Detecção de Insultos Racistas no Twitter

Gabriela Vitória F. de Andrade Sudo[1], **Yuri de Almeida Malheiros Barbosa**[1]

[1]Departamento de Ciências Exatas (DCX) – Universidade Federal da Paraíba (UFPB)
Cep 58297-000 – Rio Tinto – PB – Brasil

{gabriela,yuri}@dcx.ufpb.br

Abstract. *It is undeniable that racism is a serious problem in Brazilian society and that social networks increase the reach of a racist comment. Even knowing that racial injury is a crime, users explicitly expose their bias on the Internet. Given this fact, the present work aims to identify racist messages on Twitter through three supervised machine learning algorithms: Support Vector Machine, Naive Bayes and Logistic Regression.*

Resumo. *É incontestável que o racismo é um grave problema na sociedade brasileira e que as redes sociais aumentam o alcance de um comentário racista. Mesmo sabendo que injúria racial é crime, usuários expõem seus preconceitos de forma explícita na Internet. Diante desse fato, o presente trabalho tem como objetivo identificar mensagens racistas no Twitter através de três algoritmos de aprendizado de máquina supervisionado: Support Vector Machine, Naive Bayes e Regressão Logística.*

1. Introdução

A Internet mudou a forma do ser humano se comunicar. A rapidez da propagação de uma informação e amplitude que ela pode tomar permite a conexão de bilhões de pessoas de diferentes culturas e opiniões em todo o mundo.

Segundo relatório "Digital in 2019: The Americas"divulgado pelas empresas We are Social e Hootsuite [1] [2] , cerca de 45% da população mundial são agora usuários de mídias sociais: um total de 3,5 bilhões de pessoas; no Brasil, cerca de 66% da população é ativa nas mídias sociais. Em 2006, a plataforma Twitter foi criada, sendo hoje uma das maiores redes sociais onde as pessoas podem divulgar qualquer tipo de informação em tempo real. Com posts de até 280 caracteres com suporte para fotos e vídeos, os usuários ficam por dentro dos assuntos mais comentados do momento e o que está acontecendo ao redor do mundo (Twitter, 2019). Apesar de ter como intuito promover a comunicação e disseminação de conteúdos, nos últimos anos, pode-se observar a manifestação de um lado obscuro, onde usuários cometem atos ilícitos, propagam mensagens de conteúdo prejudicial e violam direitos fundamentais dos demais usuários [da Silva et al., 2011].

Muitos dos crimes vistos na Internet já eram e continuam sendo os mesmos praticados no mundo real. A diferença é a sensação de liberdade que a Internet provê, onde

[1]Disponível em: https://wearesocial.com/global-digital-report-2019 Acesso em: 14/jul/2019

[2]"Trabalho de conclusão de curso, sob orientação do professor Yuri de Almeida Malheiros Barbosa submetido ao Curso de Bacharelado em Sistemas de Informação do Centro de Ciências Aplicadas e Educação (CCAE) da Universidade Federal da Paraíba, como parte dos requisitos necessários para obtenção do grau de BACHAREL EM SISTEMAS DE INFORMAÇÃO."

muitos acreditam que seu ato passará impune, como também os que se escondem atrás de perfis falsos acreditando no anonimato. É importante que qualquer tipo de conteúdo que fere os direitos humanos seja denunciado. Em 2018, a ONG SaferNet Brasil processou 128.332 denúncias anônimas de 10 tipos de crimes diferentes; desse total o racismo fica em 5º lugar com 7.959 denúncias envolvendo 2.978 páginas (URLs) distintas das quais 503 foram removidas.

Este trabalho tem como objetivo identificar mensagens racistas na língua portuguesa através de três algoritmos de aprendizagem de máquina supervisionada: Support Vector Machine, Naive Bayes e Regressão Logística. Para isso foi utilizada a plataforma de desenvolvedores do Twitter.

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta a fundamentação teórica com assuntos essenciais para compreensão deste trabalho. A Seção 3 apresenta a metodologia aplicada, descrevendo o processo da coleta dos dados, classificação dos dados e treinamento dos algoritmos. A Seção 4 apresenta os resultados obtidos na análise dos algoritmos. Por fim, a Seção 5 apresenta as conclusões.

2. Fundamentação Teórica

Nesta Seção são abordados assuntos e conceitos utilizados no desenvolvimento deste trabalho que são base para o entendimento do mesmo, tais como: Racismo nas Redes Sociais, A plataforma Twitter e sua API, Aprendizagem de Máquina e Classificadores de Aprendizagem de Máquina Supervisionada.

2.1. Racismo nas Redes Sociais

O racismo segundo Eugênio e Vala (2004, p. 402) "constitui-se num processo de hierarquização, exclusão e discriminação contra um indivíduo ou toda uma categoria social". No Brasil, país miscigenado e com uma população majoritariamente afrodescendente, pode-se observar uma cultura racista advinda da época colonial e escravocrata. Essa marca cultural define padrões de comportamento que implica em uma percepção de sujeito (indivíduo ou grupo) só por ter a cor da pele negra.

No Dossiê Intolerâncias visíveis e invisíveis no mundo digital[3] acredita-se que a popularização das redes sociais ajudou a trazer a discriminação ainda mais a tona. Comentários racistas são ditos diariamente até mesmo de uma forma inconsciente pois estão presentes no vocabulário popular e enraizados na cultura. Frases como "não fala assim comigo, que não sou suas negas", "chuta que é macumba", "tão bonita que nem parece negra" ou "cabelo ruim" (sobre os cabelos crespos) são comuns nas redes sociais.

Diante disso, nota-se que as pessoas só estão reforçando e reafirmando um preconceito que já possuem fora da rede quando propagam algum discurso de ódio na Internet. Por mais que o racismo já existisse antes das redes sociais, elas têm um enorme papel na reprodução e alcance desses discursos de ódio.

2.2. Twitter

O Twitter é uma rede social em formato de microblogging, ou seja, usuários compartilham o cotidiano, como o próprio *slogan* do Twitter sugere (O que está acontecendo?) e é muito

[3]Disponível em: https://www.comunicaquemuda.com.br/dossie/racismo Acesso em: 24/jul/2019

utilizado para fazer broadcast (envio múltiplo) de informação de maneira rápida e a quem interessa.

No Twitter os usuários criam um perfil e a partir dele podem fazer postagem(tweets). Os perfis podem fazer interações no tweet que o usuário publicou por meio de retweets(compartilhar um tweet em seu perfil), comentários e etc. O efeito cascata que ocorre no Twitter ajuda na propagação também do discurso de ódio, onde o Twitter tem trabalhado em cima de uma política contra propagação de ódio e investindo em machine learning e deep learning para tentar ser mais pró-ativo nesse sentido. Jack Dorsey, cofundador do Twitter, em uma entrevista para a série de conferências TED[4] afirma que ”cerca de 38% dos tweets insultantes agora são identificados proativamente por algoritmos de aprendizado de máquina para que as pessoas não precisem denunciá-los. Mas aqueles que são identificados ainda são revisados por humanos[...]”

2.2.1. Twitter Developer Apps

O Twitter possui uma plataforma destinada a desenvolvedores, que contém diversas ferramentas que possibilitam o acesso aos dados. Para acessar as APIs do Twitter é necessário ter uma conta desenvolvedor no Twitter e criar um aplicativo via *Twitter Developer Apps*, onde irá responder um questionário explicando como usará os dados acessados. Por padrão os aplicativos só podem acessar informações públicas no Twitter. Uma vez autorizado, os desenvolvedores tem acesso aos tweets e respostas que podem ser procurados por palavras-chave ou solicitando uma amostra de Tweets de uma conta específica.

2.3. Aprendizagem de Máquina

Aprendizagem de Máquina é um subcampo da Inteligência Artificial, que tem como objetivo construir modelos em que a máquina consegue aprender a partir da experiência, reconhecendo padrões através de dados. Mitchel (1997, p.2) define aprendizagem de máquina como: “Um programa de computador que aprende a partir da experiência E, em relação a uma classe de tarefas T, com medida de desempenho P, se seu desempenho T, medido por P, melhora com a experiência E.”

Segundo Norvig e Russell (2014, p.804) existem três principais tipos de aprendizagem de máquina: Aprendizagem Supervisionada, Aprendizagem não supervisionada e Aprendizagem por reforço. A seguir uma breve explicação sobre cada um deles.

Na aprendizagem Supervisionada tem-se um conjunto estabelecido de dados de treinamento, onde cada exemplo nesse conjunto é composto por um vetor de características e uma resposta associada a ele. Ao observar esse vetor, aprende-se uma função que faz o mapeamento da entrada e saída. [Norvig and Russell, 2014]

Na aprendizagem não supervisionada o agente aprende padrões na entrada, embora não seja fornecido nenhum rótulo ou feedback explícito. [Norvig and Russell, 2014]. Esse tipo de aprendizagem é usada por plataformas para sugerir contéudos para o usuário de acordo com um perfil traçado.

Na Aprendizagem por reforço a máquina aprende a partir de uma série de reforços-recompensas ou punições. [Norvig and Russell, 2014]. A máquina tenta aprender qual a

[4]Disponível em: https://www.ted.com/talks/jack_dorsey_how_twitter_needs_to_change

melhor decisão a ser tomada em determinadas circunstâncias. Esse tipo de aprendizagem é usados em carros autônomos.

A técnica utilizada neste trabalho é a aprendizagem de máquina supervisionada, pois os dados de entrada utilizados no treinamento dos algoritmos estão previamente rotulados.

Na Aprendizagem de máquina supervisionada existem vários algoritmos dos quais foram escolhidos três - Support Vector Machine, Naive Bayes e Regressão Logística - com o objetivo de avaliar a classificação dos tweets.

3. Metodologia

Nesta Seção serão apresentados os passos metodológicos realizados para alcançar os objetivos deste trabalho. Foram 5 etapas: I- criação de uma lista com palavras-chave para a busca dos tweets, II- coleta dos dados, III- Classificação dos dados, IV- Treinamento dos algoritmos e V- Análise e validação dos algoritmos.

3.1. Coleta dos Dados

Os dados foram coletados entre o período de fevereiro 2019 à julho de 2019 utilizando a API do Twitter. Para a coleta foi utilizada uma lista com 22 possíveis palavras ou conjunto de palavras que podem se enquadrar em um contexto racista. O critério utilizado para montar essa lista foram as palavras usadas como ofensas ao pesquisar sobre racismo no Brasil. Dado essa lista, ao todo foram coletados 20.271 tweets. A quantidade de tweets coletados por cada palavra-chave é mostrada na Tabela 1.

Tabela 1. Quantidade de tweets coletados por Palavras-chave

Palavras-chave	Quantidade de tweets	Palavras-chave	Quantidade de tweets
morena	2165	moreno	1674
nego	2050	nega	1831
negra	1972	negro	1408
macaco	1925	macaca	199
preto	187	preta	994
cabelo ruim	1506	cabelo duro	527
moreninha	79	moreninho	31
aquele nego	257	aquela nega	115
so podia ser nego	3	podia ser negro	2
negão	1039	favelado	274
aquela negra	343	aquele negro	1

Pode-se observar que a busca tem um desempenho menor quando a palavra-chave é composta, por exemplo, ”cabelo ruim”, ”podia ser negro”e etc, pois a API procura tweets que contenham as palavras independente da ordem. Na tabela 2 tem-se exemplos de tweets e as palavras-chave correspondentes.

Tabela 2. Exemplo de tweets coletados e palavras-chave correspondentes

aquele negro	ATLÉTICO X BOCA Atleticano, está chegando a hora de soltar aquele grito, de colocar a emoção pra fora. A missão é clara: vestir RUBRO-NEGRO e CANTAR os 90 minutos. https://t.co/vLCUSH9umZ	aquele nego	@anagoncalvezz aquele nego feio kkk meu deus
cabelo ruim	branco de dread e cabelo crespo vcs: linde alternative amei quebrou os padroes negro de dread e cabelo crespo vcs: cruzes vai pentear esse cabelo ruim cara q coisa feia	cabelo ruim	eu pintei o cabelo de preto e nem ficou tão ruim
negro	Mas não basta o dia ser bosta, tenho que ler bosta de preto falando que a culpa do negro estar onde estar é culpa dele mesmo por ser vitimista.	negra	Que negra maravilhosa essa que tá no Altas Horas...

3.2. Classificação dos Dados

A classificação das mensagem como racista ou não racista foi feita manualmente no período de agosto de 2019 e para ajudar na classificação foi utilizada uma lista de apoio com palavras consideradas ofensivas no geral [5]. Ao todo 3.327 tweets foram analisados. Tendo em vista que algumas das palavras-chave possuem outro significado na língua portuguesa os tweets foram analisados de acordo com 4 categorias: “Positivo”, “Negativo”, “Não se enquadra” e “Gírias” explicadas na Tabela 3.

[5]https://gist.github.com/rogersdepelle/d06c25844bcbe5d8c53299eaa795d1a2

Tabela 3. Descrição das categorias e quantidade de tweets em cada uma

Categorias	Descrição	Quantidade de Tweets
Positivo	Nessa categoria encontra-se os tweets que não estão dentro de um contexto racista.	940
Negativo	Nessa categoria encontra-se os tweets que estão dentro de um contexto racista.	126
Não se enquadra	Nessa categoria estão os tweets que a palavra-chave (query) está no usuário "@", a palavra possui outro significado na língua portuguesa ou está se referindo a um personagem	1908
Gíria	Nessa categoria estão os tweets que a palavra-chave (query) está sendo usada como gíria e não se referindo a raça e expressões que são ditas como ditados populares.	353

Tabela 4. Exemplos dos tweets por categoria

Categorias	Tweets
Positivo	Não precisa ser negro para combater o racismo
	O homem negro não foi feito para arrasta correntes e sim para voar no meio da sociedade.
Negativo	seu negro da merda devias morrer queimado cabrao do caralho espero que a puta da tua mae tenha cancro e morra desmembrada corno da merda filho da puta bebe lixivia cabrao https://t.co/moOHkr0SMb
	@InnocentGhoster vc e preto negro fedido
Não se enquadra	@negro_bryam
	qualquer uniforme preto fica a coisa mais linda https://t.co/pyht2jmd2O
	use a 1º letra do seu nome animal: Macaco https://t.co/zWv3BvMBVC
	Ganhei ingresso pra macaca de camarote kkklk
Gíria	Nego acha q os outros e idiota, só olha pra si msm, nada passa batido n , q dar de maluca, mas to vendo tudo, só vou espera tá no erro !
	eu respondendo aquela parada dos 7 pontos errada sem querera menina vai pensar que eu tenho raiva dela e vivo falando KKKK o auge nega

Essa separação em categorias foi feita para que apenas tweets contendo um sentido racial (sendo racista ou não racista) fosse levado em consideração na hora do treinamento dos algoritmos e os demais fossem descartados. Com isso, apenas 32% dos tweets ana-

lisados foram classificados entre positivo e negativo. A figura 1 mostra o percentual das categorias.

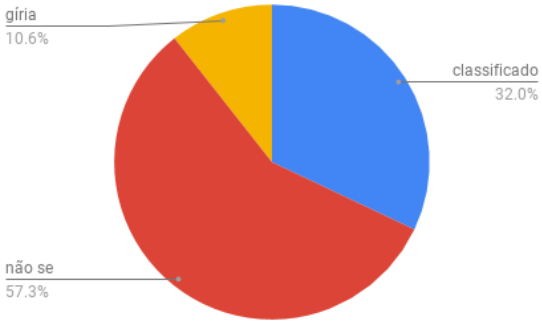


Figura 1. Porcentagem das categorias nos tweets analisados

Ao final da análise dos dados foi constatado um desbalanceamento entre os tweets racista e não racista. Mais de 80% dos tweets foram classificados como positivo, mostrado na figura 2. Esse desbalanceamento segundo Santos (2016) pode influenciar no desempenho do modelo de classificação criado por um sistema de aprendizado supervisionado.

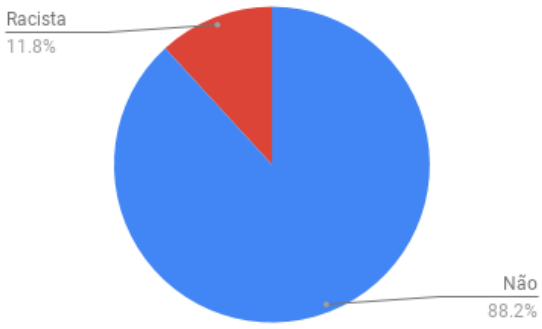


Figura 2. Tweets Racista x Não Racista

3.3. Treinamento dos Algoritmos

Devido a grande diferença dos dados mostrado anteriormente (Figura 2) os algoritmos tornam-se tendenciosos valorizando a classe predominante, neste caso os não racista, afetando no resultado final. Para resolver esse problema foi utilizado o método *under-sampling* que consiste em balancear o conjunto de dados pela eliminação de exemplos da classe majoritária (Santos, 2016). Após o balanceamento dos dados foram treinados três algoritmos de aprendizagem de máquina supervisionada: SVM, Naive Bayes e Regressão Logística.

Os algoritmos citados acima foram implementados utilizando a biblioteca do scikit-learn[6], que consiste em módulo do Python que integra uma ampla gama de algoritmos de aprendizado de máquina levando esse conhecimento para não especialistas, usando uma linguagem de alto nível para uso geral. [Pedregosa et al., 2011]

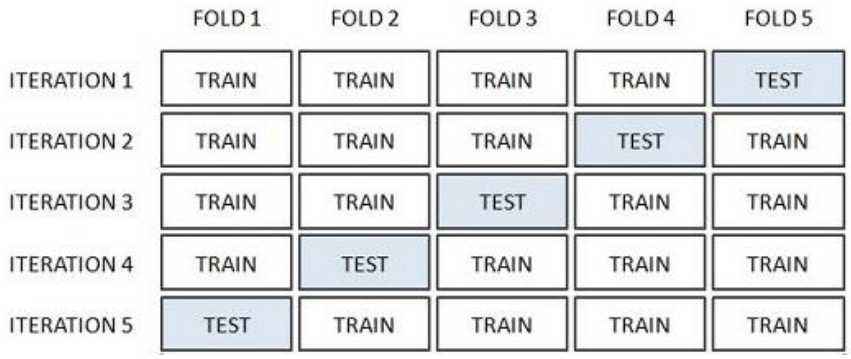
[6]Disponível em: https://scikit-learn.org/stable/ Acesso em: 26/ago/2019

Depois do treinamento é preciso validar os resultados e para isso foi utilizado o método de validação cruzada.

3.4. Validação Cruzada

A validação cruzada consiste em dividir os dados em treinamento e teste. Para validar os algoritmos deste trabalho, a validação cruzada foi executada cinco vezes em cada classificador. Para cada iteração os dados são divididos em cinco subgrupos de forma aleatória, mas mantendo as proporções de cada classe nos subgrupos, onde quatro são para o treino do algoritmo e um para o teste de validação. A figura 3 ilustra como os dados de treinamento e teste são alterados a cada iteração da validação.

Figura 3. Validação Cruzada



Fonte: Revista online How To Dou [7]

4. Resultados

Na Tabela 5, pode ser observado que as porcentagens de acerto dos algoritmos varia entre 48% e 86%. Ao calcular a média da taxa de acerto para cada algoritmo foi obtida a porcentagem de 62.29% para o Naive Bayes e o mesmo valor de 63.14% para Regressão Logística e SVM.

Tabela 5. Porcentagem de acerto de cada algoritmo

Classificadores	1ª iteração	2ª iteração	3ª iteração	4ª iteração	5ª iteração	Média de acerto
Naive Bayes	0.63461538	0.5	0.62	0.84	0.52	62.29%
Regressão Logística	0.57692308	0.5	0.66	0.86	0.56	63.14%
SVM	0.57692308	0.54	0.7	0.86	0.48	63.14%

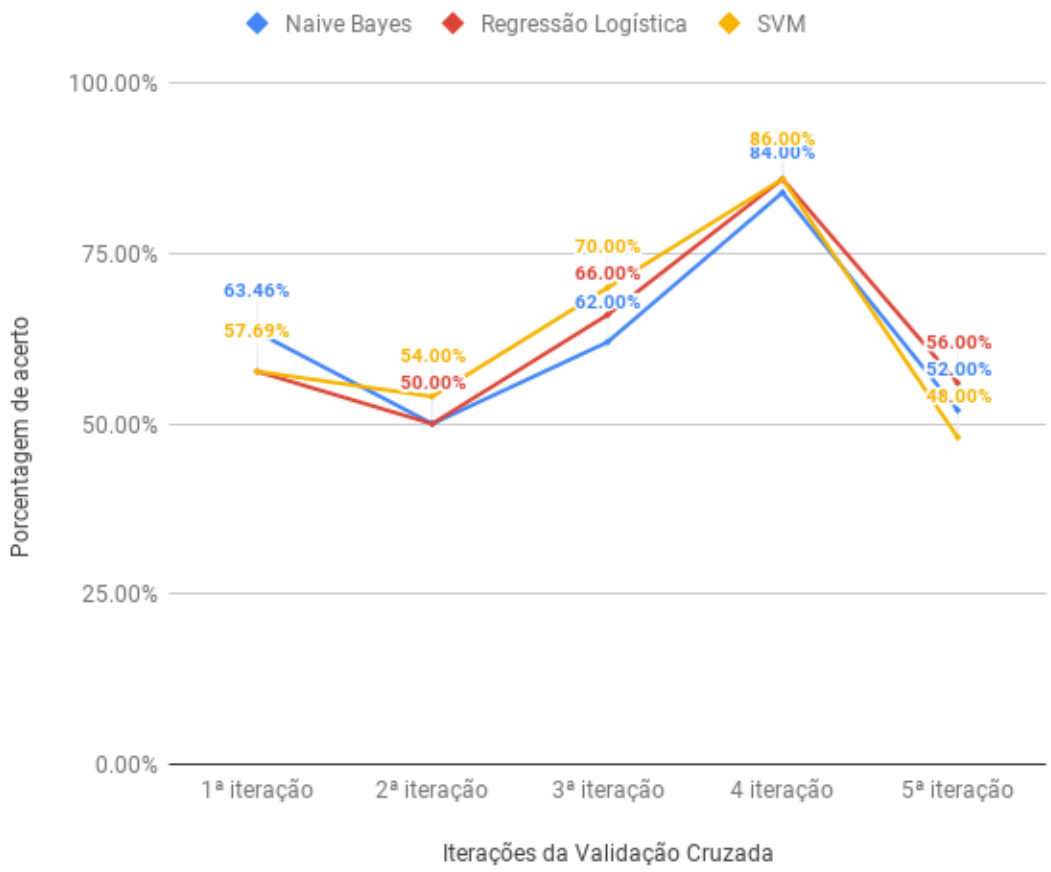


Figura 4. Porcentagem de cada algoritmo a cada iteração

Baseado na taxa de acerto mostrada anteriormente os algoritmos SVM e Regressão Logística se sairam melhores com relação a base de dados, esse resultado poderia variar caso o número de dados fosse maior, uma vez que aumentar a quantidade torna o treinamento mais eficiente. A base de treinamento serve como uma prova conceitual do funcionamento dos algoritmos, em contrapartida ela se mostra limitada para o aprendizado dos classificadores. Na figura 5 pode-se observar o resultado para um teste feito manualmente em cada algoritmo onde a saída não condiz com o resultado experado.

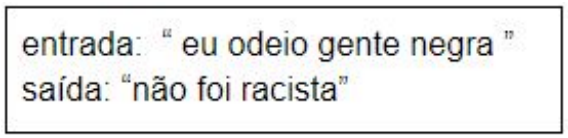


Figura 5. Saída obtida dos classificadores

5. Conclusão

Este trabalho teve como objetivo identificar mensagens racistas em português no Twitter através de três algoritmos de aprendizagem de máquina supervisionado. Foi feita a

coleta dos dados utilizando a API de busca do Twitter com uma lista de palavras-chave que possívelmente se enquadram em um contexto racista no Brasil, porém, foi visto que de 3.327 tweets analisados, 68% foram descartados. No treinamento dos algoritmos, devido ao grande desbalanceamento dos tweets racista e não racista e ao método de solução *under-sampling* usado para que esse desbalanceamento não comprometesse o desempenho dos classificadores, apenas 23.6% dos tweets classificados foram aproveitados. Os algoritmos utilizados foram: Support Vector Machine, Naive Bayes e Regressão Logística, todos eles fazem parte de aprendizagem de máquina supervisionada e foram implementados pela biblioteca Scikit-Learn.

A porcentagem de acertos dos algoritmos Regressão Logística e SVM foi igual e em relação ao Naive Bayes foi bem próxima, porém todos com menos de 65% de acerto, o que pode ser concluído é que a base de dados utilizada para o treinamento desses classificadores possa ter sido pequena afetando assim no desempenho dos mesmos.

Como trabalho futuro, pretende-se repetir o processo de coleta de dados, porém com uma lista aprimorada para a busca dos tweets, devido a variação do significado das palavras na língua portuguesa e a linguagem informal do Twitter. Como possível melhoria futura, é pretendido que a análise e classificação dos dados sejam realizadas em um período de tempo maior e com ajuda de pessoas para a classificação e validação dos dados. Por fim, disponibilizar um dataset sobre racismo na língua portuguesa para que outros trabalhos possam tomar como referência.

Referências

Rosane Leal da Silva, Andressa Nichel, Carlise Kolbe Borchardt, and Anna Clara Lehmann Martins. Discurso de ódio em redes sociais: jurisprudência brasileira. *Revista direito GV*, 7(2):445–467, 2011.

Marcus Eugênio Oliveira Lima and Jorge Vala. As novas formas de expressão do preconceito e do racismo. *Estudos de psicologia (Natal)*, 2004.

Tom M Mitchell. Machine learning, 1997.

Peter Norvig and Stuart Russell. *Inteligência Artificial: Tradução da 3a Edição*, volume 1. Elsevier Brasil, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rodrigo Magalhães Mota dos Santos. Técnicas de aprendizagem de máquina utilizadas na previsão de desempenho acadêmico. 2016.