Detecção de Discurso de Ódio e Viéses nas Redes

Trabalho de Conclusão - Demoday

Data Science & Machine Learning – Tera

Bruno Donato, Dagna Chagas, Daniela Nomura,

José Valdeir, Thiago Paim

Discurso de ódio – Definição

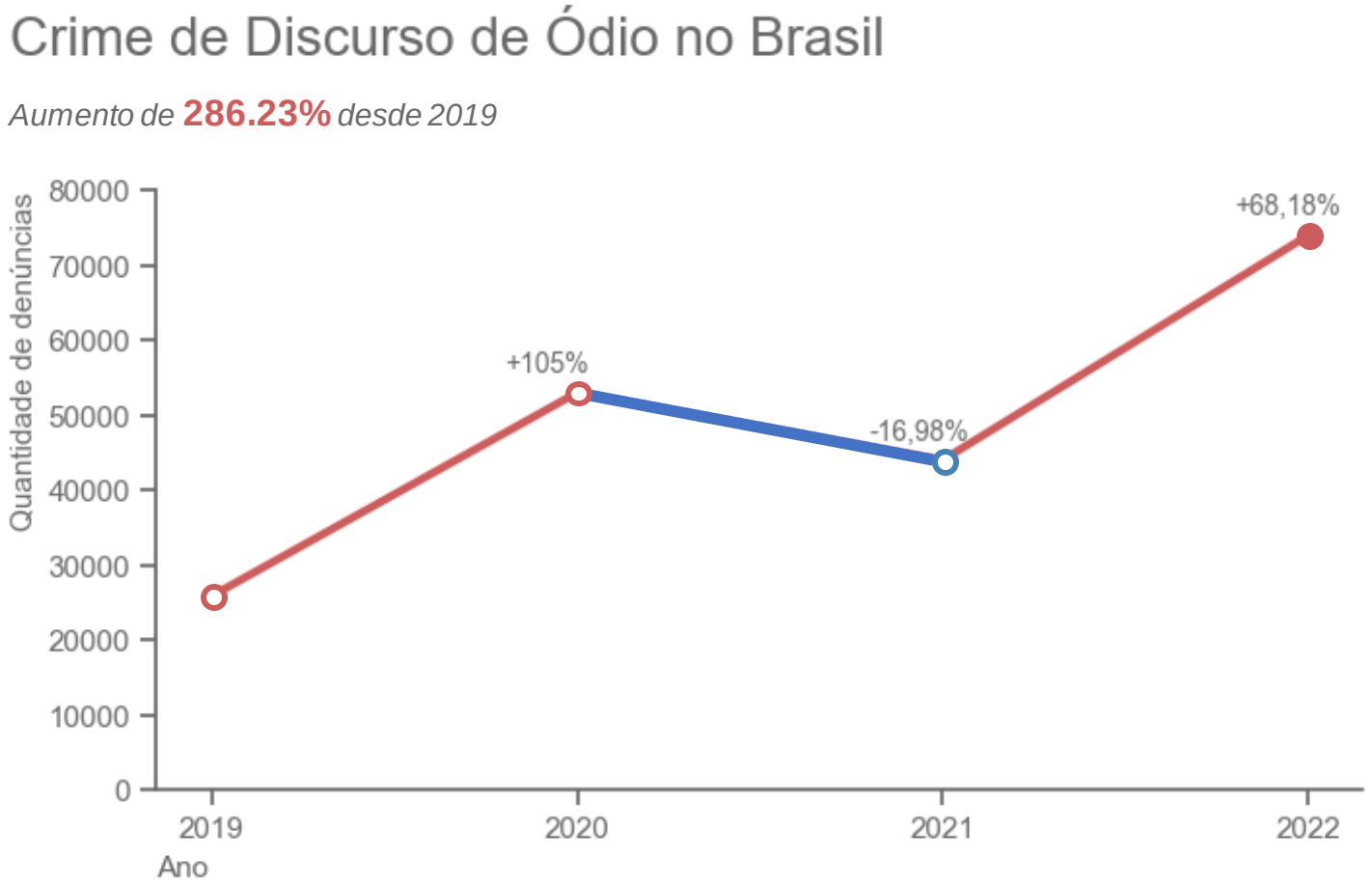
“...muito vinculado à utilização de palavras. Não é só uma violência física, mas virtual e verbal que tende a insultar, intimidar ou assediar pessoas em virtude da sua raça, cor, etnicidade e assim por diante... potencialidade ou a capacidade de instigar violência, o ódio ou discriminação contra as pessoas.”

Cristina Godoy Bernardo de Oliveira

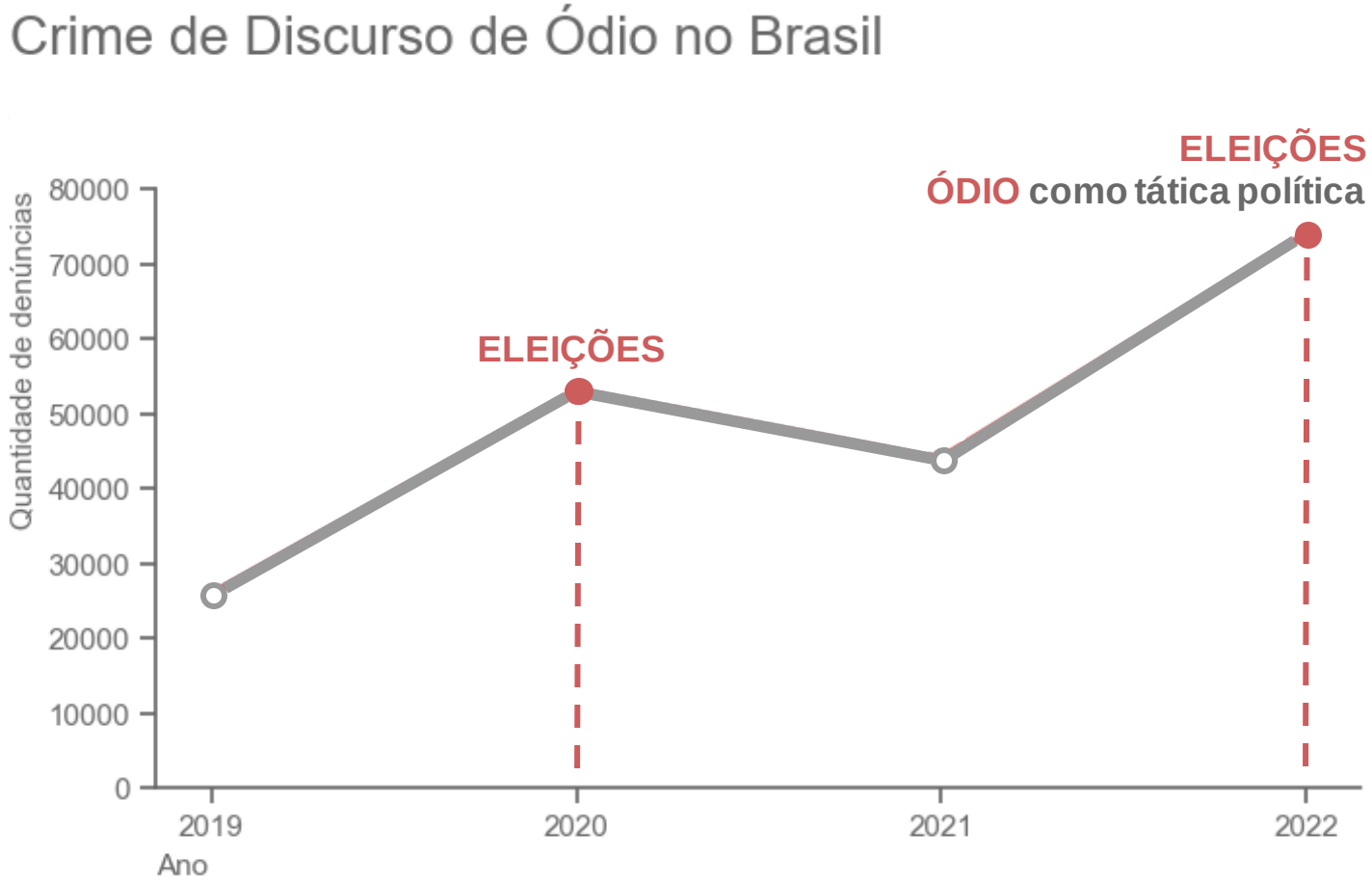
Professora da Faculdade de Direito de Ribeirão Preto - USP

Grupo de Estudos Direito e Tecnologia do Instituto de Estudos Avançados Polo Ribeirão Preto - USP

Crimes de discurso de ódio



Crimes de discurso de ódio



Crimes de discurso de ódio

Facebook - 3° trimestre de 2020:

Até 0.11% discursos de ódio a cada 10 mil visualizações

Crimes de discurso de ódio

Neonazismo

↑61%

De 2020 a 2021

Intolerância religiosa

↑456%

De 2021 a 2022

Xenofobia

↑874%

De 2021 a 2022

Crimes de discurso de ódio

Misoginia

↑251%

De 2021 a 2022

Maior número de feminicídios desde que a lei entrou em vigor no Brasil

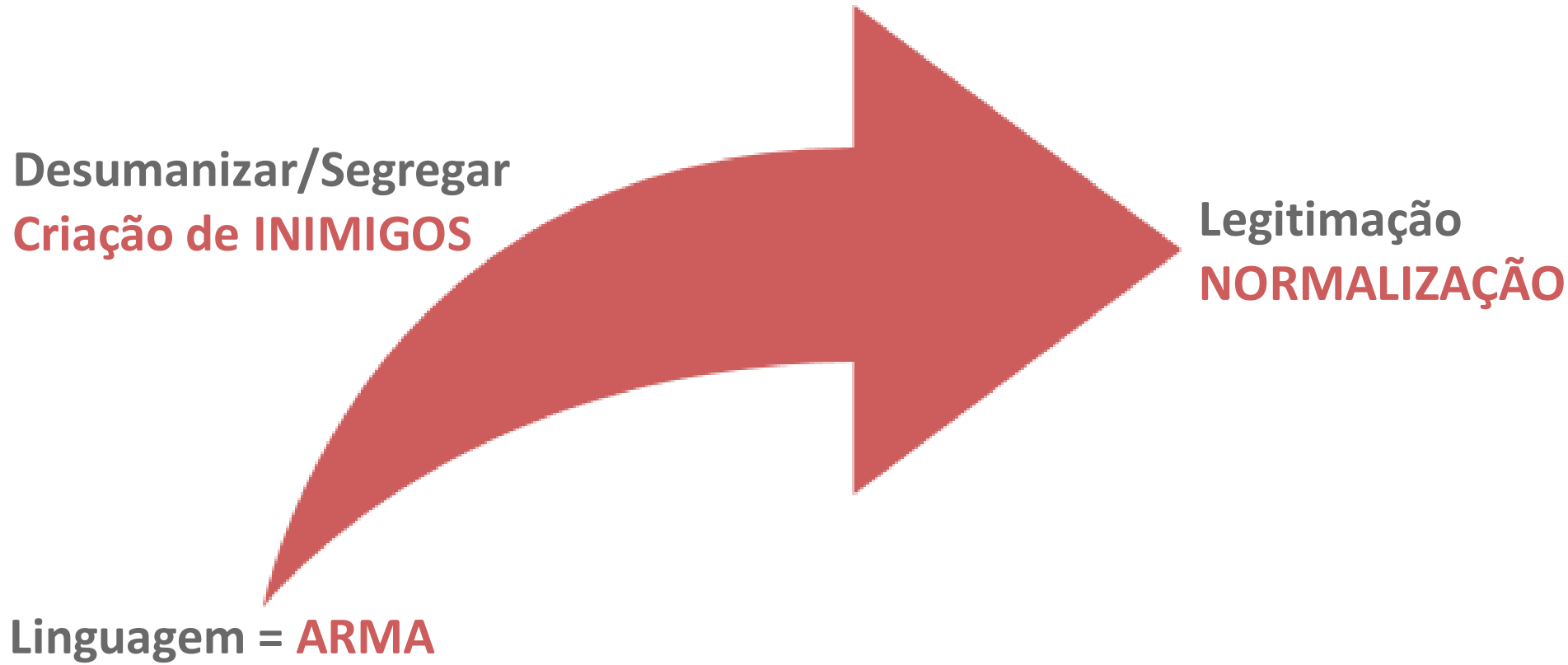
Racial + LGBTfobia

↑+250%

De 2021 a 2022

↑20% nas ocorrências de crimes

Consequências



Consequências

- Forte correlação
  - Abuso (físico e verbal) **X** sintomas de stress pós-traumático
  - Discriminação racial on-line **X** ansiedade/depressão

- LGBTQIA+/Homofobia on-line
  - Mais propensos a sofrer crimes de ódio
  - Raiva, ansiedade, depressão, stress, vergonha, culpa
  - Isolamento social e pensamentos suicidas

Chakravarthi et al 2022

Stefanita e Buf 2021

Consequências

- Curto e longo prazo
  - Ansiedade, pânico, vergonha e medo
  - Stress, depressão e alcoolismo

- Preocupação com possível materialização do ódio

Chakravarthi et al 2022

Stefanita e Buf 2021

Consequências

CORREIO BRAZILIENSE

RIO DE JANEIRO

Ex-jogadora de vôlei agride entregador negro com coleira de cachorro no RJ

EL PAÍS

Internacional

Morrer por ser gay: o mapa-múndi da homofobia

Onze países ainda punem com a morte as relações homossexuais. Um em cada três países condena a homossexualidade. Mais de 50% dos LGBT dizem ter sofrido alguma violência desde as eleições no Brasil

EXTRA

FOTOS Extra Digital Promoções

NOTÍCIAS ECONOMIA E FINANÇAS EMPREGO POLÍCIA FAMOSOS

Notícias Brasil Eleições 2022

02/10/22 06:18 02/10/22 06:19 Tweetar

Casos de homicídio por motivação política marcaram reta final da eleição; relembre

Discurso de ódio

Como lidar?

Quais caminhos possíveis?

PL das Fake News

Projeto de Lei Nº 2.630 (2020)

"...previsão de adoção de procedimentos de moderação, assegurando aos usuários o direito de reparação por dano individualizado ou difuso aos direitos fundamentais e o direito de recorrer da indisponibilização de conteúdos e contas."

Regulação das redes

- Decisões unilaterais
  - Sem transparência

- Falta de ação e responsabilidade das bigtechs
  - Afeta usuários e criadores de conteúdo
  - Propagação de ideias reprováveis

- Monetização, investimentos e ação dos algoritmos
  - Auxiliam disseminação

Outros caminhos!

- Disponibilizar ferramentas aos usuários
- Promover autonomia

Objetivos

- Uso de modelos de NLP para detecção de discurso de ódio

- Características e achados de discursos em diferentes grupos:
  - Sexualidade e gênero
  - Raça/etnia
  - Orientação política

Objetivos

- Proposta de ferramenta acessível aos usuários
- Fins de pesquisa e orientação

Metodologia – Extração de Dados

- Twitter
  - Caracteres limitados,
  - Facilidade inicial de acesso (API aberta)
  - Ferramentas de scraping

- Portal de Dados Abertos do TSE
  - Arquivos abertos com informações de deputados

Metodologia – Seleção do Modelo

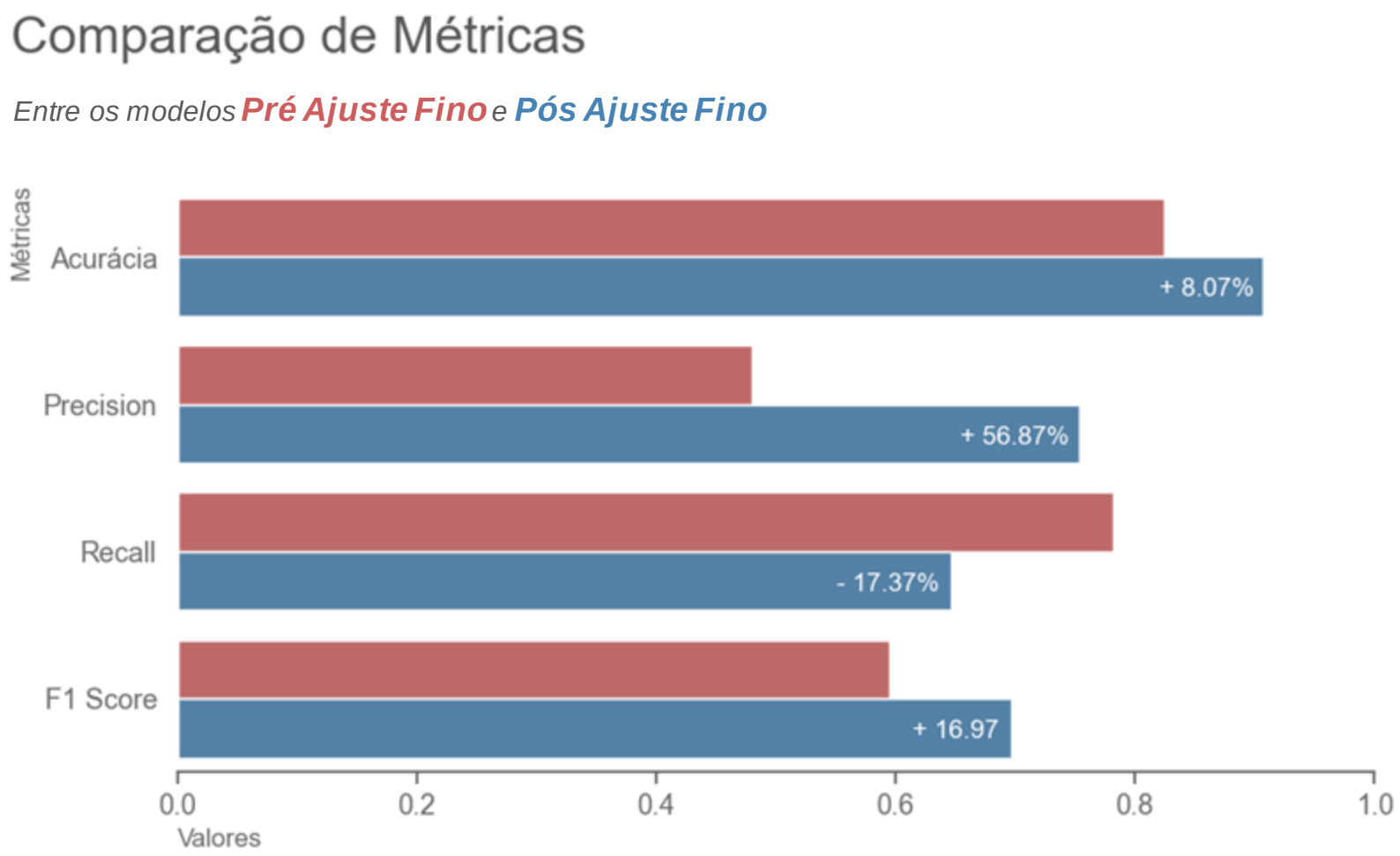
-  Huggingface – Classificador binário
  - Sentiment analysis + Hatespeech + Língua Portuguesa
  - "ruanchaves/bert-base-portuguese-cased-hatebr"

- Word2Vec + PCA
  - Contextualização do discurso de ódio

Metodologia – Avaliação de Performance

Metodologia – Ajuste Fino

Resultados – Ajuste fino



Resultados – Similaridade entre termos

- Candidata selecionada: Erika Hilton
  - Mulher, trans, negra

- Conteúdo ofensivo estava direcionado aos candidatos à presidência e não à candidata

Resultados – Similaridade

"bolsonaro"

	Termos	Score
0	mentiu	0.9906
1	sonegadores	0.9633
2	mente	0.9527
3	censurou	0.9490
4	janones	0.9285
5	novamente	0.9252
6		0.8265
7	criminoso	0.7908
8	pistoleira	0.7818
9	odeia	0.7736

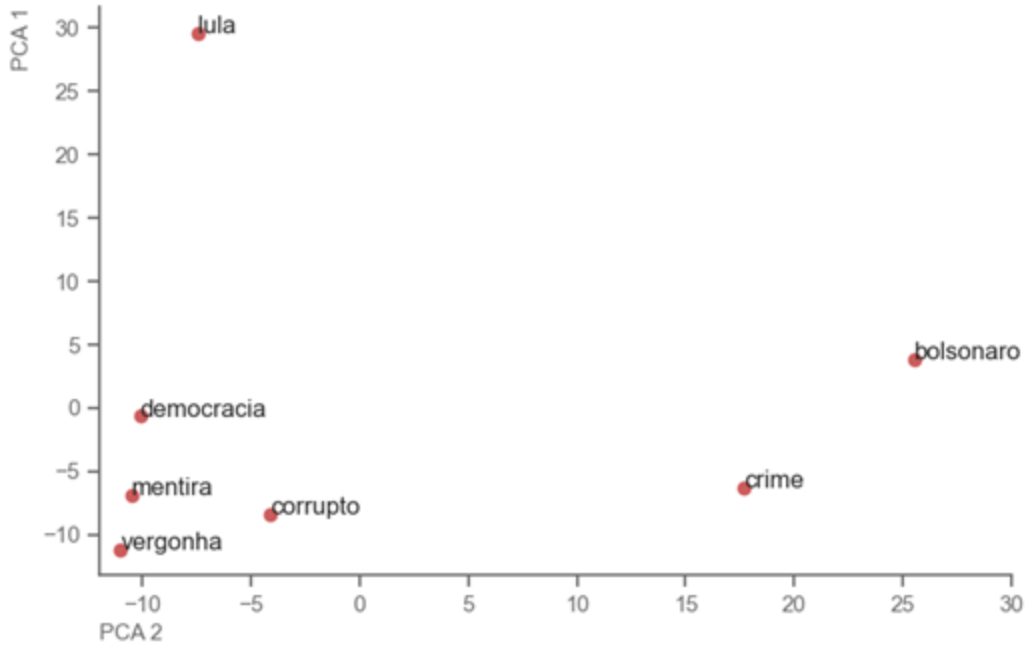
"lula"

	Termos	Score
0	venceu	0.9821
1	macetando	0.9800
2	criminoso	0.9789
3	mentiroso	0.9768
4	debate	0.9759
5	nordeste	0.9725
6	@andrejanonesadv	0.9724
7	favelado	0.9713
8	perdeu	0.9709
9	live	0.9708

Resultados – Word2Vec + PCA

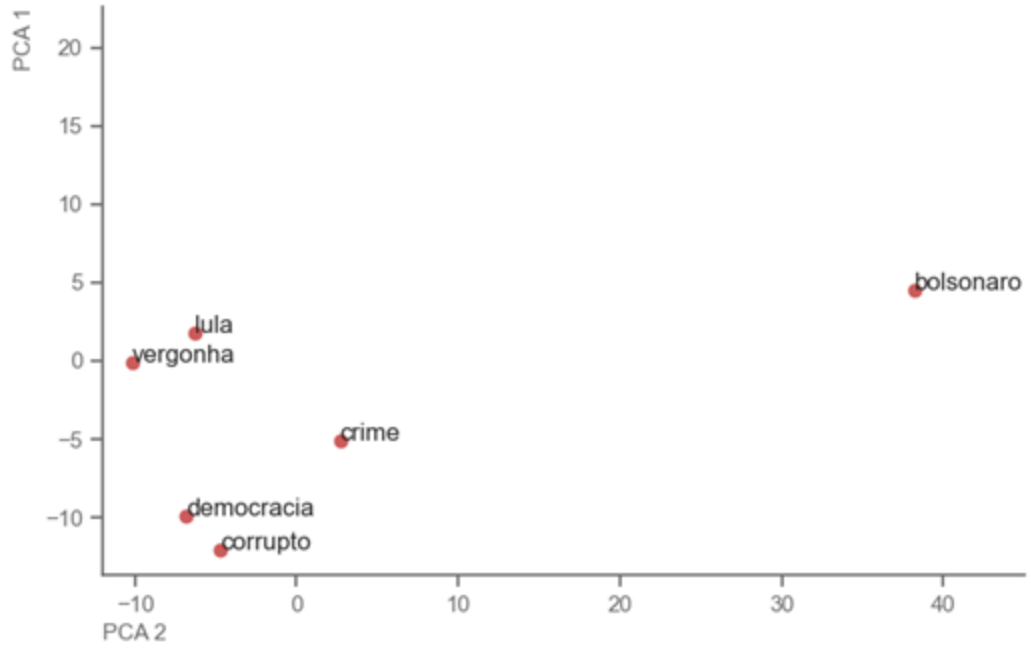
Contexto dos Tweets: Erika Hilton

Todos os Tweets e relação entre palavras selecionadas



Contexto dos Tweets Ofensivos: Erika Hilton

Tweets ofensivos com score acima de 0.75 e relação entre palavras selecionadas



Resultados

Até 0.11%

10mil visualizações

Facebook 3° trimestre de 2020

Até 14%

Twitter 4° Trimestre 2022

Limitações e Próximos Passos

“Fui atacada várias vezes e li discursos carregados de preconceito, afirmando eu jamais seria mulher, que não poderia ser tratada pelo pronome feminino, que nunca poderia gerar um filho. Também sofri ameaças de agressão e até de morte. Essas mensagens são aterrorizantes, tentam nos ofender, nos reduzir”

Madu Duarte (26 anos, artista e travesti)
Vítima de comentários de ódio nas redes sociais

https://www.diariodepernambuco.com.br/noticia/vidaurbana/2022/02/discursos-de-odio-na-internet-aumentam-60-em-um-ano.html

Limitações e Próximos Passos

- Como detectar minúcias dos discursos
  - Ironias/sarcasmos
  - Subtextos
  - Contextos culturais
- Diversidade
  - Construção de modelos pré-treinados

Referências

- Agência Brasi, Elaine Patrícia Cruz. Disponível em <https://agenciabrasil.ebc.com.br/direitos-humanos/noticia/2023-02/denuncias-de-crimes-na-internet-com-discurso-de-odio-crescem-em-2022> Acesso em: 25/07/2023
- Jornal Nacional. Disponível em <https://g1.globo.com/jornal-nacional/noticia/2023/05/01/denuncias-de-crimes-envolvendo-discurso-de-odio-nas-redes-sociais-triplicaram-nos-ultimos-6-anos-aponta-levantamento.ghtml> Acesso em: 23/07/2023
- Thais Cardoso. Disponível em <https://www.ecycle.com.br/especialistas-analisam-discurso-de-odio-e-as-consequencias-dessa-pratica/> Acesso em 23/07/2023
- Diário de pernambuco. Disponível em <https://www.diariodepernambuco.com.br/noticia/vidaurbana/2022/02/discursos-de-odio-na-internet-aumentam-60-em-um-ano.html> Acesso em: 23/07/2023
- DW, Edison Veiga. Disponível em: <https://www.dw.com/pt-br/o-que-faz-o-brasil-ser-l%C3%ADder-em-viol%C3%AAncia-contra-pessoas-trans/a-58122500> Acesso em: 21/07/2023
- Marcela Penna. Disponível em: <https://canaldedenuncias.com.br/homofobia-e-discurso-de-odio-online-entenda-a-importancia-do-monitoramento/> Acesso em: 21/07/2023
- Jornal da USP, Ricardo Alexino Ferreira. Disponível em: <https://jornal.usp.br/atualidades/os-discursos-homofobicos-e-a-violencia-contra-segmentos-da-diversidade/> Acesso em: 22/07/2023
- El País, Ana Alfageme. Disponível em: <https://brasil.elpais.com/brasil/2019/03/19/internacional/1553026147_774690.html> Acesso em:25/07/2023
- Mariana Gonzalez. Disponível em: <https://www.uol.com.br/universa/noticias/redacao/2021/07/15/lgbtfobia-cresce-20-no-brasil-numero-ainda-e-subnotificado-diz-advogada.html> Acesso em: 23/07/2023
- Malu Pinheiro. Disponível em: <https://glamour.globo.com/lifestyle/noticia/2021/08/como-o-discurso-de-odio-nas-redes-sociais-afeta-saude-mental-de-criancas-e-adolescentes.ghtml> Acesso em: 20/07/2023

Referências

- Correio Braziliense, Cecília Sóter. Disponível em: <https://www.correiobraziliense.com.br/brasil/2023/04/5086498-ex-jogadora-de-volei-agride-entregador-negro-com-coleira-de-cachorro-no-rj.html> Acesso em: 25/07/2023
- UOL. Disponível em: <https://noticias.uol.com.br/politica/ultimas-noticias/2023/02/15/carta-george-washington-bolsonaro.htm> Acesso em: 25/07/2023
- G1. Diponível em: <https://g1.globo.com/ce/ceara/noticia/2022/09/26/homem-e-assassinado-no-ceara-apos-bate-boca-policia-investiga-motivacao-politica.ghtml> Acesso em: 24/07/2023
- Instituto Ethos. Disponível em:< https://www.ethos.org.br/cedoc/discurso-de-odio-e-um-dos-sinais-de-alerta-de-genocidio-e-de-outros-crimes-alerta-guterres/> Acesso em: 23/07/2023
- Casseli, T et al. HateBERT: Retraining BERT for Abusive Language Detection in English. Fifth Workshop on Online Abuse and Harms, pages 17–25 August 6, 2021.
- Stefanita, O; Buf, DM Hate Speech in Social Media and Its Effects on the LGBT Community: A Review of the Current Research. Romanian Journal of Communication and Public Relations vol. 23, no. 1 (52) / April 2021, 47-55
- Chakravarthi et al. How can we detect Homophobia and Transphobia? Experiments in a multilingual code-mixed setting for social media governance. International Journal of Information Management Data Insights Volume 2, Issue 2, November 2022
- Janah, S; Oussalah, M. A systematic review of hate speech automatic detection using natural language processing. Neurocomputing Volume 546, 14 August 2023
- Sarfraz et al. Influence of Internet Language Hate Speech on Young Adults Mental Health and its Detection Method. Pakistan Journal of Medical & Health Sciences Vol. 16 No. 07 (2022)
- Tavares, RCL; Sousa, RSN. Hate Speech in Social Media and DISCURSOS SOBRE A CRIMINALIZAÇÃO DA HOMOFOBIA E DA TRANSFOBIA NO PORTAL DE NOTÍCIAS O ANTAGONISTA. Trabalhos em Linguística Aplicada 61 (2),May-Aug 2022