Detecção de Discurso de Ódio no Twitter

Guilherme Marcato Mendes Justiça[1], Leila Cristina C. Bergamasco[2]
[1] DEPT. CIÊNCIA DA COMPUTAÇÃO, CENTRO UNIVERSITÁRIO FEI
[2] DEPT. CIÊNCIA DA COMPUTAÇÃO, CENTRO UNIVERSITÁRIO FEI

47

SICFEI XIII Simpósio de Iniciação Científica, Didática e de Ações Sociais de Extensão da FEI

Sobre o projeto de pesquisa:

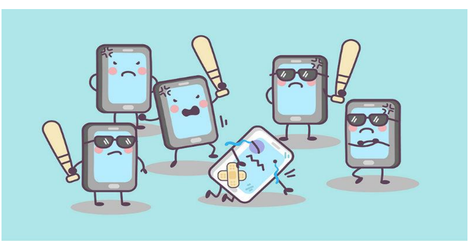
Esse projeto está sendo desenvolvido em parceria com o MPF com o intuito de contribuir para a segurança cibernética por meio da criação de técnicas computacionais.

Para isso, usaremos técnicas de Processamento de Linguagem Natural, métodos de Aprendizado de Máquina, ferramentas para análise e avalição de resultados .

Mas, o que significam esses termos? Para que servem?

O que é discurso de ódio?

A ONU define como: "Qualquer tipo de comunicação verbal, escrita ou comportamental que ataque ou use termos discriminatórios para se referir a uma pessoa ou grupo.".



O Aprendizado de Máquina:

O Aprendizado de máquina consiste em ensinar a inteligência artificial, a partir dos dados fornecidos, a encontrar o melhor caminho para identificar padrões com o mínimo de intervenção humana.

- **Recomendação de Conteúdo**
- **Tradução de textos**
- **Reconhecimento de fala/facial**

O Processamento de Linguagem Natural (PLN):

Como conhecido, o PLN é um ramo da inteligência artificial que permite com que os computadores entendam os humanos.

O objetivo do processamento de linguagem natural é facilitar a extração de sentido do texto.

A tarefa da Mineração de Texto:

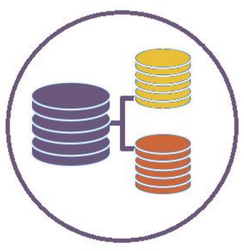
A Mineração de Texto tem como objetivo "escavar" todo um documento em busca de padrões ocultos e utéis dentro dos texto.

E COMO ISSO SE APLICA NESSE PROJETO?

DATASET

Dataset é um conjunto de dados coletados para análise. Utilizamos para desenvolvimento do projeto o dataset: "*A Hierarchically Labeled Portuguese Hate Speech Dataset*". Retirado no Kaggle, ele possui 5670 tweets sendo 1788 classificados como "0" e 3882 classificados como "1".

Texto	*Class*
Que dia lindo	**0**
Você é #*!!$@	**1**



Frase Original:
Bem vindos a SICFEI, aproveite os artigos!

Frase Pré-processada:
[bem, vindos, sicfei, aproveite, artigos]

PRÉ-PROCESSAMENTO

Nessa etapa ocorre a filtragem do texto do dataset, para facilitar a análise dos dados utilizamos técnicas como:

-Remoção de stopwords/valores nulos/acentos/caracteres específicos não desejados.

-Lowercase.

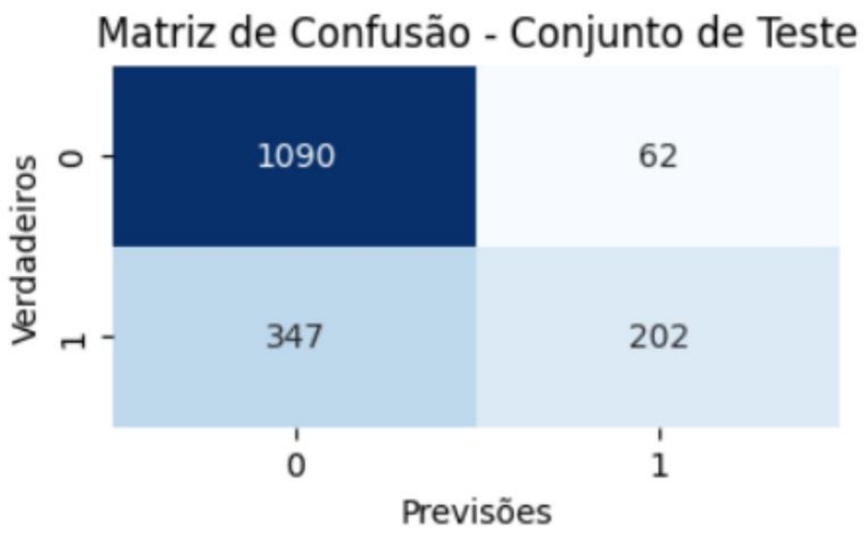
-Tokenização.

TREINAMENTO

Nesse processo, o aprendizado de máquina e a mineração de texto são utilizados.

Após a limpeza do dataset, começa o treinamento da IA para achar os padrões dos textos e, sozinha, aprender a identificar e classificar cada frase.

O treinamento funciona da seguinte forma:
Primeiro transformamos o dataset em vetores, utilizamos dois: *CountVectorizer* e *TFIDVectorizer*.
Escolhemos dois métodos de treinamento, *DecisionTree* e *RandomForest*, para comparação do resultado.
A partir disso, definimos as variáveis para teste, assim como o *test_size* do dataset. Outras técnicas utilizadas foram o *K-fold* e validação cruzada para melhor performance do treinamento.



AVALIAÇÃO

Depois do treinamento, a IA testa o que aprendeu com o restante do dataset. E assim, por meio de outras técnicas, conseguiremos entender se a IA classificou os textos corretamente.
Matriz de confusão, F1-score, entre outras, são formas de visualizar a precisão da nossa IA.

E assim, com um resultado satisfatório de avaliação, usaremos as mesmas técnicas utilizadas com a base de dados fornecida pelo MPF.