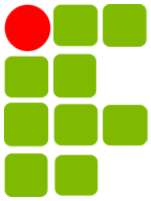


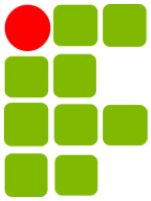
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior de Bacharelado em Engenharia de Computação

Análise de Engajamento de Usuários em Publicações sobre Política Eleitoral Contendo Discurso de Ódio na Rede Social Twitter

José Aurélio Epaminondas de Carvalho

Orientador: Dr. Marcelo José Siqueira Coutinho de Almeida

Campina Grande, Dezembro de 2022



Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
Campus Campina Grande
Coordenação do Curso Superior de Bacharelado em Engenharia de Computação

Análise de Engajamento de Usuários em Publicações sobre Política Eleitoral Contendo Discurso de Ódio na Rede Social Twitter

José Aurélio Epaminondas de Carvalho

Monografia apresentada à Coordenação do Curso Superior de Bacharelado em Engenharia de Computação do IFPB - Campus Campina Grande, como requisito parcial para conclusão do curso de Bacharelado em Engenharia de Computação.

Orientador: Dr. Marcelo José Siqueira Coutinho de Almeida
Campina Grande, Dezembro de 2022

Análise de Engajamento de Usuários em Publicações sobre Política Eleitoral Contendo Discurso de Ódio na Rede Social Twitter

José Aurélio Epaminondas de Carvalho

Dr. Marcelo José Siqueira Coutinho de Almeida
Orientador

Elmano Ramalho Cavalcanti

Igor Barbosa da Costa

Campina Grande, Paraíba, Brasil
Dezembro/2022

C331a Carvalho, José Aurélio Epaminondas de.
Análise de engajamento de usuários em publicações sobre política eleitoral contendo discurso de ódio na rede social *Twitter*. - Campina Grande, 2022.
41 f. : il.

Trabalho de Conclusão de Curso (Curso de Graduação em Engenharia de Computação) - Instituto Federal da Paraíba, 2022.
Orientador: Prof. Dr. Marcelo José Siqueira Coutinho de Almeida.

1.Rede social- *Twitter* 2. Discurso de ódio 3. Métricas de engajamento I. Almeida, Marcelo José Siqueira II. Título.

CDU 004

Nosso conhecimento nos fez cínicos, nossa inteligência nos fez cruéis e severos.
Nós pensamos muito e sentimos pouco.
Mais do que máquinas, nós precisamos de humanidade.
Mais do que inteligência, nós precisamos de carinho e bondade
Sem essas qualidades a vida será violenta, e tudo será perdido.
Charles Chaplin - O Grande Ditador.

Agradecimentos

Primeiramente a Deus por me abençoar, me iluminar e me guiar em uma vida de sangue, suor e lágrimas. E pela proteção que peço diariamente aos meus familiares e a todas as pessoas importantes para mim.

Ao professor orientador Dr. Marcelo José Siqueira Coutinho de Almeida, pela sugestão de um tema bastante importante, onde pude trabalhar e conhecer novas ferramentas e conceitos. Mesmo com problemas de saúde, com sua ajuda, possibilitou a conclusão deste trabalho.

A minha mãe professora Sebastiana Epaminondas de Barros Pereira, por ser meu maior exemplo de mulher, batalhadora e guerreira que sempre foi. Desde seus 14 anos - até mais nova do que isso, quando perdeu seus pais, trabalhou para que podesse ter uma vida melhor e que meus irmãos e eu podemos usufruir e viver uma vida que nos dê orgulho. Nunca se deixou abalar por nada nesse mundo.

Ao meu pai Norberto Luiz Pereira de Carvalho a quem eu sigo como maior exemplo. Sempre teve seus problemas, principalmente por causa do trabalho como oficial de justiça, mas nunca deixou de estar ao meu lado, me apoiando com todas as suas suas forças. Que pode ter certeza que ficaria sem comer se significasse uma vida melhor para meus irmãos e eu.

Aos meus irmãos, Norberto Luiz Pereira de Carvalho Júnior e Emmanuel Natan Epaminondas de Carvalho. Pois por ser o irmão mais novo, sempre tive a proteção deles.

A minha namorada e se Deus permitir a mulher da minha vida, Mariana Lopes Feitosa de Sousa, por ter aparecido em um momento da minha vida onde eu estava desacredidato. Ela é o motivo, mesmo em dias tristes e sombrios, do meu sorriso e meu coração aquecido de amor e carinho. Com todas as suas dificuldades, sempre se preocupa com o meu bem-estar. Que Deus sempre lhe abençoe.

A todos os meus familiares e amigos que sempre acrescentaram qualidade, amor e perserverança na caminhada da minha vida.

Aos meus professores que acrescentaram valor e conhecimento a minha vida para alcançar meus objetivos.

E finalmente, agradeço a mim mesmo, o primeiro Engenheiro de Computação na minha família. Por nunca ter desistido de nada na vida. De ficar trancado em casa estudando e trabalhando dias e mais dias com Deus e os meus pensamentos. E finalmente, deixo uma mensagem para o "eu" do futuro: continue sempre assim e se for para mudar, mude para o melhor, a vida é tão curta, aproveite ela, com amor, carinho e compaixão. Não viva de

intrigas, brigas e violência. Ame ao próximo como Jesus nos amou.

José Aurélio

Resumo

As redes sociais foram criadas para que as pessoas ao redor do mundo pudessem se comunicar, divulgar ideias, eventos, suas vidas, etc. No entanto, existe um grande problema associado às redes sociais, que é o discurso de ódio. As redes sociais trouxeram uma grande capacidade de propagação de discursos de ódio que prejudicam a vida de muitas pessoas, principalmente grupos minoritários. No entanto, o discurso de ódio não é apenas sobre pessoas, assuntos extremamente importantes como a política eleitoral do país e as pessoas a ela associadas são atacados diariamente. Para evitar a propagação do discurso de ódio, as redes sociais contam com algoritmos de aprendizado de máquina e aprendizado profundo para automatizar o processo de detecção e exclusão de postagens, mas esses algoritmos não são perfeitamente eficazes, permitindo que algumas postagens que contenham discurso de ódio permaneçam em seu meio.

Nesse contexto, serão analisadas publicações com discurso de ódio no contexto de política eleitoral da rede social Twitter, por meio da construção de um software capaz de obter essas publicações, utilizar algoritmos de aprendizado de máquina para detectar discurso de ódio, avaliar suas performances e as publicações que tenham o presença de discurso de ódio serão obtidas métricas de engajamento que serão o foco da análise das publicações e seus possíveis comportamentos.

Palavras-chave: Discurso de Ódio; Aprendizado de Máquina; Métricas de Engajamento.

Abstract

Social networks were created for the purpose that people around the world could communicate, disseminate ideas, events, their lives, etc. However, there is a big problem associated with social media, which is hate speech. Social networks have brought a great ability to propagate hate speeches that harm the lives of many people, mainly minority groups. However, hate speech is not just about people, extremely important issues such as the country's electoral politics and the people associated with it are attacked daily. To prevent the spread of hate speech, social networks rely on machine learning and deep learning algorithms to automate the process of detecting and deleting posts, but these algorithms do not are perfectly effective, allowing some posts that contain hate speech remain in their midst.

In this context, publications with hate speech in the context of elected politics of the social network Twitter will be analyzed, through the construction of a software capable of obtaining these publications, use machine learning algorithms to detect hate speech, evaluate their performances and the publications that have the presence of hate speech, engagement metrics will be obtained that will be the focus of the analysis of publications and their possible behaviors.

Keywords: Hate Speech; Machine Learning; Engagement Metrics.

Sumário

Lista de Abreviaturas

API	*Application Programming Interface*
STF	*Supremo Tribunal Federal*
CSV	*Comma Separated Values*
XML	*eXtensible Markup Language*
JSON	*JavaScript Object Notation*
HTTP	*Hypertext Transfer Protocol*
NLTK	*Natural Language Toolkit*
PLN	*Processamento de Linguagem Natural*
SVM	Support Vector Machine
ML	Machine Learning
SVC	*Support Vector Classification*

Lista de Figuras

Capítulo 1

Introdução

As redes sociais têm como objetivo conectar o mundo e possibilitar a comunicação entre pessoas, permitindo a difusão de suas opiniões. As interações possibilitam um maior engajamento entre indivíduos e/ou instituições. Nos EUA, mais de 170 milhões usuários gastaram mais de 121 bilhões de minutos nas redes sociais e 47% dos usuários se engajaram apenas no mês de julho de 2012. Quantificar as interações positivas e negativas possibilita a utilização do engajamento como indicador da qualidade de marketing a organizações políticas, sociais, empresariais, etc [Soares e Monteiro 2015].

Porém, pessoas aproveitam do poder das redes sociais para propagar o racismo, a xenofobia, o antissemitismo, entre outras formas de preconceito. Muitas vezes essas pessoas são levadas pelo pretexto da liberdade de expressão, mas estão apenas praticando crimes [Dantas e Netto 2021].

Qualquer tipo de expressão de ideias que incitem a discriminação é considerado como discurso de ódio [Schäfer, Leivas e Santos 2015]. Ele pode ser apresentado de diversas formas, tais como textos, imagens e/ou vídeos, utilizando termos pejorativos para se referir a um grupo de pessoas, muitas vezes minorias [Allan, Richard 2017]. O poder da Internet e das Redes Sociais possibilita que essas mensagens de ódio sejam divulgadas para pessoas ao redor do mundo.

O discurso de ódio pode ser considerado um dos males do século XXI, pois o mesmo tem causado inúmeros danos e prejudicado a vida de várias pessoas, como exemplo, casos de pessoas que postam imagens em suas redes sociais, porém por não terem o "padrão" de beleza acabam recebendo xingamentos, podendo evoluir para casos de ameaças de agressão e de morte [Baggs, Michael 2021].

É dever dos matenedores das redes sociais transformar o seu ambiente em um lugar seguro aos seus usuários, retirando qualquer tipo de mensagem que possa conter discurso de ódio. A Inteligência Artificial é uma ferramenta muito poderosa na detecção automática dessas mensagens. As técnicas de aprendizado de máquina e mais recente o aprendizado profundo aumentam ainda mais a rapidez e a probabilidade de detecção do discurso de ódio.

Porém, os algoritmos de aprendizado de máquina e aprendizado profundo não possuem

uma acurácia perfeita, isso possibilita que algumas mensagens permaneçam por um tempo circulando pela rede e, consequetemente, alcansem um maior número de pessoas.

1.1 Justificativa e Relevância do Trabalho

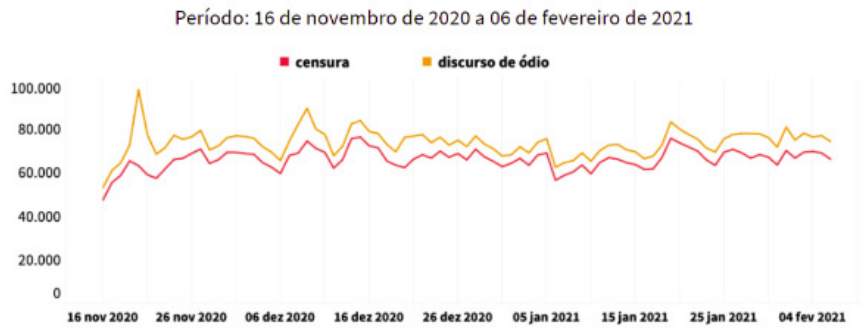
Embora seja um problema que sempre esteve presente nas redes sociais, o discurso de ódio acabou tomando novas proporções, como no caso da *COVID-19*, no ano de 2020, quando grande parte da população viu a necessidade de se confinar em suas residências para evitar a propagação do vírus, o que possibilitou um acesso maior e por mais tempo às redes sociais fazendo com que houvesse um aumento nos casos de discurso de ódio, como também o aumento sobre a discusão do problema [Baggs, Michael 2021].

As figuras 1.1 e 1.2 ilustram o aumento do fluxo de debate sobre o discurso ódio no período de novembro de 2020 até fevereiro de 2021. Isso implica que cada vez mais pessoas estão procurando sobre o assunto e entendendo a importância de seu debate. As figuras também ilustram a censura de contas responsáveis pela propagação do discurso de ódio [Ruediger *et al.* 2021].

O aumento do casos de discurso de ódio não fica recluso a ataques a pessoas e/ou grupos de pessoas, assuntos de extrema importância são atacados constantemente e alguns desses assuntos são a política e as eleições. Como no caso das eleições de 2018 que o números de denúncias mais do que dobrou em relação a de 2014, as denúncias foram de 14.653 para 39.316 [Mesquita, Lígia 2018].

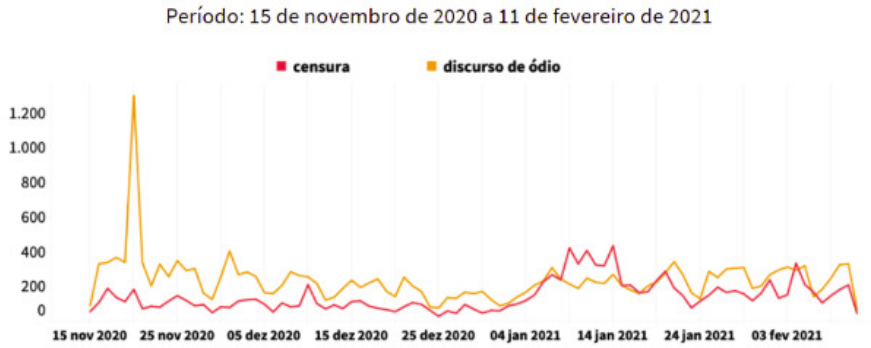
Então este trabalho terá a importância em demonstrar o poder de propagação quando usuários das redes sociais reagem às publicações que contenham discurso de ódio com *likes*, *retweets*, *replies* e *quotes* - principais métricas de engajamento do Twitter.

Figura 1.1: *Gráfico do Volume de Interações das Categorias no Twitter*



Fonte: Portal FGV - Fundação Getúlio Vargas, 2021.

Figura 1.2: *Gráfico da Evolução do Debate de Discurso de Ódio e Censura no Facebook*



Fonte: Portal FGV - Fundação Getúlio Vargas, 2021.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo deste trabalho é analisar o engajamento de publicações no contexto da política eleitoral que contenham discurso de ódio na rede social Twitter, por meio do desenvolvimento de um *software*.

1.2.2 Objetivos Específicos

Para alcançar o objetivo geral, os objetivos específicos foram definidos da seguinte forma para analisar publicações em linguagem pt-br e durante o período de 6 meses:

- Obtenção das publicações oriundas da rede social Twitter;
- Identificar os *tweets* que possuem o discurso de ódio;
- Dos *tweets* identificados com discurso de ódio, será obtido e aferida as métricas de engajamento;
- Geração de relatórios com informações gráficas;
- Divulgação dos resultados;

1.3 Metodologia

A metodologia desse projeto tem como abordagem uma pesquisa quantitativa e logo abaixo encontram-se os passos para o desenvolvimento do software capaz de obter dados de uma rede social, detectar a presença do discurso de ódio e calcular as métricas de engajamento, gerando assim dados para serem analisados. As atividades serão descritas a seguir:

- Estudo de bases bibliográficas que utilizam as técnicas de aprendizado de máquina e as ferramentas - como o Twitter API e tweepy - que serão utilizadas neste projeto;
- Desenvolvimento do código em Python, utilizando o Jupyter Notebook;
- Estudo da documentação da Twitter API e tweepy para obtenção dos *tweets*, tratamentos e obtenção das métricas de engajamento;
- Estudo da documentação do Scikit Learn e do NLTK para o PLN e a detecção do discurso de ódio;
- Análise das métricas dos algoritmos de aprendizado de máquina, como a acurácia, precisão, recall e o *F-Measure*;
- Análise das métricas de engajamento;
- Obtenção e geração gráfica dos resultados obtidos;

1.4 Organização do Documento

O restante do documento está organizado da seguinte forma. No capítulo 2 será apresentado a fundamentação teórica, que abordará temas e conceitos sobre discurso de ódio, Engajamento, Twitter API, o aprendizado de máquina e a biblioteca tweepy. No capítulo 3 será demonstrado todo o processo de desenvolvimento, com a utilização de todas as ferramentas necessárias e aplicação dos conceitos discutidos no capítulo 2. O capítulo 4 será focada na exibição dos resultados obtidos, todos os dados das métricas de engajamento e as acurácias dos algoritmos de Aprendizado de Máquina. No capítulo 5, será referente às considerações finais e sugestões para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo será apresentado toda a teoria que envolve a construção do trabalho. Serão abordados as definições de discurso de ódio e engajamento, como também descrição do estado da arte do aprendizado de máquina para análise e detecção do discurso de ódio e das ferramentas e bibliotecas utilizadas para obtenção e tratamentos dos dados.

2.1 Discurso de Ódio

O discurso de ódio se caracteriza por "palavras que tendam a insultar, intimidar ou assediar pessoas em virtude de sua raça, cor, etnicidade, nacionalidade, sexo ou religião, ou tem a capacidade de instigar a violência, ódio ou discriminação contra tais pessoas" [Brugger, Winfried 2007], ou seja, todo e qualquer ato que tenha a intenção de insultar e instigar a uma pessoa ou grupo de pessoas pode ser classificado como discurso de ódio, porém esse atos não se reprimem aos grupos citados acima, outros grupos que podem ser citados são idosos, pessoas com diferentes tipos de orientação sexual, pessoas com deficiência, entre outros.

2.2 Engajamento em Redes Sociais

O engajamento nas redes sociais pode ser definida como a quantidade de vezes que os usuários interagiram com determinada publicação [Wadhwa *et al.* 2017]. Cada clique feito na rede social, neste caso o Twitter, como os *retweets*, *replies*, *follow*, *likes*, *links*, *hashtags*, *cards*, páginas navegadas, etc, pode ser considerado como engajamento. Com o uso da API, o desenvolvedor poderá aferir todas essas infromações e realizar a sua análise.

O engajamento é de extrema importância para uma rede social, pois são por essas métricas que se consegue ver todas as interações dos seus usuários, possibilitando melhorias e assim uma maior permanência dos seus usuários [dos Santos, R. O. 2022]. Quando se passa mais tempo na rede, os algoritmos nela presentes conseguem obter mais dados dos seus usuários. Assim, com todas essas informações, a rede social pode oferecer anúncios mais agradáveis e que chamem mais atenção, consequentemente haverá um aumento de seu faturamento e das

pessoas/empresas que contratam os seus serviços.

2.3 Twitter

O Twitter é uma rede social, lançada em 21 de março de 2006, cujos os seus usuários podem enviar e/ou receber informações de outros usuários. Criado com os princípios da liberdade de expressão e diversidade de usuários, ideias, perpectivas e informações [Twitter 2016]. Desde sua criação aos dias atuais, o Twitter conta com sua grande quantidade de usuários diários ativos e monetizáveis [Braun, Daniela 2022], sendo uma das redes sociais mais utilizadas no mundo, gerando assim uma grande quantidade de dados diários, valiosos para estudos.

2.3.1 Twitter API

A API do Twitter é composta por Web Services que são baseados na arquitetura REST e que disponibiliza métodos para a obtenção dos dados necessários [XAVIER, O. C 2015]. A API pode ser dividida em três partes:

- *REST API*: responsável por manipular os dados de usuários e conexões entre eles;
- *Search API*: responsável por disponibilizar serviços de pesquisa de mensagens e usuários;
- *Streaming API*: responsável por gerar conexões persistentes que serão utilizadas para uma troca de informações de forma síncrona.

2.3.2 Métricas de Engajamento

As métricas de engajamento são os principais fatores para o cálculo da interação do usuário com os conteúdos postados na rede social e podem ser divididos da seguinte maneira [Twitter 2022]:

- *Public Metrics* (Métricas Públicas) : São métricas utilizadas para medir o engajamento de *tweets* de um ou vários usuários. Essas métricas não necessitam da permissão do usuário para que terceiros possam acessar, sendo elas as quantidades de *retweets, replies, likes* e quotes;
- *Non-Public Metrics* (Métricas Não-Públicas) : Também são métricas para medição de engajamento. Porém, ao contrário das métricas públicas, ela necessita da autenticação dos usuários.

Quanto a maior quantidade de métricas a serem aferidas, melhor e mais detalhada será a análise, pois diferentes tipos de pessoas levam a diferentes tipos de reações. A exemplo

de uma empresa, em que existe um produto que tenha pouco *likes*, porém há uma grande quantidade de cliques em sua url, a empresa pode mudar sua estratégia de *marketing*, em vez de retirar o produto do seu estoque, pois as publicações estão sendo visualizadas pelos usuários, porém o produto não está tendo as reações desejadas.

2.3.3 Métricas Públicas

As métricas públicas que serão utilizadas para o trabalho podem ser definidas da seguinte forma [Twitter 2016]:

- *Retweet*: é a republicação do *tweet* de outros usuários;
- *Like*: é a indicação que o usuário gostou de determinado *tweet*;
- *Reply*: são as respostas que os usuários fazem aos *tweets*, possibilitando a conversa entre usuários;
- *Quote*: é a citação de um *tweet*. Pode ser explicada como a republicação de uma postagem com os comentários de um usuário diferente.

2.4 Tweepy

O Tweepy é uma biblioteca implementada em Python que permitirá o acesso de aplicativos de terceiro aos *endpoints* da API do Twitter. Entre algumas das funções podem ser citadas a obtenção de *tweets* da linha do tempo, postar novas publicações, seguir e desseguir usuários, etc. A praticidade da API, junto com o poder do Python, permitirá ao desenvolvedor a obtenção de dados confiáveis e com grande eficiência [Tweepy 2022].

2.5 Aprendizado de Máquina

O aprendizado de máquina trata-se de um ramo evoluído dos algoritmos computacionais que são designados a emular a inteligência humana pelo aprendizado de seu ambiente ao seu redor [Naqa e Murphy 2015]. A aplicação do aprendizado de máquina pode ser bastante variado desde o reconhecimentos de características de um cliente e mostrar ao mesmo produtos que podem lhe interessar, aumentando assim as chances de compra, como também automatizar toda uma linha de produção, entre outras. O aprendizado de máquina é uma grande ferramenta para o mundo moderno na análise de dados, pois dado a grande quantidade de dados e pelos seus diferentes tipos que são gerados diariamente - dados geoespaciais, financeiros, médicos, etc - seria praticamente impossível utilizar métodos e/ou algoritmos convencionais para analisar todo esse volume.

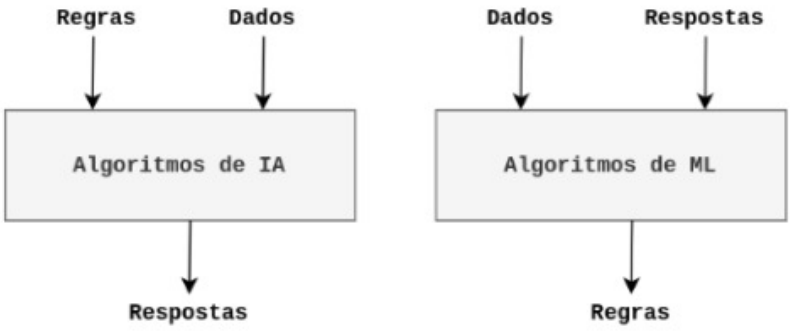
2.5.1 O Paradigma do Aprendizado de Máquina

Para obter soluções com o uso dos algoritmos de aprendizado de máquina é necessário que eles aprendam através dos dados e não da lógica [Pacheco, André 2021], como ilustra na figura 2.1. Para explicar melhor esse paradigma pode citar um exemplo de um algoritmo de visão computacional cujo mesmo consegue classificar fotos como cachorros ou porcos.

Antes de colocar o algoritmo a prova é necessário treiná-lo. Pode se obter os dados de treinamento a partir das características dos animais - *features* - em que o desenvolvedor pode definir as características do cachorro como um animal que tem orelha grande, pelo longo e rabo longo e para o porco um animal que tem orelha grande, pelo curto e rabo curto. Agora para verificar a sua eficiência é necessário analisar novos dados e tentar informar se o animal é um cachorro ou um porco [Cyberjcts 2020].

Dado o exemplo ilustrado na figura 2.2, o algoritmo irá analisar o animal 3, que tem orelha pequena, pelo longo e rabo longo, qual seria a resposta? O algoritmo, com os dados de treinamento, pode inferir que o terceiro animal é um cachorro, pois ele possui o pelo longo e o rabo longo e o porco não. Porém, e se a foto for a de um javali? que pode corresponder às características do animal 3. Por isso, quanto mais dados de treinamento e mais específicos referentes às suas características, melhor será sua eficiência.

Figura 2.1: *Diferença entre o Paradigma de Inteligência Artificial e Machine Learning*



Fonte: Computação Inteligente, 2021.

2.5.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural trata computacionalmente as diversas característica da comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos [Gonzalez e Lima 2003]. Para simplificar, o principal objetivo do PLN é fazer com o que o computador possa se comunicar em linguagem humana [Felippo *et al.* 2021], atuando na formação de possíveis relações semânticas com a união de palavras e frases e uma vez que são determinadas, algumas delas serão descartadas, pois não irão fazer sentido.

Figura 2.2: *Análise de Features de Animais*

Fonte: Autoria própria.

O texto bruto obtido das redes sociais passam por técnicas de pré-processamento, como a utilização da técnica Bag-of-Words cuja idéia principal é verificar a quantidade de ocorrências de palavras em um texto, não se importando com a ordem ou estrutura textual [Brownlee, Jason 2017], para que aí sim possa servir como entrada de um modelo de aprendizado de máquina ou aprendizado profundo.

2.5.3 Análise de Sentimentos

A análise de sentimentos tem como principal objetivo a definição de técnicas automáticas capazes de extrair informações subjetivas de texto em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser utilizados por um sistema de apoio ou tomador de decisão [Benevenuto, Ribeiro e Araújo 2015]. Os termos a seguir são comumentes associados à análise de sentimentos:

- Polaridade: é definida como o grau de positividade e negatividade de uma mensagem. Que no caso deste trabalho é classificar se um texto contém o discurso de ódio ou não.
- Força de sentimento: é a representação da intensidade da polaridade de uma mensagem. Por exemplo, se a mensagem for classificada como positivo - contém o discurso de ódio, ela pode receber uma intensidade e os com maior intensidade serão focados, pois mesmo que seja positivo, porém recebe uma baixa intensidade, ela será marcado como um falso positivo, necessitando uma maior análise para decidir se a mensagem contém ou não o discurso de ódio.
- Sentimento/Emoção: é a indicação do sentimento específico contido em uma mensagem. Essa representação possibilita o foco aprofundado da pesquisa, pois dado o exemplo deste trabalho, o foco seria a análise de sentimentos negativos como a raiva, o nojo, o medo e etc.

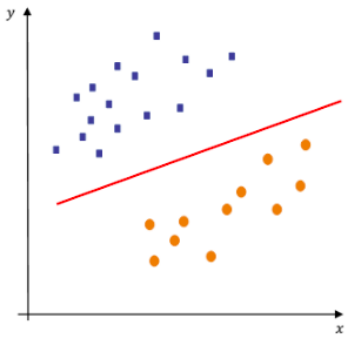
- Subjetividade e Objetividade: uma mensagem objetiva em geral contém fatos ou informações, já uma mensagem subjetiva contém sentimentos pessoais e opiniões, são essas mensagens subjetivas que mais estão presentes nas redes sociais.

2.5.4 Máquina de Vetor de Suporte

A Máquina de Vetor de Suporte ou *Support Vector Machine*, em inglês, é um conjunto de métodos que utilizam o aprendizado indutivo não-supervisionado e que são utilizados para classificação, regressão e detecção de anomalias [Lorena e Carvalho 2007]. Nesta técnica são colocados todos os dados em um plano n-dimensional (n é a quantidade de dados possuídos). Com a execução da técnica são encontrados os hiperplanos que melhor conseguem diferenciar as classes. Na figura 2.3 pode ser visualizado uma representação gráfica da atuação do SVM em dois conjuntos de dados - quadrados azuis e círculos laranjas - cujo mesmo irá encontrar a opção mais otimizada para poder separar as classes.

Os vetores de suporte são os pontos mais próximos às linhas verdes - margens, ilustrado na figura 2.4, eles empurram as linhas das margens para próximas das outras classes, sendo esse o foco principal do SVM que é observar os vetores de suporte que estão nos extremos dos dados e que são os mais próximos de outras classes. A margem máxima é a soma de todos os vetores de suporte mais próximos do hiperplano [Khoong, Wei Hao 2021].

Figura 2.3: *Representação do SVM*



Fonte: Towards Data Science, 2021.

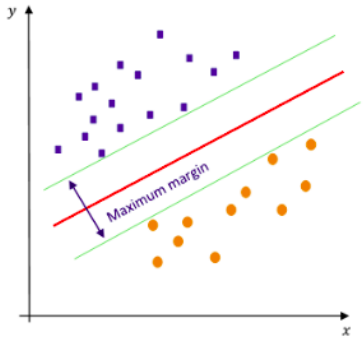
2.5.5 SVC

O SVC é uma classe de classificação que utiliza o SVM. De acordo com [Pedregosa *et al.* 2011] e [Buitinck *et al.* 2013], que dado um vetor $x_i \in R^p, i = 1, \cdots, n$, em duas classes, e um vetor $y \in \{1, -1\}^n$, o objetivo é encontrar $w \in R^p$ e $b \in R$ tal que a previsão dada por $sign\left(w^T\phi(x) + b\right)$ é correta para a maioria das amostras.

O SVC resolve o seguinte problema primal:

$$\underset{w,b,\zeta}{min}\frac{1}{2}w^T w + C\sum_{i=1}^{n}\zeta_i \tag{2.1}$$

Figura 2.4: *Margem Máxima*



Fonte: Towards Data Science, 2021.

$$\text{sujeito a, } y_i \left(w^T \phi(x) + b\right) \geq 1 - \zeta_i \tag{2.2}$$

$$\zeta \geq 0, i = 1, \cdots, n \tag{2.3}$$

De forma intuitiva, tenta-se maximar as margens - minimizando $\|w\|^2 = w^T w-$ enquanto atrai sobre si penalidades quando uma amostra é classificada incorretamente ou dentro do limite da margem. No mundo perfeito, o valor $y_i \left(w^T \phi(x) + b\right)$ teria que ser ≥ 1 para todas as amostras. Porém os problemas não são perfeitamente separados por hiperplanos. Então permite-se que algumas amostras estejam a uma distância ζ_i de suas margens limites. O termo de penalidade C controla a força desta penalidade, portanto, atua como um parâmetro de regularização inverso.

O problema dual para o primal é

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \tag{2.4}$$

$$\text{sujeito a, } y^T \alpha = 0 \tag{2.5}$$

$$0 \leq \alpha_i \leq C, i = 1, \cdots, n \tag{2.6}$$

Onde e é o vetor de todos os 1's e Q é uma matriz positiva semidefinida n por n, $Q_{i,j} \equiv y_i y_j K(x_i, x_j)$, onde $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ é o kernel. Os termos α_i são chamados de coeficientes duais e são limitiados superiormente por C. Essa representação dual destaca o fato que os vetores de treinamento são implicitamente mapeados em uma espaço dimensional maior - ou até infinito - pela função ϕ.

Já que o problema de otimização foi resolvido, a saída da função de decisão se torna, para determinada amostra x:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x_j) + b \tag{2.7}$$

e a classe prevista corresponde ao seu sinal. Apenas necessita somar os vetores de suporte, por causa que os coeficientes α_i são zero para outras amostras.

2.5.6 LinearSVC

Como o SVC, o LinearSVC é uma classe de classificação para o SVM. De acordo com [Pedregosa *et al.* 2011] e [Buitinck *et al.* 2013], o problema primal pode ser formulado equivalente como:

$$\underset{w,b}{min} \frac{1}{2} w^T w + C \sum_{i=1}^{n} max(0, 1 - y_i(w^T \phi(x_i) + b)) \tag{2.8}$$

A única função kernel aceita é a linear, pois não envolve produtos internos entre as amostras, ao contrário da forma dual.

2.5.7 Matriz de Confusão

A matriz de confusão é uma tabela que tem como objetivo mostrar as frenquências de classificação para classe do modelo. Ela mostra a distribuição das ocorrências das classes atuais e das classes previstas [IBM 2021]. Se definirmos a classe de interesse como os *tweets* que possuem o discurso de ódio e que o modelo irá classificar duas classes - *yes* e *no* - a matriz de confusão pode ser montada da seguinte forma:

- Verdadeiro positivo - ocorre quando o modelo prediz corretamente a classe de interesse;
- Verdadeiro negativo - ocorre quando o modelo prediz incorretamente a classe de interesse;
- Falso positivo - ocorre quando o modelo prediz corretamente a classe que não é de interesse;
- Falso negativo - ocorre quando o modelo prediz incorretamente a classe que não é de interesse;

A imagem 2.5 ilustra uma exemplo da matriz de confusão. Em sua leitura na vertical, o modelo classifica que 104 exemplos de sua base de dados foram classificadas corretamente com a classe *yes* e 33 foram classificadas incorretamente, para a classe *no*, foram classificadas corretamente 91 exemplos e 12 foram classificadas incorretamente. A leitura na horizontal, 104 foram classificados corretamente para *yes* e 12 incorretamente, para a classe *no*, 91 foram classificadas corretamente e 33 incorretamente.

Figura 2.5: *Matriz de Confusão*

	No (predicted)	Yes (predicted)	Total
No	**91**	33	124
Yes	12	**104**	116
Total	103	137	**240**

Fonte: IBM, 2021

2.5.8 Métricas de Avaliação

Essas métricas são utilizadas para avaliar a performance de algoritmos de aprendizado de máquina. Dentre algumas dessas métricas estão a acurácia, a precisão, o *recall* e o *F-Measure*.

Primeiro, a acurácia pode ser definida como o número total de predições corretas dividida pelo número total de predições feitas a um conjunto de dados [Brownlee, Jason 2020]. Porém, a acurácia pode ser inapropriada para problemas de classificação desbalanceadas, pelo motivo que o grande número de exemplos da classe majoritária sobrecarregará a classe minoritária, fazendo com que modelos inábeis possam ter uma alta porcentagem em seu resultado. Como alternativa ao uso da acurácia, métricas de precisão e *recall* são requeridas.

Precisão é uma métrica que calcula o número das predições positivas corretas realizadas pelo modelo [Brownlee, Jason 2020], sendo representada pela proporção de exemplos positivos previstos corretamente dividida pelo o número total de exemplos positivos previstos pelo modelo.

Recall é a métrica que calcula o número de predições positivas corretas pelo o número total de predições positivas realizadas pelo modelo [Brownlee, Jason 2020]. Ao contrário da precisão que apenas informa o número das previções positivas corretas do número total de predições positivas, o *recall* informa o número de previções positivas perdidas.

Tanto a precisão como o *recall* sozinhos não conseguem detalhar totalmente os resultados da previsão de um modelo. Então, *F-Measure* é a métrica que irá prover um meio de combinar as propriedades da precisão e do *recall* em uma única medida [Brownlee, Jason 2020].

2.6 NLTK

O NLTK é uma biblioteca para a construção de programas - escritos em Python - que trabalhem com a linguagem humana. A biblioteca provê um vasto conjunto de bibliotecas de processamento textual para classificação, tokenização, stemização, marcação, análise e raciocínio semântico, wrappers para bibliotecas PLN de força industrial, como também mais de 50 corporas e recursos léxicos como o WordNet [NLTK 2022].

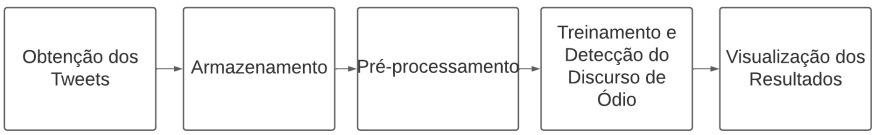
Capítulo 3

Desenvolvimento

Neste capítulo, serão apresentados os principais passos para a construção do projeto. Esses passos são: obtenção de dados, armazenamento, tratamento, detecção do discurso de ódio e exibição gráfica dos resultados obtidos.

3.1 Fases de Desenvolvimento do Projeto

Figura 3.1: *Fases de Desenvolvimento do Projeto*



Fonte: Autoria Própria.

A figura 3.1 ilustra as fases de desenvolvimento que será seguida por este projeto. As etapas dessa linha de desenvolvimento podem ser separadas em cinco. As duas primeiras são obtenção dos *tweets* e armazenamento dos dados. Em seguida, será a etapa do tratamento dos dados(pré-processamento), esse processo facilita a análise dos algoritmos de aprendizado de máquina, aumentando suas eficiências. A quarta etapa será o treinamento do algoritmo e a detecção do discurso de ódio com o uso do SVM. A última etapa será para a obtenção dos dados desejáveis que serão utilizados para análise das métricas de engajamento.

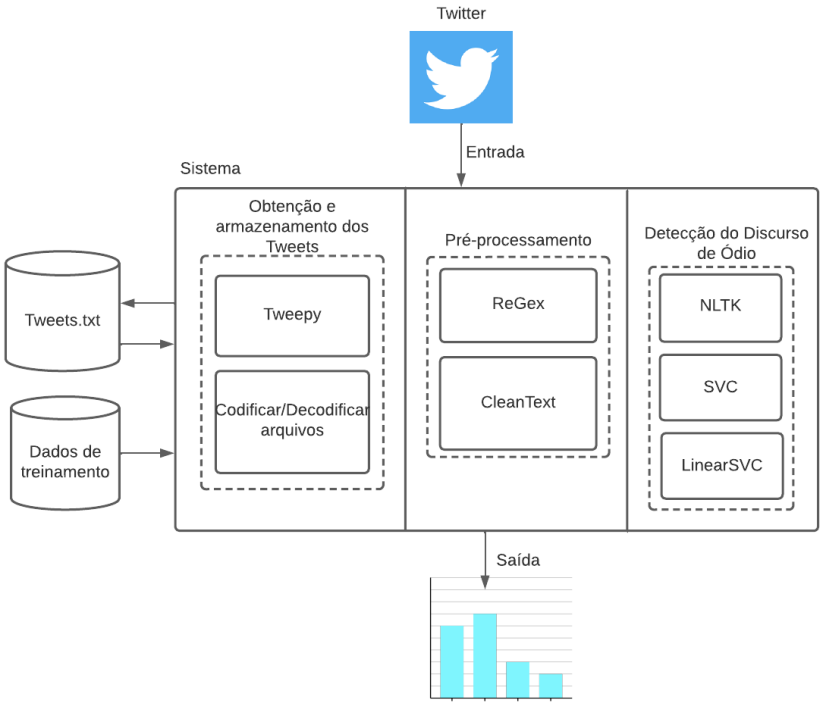
3.2 Arquitetura

A figura 3.2 ilustra a arquitetura do sistema desenvolvido. Toda a sua construção e ferramentas utilizadas serão descritas seguindo as etapas das fases de desenvolvimento discutido

no tópico 3.1 que serão explanadas nos tópicos seguintes.

Para as etapas de obtenção e armazenamento foram utilizadas: Tweepy e funções do Python para codificação e decodificação de strings em arquivos. Para a etapa de pré-processamento foram utilizadas: o pacote **ReGex** para busca de padrões que sejam de interesse a sua remoção e a biblioteca **clean-text** para remoção de emojis e acentuações. Para a etapa de treinamento do algoritmos de aprendizado de máquina e detecção do discurso de ódio foram utilizadas: o **NLTK**, o **SVC** e o **LinearSVC**. E para a última etapa: o **pandas** e o **seaborn** para geração gráfica dos resultados obtidos.

Figura 3.2: *Arquitetura do Sistema*



Fonte: Autoria própria.

3.3 Obtenção e Armazenamento da Base de Dados

O objetivo dessa seção será explicar o processo de obtenção das duas bases de dados que serão utilizadas por este trabalho. Uma para o treinamento dos algoritmos de aprendizado de máquina e outra para a validação e verificação de suas acurácias.

3.3.1 Dados de Treinamento

O conjunto de dados de treinamento foram adquiridos a partir de um corpus de nome **OffComBR** que contém comentários ofensivos e não ofensivos retirados das notícias mais acessadas do site g1.globo.com [Pelle e Moreira 2017]. No estudo para construção do corpus,

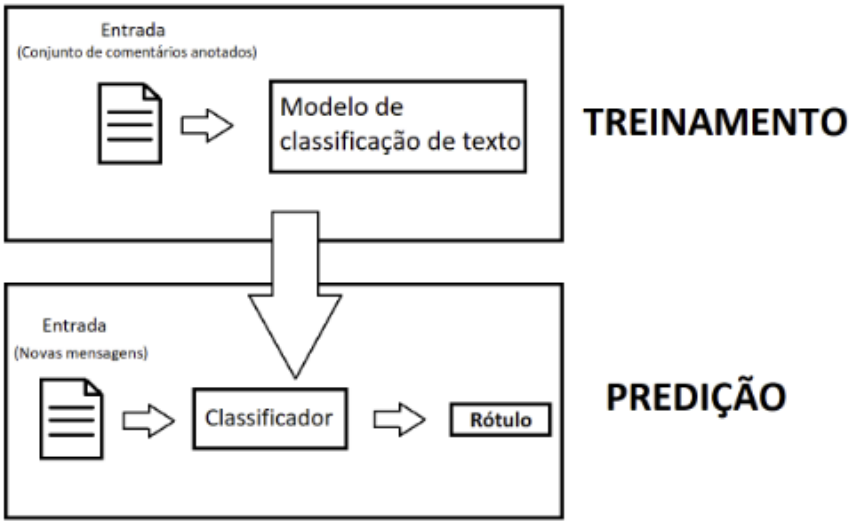
foi observado que as notícias, com categorias política e esporte, foram as que mais continham comentários ofensivos. Então a coleta dos dados foi focada a esses dois tópicos.

Para se obter a real classificação de que o comentário é ofensivo ou não, ele passará por avaliação de juízes e será medido o nível de concordância entre eles, sendo utilizado o método Fleiss'kappa [Fleiss 1971].

Com as informações do comentários, o algoritmo de aprendizado de máquina passará pelo processo de treinamento, possibilitando o reconhecimento de padrões nos comentários e, posteriormente, possibilitará a classificação de novos comentários [Paiva, Silva e Moura 2019]. A figura 3.3 ilustra a abstração de um modelo de aprendizado de máquina em seu processo de treinamento e predição de comentários.

Existem duas versões do corpus: o **OffComBR-2** e o **OffComBR-3**. Para este trabalho será utilizado o **OffComBR-2**, cujo mesmo possui 1250 comentários e 419 foram considerados ofensivos. A figura 3.4 ilustra alguns desses comentários.

Figura 3.3: *Esquema de Treinamento e Predição do Modelo*



Fonte: Paiva, Silva e Moura, 2019.

Figura 3.4: *Comentários Rotulados*

```
yes,'PEC  DA  VIDAAAA    VIDA LIVRE  DE  MAMATA   ESQUERDALHAAAAAA  kkkkkkkkk'
no,'fim da isencao de impostos para todas as igrejaschega desta bagunca'
no,'Vai taxar dinheiro de doacoes Coitados dos mendigos'
no,'Unico corte verdadeiramente na carne e acabando com o excesso de beneficios e diminuindo o salario dos
no,'Deveriam mas isso nao sou quem vai saber te responder pergunta la na PF'
no,'enquanto um presidente nao tiver coragem de propor acabar com TODAS as regalias e beneficios dos deputa
no,'Paulo Bertazzi  o que voce quer e uma ditadura de novo'
no,'Sr presidente o Brasil esta ao seu lado que venham as reformas chega de mimimi e vamos trabalhar'
yes,'Mais um pobre metido a besta ja ja fica sem dinheiro e vai sentir falta dos bons tempos'
no,'Bons tempos O Brasil TEVE bons tempos Entao me lembra ai por que eu nao lembro nao HAHAHA'
no,'Sabe o que e engracado e que SIM O POVO VAI PAGAR O PATO RECHEADO E BEM GORDO'
no,'pra que ele iria fazer isso kkk pra que um politico iria ser contra politicos receberem muito'
yes,'Martin Sales deixa de ser idiot  agora ninguem pode discordar do temer que e mortadela so nesse pais m
```

Fonte: OffComBr, 2017.

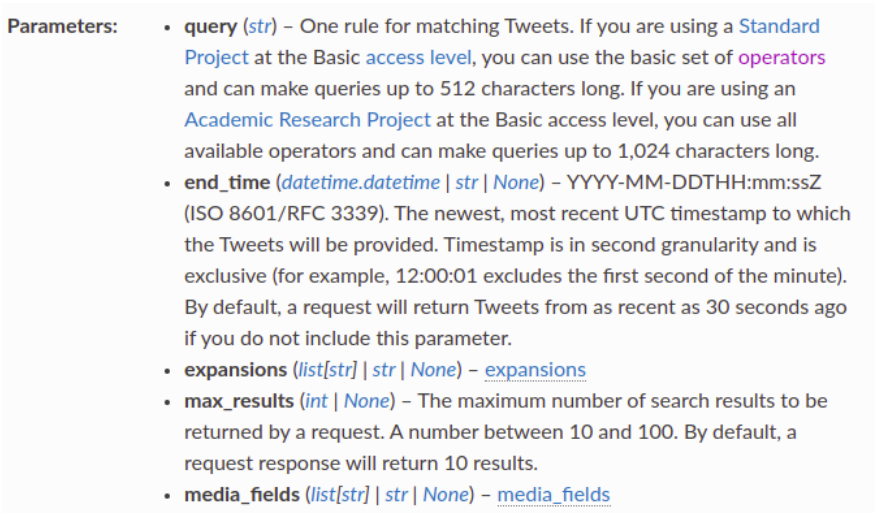
3.3.2 Dados de Validação

A obtenção dos dados de validação é realizado com o uso da biblioteca tweepy que fará a comunicação com a API do Twitter. Com o uso do método **Client.search_recent_tweets**, ilustrado na figura 3.5, será possível obter os *tweets* desejados.

O principal parâmetro do método citado anteriormente é a *query*. A *query* se trata da palavra ou conjunto de palavras que servirão como regras na busca da API, ou seja, a API só trará como resposta os *tweets* que possuam as palavras contidas na *query*. Vale ressaltar que a *query* comporta tipos de operadores que irão facilitar sua busca, evitando um grande volume desnecessário de *tweets*. Por exemplo, o operador obrigatório de conjunção **-is:retweets**, que irá trazer apenas os *tweets* originais e não os que foram retuitados.

Logo após a obtenção, os dados serão salvos em uma base textual simples, cujos mesmos conterão as informações do ID do usuário, o texto publicado e as métricas de engajamento.

Figura 3.5: *Parâmetros do Método Search Recent Tweets*



Parameters:

- **query** (*str*) – One rule for matching Tweets. If you are using a Standard Project at the Basic access level, you can use the basic set of operators and can make queries up to 512 characters long. If you are using an Academic Research Project at the Basic access level, you can use all available operators and can make queries up to 1,024 characters long.
- **end_time** (*datetime.datetime | str | None*) – YYYY-MM-DDTHH:mm:ssZ (ISO 8601/RFC 3339). The newest, most recent UTC timestamp to which the Tweets will be provided. Timestamp is in second granularity and is exclusive (for example, 12:00:01 excludes the first second of the minute). By default, a request will return Tweets from as recent as 30 seconds ago if you do not include this parameter.
- **expansions** (*list[str] | str | None*) – expansions
- **max_results** (*int | None*) – The maximum number of search results to be returned by a request. A number between 10 and 100. By default, a request response will return 10 results.
- **media_fields** (*list[str] | str | None*) – media_fields

Fonte: Tweepy, 2022.

3.4 Pré-processamento

Inicialmente os *tweets* recebidos passarão por uma série de pré-processamentos, pois os *tweets* possuem ruídos que podem dificultar a análise e por consequência aumentar a complexidade de processamento. Todos esses processos, que serão citados, são realizados para os tweets que serão obtidos a partir da API como também para a base de dados pré-treinada. Entre os processos que podem ser citados:

- Remoção de Tweets duplicados: este primeiro passo não é obrigatório, pois ele só é requisitado quando o desenvolvedor solicita várias requisições à API em curto período de tempo - o default para buscar novos tweets é de 30 segundos, como é informado pelo

parâmetro **end_time** da figura 3.5, isso faz com haja duplicadas e consequentemente mais dados desnecessários a serem analisados.

- Remoção de lixos: Este trabalho irá utilizar essa expressão "lixo" para representar *hashtags*, URLs, menções de usuários e emojis, pois os mesmos não trazem informações relevantes para a análise e muitas vezes a poluem. Então todo esse "lixo" é removido.

- Remoção de espaços vazios e quebras de linhas: como já foi dito, não existe um padrão nos dados que serão recebidos e quanto menos dados irrelevantes, melhor será a análise. Isto serve para espaços múltiplos e quebra de linhas.

Existe também uma etapa que só é utilizada para os tweets vindos da API, que é a definição das features de forma supervisionada - preparação para a aplicação do algoritmo de aprendizado de máquina.

Os tweets em sua forma mais bruta não possuem a informação de que os textos possuem ou não o discurso de ódio, isso faz com exista a necessidade de serem definidos, mesmo que erroneamente em alguns casos. A rotulação de cada *tweet* será feita de forma supervisionada pelo motivo de evitar os piores casos que podem ser obtidos se os dados de treinamentos em seus 100% de ocorrência possuírem discurso de ódio e os dados de testes não possuírem discurso de ódio em suas ocorrências e vice-versa, neste caso a eficiência do algoritmo seria de 0%. A probabilidade desse problema ocorrer é muito pequena, porém não pode ser descartada.

3.5 Detecção do Discurso de Ódio

3.5.1 Processamento de Linguagem Natural

Os processos para PLN terão como objetivo a preparação para a análise de sentimentos, pois a máquina não tem a capacidade de processar frases como os seres humanos naturalmente possuem. Para que isso ocorra são necessários procedimentos que irão facilitar a transformação dos dados para informações entendíveis para a máquina. É necessário a utilização da biblioteca NLTK para a execução dos seguintes passos:

- Remoção de *stopwords*: As stopwords podem ser classificadas como artigos, advérbios, preposições, pronomes e etc. Para este trabalho, as stopwords não trazem informações relevantes e serão elas que terão maiores frequências nos textos [dos Santos, Vinícius 2018], por exemplo, a frase "a casa está em chamas" se retirarmos o artigo "a" e a preposição "em" a frase ficaria "casa está chamas", mesmo com erros de português dá para entender o seu contexto: uma casa está pegando fogo;

- Tokenização(*Word Tokenization*): trata-se da separação de palavras de uma frase, como exemplo, a frase "Não estou bem" será separada em " 'Não' 'estou' 'bem' ". Esse

procedimento irá facilitar a polarização das palavras, por exemplo, classificar “bem” como positiva e “não” como negativa;

- Distribuição de Frequência(*Frequency Distribution*): Todas as palavras de todos os *tweets* passarão por um processo de contagem e armazenamento.

3.5.2 Análise de Sentimentos

Para que a análise de sentimentos ocorra, primeiramente é necessário criar um dicionário onde suas chaves serão as palavras mais frequentes - poderiam ser escolhidas todas as palavras, porém muitas delas têm frequências muito baixas, isso faz com que não haja tanto peso na classificação - e os valores serão booleanos - verdadeiro ou falso - da condição de que a palavra do *tweet* esteja contida na lista das palavras mais frequentes, definindo assim as *features* de cada palavra dos *tweets*. O trecho do código para que isso aconteça pode ser visualizado na figura 3.6.

Logo após o passo de definição das *features*, entrará em ação a aplicação dos algoritmos de ML. Para a classificação serão utilizados o SVC e uma de suas variantes o LinearSVC. O classificador será treinado e logo após os *tweets* oriundos da API serão validados, na figura 3.7 pode se observar o trecho de código referente ao treinamento do classificador.

Depois do treinamento, os classificadores serão colocados a teste com os dados de validação. Então será possível a criação das matrizes de confusão e a obtenção das métricas de avaliação de suas performances - acurácia, precisão, *recall* e o *F-Measure*. Os *tweets*, classificados como verdadeiros positivos, ou seja, que possuem o discurso de ódio, passarão pelo processo de obtenção das métricas de engajamento.

Figura 3.6: *Trecho de Código de Definição de Features*

```
word_feature = list(freq_words.keys())[:3000]

def find_features(document):
    words = set(document)
    features = {}
    for w in word_feature:
        #the key is the word in the 3000 most popular words
        #is gonna be the boolean value for w in words
        features[w] = (w in words)

    return features
```

Fonte: Autoria própria.

Figura 3.7: *Treinamento do LinearSVC*

```
Linear_classifier = SklearnClassifier(LinearSVC())
Linear_classifier.train(training_set)
```

Fonte: Autoria própria.

3.6 Métricas de Engajamento

Dos resultados da etapa de análise de sentimentos são retirados as métricas - *retweets, likes, replies* e *quotes* - e para cada uma das métricas é obtido a soma total de todos os *tweets*.

Para se obter uma noção da propagação do discurso de ódio pelo engajamento dos usuários, foi utilizado a métrica *retweet* que é a que tem a maior frequência - essa afirmação será demonstrada no capítulo 4 - para obter o usuário com maior quantidade de *retweets* e a sua quantidade de seguidores. Com isso poderá se obter uma porcentagem dos seguidores que contribuiram para a propagação do discurso de ódio.

Na figura 3.8 pode-se observar o trecho de código referente a obtenção dos seguidores, como também um tratamento bastante importante que é a verificação se os seguidores são **None**, pois se condição for verdadeira isso implica que a conta do usuário está desativada - seja por exclusão do próprio, bloqueio do Twitter, entre outros casos - então o trecho do código descarta esse usuário e procura o próximo usuário com mais *retweets*.

Como também foram calculadas as médias de cada uma das métricas para obter uma noção de comportamento, verificando o número das métricas que normalmente aparecem em publicações com o discurso de ódio.

Figura 3.8: *Trecho de Código para Obtenção de Seguidores com Mais Retweets*

```
def getFollwersOfMostRetweetUser(user_tweets, retweet_count):
    index_max = retweet_count.index(max(retweet_count))
    user_with_most_retweets = user_tweets[index_max].split(" ")
    id_user_with_most_retweets = user_with_most_retweets[0]

    followers = client.get_users_followers(int(id_user_with_most_retweets), max_results=1000)[0]
    if(followers == None):
        user_tweets.pop(index_max)
        retweet_count.pop(index_max)
        getFollwersOfMostRetweetUser(user_tweets, retweet_count)


    return (max(retweet_count), len(followers), user_with_most_retweets)

max_retweets, followers_count, user_most_retweets = getFollwersOfMostRetweetUser(user_tweets, retweet_count)
```

Fonte: Autoria própria.

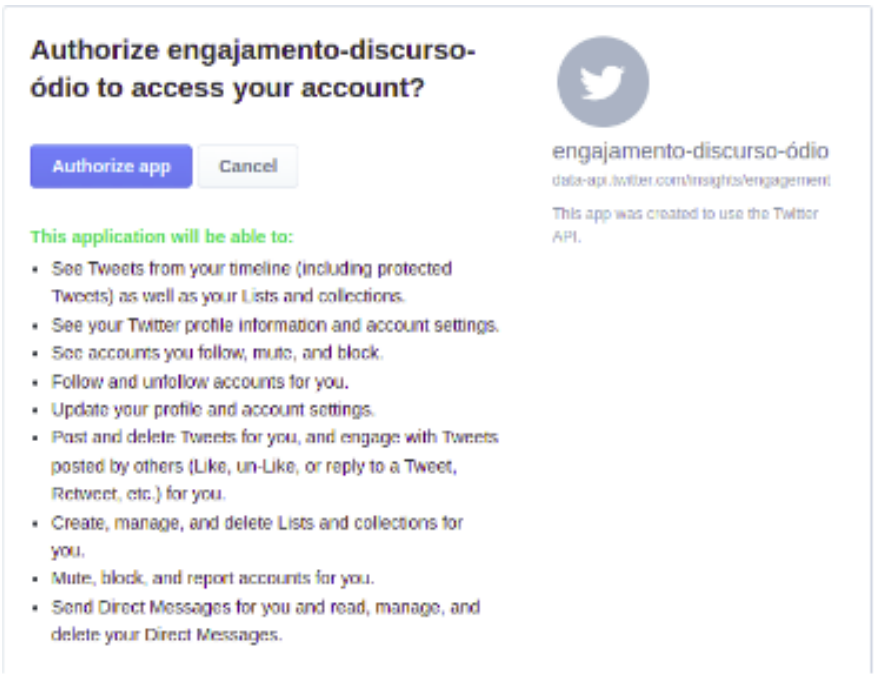
Capítulo 4

Resultados Obtidos

Com todo o processo de desenvolvimento foi possível criar um código capaz de se comunicar com uma API externa, detectar o discurso de ódio e obter as métricas de engajamento, utilizando todos os conceitos do capítulo 2 e descritos os processos no capítulo 3. O único problema encontrado foi a não obtenção das métricas de engajamento não-públicas, pois é necessário a autorização da plataforma Twitter e de seus usuários para obtê-los, como pode observar nas figuras 4.1 e 4.2. Porém isso não obstruiu a conclusão do projeto, pois o objetivo principal seria a obtenção das métricas públicas que trazem informações suficientes ao trabalho. Portanto todos os objetivos foram alcançados.

Os dados quantitativos que são o objetivo deste trabalho são as acurácias dos algoritmos SVC e LinearSVC e as métricas de engajamento.

Figura 4.1: *Requisição da Autorização das Métricas Não-Públicas*



Fonte: Twitter, 2022.

Figura 4.2: *Resultado da Requisição*

Fonte: Twitter, 2022.

4.1 Resultados das Métricas de Performance

As classes de classificação do SVM são o SVC, LinearSVC e o NuSVC, porém, o SVC e NuSVC possuem métodos muitos semelhantes e resultados próximos. Então foi optado apenas pela utilização do LinearSVC e o SVC, explicados nos tópicos 2.5.5 e 2.5.6. No total, os classificadores analisaram 2152 *tweets* e logo após passaram pelo processo de avaliação de suas performance através da acurácia, precisão, *recall* e *F-Measure*.

4.1.1 Matriz de Confusão

A figura 4.3 ilustra a matriz de confusão do modelo que utiliza o LinearSVC. Em sua leitura na horizontal, o modelo previu 190 *tweets* corretamente com a presença de discurso de ódio e 156 previu incorretamente, previu corretamente a ausência do discurso de ódio em 890 *tweets* e 916 previu incorretamente. Em sua leitura na vertical, 190 *tweets* tinham corretamente a presença e 916 tinham incorretamente, 890 tinham corretamente a ausência e 156 tinham incorretamente.

A figura 4.4 ilustra a matriz de confusão do modelo que utiliza o SVC. Em sua leitura na horizontal, o modelo previu 2 *tweets* corretamente com a presença de discurso de ódio e nenhum foi previsto incorretamente, previu correntamente a ausência do discurso de ódio em 1046 *tweets* e 1104 previu incorretamente. Em sua leitura na vertical, 2 *tweets* tinham corretamente a presença e 1104 tinham incorretamente, 1046 tinha corretamente a ausência e nenhum foi prevista a ausência incorretamente.

Figura 4.3: *Matriz de Confusão do LinearSVC*

	Yes	No	Total
Yes	190	156	346
No	916	890	1806
Total	1106	1046	2152

Autoria Própria

Figura 4.4: *Matriz de Confusão do SVC*

	Yes	No	Total
Yes	2	0	2
No	1104	1046	2150
Total	1106	1046	2152

Autoria Própria

4.1.2 Acurácia

Os resultados das acurácia do SVC e LinearSVC foram aproximadamente 48,69% e 50,18%, respectivamente. Mesmo com resultados muito próximos, o SVC mostrou ser um modelo inábil, pois, como demontrado na matriz de confusão, conseguiu prever corretamente apenas dois *tweets* com a presença do discuro ódio. O que faz com que sua acurácia chegue aos 50%, foi a grande quantidade de *tweets* que o modelo classificou corretamente com a ausência do discurso de ódio. Este exemplo monstra que a acurácia nem sempre é a melhor métrica para avaliação de performance de modelos de aprendizado de máquina.

4.1.3 Precisão

Os resultados da precisão foram obtidas das duas classes, como também uma média entre elas. Para a obtenção dessa métrica foi utilizada o módulo *metrics* da biblioteca do sklearn, este módulo também será utilizado para a obtenção do *recall* e do *F-Measure*.

A precisão do LinearSVC para classe *yes* foi de 54,91% e para classe *no* foi de 49,28% e sua média foi de 52,09%. Para o SVC, a precisão para classe *yes* foi de 100% e para classe

no foi de 48,65% e sua média foi de 74,32%. A alta precisão do SVC se dá pelo fato de que o modelo apenas classificou dois *tweets* com a presença do discurso de ódio e suas classificações foram corretas, porém, esse fato não indica uma alta qualidade do modelo.

4.1.4 Recall

Como a precisão, o *recall* foi utilizado para as duas classes e obtenção da média entre elas. Para LinearSVC, o *recall* para classe *yes* foi de 17,17%, para o *no* foi de 85,08% e sua média foi de 51,13%. Para o SVC, o *recall* para classe *yes* foi de 0,18%, para o *no* foi de 100% e sua média foi de 50.09%. Essa métrica já demonstra mais se o resultados dos modelos são satisfatórios ou não.

4.1.5 F-Measure

Para o LinearSVC, o resultado do *F-Measure* foi de 26,17% para classe *yes* , 62,41% para classe *no* e sua média foi de 44,29%. Para o SVC, o *F-Measure* foi de 0,36% para classe *yes*, 65,45% para classe *no* e sua média foi de 32,90%.

4.2 Justificativa

Este tópico será responsável por explanar algumas justificativas do porque os resultados das métricas foram baixas. A primeira possibilidade se dá pela base de dados de treinamento, pois possuí comentários ofensivos com tópicos de política e esportes, e o foco deste trabalho se dá pelo assunto do discurso de ódio na política. Comentários ofensivos nem sempre são considerados discurso de ódio. Então a obtenção de um corpus sobre o assunto e em pt-br seria crucial para o aumento da qualidade dos modelos.

Como também a definição dos *tweets* oriundos da API foram classificados pelo desenvolvedor sem contar com a ajuda de especialistas. No corpus OffComBr, os pesquisadores contavam com pelo menos dois especialistas e a partir do nível de concordância entre se definia se o comentário era ofensivo ou não. A real definição dos dados de validação também serviria para o aumento da qualidade dos modelos.

Por último, a construção da *query*. Se o trabalho contasse com a ajuda de especialistas, o desenvolvedor poderia realizar uma busca mais consistente com termos de discurso de ódio que comumente são utilizados para a política.

4.3 Resultado das Métricas

Após obtenção das métricas dos classificadores foi possível observar que o modelo que utiliza o LinearSVC é o mais hábil para se obter as métricas de engajamento. As métricas de engajamento foram retirada dos *tweets* que continham o discurso de ódio, um desses *tweets*

pode ser observado na figura 4.5. Dos 190 *tweets* se obteve a soma total de cada métrica que pode ser observado na figura 4.6, os resultados foram, respectivamente, 2433, 13, 19 e 0 para *retweets, replies, likes* e *quotes*.

A métrica com o maior volume é o *retweet*, então foi pensado em retirar o usuário com o maior número de *retweets* e verificar seu poder de engajamento. Na figura 4.7 pode se observar que o usuário possui 700 seguidores, onde 245 retuitaram o texto que contém o discurso de ódio, isso representa aproximadamente 35% de seus seguidores que repassaram a mensagem de ódio. Porém, esse número de *retweets* poderia ser maior que o número de seguidores, pois, se a conta do usuário for pública, qualquer um que visualizar a publicação poderá dar *retweet*.

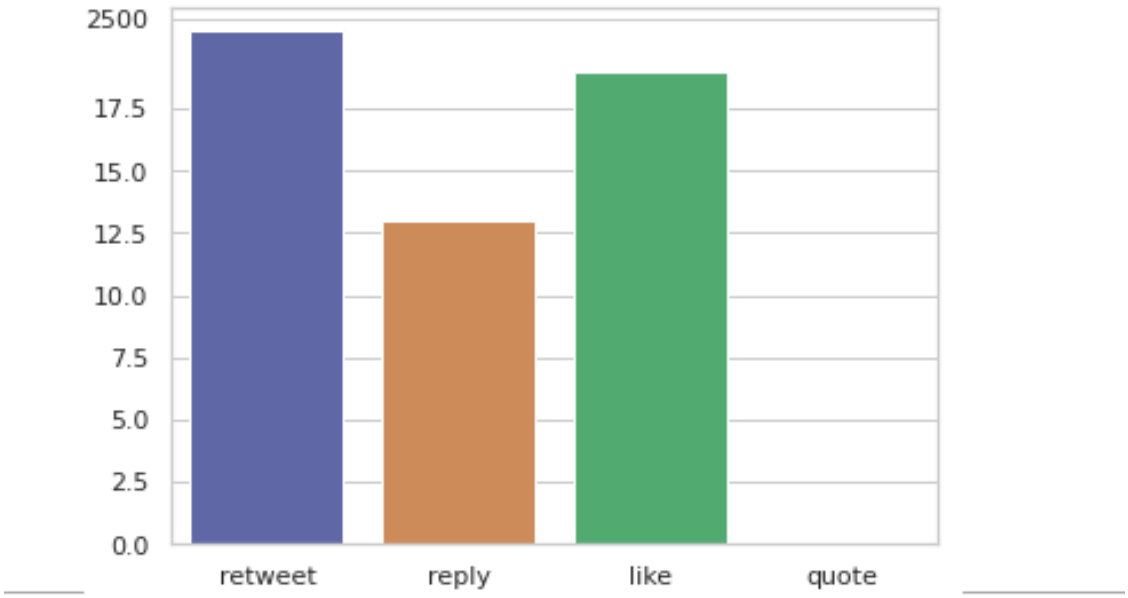
Este trabalho também tenta entender o motivo pela grande diferença entre as métricas. Para isso foi pensado em obter as médias de cada uma das métricas para se obter um comportamento. Os resultados foram, respectivamente, 12,81, 0,07, 0,1, 0,0 para *retweets, replies, likes* e *quotes*. Pode se observar que esses *tweets* podem ter como principal objetivo alcançar o maior número de pessoas, se justificando o utilização de *bots*, porém não é de conhecimento do desenvolvedor se a API do Twitter possui algum mecanismo de verificação de perfis de usuários.

Figura 4.5: *Tweet com Discurso de Ódio Detectado pelo Modelo*

```
1510633066775068690 @dilmabr Quem é mesmo Dilma Rousseff, uma mulher burra que não sabe se
expressar, não pode falar de ninguém, quem comete atrocidades contra a Democracia é o STF, que
soltou o Ladrão do Lula, chefe da maior facção do Brasil, é outra coisa os Ministros do STE e STF
deveriam estar presos.
```

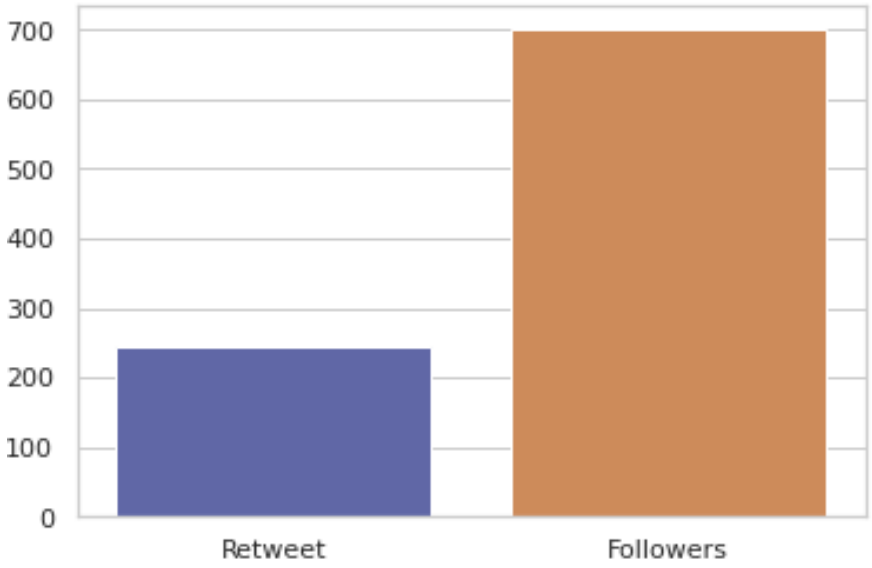
Fonte: Autoria Própria

Figura 4.6: *Plot do Total das Métricas de Engajamento*



Fonte: Autoria própria.

Figura 4.7: *Plot da Quantidade de Seguidores do Usuário com Maior Número de Retweets*



Fonte: Autoria própria.

Capítulo 5

Considerações Finais e Sugestões para Trabalhos Futuros

É notória a capacidade de comportilhamento de ideias e comunicação que as redes sociais possibilitam aos seus milhões de usuários ao redor do mundo. Então, conclui-se este trabalho demonstrando uma pequena fração desse poder quando utilizada para propagar o discurso de ódio. Pode-se ressaltar o desenvolvimento do código por ferramentas e algoritmos até então desconhecidos pelo desenvolvedor, cujo mesmo foi capaz de criar uma aplicação que utiliza uma API externa, detecta o discurso de ódio de forma automatizada e gera dados quantitativos sobre as métricas de engajamento.

Para o melhoramento deste trabalho poderão ser aplicadas técnicas de aprendizado profundo, trazendo mais eficácia na detecção de discurso de ódio. Também poderia se fazer o uso de novas ferramentas de busca para obtenção de dados de outras redes sociais, enriquecendo os resultados e possibilitando o desenvolvimento de novos trabalhos sobre o assunto. E por último, com a ajuda de especialistas, a construção de um corpus de discurso de ódio em pt-br voltado a política eleitoral, facilitando a análise dos algoritmos.

Referências Bibliográficas

[Allan, Richard 2017] Allan, Richard. Hard questions: Who should decide what is hate speech in an online global community. *Facebook newsroom*, v. 27, 2017. 1

[Baggs, Michael 2021] Baggs, Michael. *Discurso de ódio na internet aumentou durante a pandemia, aponta pesquisa*. 2021. Disponível em: <https://www.bbc.com/portuguese/geral-59300051>. Acesso em: 06 de abril 2022. 1, 2

[Benevenuto, Ribeiro e Araújo 2015] BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. *Sociedade Brasileira de Computação*, 2015. 9

[Braun, Daniela 2022] Braun, Daniela. *Brasil tem a quarta maior base de usuários do Twitter no mundo*. 2022. Disponível em: <https://valorinveste.globo.com/mercados/internacional-e-commodities/noticia/2022/04/25/brasil-tem-a-quarta-maior-base-de-usuarios-do-twitter-no-mundo.ghtml>. Acesso em: 12 de outubro 2022. 6

[Brownlee, Jason 2017] Brownlee, Jason. *A Gentle Introduction to the Bag-of-Words Modal*. 2017. Disponível em: <https://machinelearningmastery.com/gentle-introduction-bag-words-model>. Acesso em: 25 de abril 2022. 9

[Brownlee, Jason 2020] Brownlee, Jason. *How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification*. 2020. Disponível em: <https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/#:~:text=The%20precision%20for%20this%20model,Precision%20%3D%2090%20%2F%20120>. Acesso em: 18 de dezembro de 2022. 13

[Brugger, Winfried 2007] Brugger, Winfried. Proibição ou proteção do discurso do ódio? algumas observações sobre o direito alemão e o americano. *Direito Público*, v. 4, n. 15, 2007. 5

[Buitinck *et al.* 2013] BUITINCK, L. *et al.* Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013. 10, 12

[Cyberjcts 2020] CYBERJCTS. *Entre porcos e cachorros.* 2020. Disponível em: <https://cyberjcts.com/python-entre-porcos-e-cachorros/>. Acesso em: 7 de novembro 2022. 8

[Dantas e Netto 2021] DANTAS; NETTO, S. *"Não é preconceito, é a minha opinião": discurso de ódio e os contornos da Liberdade de Expressão no (des)respeito à diversidade*. 2021. Disponível em: <https://www.editoraforum.com.br/noticias/discurso-de-odio-e-os-contornos-da-liberdade-de-expressao-no-desrespeito-a-diversidade/>. Acesso em: 05 de agosto 2022. 1

[dos Santos, R. O. 2022] dos Santos, R. O. Algoritmos, engajamento, redes sociais e educação. *Acta Scientiarum. Education*, v. 44, p. e52736–e52736, 2022. 5

[dos Santos, Vinícius 2018] dos Santos, Vinícius. *Como remover stopwords em Python*. 2018. Disponível em: <https://www.computersciencemaster.com.br/como-remover-stopwords-em-python/>. Acesso em: 04 de julho de 2022. 18

[Felippo *et al.* 2021] FELIPPO, A. D. *et al.* Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In: SBC. *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. [S.l.], 2021. p. 335–343. 8

[Fleiss 1971] FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin*, American Psychological Association, v. 76, n. 5, p. 378, 1971. 16

[Gonzalez e Lima 2003] GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento da linguagem natural. In: *XXIII Congresso da Sociedade Brasileira de Computaçao*. [S.l.: s.n.], 2003. v. 3, p. 347–395. 8

[IBM 2021] IBM. *Visualização da Matriz de Confusão*. 2021. Disponível em: <https://www.ibm.com/docs/pt-br/db2/10.5?topic=visualizer-confusion-matrix-view>. Acesso em: 18 de dezembro de 2022. 12

[Khoong, Wei Hao 2021] Khoong, Wei Hao. *Support Vector Machines In Under 5 minutes: A Brief Introduction To Key Concepts*. 2021. Disponível em: <https://towardsdatascience.com/support-vector-machines-in-under-5-minutes-3074762a49bf>. Acesso em: 30 de junho de 2022. 10

[Lorena e Carvalho 2007] LORENA, A. C.; CARVALHO, A. C. D. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. 10

[Mesquita, Lígia 2018] Mesquita, Lígia. *Denúncias de discurso de ódio online dispararam no 2º turno das eleições, diz ONG*. 2018. Disponível em: <https://epocanegocios.globo.com/Brasil/noticia/2018/11/denuncias-de-discurso-de-odio-online-dispararam-no-2-turno-das-eleicoes-diz-ong.html>. Acesso em: 26 de abril 2022. 2

[Naqa e Murphy 2015] NAQA, I. E.; MURPHY, M. J. What is machine learning? In: *machine learning in radiation oncology*. [S.l.]: Springer, 2015. p. 3–11. 7

[NLTK 2022] NLTK. *Documentation Natural Language Toolkit*. 2022. Disponível em: <https://www.nltk.org/>. Acesso em: 26 de abril de 2022. 13

[Pacheco, André 2021] Pacheco, André. *O que é machine learning?* 2021. Disponível em: <http://computacaointeligente.com.br/conceitos/o-que-e-machine-learning/>. Acesso em: 26 de abril de 2022. 8

[Paiva, Silva e Moura 2019] PAIVA, P. D.; SILVA, V. M. da; MOURA, R. S. Detecção automática de discurso de ódio em comentários online. In: SBC. *Anais da VII Escola Regional de Computação Aplicada à Saúde*. [S.l.], 2019. p. 157–162. 16

[Pedregosa *et al.* 2011] PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011. 10, 12

[Pelle e Moreira 2017] PELLE, R. P. de; MOREIRA, V. P. Offensive comments in the brazilian web: a dataset and baseline results. In: SBC. *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. [S.l.], 2017. 15

[Ruediger *et al.* 2021] RUEDIGER, M. A. *et al.* Discurso de ódio em ambientes digitais. 2021. 2

[Schäfer, Leivas e Santos 2015] SCHÄFER, G.; LEIVAS, P. G. C.; SANTOS, R. H. dos. Discurso de ódio: da abordagem conceitual ao discurso parlamentar. *Revista de informação legislativa*, Senado Federal, v. 52, n. 207, p. 143–158, 2015. 1

[Soares e Monteiro 2015] SOARES, F. R.; MONTEIRO, P. R. R. Marketing digital e marketing de relacionamento: interação e engajamento como determinantes do crescimento de páginas do facebook. *NAVUS-revista de gestão e tecnologia*, Serviço Nacional de Aprendizagem Comercial, v. 5, n. 3, p. 42–59, 2015. 1

[Tweepy 2022] TWEEPY. *Tweepy Documentation*. 2022. Disponível em:<https://docs.tweepy.org/en/stable/>. Acesso em: 18 de agosto de 2022. 7

[Twitter 2016] TWITTER. *About Twitter*. 2016. Disponível em: <https://about.twitter.com/pt/who-we-are/our-company>. Acesso em: 12 de outubro 2022. 6, 7

[Twitter 2022] TWITTER. *Engagement API*. 2022. Disponível em: <https://developer.twitter.com/en/docs/twitter-api/enterprise/engagement-api/overview>. Acesso em: 12 de outubro 2022. 6

[Wadhwa *et al.* 2017] WADHWA, V. *et al.* Maximizing the tweet engagement rate in academia: analysis of the ajnr twitter feed. *American Journal of neuroradiology*, Am Soc Neuroradiology, v. 38, n. 10, p. 1866–1868, 2017. 5

[XAVIER, O. C 2015] XAVIER, O. C. *Utilizando a API do Twitter no desenvolvimento de aplicações web com PHP e cURL.* 2015. Disponível em: <http://www.linhadecodigo.com.br/artigo/3471/utilizando-a-api-do-twitter-no-desenvolvimento-de-aplicacoes-web-com-php-e-curl.aspx>. Acesso em: 9 de junho 2022. 6

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

Monografia do TCC com Ficha catalográfica

Assunto: Monografia do TCC com Ficha catalográfica
Assinado por: Paulo Ribeiro
Tipo do Documento: Dissertação
Situação: Finalizado
Nível de Acesso: Ostensivo (Público)
Tipo do Conferência: Documento Original

Documento assinado eletronicamente por:

- **Paulo Ribeiro Lins Junior**, **COORDENADOR DE CURSO - FUC1 - CCEC-CG**, em 05/02/2023 08:02:14.

Este documento foi armazenado no SUAP em 05/02/2023. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 735749
Código de Autenticação: 3e3f19378d