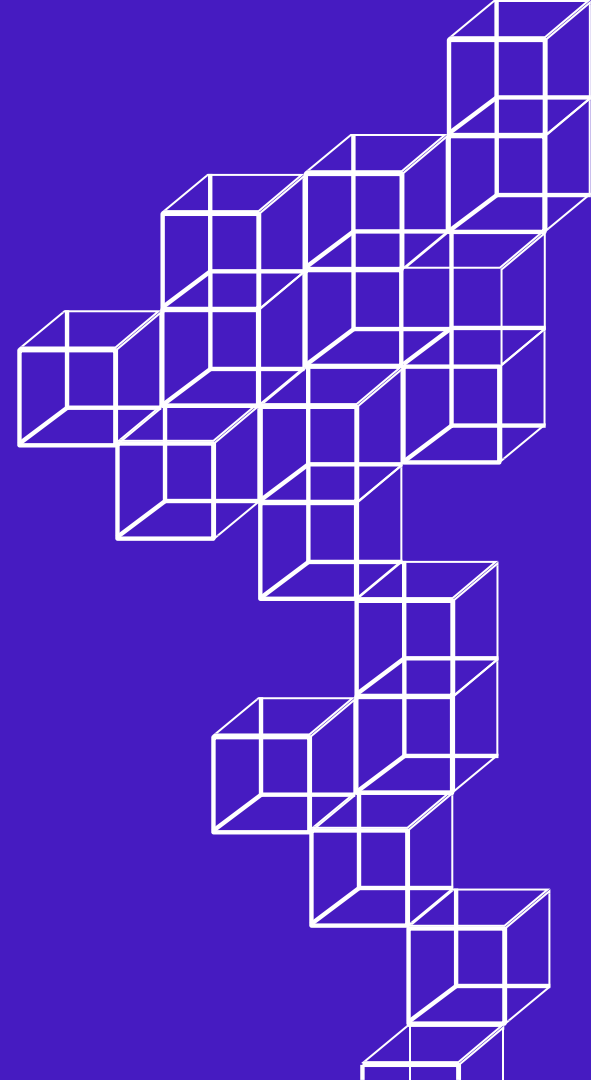


Construindo e avaliando uma ferramenta de detecção de violência eleitoral na Web

Henrique S. Xavier, Ana E. Duarte, Beatriz R. Corrales *et al.*
Rio de Janeiro, 20 de abril de 2023



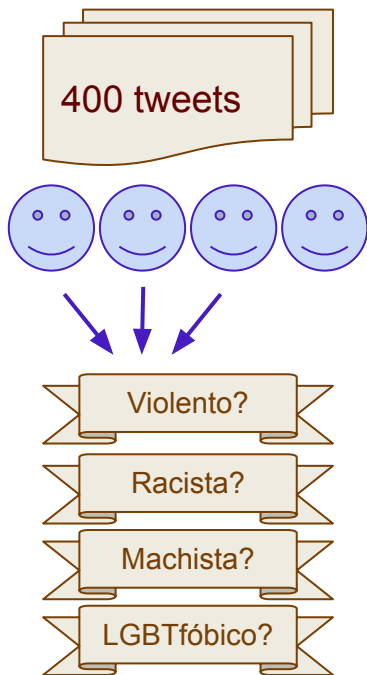
Construção do modelo

- BERT pré-treinado em MLM, NSP com brWaC (3,5M páginas)
- Ajustado a Pelle & Moreira (2017) + Fortuna *et al.* (2019): 926 exemplos e 4713 contraexemplos.

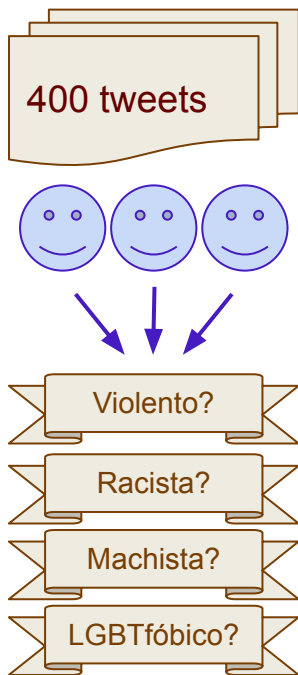
Métrica	BERTimbau	Bag-of-words	Aleatório
Acurácia	0,901	0,872	0,510
F1	0,596	0,471	0,281
Precisão	0,689	0,736	0,185
Revocação (Recall)	0,525	0,346	0,584

Anotação de tweets direcionados a candidaturas

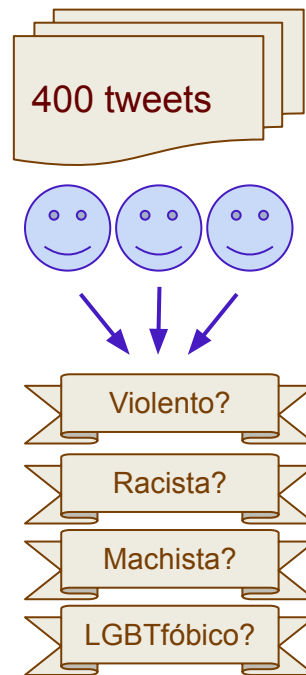
Pessoas negras



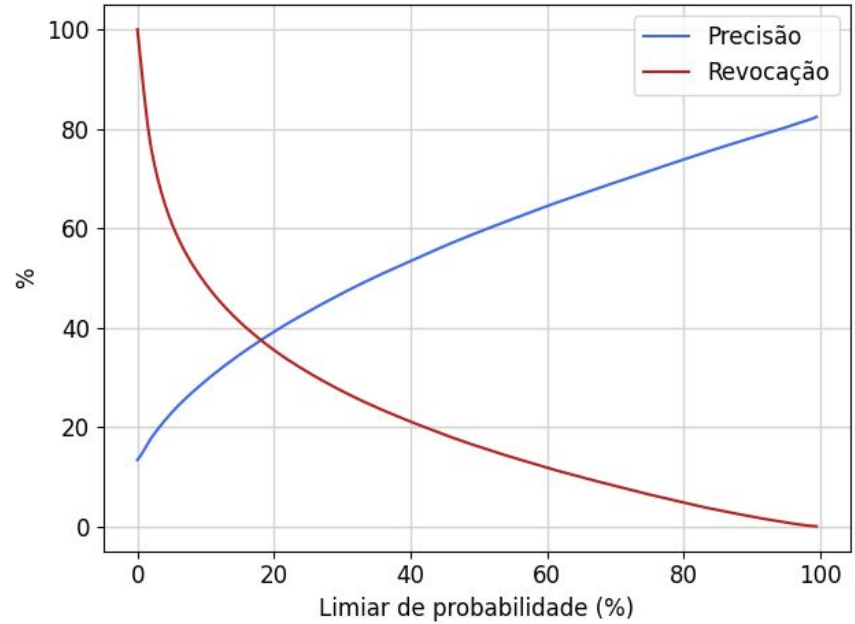
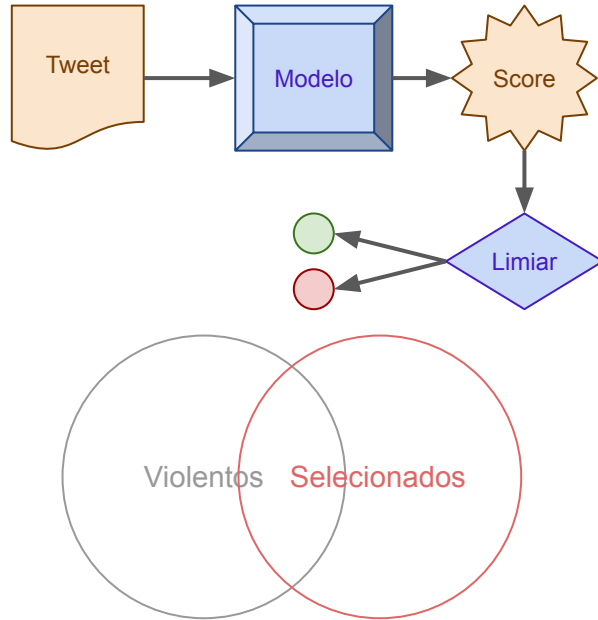
Mulheres



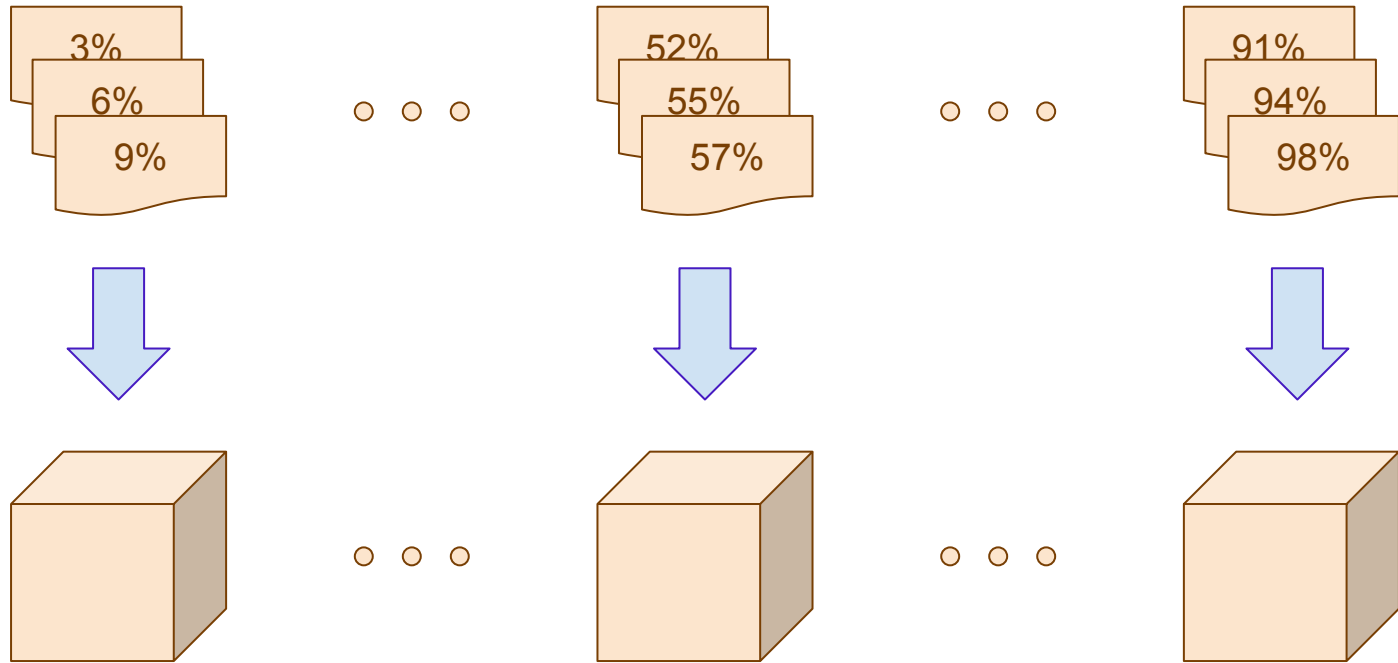
LGBT



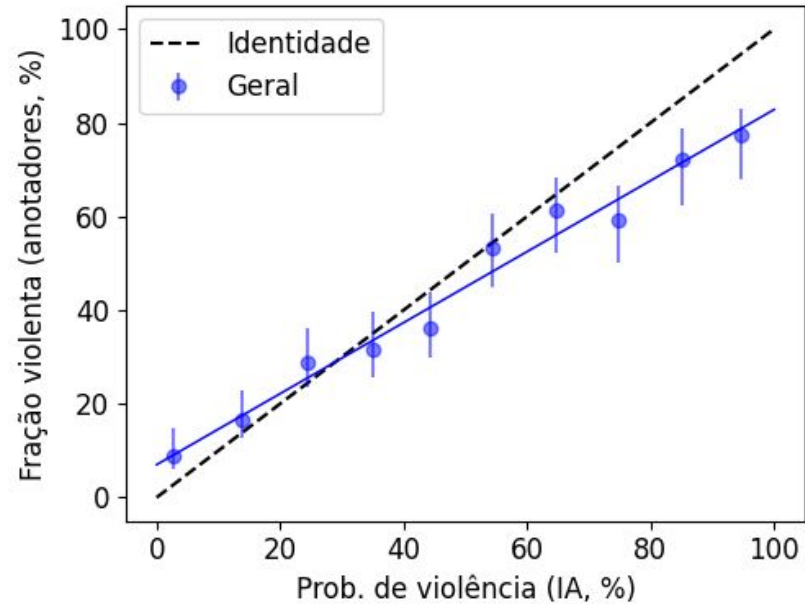
Grau de acerto em tweets individuais




Estimativa de fração violenta de um corpus




Estimativa geral da probabilidade de violência




Avaliação da violência por objeto do comentário

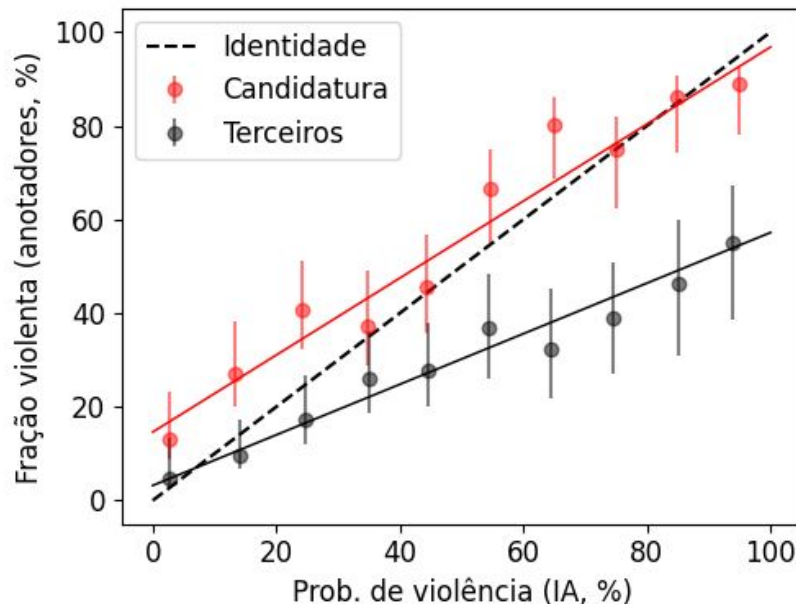
 ...
Cansou de passar vergonha na câmara, agora passa na assembleia kkkkkkk que catracada

Comentário sobre a candidatura

 ...
Impressionante como vc se preocupa com twitter ao invés de trabalhar
Deputado vag@bonds vc cara...

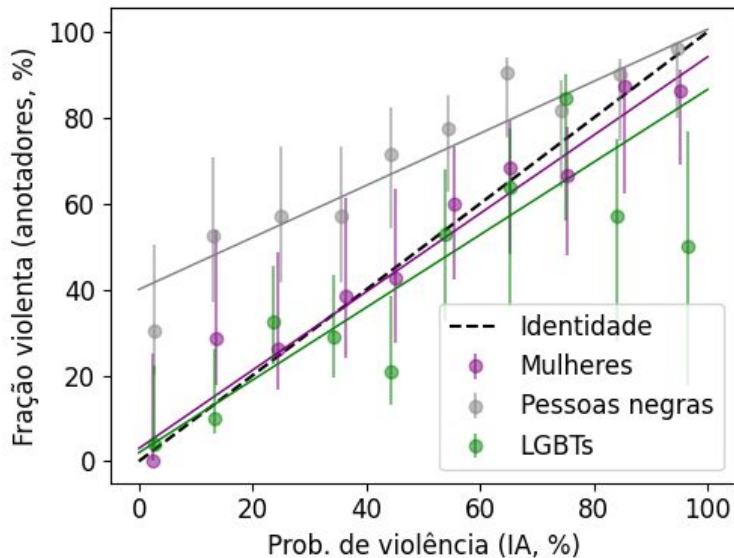
Comentário sobre terceiros

 ...
eu depois dessa ia dormir 😴

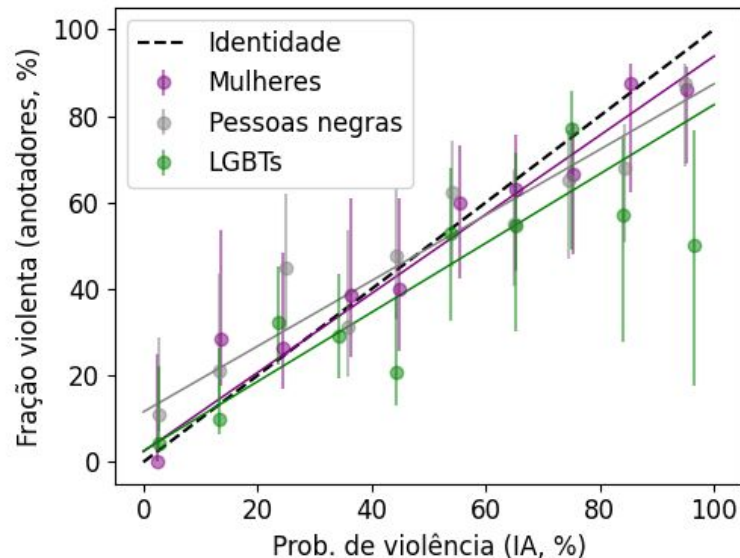


Avaliação da violência por grupo social

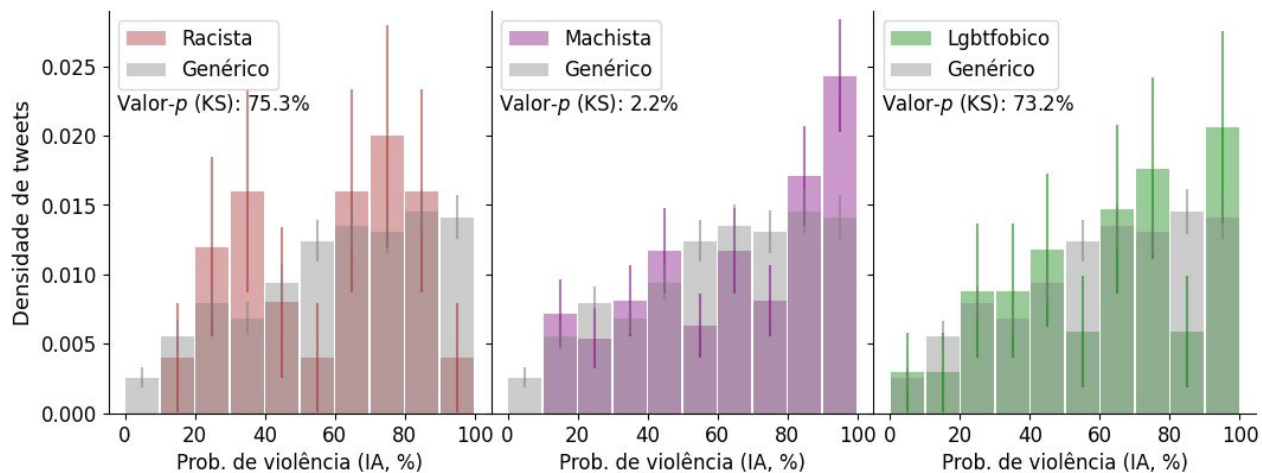
Trio original



Trio alternativo



Avaliação por tipo de violência



Conclusões: desafios

- **Obstáculos ao modelo**

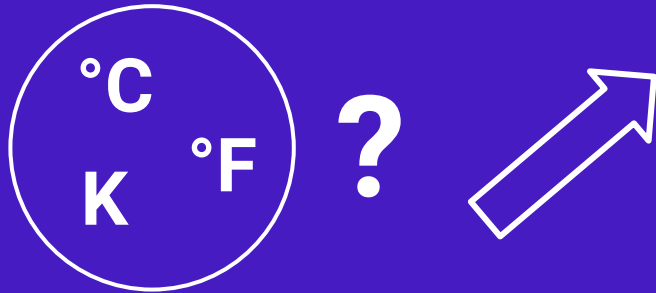
Subjetividade, falta de contexto, ironia (apesar de rara), descarte de imagens e vídeos, ataques indiretos

- **Incapacidade de distinguir ataque de reação**

- **Revisão humana continua indispensável**

Conclusões: oportunidades

- Modelo pode funcionar como “termômetro” da violência



- PS: corpus em comparação devem ter componentes similares

ceweb.br nic.br cgi.br

OBRIGADO



Henrique S. Xavier

E-mail: hxavier@nic.br

<http://www.ceweb.br>

Design Beatriz Rocha