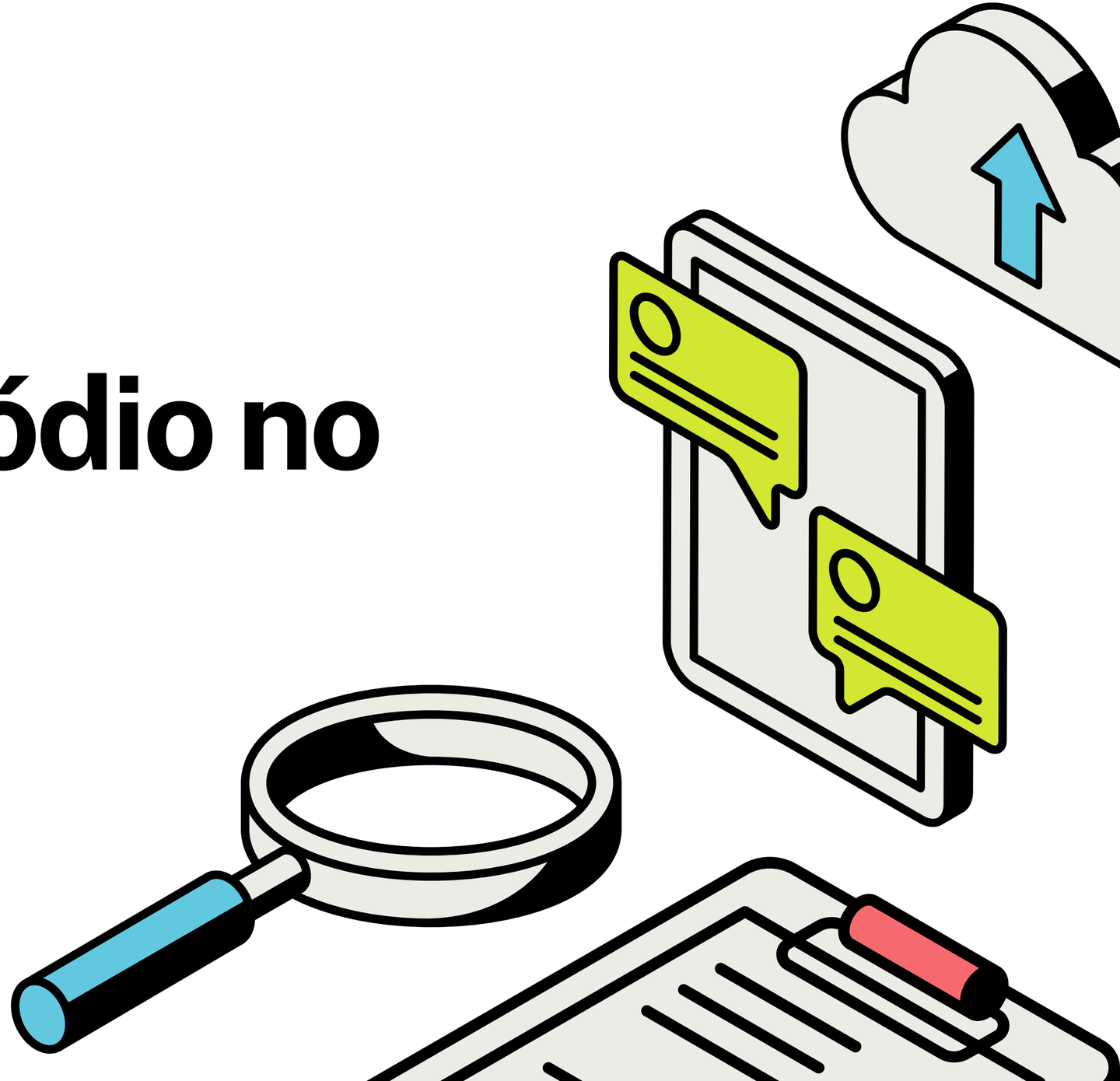
Análise de discursos de ódio no Twitter/X

Alex Alves Cardoso
Guilherme César Athayde
Julia Campanelli Granja
Yara dos Santos Rodrigues

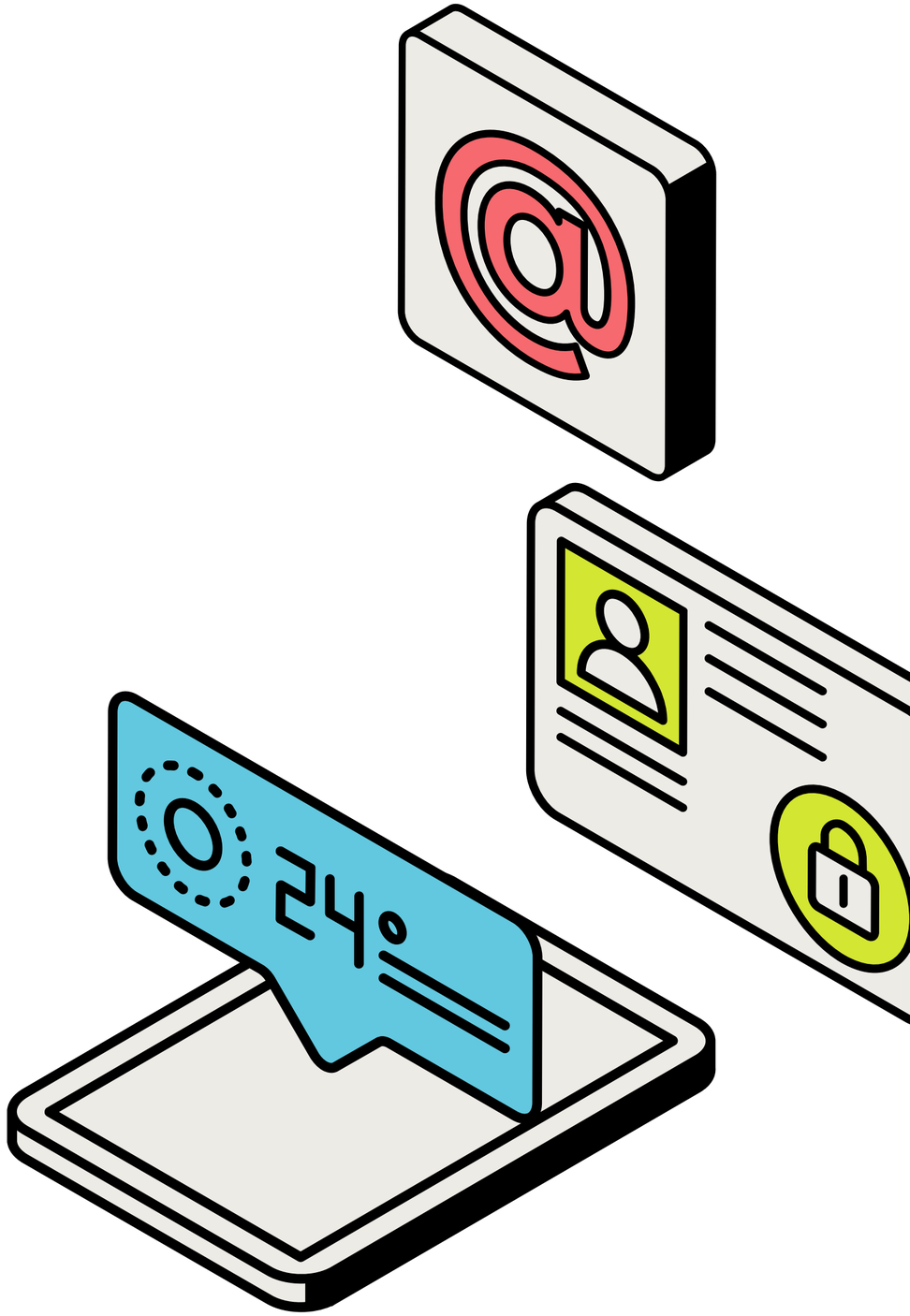


Sumário:

1. Introdução
2. Contextualização
3. Motivação do Problema
4. Aplicação Escolhida
5. Referências

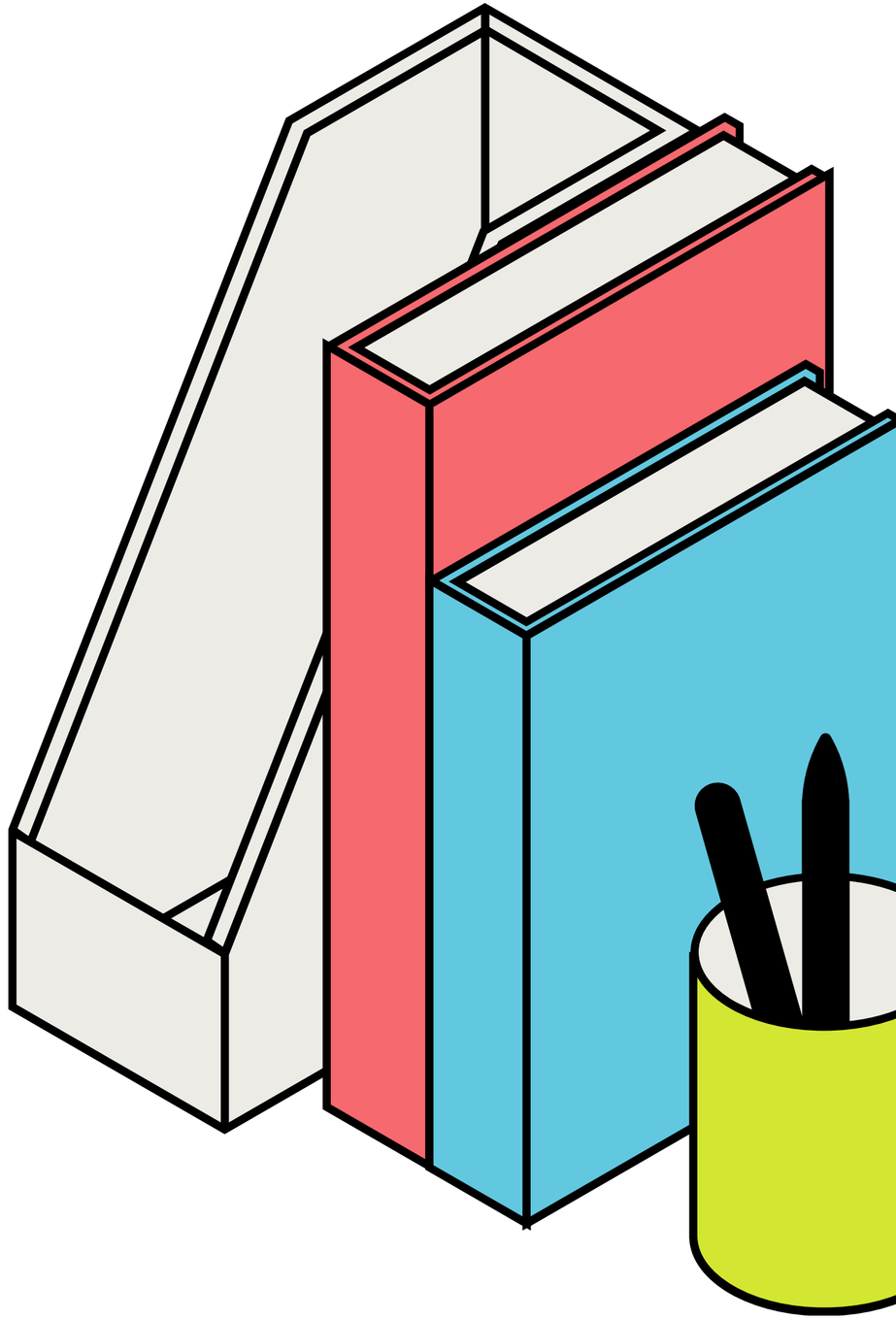
Introdução: Mídias Sociais e Discursos de Ódio

- As mídias sociais conectam milhões globalmente, mas enfrentam desafios como o discurso de ódio.
- No X/Twitter, essas mensagens aparecem como ofensas, ameaças e incitação à violência.
- Consequências: danos psicológicos, polarização social e marginalização de comunidades.
- O debate sobre liberdade de expressão e moderação permanece intenso.

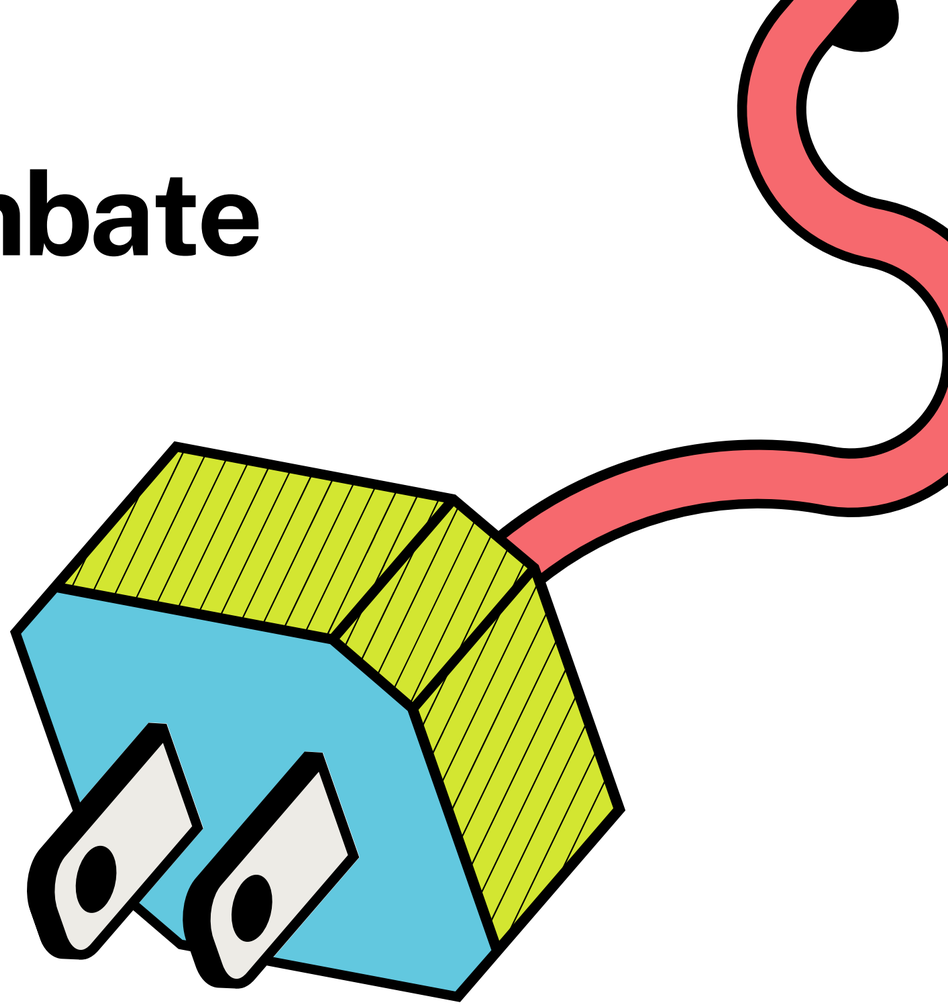


Contextualização: Impactos do Discurso de Ódio Online

- Alvos: minorias raciais, mulheres, LGBTQIA+, religiosos e pessoas com deficiência.
- Consequências: isolamento, violência real e ambientes digitais hostis.
- Plataformas enfrentam desafios na moderação devido à subjetividade e códigos disfarçados.
- Soluções urgentes buscam criar espaços online mais seguros e inclusivos.



Motivação do Problema: Combate ao Discurso de Ódio com IA

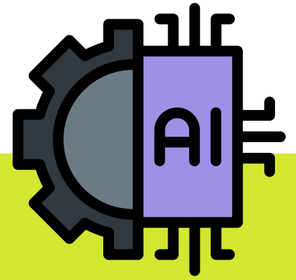


- IA pode identificar padrões e mitigar discursos de ódio em tempo real.
- Ferramentas avançadas analisam grandes volumes de dados e apoiam decisões.
- Objetivo: promover segurança, respeito e inclusão em plataformas digitais.

Aplicação Escolhida

Data Mining

Banco de Tweets



Algoritmos

Data Mining

- **Processo de exploração de grandes volumes de dados.**
- **Identifica padrões, tendências e informações úteis.**
- **Combina estatística, aprendizado de máquina e bancos de dados.**

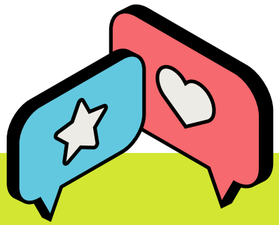
Coleta em escala: Tweets em grande volume.

Identificação de padrões: Reconhece palavras-chave e hashtags relevantes.

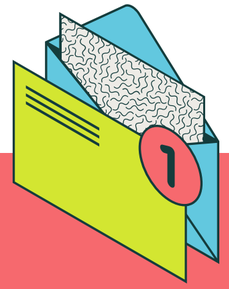
Automação: Reduz tempo de análise manual.

Base para IA: Treina modelos para análise de discurso de ódio.

Banco de Tweets



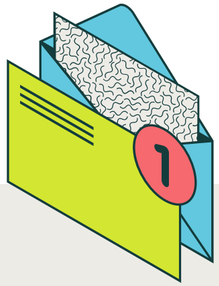
Utilização de diversas bases de dados de tweets, com categorias como discurso de ódio, linguagem tóxica, ofensiva e neutra.



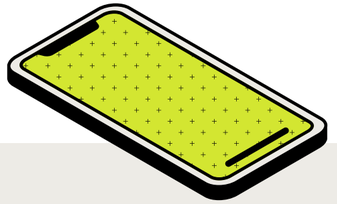
Dados disponíveis em diferentes idiomas, incluindo inglês e português.

Proporciona uma análise abrangente e comparativa, essencial para os testes.

Algoritmos

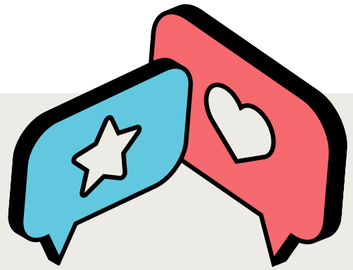


Classificação de Sentimentos:
Naive Bayes, SVM, Random Forest: Para categorizar textos como discurso de ódio ou neutros.

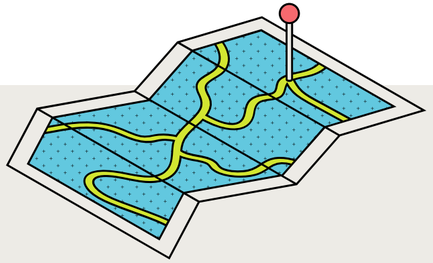


Detecção de Linguagem Abusiva: TF-IDF e Word Cloud: Identifica palavras-chave e termos associados a discurso de ódio.

Algoritmos



Análise de Redes Sociais:
Análise de Grafos - Mapeia a propagação e influência de contas que disseminam discurso de ódio.



Deep Learning: LSTM, BERT/BERTweet: Modelos para capturar contexto e sarcasmo no discurso de ódio.

Referências:

LEE, Jaeyoung; PITTALUGA, Francesco; MENG, Yuhao. Intersectional Bias in Hate Speech and Sentiment Analysis: A Case Study on Twitter. 2023. Disponível em: https://github.com/jaeyk/intersectional-bias-in-ml. Acesso em: 16/12/2024.

TEIXEIRA, Thiago. Toxic Language PTBR Classification. Disponível em: https://github.com/thiagot35/Toxic_Language_PTBR_Classification. Acesso em: 17/12/2024.

Obrigado!