PROPOSTA DE PROJECTO PARA IA

Curso: Engenharia Informática

Turno: Diurno

Daniela Filipe - 30009692 Igor Paulo – 30009443 Ricardo Lourenço-30011111

Tema: Detecção de discurso de ódio em redes sociais

1.Declaração do Problema e Definição da Tarefa

Nosso sistema propõe-se a identificar automaticamente mensagens que contenham discurso de ódio em redes sociais. A entrada do sistema serão mensagens ou postagens textuais retiradas de plataformas como Twitter ou Facebook, enquanto a saída será uma classificação binária: “Discurso de Ódio” ou “Não É Discurso de Ódio”.

O problema do mundo real que buscamos resolver é a proliferação de conteúdo prejudicial online, que pode gerar impacto negativo na sociedade, como violência, discriminação e cyberbullying. Este sistema será uma ferramenta útil para moderação de conteúdo, auxiliando as plataformas a criar um ambiente mais seguro para os usuários. O âmbito é razoável: não buscamos identificar todos os tipos de conteúdo ofensivo, mas focar especificamente no discurso de ódio.

2.Comportamento de Entrada/Saída

As entradas são mensagens textuais em formato estruturado (ex.: JSON ou CSV), contendo informações como autor, conteúdo da mensagem e data. Exemplos concretos incluem:

- Entrada: “Você não deveria estar aqui. Ninguém gosta de pessoas como você.”
- Saída: “Discurso de Ódio”

Outro exemplo:

- Entrada: “Hoje é um dia maravilhoso para todos!”
- Saída: “Não É Discurso de Ódio”

Os dados preliminares usados incluem conjuntos de dados públicos, como o Hate Speech Dataset do Twitter e o Kaggle’s Hate Speech Dataset. Estes contêm exemplos rotulados manualmente, permitindo treinar e avaliar modelos de aprendizado de máquina.

3.Métrica de Avaliação

Usaremos a precisão, a métrica F1 e a área sob a curva ROC (AUC-ROC) para avaliar o desempenho. A precisão mede a proporção de previsões corretas, enquanto a métrica F1 avalia o equilíbrio entre precisão e recall, sendo ideal para dados desbalanceados. A AUC-ROC ajudará a entender a capacidade do modelo de separar corretamente as classes. Estes métodos fornecem uma visão abrangente da eficácia do sistema.

4.Trabalhos Relacionados

Vários projetos abordaram o tema, incluindo:

- *Davidson et al. (2017)*: Utilizaram regressão logística para classificar discurso de ódio em tweets, obtendo boas taxas de precisão.
- *Badjatiya et al. (2019)*: Implementaram modelos baseados em aprendizado profundo usando embeddings como Word2Vec e Glove.

Nosso trabalho irá combinar técnicas semelhantes e explorar novas abordagem como Transformers.

5.Baseline e Oráculos

O baseline escolhido é um classificador de Naïve Bayes, que é simples e eficiente para classificação de texto. O oráculo será baseado em uma rede neural profunda (DNN) com embeddings, que representa o limite superior esperado de desempenho. Esta abordagem ajudará a avaliar a dificuldade da tarefa e fornecerá comparações claras entre soluções simples e sofisticadas.

6.Metodologia

Nosso método utiliza aprendizado de máquina supervisionado. Os passos incluem:

1. **Limpeza e Normalização de Dados**: Remoção de URLs, mencões, stopwords e normalização de palavras.
2. **Vetorizadores**: Usaremos TF-IDF e embeddings como BERT para transformar texto em representações numéricas.
3. **Modelos**: Experimentaremos com regressão logística, SVM, Random Forests e Transformers (ex.: BERT).
4. **Validação**: Dividiremos os dados em conjuntos de treino, validação e teste, usando validação cruzada para evitar overfitting.

Desafios:

- Dados desbalanceados: Precisaremos de técnicas como oversampling ou focal loss.

- Ambiguidade textual: Frases contextualmente neutras podem ser mal interpretadas.

Prós e Contras dos Métodos:

- Simples (ex.: Naïve Bayes): Fáceis de implementar, mas com precisão limitada.
- Complexos (ex.: Transformers): Alta precisão, mas exigem mais recursos computacionais.

A escolha final dos algoritmos será baseada no melhor compromisso entre precisão e eficiência computacional. Este sistema pode ser uma base robusta para moderadores de conteúdo em redes sociais e plataformas online.