Machine Learning*

O conceito de *fairness* (justiça) entrou em foco após a criação da Lei dos Direitos Civis dos Estados Unidos em 1964, que efetivamente proibiu a discriminação com base na raça, cor, religião, sexo ou nacionalidade de um indivíduo. Esta lei trouxe disposições importantes que ainda moldam a compreensão do público do que é ser justo ou injusto (HUTCHINSON; MITCHELL, 2019). Estas disposições impedia impedia as agências governamentais que recebem fundos federais (incluindo universidades) de discriminar com base na raça, cor ou nacionalidade; e impedia empregadores com 15 ou mais empregados de discriminar com base em raça, cor, religião, sexo ou nacionalidade.

A partir de então foram levantadas questões na época de que os testes usados para avaliar a capacidade e adequação na educação e no emprego eram discriminatórios em bases proibidas pela nova lei (HUTCHINSON; MITCHELL, 2019). Estimulando então uma grande quantidade de pesquisas sobre como essa injustiça poderia ser medida matematicamente nos testes educacionais e de emprego. Durante o período de 1966 e 1976 originou-se uma grande quantidade de pesquisas feitas sobre *fairness*, que se alinham bastante com as pesquisas de *fairness* em *machine learning* feitas a partir de 2010. Entretanto, durante esse período os resultados das pesquisas foram decepcionantes e a partir do final da década de 1970 esse movimento de pesquisas desapareceu em grande parte (HUTCHINSON; MITCHELL, 2019).

Com redução da barreira para a utilização do *machine learning* assim como a sua democratização, o tema de *fairness* voltou aos holofotes a partir de 2010 com uma intensificação ainda mais forte a partir de 2016, como pode ser visto na Figura 7. Entretanto, desta vez essas pesquisas, em sua maioria, estão diretamente ligadas ao quão justo são os modelos de ML criados (CATON; HAAS, 2020).

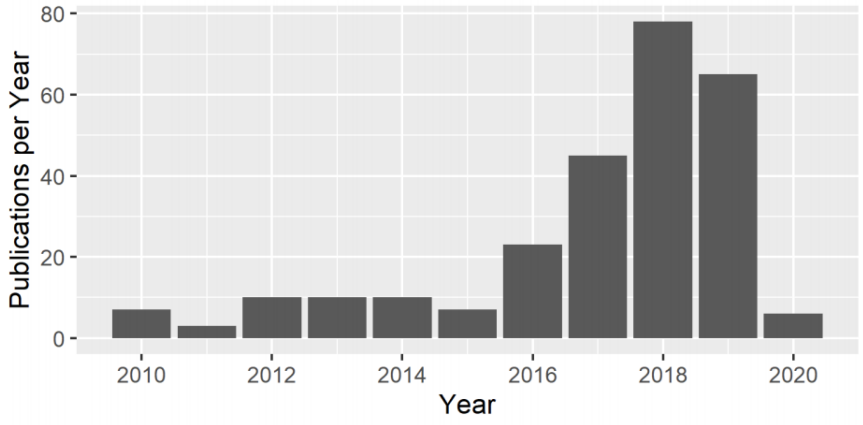

Figura 7 – Número de artigos relacionados a *fairness* em ML. Fonte: (CATON; HAAS, 2020)

De acordo com (CATON; HAAS, 2020), essa democratização dos conhecimentos de ML não está focando um aspecto muito importante, que é a inclusão de raciocínio ético nos conhecimentos e cursos ligados ao ML. Com isso, está ocorrendo a proliferação da utilização do ML nos setores, sem um devido cuidado relacionado as questões ligadas à justiça desses modelos computacionais, muitas vezes reproduzindo os vieses sociais já existentes na sociedade, sem sua devida subjetividade. Nos Estados Unidos, por exemplo, a taxa de reincidência de crimes é maior na população negra, caso seja feito um modelo utilizando apenas esses dados, esse padrão será aprendido e utilizado, sem olhar para o contexto de repressão da população negra ocorrida no país (RIBAS, 2019), fato este que

pode ser visto no algoritmo COMPAS (Perfil de Gerenciamento Corretivo de Infratores para Sanções Alternativas).

O algoritmo COMPAS, foi elaborado pela empresa Northpointe (hoje com o nome Equivant), com o intuito de realizar avaliações de riscos de pessoas reincindirem na pratica de crimes, para auxiliar a tomada de decisões dos juízes nos tribunais dos Estados Unidos, mitigando riscos futuros. Entretanto, em 2016, o jornal ProPublica (jornal de cunho investigativo) divulgou um estudo sobre este algoritmo, que colocou em dúvida o seu uso, pois foi constatado que o algoritmo era racialmente enviesado, dado que o *score* de avaliação de risco de pessoas negras era maior que o de pessoas brancas, durante este estudo, foi-se observado que os dados eram viciados com informações anteriores, influenciando negativamente as decisões (VIEIRA, 2019).

Por fim, (CATON; HAAS, 2020) nos diz que existem problemas de *fairness* em todas as partes do processo de ML, principalmente no pré-processamento dos dados e na própria construção dos modelos de aprendizado, que necessitam da utilização conhecimentos avançados e custosos para se tornarem mais justos sem piorar a acurácia do algorítmo. Trazendo então desafios gigantescos que ainda não estão sendo amplamente observados na criação dos algorítmos de ML.

4 Metodologia

Este capítulo do trabalho, tem como meta, expor as metodologias utilizadas na pesquisa (seção 4.1), o planejamento da pesquisa (seção 4.2) e as ferramentas utilizadas (seção 4.3). A fim de atingir os objetivos gerais e específicos do trabalho.

4.1 Metodologias de Pesquisa

Segundo (PRODANOV; FREITAS, 2013), dado a composição ideológica presente em todas as pesquisas, elas tem a necessidade de um fundamento teórico e metodológico para terem êxito.

Podem existir diversos tipos de pesquisas, cada tipo com suas próprias singularidades. Segundo (PRODANOV; FREITAS, 2013), existem quatro formas de classificação de pesquisa. São pesquisa quanto à natureza de pesquisa, quanto aos seus objetivos, quanto aos seus procedimentos e quanto à abordagem do problema, sendo os três primeiros tipos citados como formas clássicas de classificação de pesquisa, conforme figura 8.

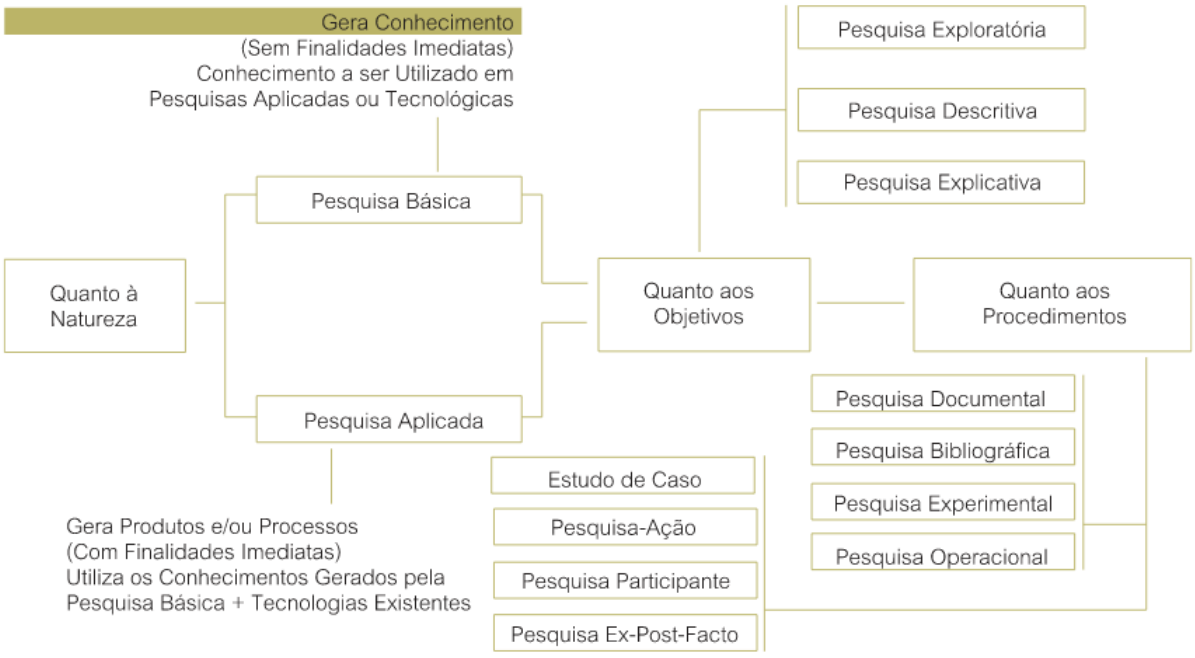

Figura 8 – Formas clássicas de classificação científica (PRODANOV; FREITAS, 2013).

Do ponto de vista da sua natureza a pesquisa pode ser básica ou aplicada. Uma pesquisa básica tem o objetivo de gerar novos conhecimentos úteis para o avanço da ciência sem a aplicação prática, já a pesquisa aplicada tem o objetivo de gerar conhecimentos para

aplicação prática focado na solução de problemas específicos (PRODANOV; FREITAS, 2013).

Do ponto de vista de seus objetivos a pesquisa pode ser exploratória, descritiva ou explicativa. A pesquisa exploratória tem como finalidade adquirir mais informações sobre o assunto que será investigado, possibilitando sua definição e delineamento, ou seja, tem o objetivo de facilitar a delimitação do tema da pesquisa, a pesquisa descritiva busca registrar e descrever fatos sem a interferência do autor e a pesquisa explicativa visa explicar o motivo das coisas e suas causas, através do registro, análise, classificação e interpretação dos fenômenos observados, ou seja, tem o objetivo de identificar fatores que determinam ou contribuem para a ocorrência de fenômenos (PRODANOV; FREITAS, 2013).

Em relação aos procedimentos técnicos, ou seja, o modo como são obtidos os dados necessários para a elaboração da pesquisa, segundo (PRODANOV; FREITAS, 2013) é necessário traçar um modelo conceitual e operativo, denominado *design*. O elemento mais importante para a identificação deste *design* é o procedimento adotado para a coleta de dados. Existem dois grandes grupos de *design*: os que utilizam de fontes de papel (pesquisa bibliográfica e documental) e os que utilizam dados fornecidos por pessoas (pesquisa experimental, pesquisa ex-postfacto, o levantamento, o estudo de caso, a pesquisa-ação e a pesquisa participante).

Do ponto de vista da abordagem do problema, a pesquisa pode ser quantitativa ou qualitativa. A pesquisa quantitativa parte da ideia de que tudo pode ser quantificável, ou seja, utiliza da tradução de opiniões e informações em números para classificá-las e analisá-las através de técnicas estatísticas, já a pesquisa qualitativa diz que existe uma subjetividade no sujeito que não pode ser quantificável, ou seja, traduzido em números (PRODANOV; FREITAS, 2013).

O presente trabalho se classifica como uma pesquisa aplicada do ponto de vista de sua natureza, exploratória do ponto de vista de seu objetivo, utiliza como procedimentos técnicos a pesquisa bibliográfica e o estudo de caso e tem tanto características qualitativas quanto quantitativas do ponto de vista da abordagem do problema.

4.2 Planejamento da Pesquisa

Segundo (PRODANOV; FREITAS, 2013), pesquisa é a construção de conhecimento original de acordo com as exigências científicas. Para que sejam cumpridas essas exigências, são observados alguns critérios, tais como, de coerência, exigência, consistência, originalidadde e objetivação. Uma pesquisa deve ter uma pergunta a que devemos responder, dados para se chegar à resposta e a indicação da confiabilidade da resposta.

Uma pesquisa é constituida trê fases (PRODANOV; FREITAS, 2013):

1. Fase decisória: referente à escolha do tema, à definição e à delimitação do problema de pesquisa;

2. Fase construtiva: referente ao planejamento e à execução da pesquisa;

3. Fase redacional: referente à análise dos dados e das informações obtidas na fase construtiva.

Para a elaboração do planejamento das atividades deste trabalho foi elaborado um fluxo de trabalho (figura 11), para elaboração deste fluxo foi preciso identificar os fluxos de trabalho da mineração de opinião (Figura 9) e do aprendizado de máquina (Figura 10).

De acordo com (HEMMATIAN; SOHRABI, 2017), temos o seguinte ciclo de vida de um projeto de mineração de opinião:

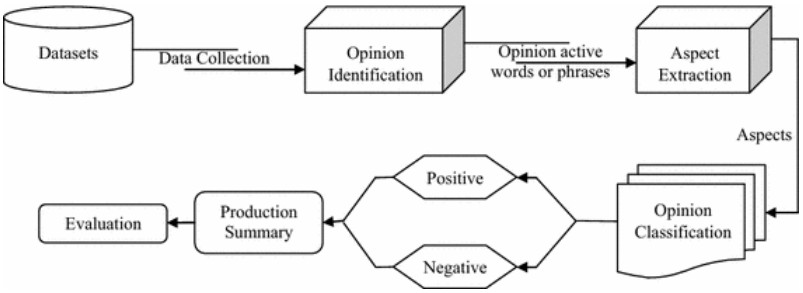

Figura 9 – Fluxo de trabalho para a mineração de opnião (HEMMATIAN; SOHRABI, 2017).

De acordo com os autores (BRINK JOSEPH RICHARDS, 2016) temos o seguinte fluxo básico para o aprendizado de máquina.

A partir dos fluxos de trabalho 9 e 10 e das fases de pesquisa construtiva e redacional, foi elaborado o fluxo de trabalho que será utilizado na elaboração do modelo e da análise da fase redacional.

A fase decisória do trabalho é referente à escolha do tema, à definição e à delimitação do problema de pesquisa deste documento. Já os fluxos da fase construtiva e redacional podem ser visto nas fíguras 12 e 13.

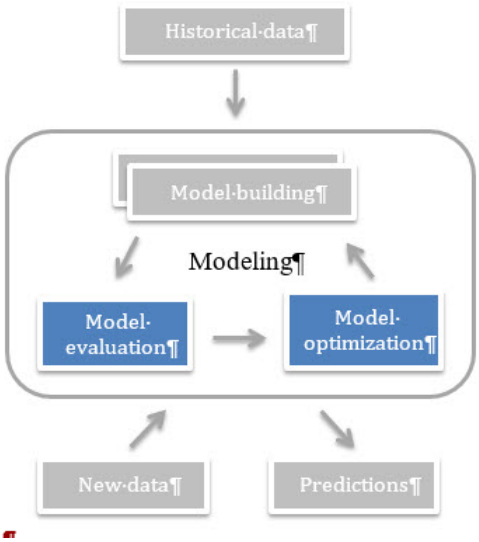

Figura 10 – Fluxo de trabalho para o aprendizado de máquina no mundo real (BRINK JOSEPH RICHARDS, 2016).

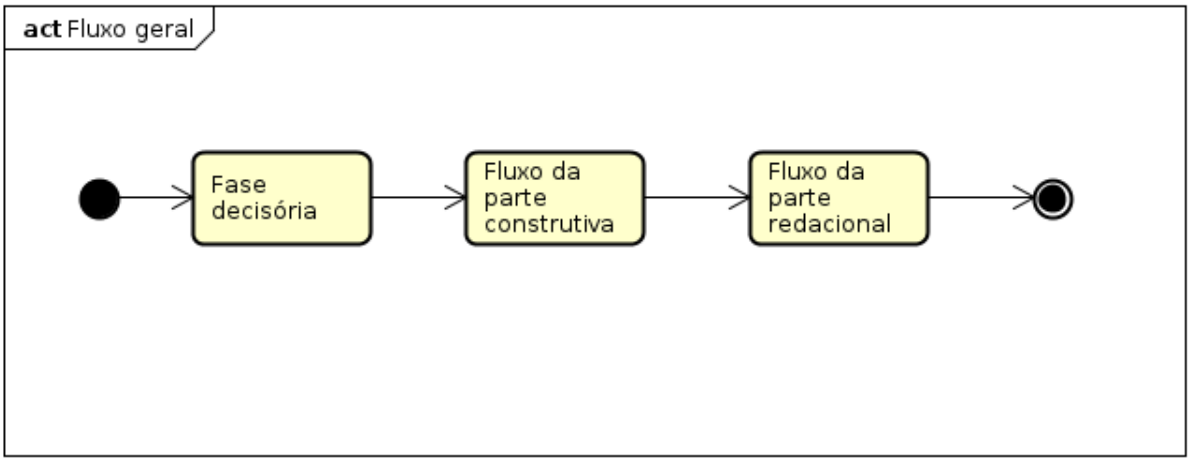

Figura 11 – Fluxo geral do trabalho.

4.2.1 Fluxo da elaboração do modelo

- Entendimento do contexto: Esta atividade é referente à análise do objetivo ao qual o modelo vai ter que satisfazer e identificar os possíveis problemas.
- Análise e escolha das ferramentas: A partir da identificação dos modelos e estratégias já utilizadas para problemas semelhantes, serão selecionadas ferramentas para satisfazer o processo de criação e execução do modelo.
- Estabelecer acesso aos dados: Após a escolha de uma ferramenta para acessar os dados, é necessário montar uma estratégia e algoritmos para começar a coletar os dados, assim como definir o tempo e quantidade de dados necessários para se

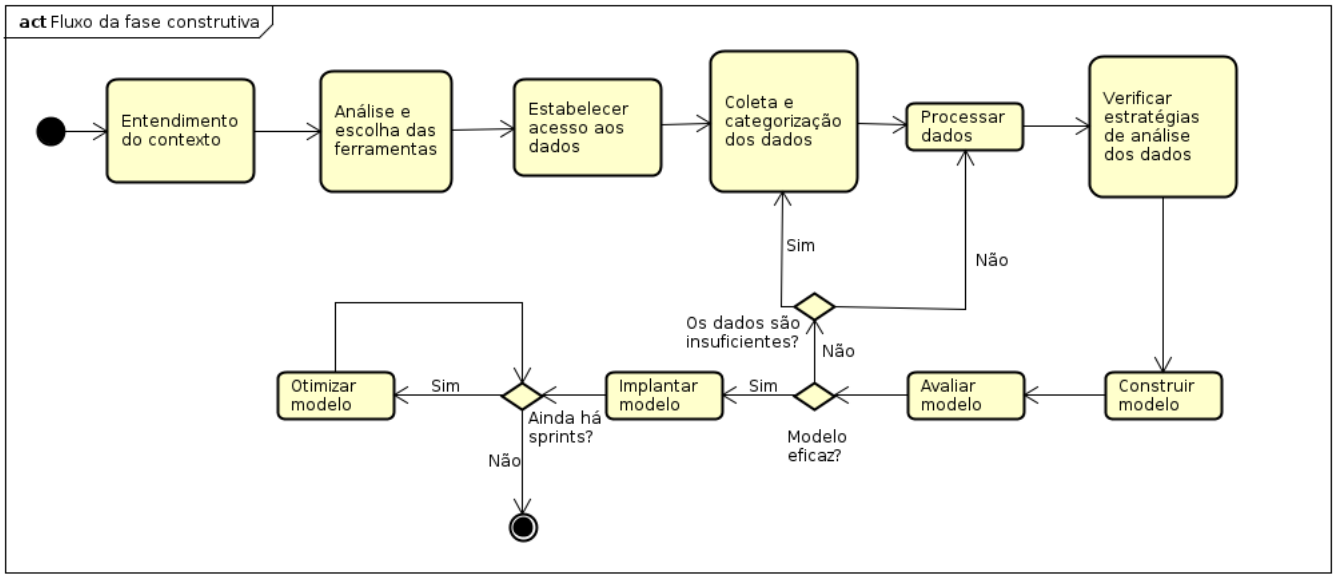

Figura 12 – Fluxo da elaboração do modelo.

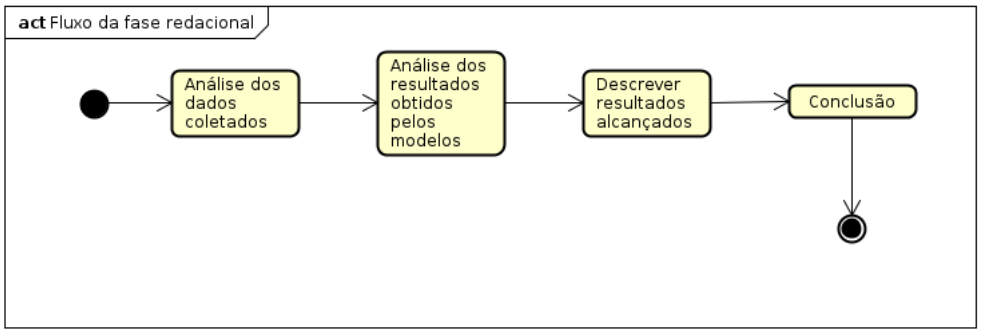

Figura 13 – Fluxo da fase redacional.

começar a treinar os modelos.

- Coleta e categorização dos dados: Esta atividade será para a execução da coleta dos dados durante o tempo definido na fase anterior, após a coleta os dados precisam ser filtrados e categorizados à mão para se aplicar os modelos.

- Processar dados: A partir do momento em que se tiver conjunto de dados categorizados é necessário prepará-lo, através de estratégias de limpeza dos dados, com o objetivo de facilitar a análise.

- Verificar estratégias de análise dos dados: Nesta atividade será feita a análise de quais estratégias de categorização serão mais eficazes para o tipo de dado obtido.

- Construir modelo: Nesta atividade serão construido diversos modelos de acordo com a estratégia de análise definida.

- Avaliar modelo: Nesta atividade será feita uma análise se o modelo satisfaz o objetivo. Caso não satisfaça é necessário identificar os possíveis motivos, caso seja identificado que a quantidade de dados não é o suficiente é necessário coletar mais dados, caso existam dados o suficiente é necessário voltar para o processamento e avaliar mudanças nas estratégias de análise e fases posteriores. Entretanto, segundo (PETEIRO-BARRAL; GUIJARRO-BERDIÑAS, 2013) quanto maior a quantidade de dados melhor será o resultado do modelo, logo para não cair nesse ciclo de melhorar o modelo através do aumento do conjunto de dados, deverá ser observado o conjunto de dados utilizados em projeto similares que obtiveram modelos com um bom desempenho.

- Implantar modelo: Com o sucesso do modelo, é possível começar a fazer uma análise dos *tweets* em tempo real de perfis selecionados e verificar a taxa de acerto dos mesmos.

- Otimizar modelo: A partir do implantação do modelo, novos dados serão avaliados e servirão de insumo para melhoria do modelo.

4.2.2 Fluxo da fase redacional

- Análise dos dados coletados: Nesta atividade será feito uma análise das estratégias utilizadas para coleta, categorização e processamento dos dados.

- Análise dos resultados obtidos pelos modelos: Nesta atividade será feita uma análise dos modelos construidos e seus resultados, assim como seu processo de construção.

- Descrever resultados alcançados: Verificar as melhores estratégias observadas para satisfazer os objetivos.

- Conclusão: Descrever os resultados obtidos, assim como quais objetivos foram alcançados e concluir o trabalho.

4.2.3 Kanban

A metodologia que será utilizada durante os fluxos de elaboração do modelo e redacional é o Kanban. O kanban tem o objetivo de dar visibilidade sobre o desenvolvimento das tarefas, assim como visibilidade ao *backlog* de tarefas com o objetivo de otimizar o tempo de produção (MARIOTTI, 2012). Será contruído um quadro físico onde existirão as atividades necessárias, em andamento e finalizadas, o kanban deve estar sempre atualizado garantindo assim os princípios da metodologia Kanban.

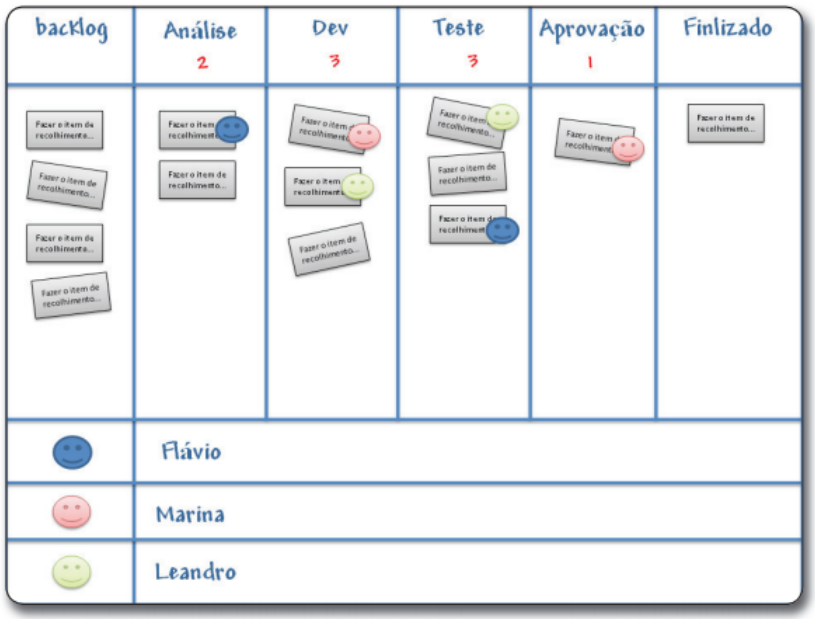

Figura 14 – Ilustração de um Kanban. (MARIOTTI, 2012)

Segundo (MARIOTTI, 2012), o desenvolvimento de um quadro Kanban pode variar de acordo com a equipe, ou seja, a quantidade de fases é variável.

O Kanban utilizado neste trabalho terá as seguintes fases: Backlog, Dev e Finalizado. As atividade dos fluxos da elaboração do modelo e da fase redacional serão restreadas para itens de backlog onde cada um destes itens pode ser quebrado em itens menores.

4.2.4 Licença do projeto

Todo trabalho será realizado como software livre, sob a licença GPL (General Public License) da GNU, versão 3. Disponível no github em maugustoo/twitter_analysis.

4.3 Contexto

Para o desenvolvimento de qualquer trabalho referente ao aprendizado de máquina, o primeiro passo é conseguir acesso aos dados. Neste trabalho o acesso aos dados será feito através da rede social *Twitter*, entretanto neste ponto já são encontrados alguns problemas, como:

- Como não foi encontrado um conjunto de dados em relação ao tema já classificados, vai ser necessário classificar os dados à mão.
- É necessário uma definição bem embasada do que vai ser considerado racismo e não racismo na categorização dos dados.

- Como a categorização vai ser feita por apenas uma pessoa é possível *tweets* serem classificados erroneamente como racistas ou não racistas.

Para mitigar esses problemas, existem fases no processo de desenvolvimento do modelo específicas para desenvolver coisas referentes a estes itens.

Com o acesso devido aos dados, é possível então fazer o processamento dos dados, análisar as possíveis estratégias, aplicá-las e então escolher o modelo que teve o melhor resultado. A partir dai, existe a implantação deste modelo para classificar novos dados. Após a classificação destes dados será feito uma validação do resultado obtido, neste ponto podemos identificar o problema:

- Devido a grande quantidade de dados, será feito a validação apenas em uma quantidade limitada de dados, devido ao tempo para realização da tarefa.

4.4 Ferramentas

As ferramentas estão divididas em dois grupos: As ferramentas auxiliares e as de desenvolvimento. As **ferramentas auxiliares** são ferramentas utilizadas como um apoio para a escrita e mantimento dos insumos necessários para o desenvolvimento do trabalho escrito e da aplicação prática, já as **ferramentas de desenvolvimento** são as ferramentas utilizadas para o desenvolvimento da aplicação prática.

4.4.1 Ferramentas auxiliares

- LaTeX: O LaTeX é um software livre de formatação de texto de alta qualidade que contém recursos projetados para a produção de documentação científica. O LaTeX é o padrão de fato para a comunicação e publicação de documentos científicos. O LaTeX foi utilizado para a construção deste documento.
- Overleaf: Um editor de LaTeX online e fácil de usar. Sem instalação e conta com controle de versões,
- Git: O Git é um sistema de controle de versão distribuído de código aberto e gratuito, projetado para lidar com tudo, de projetos pequenos a grandes, com velocidade e eficiência. Além de ser uma ferramenta de fácil utilização e ótimo desempenho.
- Github: GitHub é uma plataforma de hospedagem de código-fonte com controle de versão usando o Git. Ele permite que programadores, utilitários ou qualquer usuário cadastrado na plataforma contribuam em projetos privados e/ou Open Source de qualquer lugar do mundo. Ela foi escolhida pela visibilidade que o projeto pode ter através da mesma.

- Astah Community: É uma ferramenta para modelagem de processos que foi utilizada para a modelagem dos processos dispostos nesse documento.

4.4.2 Ferramentas de desenvolvimento

- Python: Esta linguagem de programação foi escolhida como linguagem de programação pois já possui uma grande quantidade de bibliotecas prontas que auxiliam na implementação de modelos de Machine Learning e na coleta dos dados do Twitter.
- Scikit-learn: Esta biblioteca para python provem uma série de métodos de aprendizado de máquina, contendo todos os algoritmos que serão utilizados na aplicação prática deste trabalho.
- Tweepy: É uma biblioteca open source para python que provém uma maneira simples e rápida para acesso à API do Twitter.

5 Resultados Finais

Este capítulo tem o objetivo de descrever os resultados obtidos na aplicação da metodologia descrita na seção 4.2.1.

5.1 Objetivo: Implementação de um modelo de categorização de tweets

O objetivo geral do trabalho é fazer a implementação de um modelo de categorização de *tweets*, de forma a auxiliar na rápida identificação de publicações que estejam promovendo discurso de ódio. Para isto, foi necessário implementar quatro modelos distintos de aprendizado de máquina e então análisar qual modelo obteve o melhor resultado de acurácia, como os dados são rotulados esses modelos são caracterizados como modelos de aprendizado supervisionado, e como os valores destes rótulos são discretos, o método utilizado foi de uma classificação.

Para a conclusão do objetivo geral, foram estipulados alguns objetivos específicos, tais como: Obter e rotular um conjunto de dados de tweets em relação ao teor racista, Tratar o conjunto de dados de forma a extrair apenas as informações necessárias para o treinamento dos algoritmos, Aplicar diferentes estratégias de classificação ao conjunto de dados categorizados e Analisar os resultados das diferentes estratégias utilizadas.

O desafios decorrentes de cada um dos objetivos específicos serão falados nos subcapítulos a seguir.

5.1.1 Obter e rotular um conjunto de dados de *tweets*

5.1.1.1 Coleta dos dados

Os dados foram coletados utilizando a biblioteca *tweepy* disponível na linguagem de programação *python*, durante o período de 30 de maio de 2019 a 12 de junho de 2019, como disposto na Figura 15. Os *tweets* foram coletados a partir da observação de alguns perfis de pessoas públicas negras e políticos, observando tanto os *posts* dessas pessoas, assim como a interação dos seus seguidores com esses *posts*. Foram coletados 101737 *tweets* ligados a estes perfis.

A partir desses *tweets* coletados foi feito uma segunda análise para verificar quais *posts* estão dentro do contexto do trabalho. O critério para definir se o *tweet* está dentro do contexto do trabalho foi feito a partir de uma lista de palavras usadas em momentos

onde são feitas ofensas racistas no Brasil, esse conjunto de 34 palavras pode ser visto na figura 16.

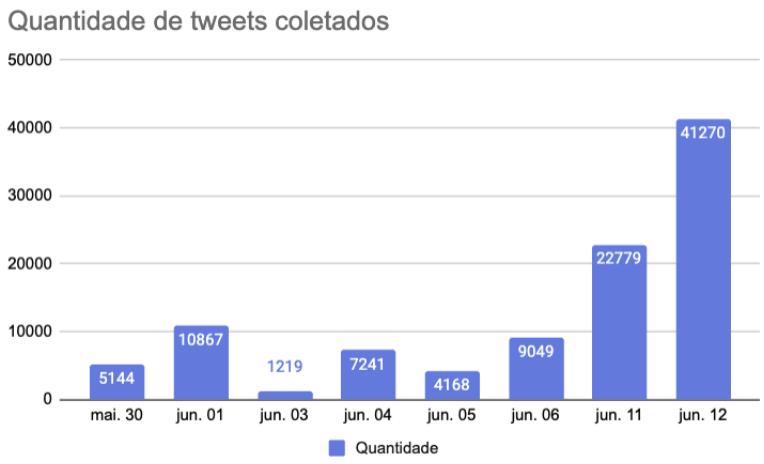

Figura 15 – Quantidade de *tweets* coletados por dia. Fonte: Autor

```
['negraiada', 'negaiada', 'ladr', 'bandid', 'saci', 'racista',
 'racismo', 'escrav', 'senzala', 'neg', 'negr', 'neguin', 'neguim',
 'criol', 'crioul', 'moren', 'morenin', 'morenim', 'preto', 'preta',
 'pixaim', 'pixain', 'macaco', 'macaca', 'urubu', 'cabelo ruim', 'mulat',
 'empregadinha', 'africa', 'nariz', 'beiço', 'crespo', 'cota']
```

Figura 16 – Palavras utilizadas para filtragem do contexto racista. Fonte: Autor

Para esta etapa, apenas um subconjunto dos dados coletados inicialmente foi utilizado, como é possível ver na figura 17.

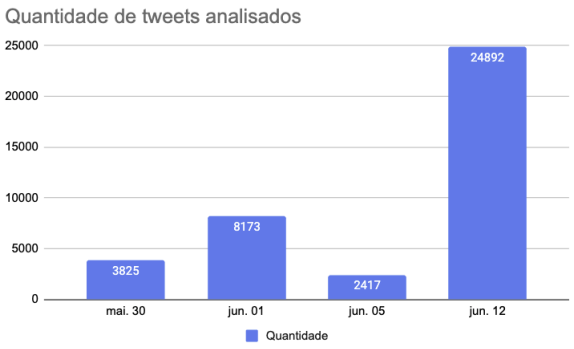

Figura 17 – Quantidade de *tweets* que foram analisados de acordo com o contexto do trabalho. Fonte: Autor

Dos 39307 analisados, foi constatado que 1271 estão dentro do contexto de racismo analisado no trabalho, como pode ser visto na figura 18. Esta filtragem de dados foi feita para que apenas *tweets* contendo um sentido ~ racial (sendo racista ou nao racista), fossem rotulados e considerados na hora do treinamento dos algorítmos.

Figura 18 – Quantidade de *tweets* que foram verificados como dentro do contexto do racismo. Fonte: Autor

5.1.1.2 Rotulagem dos dados

A partir deste conjunto de dados coletados e verificados como dentro do contexto do trabalho visto na figura 18, foi feito um trabalho de rotulagem dos dados à mão. Os *tweets* foram rotulados como racistas ou não racistas de acordo com a descrição da tabela 1. Dos 1271 *tweets* que foram analisados como dentro do contexto do tema racismo, foram encontrados 118 *tweets* racistas, figura 19.

Categoria	**Descrição**	**Quantidade**
Negativo	Nessa categoria encontra-se os tweets que não estão dentro de um contexto racista.	1153
Positivo	Nessa categoria encontra-se os tweets que estão dentro de um contexto racista.	118

Tabela 1 – Descrição das categorias e quantidade de *tweets* em cada uma

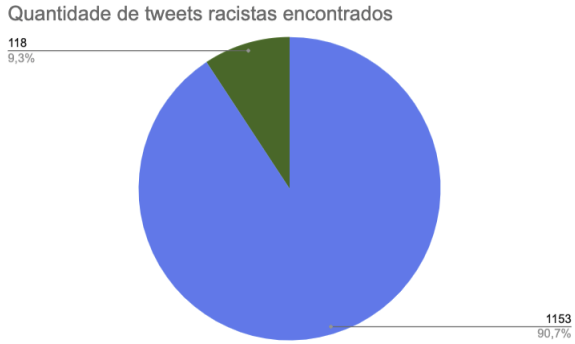

Figura 19 – Quantidade de *tweets* explicitamente racistas encontrados. Fonte: Autor

Alguns exemplos dos dados rotulados podem ser vistos na tabela 2, de acordo com a descrição vista na tabela 1.

Categoria	Tweet
Negativo	"Um uruguaio me chamou de macaco, denunciei o tweet e o que o Twitter fez? Agradeceu pela denúncia .... E nada mais
	"@Oliveiralucas66 @Karine_b24 Tá ligado eu te falo desde cedo, meu cabelo não é bombril é cabelo crespo"
Positivo	@jose_simao Esse macaco do pode ser retardado
	@_Bergonzini volta pra senzala fdp

Tabela 2 – Exemplos de *tweets* rotulados

Ao final da rotulagem dos dados é possível verificar na figura 19 um grande desbalanceamento entre as classes rotuladas, onde mais de 90% dos dados foram rotulados com a categoria negativa, dentro do contexto racismo. Segundo (SANTOS, 2016), esse desbalanceamento de classes pode influenciar o desempenho do modelo criado utilizando aprendizado supervisionado, e para este problema é possível aplicar as seguintes técnicas 3.3.1.1.(i) ou 3.3.1.1.(ii).

5.1.2 Tratar o conjunto de dados de forma a extrair apenas as informações necessárias para o treinamento dos algoritmos

Como foi verificado um grande desbalanceamento entre as classes rotuladas no conjunto de dados (Figura 19), foi preciso aplicar a técnica de *undersampling* citada em 3.3.1.1.(i). Após ser feito o *undersampling*, o conjunto de dados ficou da forma vista na figura 20, consistindo em 123 *tweets* classificados como não racistas e 118 classificados como racistas.

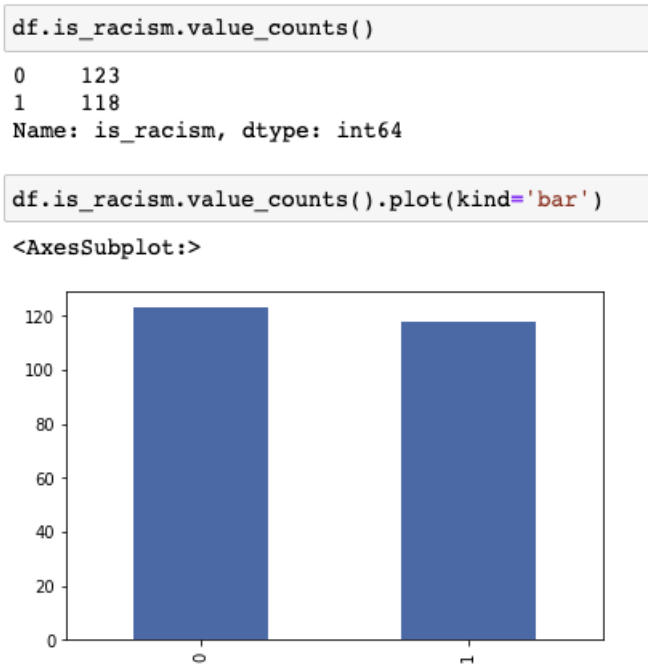

Figura 20 – Conjunto de dados após aplicação do *undersampling*. Fonte: Autor

Como visto em 3.3.2.1.(i), em redes sociais é muito comum a republicação de informações, no caso da *API* do *Twitter* nós temos algumas informações no *JSON* retornado que indicam se essa informação é um *retweet*, uma citação ou extensão a outro *tweet*. Indicados pelas chaves *retweeted_status*, *quoted_status* e *extended_tweet*. Todas essas chaves retornam seus respectívos textos, com essas informações foi possível aplicar a técnica de remoção de dados duplicados juntamente com a coleta dos dados de forma a não guardar os *tweets* duplicados no nosso conjunto de dados. Não sendo necessário aplicar novamente esta técnica na hora de tratar os dados.

O NLTK (*Natural Language Toolkit*) é uma biblíoteca que contém pacotes para fazer a implementação de algumas das técnicas citadas em 3.3.2.1. Entretanto, para o propósito do trabalho apenas as técnicas de remoção de *stopwords*, remoção de caracteres indesejados e *tokenização* foram utilizadas. A técnica de *lemmatization* não foi utilizada pois não estava disponível para o português e a técnica de *stemming* não foi utilizada, pois no português o sufixo das palavras pode ser alterado como forma de ofensa.

Para a remoção das *stopwords* primeiramento foi utilizado o trecho de código mostrado na Figura 21. Entretanto, ao analisar os textos gerados após a remoção das *stopwords* pela biblioteca *NLTK*, foi verificado que algumas palavras que poderiam modificar o significado do texto foram excluidas, como as palavras: Não e palavras relacionadas aos verbos ser e estar. Logo foi preciso visualizar quais eram as palavras utilizadas pela biblioteca (Figura 22 e criar um conjunto que seja adequado ao propósito do trabalho (Figura 23).

```
# Remove Stop Words from database
nltk.download('stopwords')

def removeStopWords(instance):
    stopwords = set(nltk.corpus.stopwords.words('portuguese'))
    words = [ i.lower() for i in instance.split() if not i in stopwords ]
    return ( " ".join(words))

tweets = [ removeStopWords(tweet) for tweet in tweets_with_stopwords ]
```

Figura 21 – Fragmento de código para remoção das *stopwords*. Fonte: Autor

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

Figura 22 – *Stopwords* utilizadas pela biblioteca NLTK. Fonte: Autor

{'estão', 'tivermos', 'tiveram', 'ao', 'houvera', 'isto', 'tenhamos', 'esse', 'num', 'lhe', 'teria', 'muito', 'houver
em', 'estivéramos', 'era', 'aquele', 'houver', 'houverei', 'deles', 'pelo', 'te', 'estivera', 'isso', 'fora', 'qual',
'houvesse', 'houvessem', 'esteja', 'sua', 'tivesse', 'suas', 'das', 'houvermos', 'aqueles', 'nem', 'estejamos', 'éram
os', 'estivermos', 'tivéramos', 'nós', 'tua', 'houveria', 'houvemos', 'fomos', 'tenha', 'este', 'tinha', 'forem', 's
ó', 'haja', 'terão', 'também', 'um', 'teus', 'minhas', 'terei', 'tínhamos', 'uma', 'teve', 'tuas', 'o', 'ou', 'houver
emos', 'fosse', 'tiver', 'minha', 'mesmo', 'numa', 'da', 'meu', 'esteve', 'do', 'for', 'ele', 'os', 'já', 'ela', 'del
a', 'quem', 'nosso', 'me', 'estiveram', 'em', 'seu', 'aquilo', 'hão', 'até', 'houverá', 'houverão', 'mas', 'houvéramo
s', 'estive', 'depois', 'elas', 'tivera', 'delas', 'nossas', 'foram', 'nossa', 'houveram', 'aos', 'eles', 'hajamos',
'pelos', 'esta', 'fossem', 'estas', 'dos', 'entre', 'tenho', 'pelas', 'estivemos', 'é', 'tivemos', 'tivéssemos', 'ser
ia', 'pela', 'tenham', 'mais', 'temos', 'sem', 'eu', 'e', 'no', 'quando', 'teu', 'havemos', 'com', 'houveriam', 'ess
a', 'essas', 'tive', 'tu', 'nos', 'nossos', 'seus', 'dele', 'vocês', 'teriam', 'você', 'às', 'aquelas', 'estivéssemo
s', 'as', 'estes', 'há', 'houveríamos', 'será', 'como', 'à', 'tinham', 'hei', 'têm', 'terá', 'se', 'sejam', 'formos',
'a', 'que', 'foi', 'estava', 'de', 'tiverem', 'está', 'hajam', 'seja', 'estiver', 'lhes', 'estejam', 'eram', 'tem',
'teremos', 'meus', 'por', 'houve', 'teríamos', 'para', 'aquela', 'tivessem', 'vos', 'houvéssemos', 'na', 'nas', 'esse
s'}

Figura 23 – *Stopwords* utilizada para tratar os dados. Fonte: Autor

Após a remoção das *stopwords* foi aplicado a técnica de remoção de links e caracteres indesejados, que pode ser vista na figura 24. Na tabela 3, pode ser visto alguns exemplos de *tweets* antes e depois do tratamento dos dados.

```
# Remove caracteres
def clean_data(instance):
    instance = re.sub(r'http\S+', '', instance)
    instance = re.sub(r'@\S+', '', instance)
    instance = re.sub('\?|\.|\!|\/|\;|\:|\—|\]|\[|\,', '', instance)

    return instance.strip()

tweets = [ clean_data(tweet) for tweet in tweets ]
```

Figura 24 – Fragmento de código para remoção de *links* e caracteres indesejados. Fonte: Autor

	Tweet
Antes	'grande abraço pros fãs do macaco do Michael jackson' '@dioliveira000 Pelo menos não sou racista' '@nalbuquerqueg então pq vc rt que queria 12 desses pastéis sua vaca preta' '@BonitaPreta @PauloTruglio NOJO DESSA LACAIA. É DA MESMA LAIA DO PRETO ESCRAVO DO BOZO.'
Depois	'grande abraço pros fãs macaco michael jackson' 'pelo menos não sou racista' 'então pq vc rt queria 12 desses pastéis vaca preta' 'nojo dessa lacaia é da mesma laia do preto escravo do bozo'

Tabela 3 – Exemplos de *tweets* após a aplicação de técnicas de tratamento dos dados

Também foi feito uma *tokenização* do conjunto de dados. Para fazer a *tokenização* foi utilizado a biblioteca *nltk.tokenize* que contém uma classe especifica para fazer a *tokenização* de *tweets*, figura 25. Também foi o *CountVectorizer* que fornece uma maneira simples de *tokenizar* uma coleção de documentos de texto e criar um vocabulário de palavras conhecidas. Essa técnica nos retorna um vetor codificado que contém o comprimento de todo o vocabulário e uma contagem do número de vezes que cada palavra apareceu no documento. Dessa forma, cada palavra da nossa base se tornou uma coluna da matriz gerada, figura 26.

O modelo foi avaliado com dois *vectorizers* diferentes, figura 25. O primeiro é um *Unigram*, ou seja, apenas palavras únicas são consideradas, o segundo é um modelo *Unigram + Bigrams*, onde além das palavras únicas os pares de palavras também são considerados.

```
tweet_tokenizer = TweetTokenizer()

#vectorizer = CountVectorizer(analyzer="word", tokenizer=tweet_tokenizer.tokenize)
vectorizer = CountVectorizer(ngram_range=(1,2), tokenizer=tweet_tokenizer.tokenize)

freq_tweets = vectorizer.fit_transform(tweets)
```

Figura 25 – Fragmento de código para a tokenização dos dados. Fonte: Autor

```
freq_tweets.shape

(241, 4026)
```

Figura 26 – Contagem de linhas e colunas na matriz gerada pela *tokenização*. Fonte: Autor

5.1.3 Aplicar diferentes estratégias de classificação ao conjunto de dados categorizados

Em uma primeira etapa de criação de modelos de classificação foram criados três modelos simples de classificação utilizando três algoritmos de aprendizagem de máquina supervisionada: o *Naive bayes*, o Support vector machine (SVM) e o *Logistic regression*, que podem ser vistos nas fíguras 27, 28 e 29 respectivamente. E foram analisados os seus resultados utilizando dois *vectorizers* diferentes, um *Unigram* e outro *Unigram + Bigrams*.

```
# Naive Bayers Model

model = MultinomialNB()
model.fit(freq_tweets, classes)
```

Figura 27 – Fragmento de código para o modelo *Naive bayes*. Fonte: Autor

```
# SVM Model

model = svm.SVC()
model.fit(freq_tweets, classes)
```

Figura 28 – Fragmento de código para o modelo *Support vector machine*. Fonte: Autor

```
# Logistic Regression Model

model = LogisticRegression()
model.fit(freq_tweets, classes)
```

Figura 29 – Fragmento de código para o modelo *Logistic regression*. Fonte: Autor

Também utilizado um modelo conténdo dois modelos, o *DistilBERT* e um modelo básico de *Logistic regressions*. O *DistilBERT* processa a frase e passa algumas informações extraídas dela para o próximo modelo. O *DistilBERT* é uma versão menor, rápida, leve e de código aberto feita pela equipe da *HuggingFace* baseando-se na arquitetura *Bidirectional Encoder Representations* (BERT), que tem um desempenho que se aproxima do BERT. O próximo modelo, é o mesmo modelo básico utilizado anteriormente e visto na figura 29, que pegará o resultado do processamento do *DistilBERT* e classificará a frase como positiva ou negativa.

5.1.4 Analisar os resultados das diferentes estratégias utilizadas

Para analisar os resultados de cada uma das estratégias foi utilizado uma validação cruzada do tipo *k-fold cross-validation*, que consiste em separar os dados em de treinamento e de teste. A *k-fold cross validation* consistiu em separar os dados em 10 subgrupos diferentes, onde são feitas 10 iterações, cada iteração pega um destes subgrupos e os utiliza como dado de teste para o restante, como pode ser visto na figura 30.

Foi utilizado a biblioteca *cross_val_predict* do *sklearn.model_selection* para a criação da validação, como pode ser visto na figura 31.

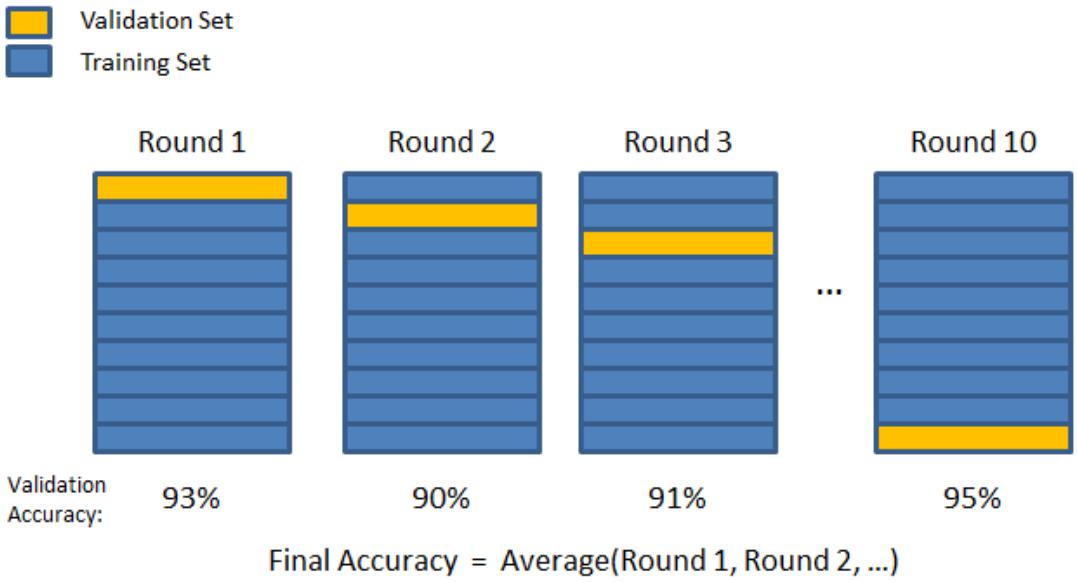

Figura 30 – Ilustração da *k-fold (10-fold) cross validation*. Fonte: Machine Learning for Protein Function - Scientific Figure on ResearchGate Available from: https://www.researchgate.net/ [accessed 9 Apr, 2021]

```
# Cross Validation

results = cross_val_predict(model, freq_tweets, classes, cv=10)

print(metrics.accuracy_score(classes, results))
```

Figura 31 – Fragmento de código para a criação da validação cruzada. Fonte: Autor

A partir do resultado obtido na validação foi criado a tabela 4 contendo os resultados da média das 10 iterações da validação cruzada realizada. É possível perceber que o valor obtido varia entre 65% a 81.74%.

Modelo	**Acurácia Unigram**	**Acurácia Unigram + Bigrams**
Naive Bayes	0.7800	0.6929
SVM	0.7676	0.7303
Logistic Regression	0.8174	0.8133
DistilBERT + Logistic Regression	0.65	

Tabela 4 – Média de acurácia da validação cruzada. Fonte: Autor

Baseado nesta taxa de acerto, o algorítmo de regressão logística simples teve o melhor resultado tanto em *Unigram*, como em *Unigram + Bigrams*, tendo em vista a nossa base de dados. É importante ressaltar que essas porcentagens de acurácia dos modelos criados pode mudar conforme o número de dados cresça, uma vez que aumentar a quantidade torna o treinamento mais eficiente. Esta base de treinamento é uma prova para demonstrar concentualmente o funcionamento dos modelos, entretanto é possível perceber que há uma grande limitação para o aprendizado dos classificadores, pois ao se testar novos exemplos é possível perceber que o modelo ficou tendencioso a certas palavras, independente do contexto, vide figura 32, onde na primeira linha dos resultados temos a classificação da frase (1 - Racista, 0 - Não racista) e na segunda linha temos o que levou ao algorítmo a classificar isso, onde segundo o modelo, quanto mais próximo de um a coluna estiver, maior é a chance dessa entrada ter tal classificação. Logo, ainda podemos verificar que o modelo não só está classificando os dados erroneamente, como está dando uma alta probabilidade (0.9 e 0.86) das frases serem racistas.

```
tests = [
    'eu vi um macaco subindo na árvore',
    'meu cabelo é preto'
]

freq_testes = vectorizer.transform(tests)

print(model.predict(freq_testes))
model.predict_proba(freq_testes).round(2)
```

```
[1 1]

array([[0.1 , 0.9 ],
       [0.14, 0.86]])
```

Figura 32 – Saidas para novos exemplos pro modelo *Logistic regression* treinado. Fonte: Autor

5.1.5 Problemas na construção

Durante a construção do modelo houve alguns fatos que dificultaram sua construção ou até mesmo impediram. Estes fatores podem estar diretamente ligados a alguns problemas de *fairness* que são citados na literatura.

O primeiro grande problema encontrado foi a falta de uma base de dados para o tema, fato este que demonstra mais uma vez a falta de relevância que a sociedade o tema. Uma vez que não existam base pré-rotuladas de comentários racistas, a construção de modelos pra abordar o tema se torna muito mais difícil. Reforçando o racismo estrutural que modela a sociedade até hoje, onde a negação do tema não só não contribui para a solução do problema assim como ainda o dificulta. E esta negação do tema se dá tanto quando as pessoas falam que o racismo não existem, assim como a partir do momento em que as pessoas não buscam juntar dados para estudar e expor os problemas inerentes à isto.

Além dos problemas ligados ao racismo, ainda temos problemas ligados à própria lingua portuguesa que podem dificultar a evolução de trabalhos de aprendizado de máquina, estes problemas são relacionados às técnicas de tratamento dos conjuntos de dados que não conseguem abrangir bem o idioma *pt-BR*, onde a técnica de *stemming* para o contexto de analise sentimental dos comentários no português pode mudar o sentido da palavra analisada a partir da sua redução ao radical da palavra em que pode alterar o sentido de ofensas que são feitas usando o diminutivo de algumas palavras e a técnica *lemmatization* que não estava disponível para o português na biblioteca nltk.

6 Conclusão

O objetivo deste trabalho foi identificar mensagens racistas em português no *Twitter* através da utilização de algoritmos de aprendizado de máquina supervisionados. Para isto, foram utilizados três algorítmos: o SVM, o *Naive Bayes* e regressão logística.

Quanto ao que se refere à parte da obtenção dos dados, a API do *twitter* se mostrou uma forma simples e *friendly* de se coletar dados, onde é possível fazer a filtragem por perfís, palavras chaves, linguagem ou país. Entretanto, como não existiu uma base de dados pré-rotulada para se utilziar no trabalho, a maior dificuldade encontrada foi em se fazer a rotulagem à mão dos dados. Este problema, é o que mais afetou no resultado final do trabalho, pois não se teve uma base de dados grande o suficiente para se obter um resultado não enviesado na base de treinamento.

Em relação ao tratamento dos dados, a linguagem de programação *python* auxiliou bastante com sua facilidade de tratamento de *strings* e juntamento com a biblioteca NLTK foi possível utilizar algumas técnicas de tratamento dos dados, como a tokenização e remoção de *stopwords*. Entretanto, neste ponto é importante relatar que existe uma dificuldade adicional para se utilizar algumas das técnicas da biblioteca no idioma Português, o que demonstra um pouco mais o descompromisso de se construir projetos para outros idiomas diferentes do inglês.

A utilização do *undersampling* após o tratamento inicial dos dados conseguiu desenviesar um pouco o modelo para o conjunto de exemplos, pois com uma grande quantidade de dados com resposta positivas no modelo, enviesam o modelo a um ponto que ainda que ele erre todos os dados negativos, existirá uma taxa de acurácia grande. Dessa forma, foi observado que a aplicação desta técnica foi satisfatória para o propósito do trabalho.

Para o treinamento dos modelos, foi utilizado a biblioteca do *Scikit-Learn*. O *Scikit-Learn* se provou uma biblíoteca bem completa, onde foi possível utilizá-la para a construção de diversos modelos de aprendizado de máquina, tanto utilizando uma estratégia Unigram, quanto Unigram + Bigrams. Além da construção do modelo, a própria biblioteca do *Scikit-Learn* nos dá uma maneira fácil de se utilizar diferentes modos de validação do modelo e também verificar sua acurácia, onde foi possível utilizar a *k-fold cross-validation* com 10 subconjuntos de treinamentos.

Os algoritmos Naive Bayes, SVM e Logistic Regression mostraram Unigram um desempenho satisfatório, todos com uma taxa de acurácia acima de 76%, como pode ser observado na tabela 4. Entretanto, ao se fazer testes com novas entradas o modelo se mostrou bem enviesado pelos dados de treinamento. Logo, apesar da taxa de acurácia ter

sido satisfatória para o auxílio da identificação de comentários racistas nas redes sociais, o que pode ser concluido é que existe um problema na comunidade de software que não se atenta ao tema do racismo e nem quer auxiliar a buscar formas de demonstrar que o problema existe e deve ser estudado, tanto pela polêmica envolvida, quanto pelo público que se beneficia deste tema ainda não estar amplamente inserido neste meio social. Com isso, não se desenvolvem trabalhos sobre o tema, não existindo assim uma quantidade de base de dado rotuladas para se fazerem estudos e com está não existência de conjunto de dados a criação de modelos para expor esses conteúdos é desestimulada.

Referências

AGARWAL B. XIE, I. V. O. R. A.; PASSONNEAU, R. Sentiment analysis of twitter data. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop on Languages in Social Media*. [S.l.], 2011. p. 30–38. Citado na página 19.

AGARWAL, F. B. A.; MCKEOWN, K. R. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL '09*. [S.l.], 2009. p. 24–32. Citado na página 19.

BOSE, I.; MAHAPATRA, R. K. Business data mining—a machine learning perspective. *Information & management*, Elsevier, v. 39, n. 3, p. 211–225, 2001. Citado na página 26.

BRASIL. Lei nº 12.737, de 30 de nov. de 2012. nov 2017. Citado na página 17.

BRINK JOSEPH RICHARDS, M. F. H. *Real-World Machine Learning. 1st. ed.* [S.l.]: Manning Publications Co., 2016. Citado 6 vezes nas páginas 9, 29, 30, 31, 37 e 38.

BRUGGER, W. Proibição ou proteção do discurso do Ódio? algumas observações sobre o direito alemão e o americano. *Direito Público*, v. 4, n. 15, 2010. ISSN 2236-1766. Disponível em: <https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/1418>. Citado na página 21.

CATON, S.; HAAS, C. *Fairness in Machine Learning: A Survey*. 2020. Citado 3 vezes nas páginas 9, 32 e 33.

CAVNAR W.B. E TRENKLE, J. M. N-grambased text categorization. In: LAS VEGAS, NV. *3rd Annual Symposium on Docu-ment Analysis and Information Retrieval*. [S.l.], 1994. p. 161–175. Citado na página 25.

HEMMATIAN, F.; SOHRABI, M. K. A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, Dec 2017. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-017-9599-6>. Citado 2 vezes nas páginas 9 e 37.

HU, M.; LIU, B. Determining the sentiment of opinions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 20th International Conference on Computational Linguistics, COLING '04*. [S.l.], 2004. Citado na página 19.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04*. [S.l.], 2004. p. 168–177. Citado na página 19.

HU, M.; LIU, B. Recognizing contextual polarity in phrase-level sentiment analysis. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*. [S.l.], 2005. p. 347–354. Citado na página 19.

HUTCHINSON, B.; MITCHELL, M. 50 years of test (un)fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, Jan 2019. Disponível em: <http://dx.doi.org/10.1145/3287560.3287600>. Citado 2 vezes nas páginas 31 e 32.

IBGE. Pesquisa nacional por amostra de domicílios contínua. *IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA*, 2017. Citado na página 17.

LAROSE, D. T. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2014. Citado na página 28.

MARIOTTI, F. S. Kanban: o ágil adaptativo - revista engenharia de software magazine 45. 2012. Citado 3 vezes nas páginas 9, 40 e 41.

MARTINS, I. C. *O racismo nas redes sociais: o mundo virtual é feito por pessoas de carne e osso*. 2014. [Online; accessed 25-Maio-2019]. Disponível em: <https://www.vvale.com.br/geral/racismo-redes-sociais/>. Citado 2 vezes nas páginas 17 e 21.

MARTINS, R. S. e Andressa Nichel e Carlise Borchardt e A. C. Discurso de ódio em redes sociais: jurisprudência brasileira. *Revista Direito GV*, v. 7, n. 2, p. 445–467, 2011. ISSN 2317-6172. Disponível em: <http://bibliotecadigital.fgv.br/ojs/index.php/revdireitogv/article/view/23964>. Citado 2 vezes nas páginas 21 e 22.

MINER, G. et al. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Elsevier Science, 2012. (ITPro collection). ISBN 9780123869791. Disponível em: <https://books.google.com.br/books?id=-B6amxqygTMC>. Citado na página 25.

MITCHELL, T. M. et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, n. 37, p. 870–877, 1997. Citado na página 26.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: *Sistemas Inteligentes Fundamentos e Aplicações*. 1. ed. Barueri-SP: Manole Ltda, 2003. p. 89–114. ISBN 85-204-168. Citado 5 vezes nas páginas 27, 28, 29, 30 e 31.

NUNES, S. S. *Racism against blacks: a research about a subtle prejudice*. Tese (Doutorado) — Doctoral thesis, Institute of Psychology, University of São Paulo, 2010. Citado na página 22.

PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. jan 2010. Citado na página 19.

PAWAR, P. Y.; GAWANDE MEMBER, I. S. H. A comparative study on different types of approaches to text categorization. In: INTERNATIONAL ASSOCIATION OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY (IACSIT). *International Journal of Machine Learning and Computing, Vol. 2, No. 4*. [S.l.], 2012. Citado na página 25.

PEREIRA, V. G. *Using Supervised Machine Learning and Sentiment Analysis Techniques to Predict Homophobia in Portuguese Tweets*. Dissertação (Dissertação de mestrado) — Fundação Getulio Vargas - Escola de Matemática Aplicada, 2018. Citado 2 vezes nas páginas 18 e 19.

PETEIRO-BARRAL, D.; GUIJARRO-BERDIÑAS, B. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, v. 2, n. 1, p. 1–11, Mar 2013. ISSN 2192-6360. Disponível em: <https://doi.org/10.1007/s13748-012-0035-5>. Citado na página 40.

PILA, A. D. *Seleção de atributos relevantes para Aprendizado de máquina utilizando a abordagem Rough Sets*. Tese (Doutorado) — Masters Dissertation, University of São Paulo, 2001. Citado na página 27.

PRODANOV, C. C.; FREITAS, E. C. de. *Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico-2ª Edição*. [S.l.]: Editora Feevale, 2013. Citado 3 vezes nas páginas 9, 35 e 36.

RIBAS, P. Coragem, especial sxsw. Thoughtworksn Inc, p. 10–11, 2019. Disponível em: <https://www.thoughtworks.com/coragem/5/ia>. Citado na página 32.

RICHERT, W. *Building Machine Learning Systems with Python*. Packt Publishing, 2013. (Community experience distilled). ISBN 9781782161417. Disponível em: <https://books.google.com.br/books?id=C-yglCEcK0sC>. Citado na página 27.

SALLOUM, S. et al. A survey of text mining in social media: Facebook and twitter perspectives. *Advances in Science, Technology and Engineering Systems Journal*, v. 2, p. 127–133, 01 2017. Citado na página 25.

SALLOUM1, S. A. et al. O racismo nas redes sociais: O preconceito real assumido na vida virtual. In: UNIVERSIDADE FEDERAL DE SANTA MARIA - UFSM. *4 Congresso Internacional de Direito e Contemporaneidade*. [S.l.], 2017. p. 1–6. ISSN 2238-9121. Citado na página 21.

SALVAGNI, M. S. e Cristine Nodari e J. Disseminação do ódio nas mídias sociais: análise da atuação do social media. *Interações (Campo Grande)*, v. 19, n. 1, 2018. Disponível em: <http://www.interacoes.ucdb.br/article/view/1535>. Citado na página 17.

SANTOS, C. As diferenças entre o crime de racismo e a injúria qualificada. In: *Revista Consultor Jurídico, São Paulo, v. 1.* [S.l.: s.n.], 2004. Citado 2 vezes nas páginas 21 e 22.

SANTOS, R. M. M. dos. *Técnicas de Aprendizagem de Máquina Utilizadas na Previsão de Desempenho Acadêmico*. 2016. Citado 2 vezes nas páginas 28 e 48.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 34, n. 1, p. 1–47, mar. 2002. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/505282.505283>. Citado 2 vezes nas páginas 25 e 26.

SOARES, F. d. A. *Categorização Automática de Textos Baseada em Mineração de Textos.* Tese (Doutorado) — Doctoral thesis, PUC-Rio, 2013. Citado na página 26.

STROPPA, W. C. R. e T. Liberdade de expressÃo e discurso do Ódio: O conflito discursivo nas redes sociais. In: UNIVERSIDADE FEDERAL DE SANTA MARIA - UFSM. *3 Congresso Internacional de Direito e Contemporaneidade*. [S.l.], 2015. p. 6–21. ISSN 2238-9121. Citado na página 17.

TAN, A. hwee. Text mining: The state of the art and the challenges. In: *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. [S.l.: s.n.], 1999. p. 65–70. Citado na página 25.

TAVARES, L. G.; LOPES, H. S.; LIMA, C. R. E. Estudo comparativo de métodos de aprendizado de máquina na detecção de regiões promotoras de genes de escherichia coli. *Anais do I Simpósio Brasileiro de Inteligência Computacional*, p. 8–11, 2007. Citado na página 28.

TURNEY, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: *CoRR, vol. cs.LG/0212032, 2002.* [S.l.: s.n.], 2002. Citado na página 19.

VIEIRA, L. M. A problemÁtica da inteligÊncia artificial e dos vieses algorÍtmicos: Caso compas. Brazilian Technology Symposium, 2019. Disponível em: <https://www.lcv.fee.unicamp.br/images/BTSym-19/Papers/090.pdf>. Citado na página 33.