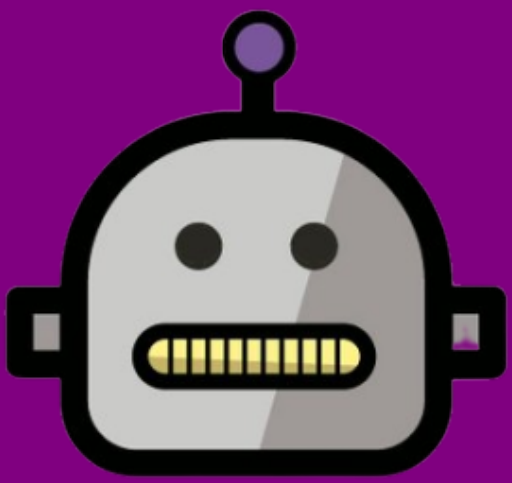
Hackathon
PegaBot

# Pesquisa e Estratégia
# Desafio 2



Grupo 2A - Abigail Lima, Ana Letícia e Pâmela Ferreira

# SUMÁRIO

# INTRODUÇÃO

## O QUE É DISCURSO DE ÓDIO

**O que significa odiar uma pessoa?**

Trata-se de um sentimento negativo em que se deseja mal ao sujeito ou objeto odiado. O ódio está relacionado com a inimizade e a repulsão. As pessoas tentam evitar ou destruir aquilo que odeiam. No caso do ódio relativamente a outro ser humano, o sentimento pode refletir-se através de insultos ou de agressões físicas..

Para entender o que é objetivamente o conceito do discurso de ódio, segue para sua definição como: a utilização de palavras "que tendem a insultar, intimidar ou assediar pessoas em virtude de sua raça, cor, etnicidade, nacionalidade, sexo ou religião" ou ainda à sua potencialidade ou "capacidade de instigar violência, ódio ou discriminação contra tais pessoas" (Winfried Brugger 2007, p. 151).

Atualmente bots estão sendo utilizados para coordenar discursos de ódio e possíveis ataques contra jornalistas e organizações da sociedade civil. Em vista disso, nos foi apresentado o seguinte desafio: desenvolver um mecanismo para monitorar ataques de ódio contra esses importantes atores no processo democrático.

Como reconhecer o discurso de ódio? Segundo a *Cartilha de Orientação para Vítimas de Discurso de Ódio*, da Defensoria Pública do Rio de Janeiro pode ser categorizado como: (I) Insulto e/ou ofensa a uma pessoa, incluindo um grupo socialmente vulnerável ao qual ela pertence; (II) Fala, gesto, expressão que instiga a violência, seja ela explícita ou implícita na fala do agressor.

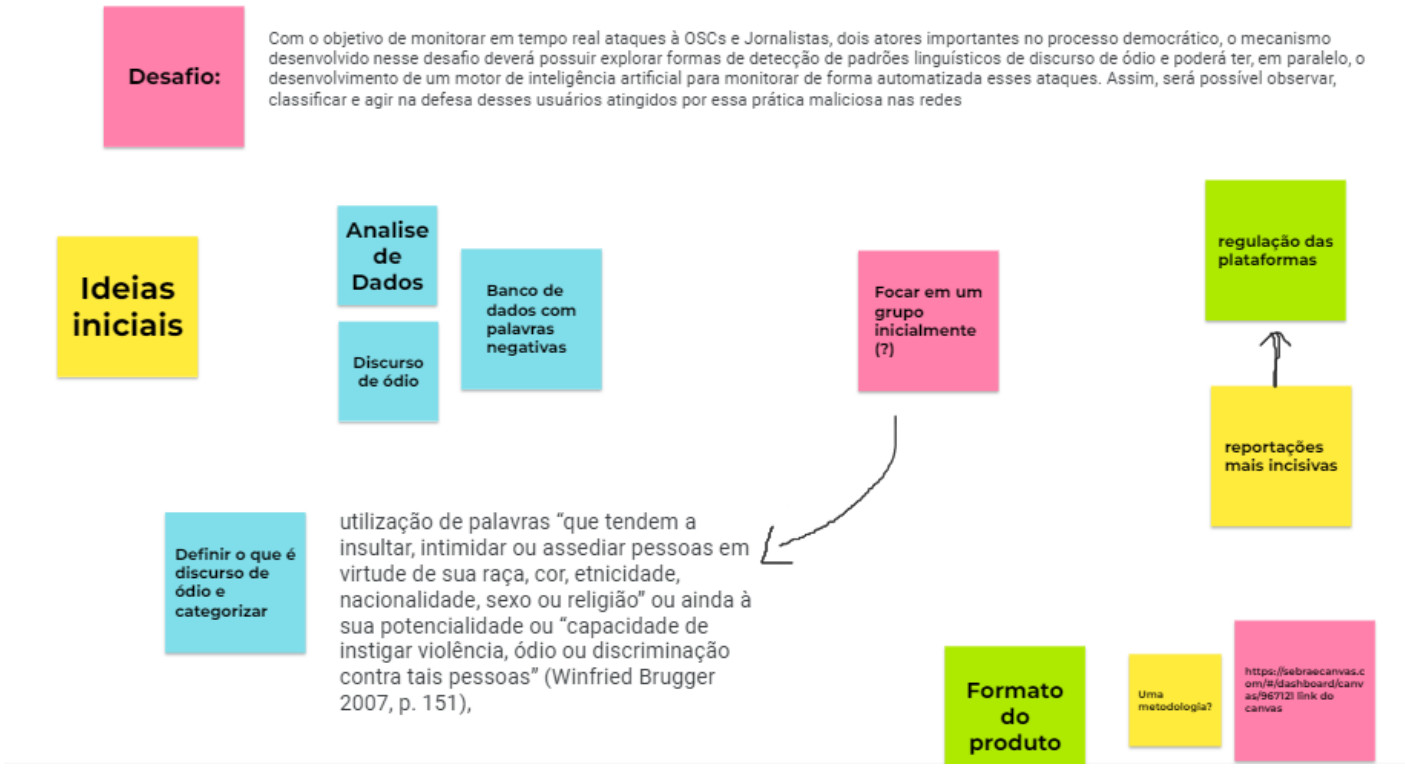
# INTRODUÇÃO

Para o levantamento de dados secundários foi utilizado como estudo o “Relatório de Consultoria para o Pegabot”, onde foi realizada uma análise sobre a identificação do que é um bot e o que não é bot, para isso foi desenvolvido alguns tipos de análises como análise temporal, análise de sentimento, análise de usuário e análise de rede.

Para a definição do que seria o Pegabot, como um software desenvolvido para identificar outros bots e incluso na categoria de Social Bots que são "softwares programados para realizar tarefas automatizadas dentro das plataformas e agir como usuários comuns" (INCTDD, 2021)

Em um dos resultados apresentados com os perfis avaliados se eram ou não bots, os que correspondiam a porcentagem igual ou acima a 70% são provavelmente bots e no relatório foram apresentados dois exemplos de bots, um que tem como funcionalidade ser replicador ou seja, foco em retweets e o outro perfil bot que foca em repetição de hashtags.

# INTRODUÇÃO



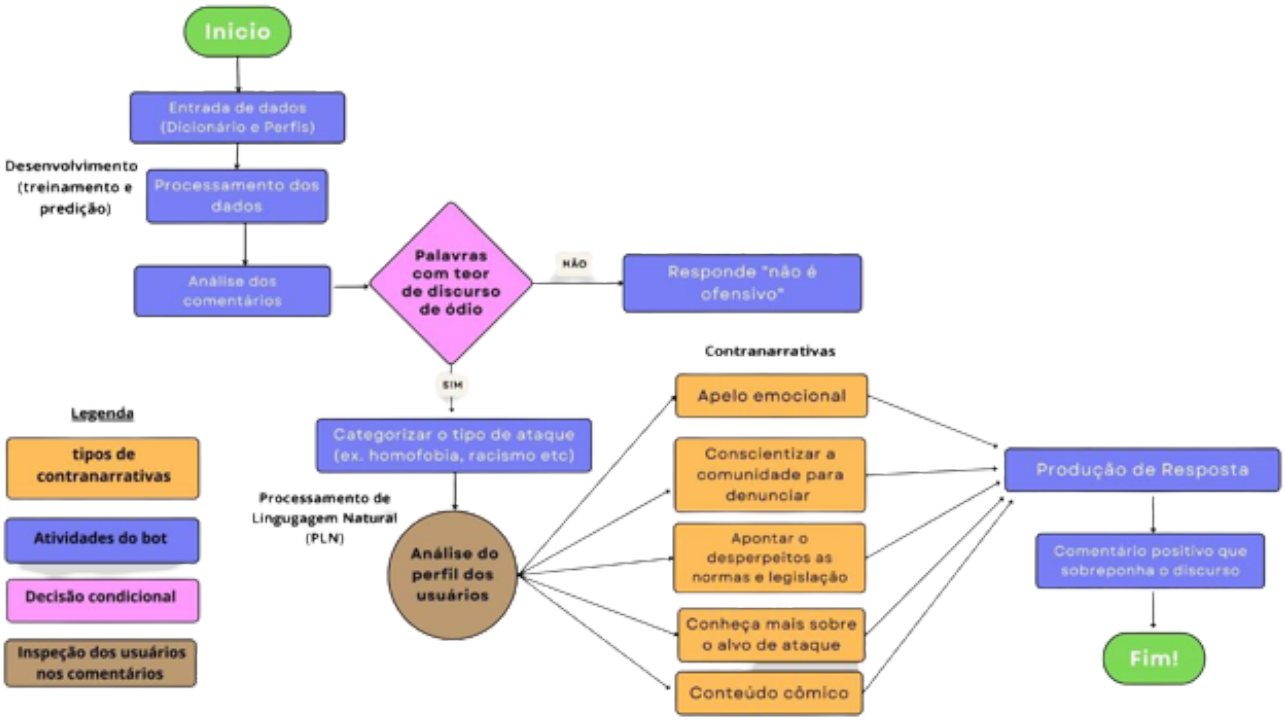

A imagem apresenta o Brainstorming inicial que tivemos para a concepção do mecanismo que seria desenvolvido frete ao desafio proposto.Pensamos nas técnicas de análise de dados para identificar discursos de ódio através de um banco de dados de palavras com conotação negativa e para definir as palavras seria necessário ter uma definição clara do que configura discurso de ódio.

Também discutimos os instrumenstos de regulamentacao que as plataformas utilizam e o que elas poderiam utilizar para inibir tais discursos.

Ademais, definimos o a metodologia utilizada e qual seria o formato de entrega do produto final.

# PROPOSTA

Para combater a disseminação desse tipo de discursos em ataques coordenados a jornalistas e OSC's, desenvolvemos a seguinte proposta de incrementando ao bot que identificaria discursos de ódio e geraria resposta para reflexão na comunidade:




Segue o link para acesso ao fluxograma:
https://www.canva.com/design/DAFLSWusONw/4nM_W0NednW4lrUpDQ7Hfw/edit?utm_content=DAFLSWusONw&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

# DESENVOLVIMENTO E METODOLOGIA

Para os próximos passos para o desenvolvimento do algoritmo, será necessário aprimorar a funcionalidade da IA do Pegabot a partir dos seguintes passos:

- Extrair os dados através da API do twitter;
- Criação de um dicionário com os comentários para o banco de dados;
- Pré processamento dos dados (com o modelo de classificação do texto, através da PLN);
- Fazer a entrada de novas mensagens que serão classificadas em categorias caso o comentário seja considerado como discurso de ódio;
- Retornar uma resposta que se seja específica a categoria.

# DESENVOLVIMENTO E METODOLOGIA

Para o desenvolvimento da ideia utilizamos algumas referências já existentes.Entre elas temos o botslayer (https://osome.iuni.iu.edu/tools/botslayer), um bot desenvolvido por pesquisadores da univerisidade de Indiana capaz de monitorar em tempo real ataques coordenados no Twitter com algoritmo que identifica anomalias através de um índice que indica a probabilidade de manipulação de conteúdos por bots.

Para a criação de um dicionário que classifique palavras de ódios e os tipos de violência relacionados a elas (homofobia, racismo, misoginia, intolerância religiosa) temos como referência os seguintes repositórios:

- Processamento de discurso de ódio violentrometro (https://github.com/cewebbr/violentometro/blob/main/dados/processados/hatespeech_fortuna3%2Boffcombr2.csv)
- HATEBASE PT (https://hatebase.org/)

# DESENVOLVIMENTO E METOLODOGIA

Para análise lexical que indique o contexto onde as palavras estão inseridas seria utilizada a ferramenta Iramutec semelhantemente ao caso apresentado no artigo "Análise Lexical por meio do software Iramuteq: Estudo do significado do Trabalho do Juiz" (FERREIRA et al., 2018).

A classificação do discurso como de ódio e a produção das respostas do bot para contrapor as narrativas de ódio seriam produzidas tomando como base legislações sobre o crimes de ódio, as próprias normas de conduta do Twitter, a Cartilha de orientação para vítimas de discurso de ódio (FGV, 2020) e a Convenção Interamericana contra Toda Forma de Discriminação e Intolerância que fornece conceitos jurídicos para conceber o que são discursos de ódio (SCHÄFER, 2015).

Por fim, a  validação da proposta será por meio da aplicação de formulários  para o público usuário de redes sociais buscando investigar o impacto das respostas do bot através de perguntas sobre Criar um formulário de perguntas, para validar o impacto da proposta, por meio de entrevista com o publico das redes sociais. (como reagiriam a uma sequencia de comentário reagiriam a uma sequencia de comentário de ódio ( se reportariam, bloqueariam, respoderiam etc)

# IMPACTOS, VIABILIDADE E CONSIDERAÇÕES FINAIS

Com o desenvolvimento de mais uma funcionalidade na IA do Pegabot, o algoritmo será aprimorado para a a estratégia de fazer análise dos comentários e classificá-los como discursos de ódio ou não, além da funcionalidade que já existe de identificar possíveis bots.

De maneira estratégica, é estudado que através da automatização dessa nova funcionalidade, o Pegabot será mais ágil em identificar discursos de ódio e sobrepor os comentários de cunho violento que viralizam para outros que tragam informações íntegras e educativas.

Assim, a proposta  se apresenta como inovadora e socialmente impactante já que adiciona a interação automatiza com os usuários para provocar reflexões e incentivar comportamentos nos usuários na contribuição da manutenção de um ambiente saudável nas redes sociais.

# REFERÊNCIAS


BRUGGER, W. PROIBIÇÃO OU PROTEÇÃO DO DISCURSO DO ÓDIO? ALGUMAS OBSERVAÇÕES SOBRE O DIREITO ALEMÃO E O AMERICANO. DIREITO PÚBLICO, V. 4, N. 15, 2007

SCHÄFER, GILBERTO; LEIVAS, PAULO GILBERTO COGO; SANTOS, RODRIGO HAMILTON DOS. DISCURSO DE ÓDIO: DA ABORDAGEM CONCEITUAL AO DISCURSO PARLAMENTAR. REVISTA DE INFORMAÇÃO LEGISLATIVA: RIL, V. 52, N. 207, P. 143-158, JUL./SET. 2015. DISPONÍVEL EM: <HTTPS://WWW12.SENADO.LEG.BR/RIL/EDICOES/52/207/RIL_V52_N207_P143

FERREIRA DA SILVA, RA; DE MORAES SOUSA, M. ANÁLISE LEXICAL POR MEIO DO SOFTWARE IRAMUTEQ: ESTUDO DO SIGNIFICADO DO TRABALHO DO JUIZ . IN: XXI SEMEAD SEMINÁRIOS EM ADMINISTRAÇÃO. SÃO PAULO: NOV. 2018. DISPONÍVEL EM: <HTTPS://WWW.RESEARCHGATE.NET/PUBLICATION/338950421_ANALISE_LEXICAL_POR_MEIO_DO_SOFTWARE_IRAMUTEQ_ESTUDO_DO_SIGNIFICADO_DO_TRABALHO_DO_JUIZ>

FGV, FUNDAÇÃO GETULIO VARGAS.CARTILHA DE ORIENTAÇÃO PARA VÍTIMAS DE DISCURSO DE ÓDIO.RIO DE JANEIRO, 2020.DISPONIVEL EM: HTTPS://BIBLIOTECADIGITAL.FGV.BR/DSPACE/HANDLE/10438/29490

DETECÇÃO DE DISCURSO DE ÓDIO UTILIZANDO VETORES DE FEATURES APLICADOS A UMA BASE NOVA DE COMENTÁRIOS EM PORTUGUÊS: HATE SPEECH DETECTION USING FEATURE VECTORS APPLIED TO A NEW COMMENT BASE IN PORTUGUESE. REVISTA DE SISTEMAS E COMPUTAÇÃO, SALVADOR, V. 10, N. 1, P. 59-68, 6 ABR. 2020. DISPONÍVEL EM: HTTPS://WWW.SEMANTICSCHOLAR.ORG/PAPER/DETEC%C3%87%C3%83O-DE-DISCURSO-DE-%C3%93DIO-UTILIZANDO-VETORES-DE-PAIVA-SILVA/C2933A4D763EEA9155FCB69037244285BB629412. ACESSO EM: 3 SET. 2022.