# DETECÇÃO AUTOMÁTICA DE DISCURSO DE ÓDIO: UM ESTUDO DE CASO NO TWITTER

*Guilherme Marcato Mendes Justiça, Leila Cristina C. Bergamasco*[2]
[1,2]*Departamento de Ciência da Computação, Centro Universitário FEI*
unifgjustica@fei.edu.br leila.cristina@fei.edu.br

***Resumo:*** A presente pesquisa tem como objetivo combater o avanço do discurso de ódio nas redes sociais, utilizando como estudo de caso o Twitter, por meio de análises e classificação de *tweets* envolvendo conceitos de mineração de texto e processamento de linguagem natural. A pesquisa tem parceria da FEI com o Ministério Público Federal com o intuito de contribuir para a segurança cibernética por meio da criação de técnicas computacionais.

## 1. Introdução

A Organização das Nações Unidas (ONU) lançou em 2019 a "Estratégia e Plano de Ação das Nações Unidas contra Discurso de Ódio", no qual define discurso de ódio como qualquer tipo de comunicação verbal, escrita ou comportamental que ataque ou use termos pejorativos ou discriminatórios para se referir a uma pessoa ou grupo com base em sua religião, etnia, nacionalidade, raça, cor, descendência ou gênero. (NATIONS, 2022)

Com os avanços da *internet*, as redes sociais foram criadas com o intuito de conectar as pessoas e de entretê-las, permitindo o compartilhamento de ideias e opiniões em forma de textos, vídeos e imagens. Dado a possibilidade do anonimato, uma parcela dos usuários se sentem seguros e invulneráveis para fazerem quaisquer tipos de comentários. O Twitter é uma das redes sociais mais acessadas atualmente e possui sua própria política contra a propagação de ódio e suas próprias regras de segurança e privacidade, cujo objetivo é manter um ambiente seguro para que todos possam dialogar no Twitter sem que haja desrespeito. Como estratégia, a empresa pede para os usuários que, ao identificarem mensagens que transmitam preconceito, assédio e as variadas formas de propagação de ódio, denunciem o *tweet*. Assim, é possível analisar cada denúncia e impor as devidas consequências para o autor da mensagem.

Adicionalmente, a Safernet é uma instituição brasileira que trabalha com a promoção da cibersegurança e desde 2005 já registrou mais de 2,5 milhões de denuncias relacionadas a crimes de ódios na *internet* (CNN, 2021). A base de dados gerada pela Safernet é consumida por diferentes órgãos governamentais, como o Ministério Público Federal (MPF) que podem utilizar as denúncias e respectivas evidências para autuação criminal dos usuários, inclusive em casos de discursos de ódio. Atualmente essa análise é feita de forma manual por especialistas criminais o que causa sobrecarga dos agentes e lentidão na ação. Em paralelo, existem métodos computacionais que podem auxiliar na detecção do discurso de ódio de forma autônoma, colaborando no combate do discurso de ódio na medida que pode realizar uma primeira filtragem nas denúncias realizadas e otimizar o processo de análise pelo especialista criminal.

Nesse sentido,o presente projeto de pesquisa tem por objetivo implementar e avaliar técnicas de mineração de texto e processamento de linguagem natural para classificar *tweets,* analisando a presença ou ausência do discurso de ódio na publicação. Após a avaliação do método implementado, será discutido com o MPF, que possui parceria com o Centro Universitário FEI, para a realização de testes com um conjunto de denúncias reais da Safernet.

## 2. Metodologia

Dado que o objetivo do presente projeto de pesquisa é implementar e avaliar técnicas de processamento de linguagem natural (PLN) para identificação de discurso de ódio em tweets, o projeto será decomposto em 6 blocos de atividades:

a) **Pesquisa Bibliográfica**: aprofundamento na literatura sobre conceitos, métodos e documentação necessárias para o desenvolvimento do projeto;

b) **Formação da base de dados**: avaliar a partir da revisão feita se existe já um base de dados em português com discurso de ódios já classificados, se poderá traduzir de uma base em inglês ou se será necessário recolher dados diretamente da plataforma Twitter.

c) **Pré-processamento**: serão aplicadas técnicas de pré-processamento da base de dados para a formatação correta dos dados e extração de características por meio de técnicas de PLN.

d) **Desenvolvimento de classificadores**: nessa etapa serão implementados classificadores para a partir das características extraídas, avaliar se o texto possui ou não discurso de ódio.

e) **Avaliação e Análise:** o método desenvolvido será avaliado considerando as métricas de acurácia, precisão e revocação. Serão feitas avaliações com bases de dados genéricas e posteriormente com a base de dados real do MPF.

f) **Documentação e Divulgação**: a partir dos resultados parciais e totais do presente projeto de pesquisa, relatórios e artigos científicos serão elaborados para divulgação em congressos científicos.

A linguagem em utilização é *Python* e bibliotecas de PLN e aprendizado de máquina como *NTK* e *scikit-learn*.

## 3. Resultados Parciais

Até o momento foram feitos testes e implementações de métodos para aprimoramento do projeto. Em primeira etapa foi realizado a coleta de dados para teste. O dataset utilizado foi o "*A Hierarchically Labeled Portuguese Hate Speech Dataset*", um arquivo .csv, que possui 5670 *tweets* para análise que podem, ou não, conter discurso de ódio (DATASET,2019)

Após coleta dos dados foi implementado os métodos de pré-processamento como *stopword*, remoção de acentos e caracteres desnecessários, assim como técnicas de *lowercase* e tokenização para filtragem e simplificação das mensagens. É possível, também, fazer extrações de *tweets* específicos para análise individual. Se quisermos entender como exatamente a palavra "amor", por exemplo, atua no dataset, o código retornaria como resultado a frequência que a palavra aparece, assim como todos os *tweets* com a palavra desejada e quem os publicou.

```
Frase Original: O dia de hoje está muito bonito!!
Frase pré-processada:[ dia, hoje, muito, bonito]
```

Figura 1 - Exemplo de pré-processamento de texto

Implementou-se uma vetorização de todo o documento utilizando duas técnicas diferentes para comparação, o *CountVectorizer* e o *TFIDVectorizer*.

O *CountVectorizer* e o *TfidfVectorizer* são duas técnicas utilizadas em PLN para pré-processar e representar texto de forma numérica, permitindo que algoritmos de aprendizado de máquina trabalhem com dados de texto. Após a limpeza e vetorização do dataset, começa o treinamento da IA para encontrar os padrões que existem no conjunto de dados para que a IA entenda e consiga classificar, automaticamente, cada mensagem como ofensiva ou não. Dois métodos de treinamento foram aplicados para uma comparação de resultado, o *DecisionTree* e *RandomForest*. Utilizou-se também a técnica de *K-fold* para melhores resultados.

*DecisionTree* é um modelo de aprendizado de máquina que representa decisões e suas consequências em uma estrutura de árvore.Os dados são divididos em subconjuntos com base nas decisões tomadas nos nós da árvore. O objetivo é chegar a uma folha da árvore que represente a classificação ou previsão final. Já o *RandomForest* é um algoritmo de aprendizado de máquina que é baseado em múltiplas árvores de decisão. A ideia por trás dele é criar uma coleção de árvores de decisão, onde cada árvore é treinada com uma amostra aleatória dos dados e usa uma seleção aleatória de recursos (variáveis) em cada divisão.

A ideia fundamental do *k-fold cross-validation* (validação cruzada k-fold) é dividir o conjunto de dados em "k" vezes aproximadamente iguais. Em seguida, o modelo é treinado e avaliado nessas k vezes, cada vez usando uma parte diferente como conjunto de teste e as outras k-1 partes como conjunto de treinamento.

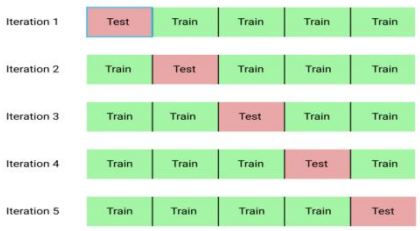


Figura 2 - Técnica de K-fold

E por fim, para conclusões do dataset, foi feita uma retirada de amostragem usando matriz de confusão e as métricas de acurácia, precisão e revocação para verificar a performance do treinamento.

| | | Valor Predito | |
|---|---|---|---|
| | | **Sim** | **Não** |
| **Real** | **Sim** | Verdadeiro Positivo (TP) | Falso Negativo (FN) |
| | **Não** | Falso Positivo (FP) | Verdadeiro Negativo (TN) |

Figura 3 - Matriz de confusão

## 4. Conclusões

O projeto, até o momento, obteve resultados razoáveis, porém não satisfatórios ao que propõem a pesquisa. Para a identificação de discurso de ódio, um resultado adequado seria com uma precisão de 85%, ou mais, de acerto do conjunto de dados. Os resultados obtidos, com as mudanças para comparações, ficam em torno de 72% e 75% de precisão. Assim, para etapas futuras, planeja-se entender o porquê desses resultados e, a partir disso, mudar e melhorar no que for necessário. O motivo do resultado abaixo do esperado pode ser questões de dataset, das técnicas de pré-processamento ou dos métodos de treinamento utilizados.

## 5. Referências

NATIONS, United. What is hate speech? | United Nations. [S.l.: s.n.], 2022. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech. [Accessed 13-Nov-2022]

CNN. Discurso de ódio nas redes sociais repete padrão de preconceitos da sociedade. [S.l.: s.n.], 2021. https://www.cnnbrasil.com.br/nacional/discurso-de-odio-nas-redes-sociais-repete-padrao-de-preconceitos-da-sociedade/. [Accessed 13-Nov-2022].

DATASET,Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A Hierarchically-Labeled Portuguese Hate Speech Dataset. Association for Computational Linguistics.

## Agradecimentos

À instituição FEI e à minha orientadora, responsável pela oportunidade e por me guiar para construção deste projeto.

[1] Aluno de IC do Centro Universitário FEI. Projeto com vigência de 04/2023 a 04/2024.