

# DETECÇÃO DE DISCURSOS DE ÓDIO

---

Abordagens e Desafios

Profa. Dra Brenda Salenave

Universidade Federal de Pelotas



01

---

## DEFINIÇÃO

O que é, o que não é e  
quais os impactos dos  
discursos de ódio

02

---

## ABORDAGENS

Como detectar esse  
tipo de discurso e o  
que vem sendo feito  
para isso

03

---

## DESAFIOS

O que nos impede  
de ir além



01

# DEFINIÇÃO

O que é, o que não é e quais os impactos dos discursos de ódio.



# Discursos simbolicamente prejudiciais

- Termos diferentes, porém semelhantes, podem ser enquadrados como discursos simbolicamente prejudiciais
  - discurso perigoso
  - discurso tóxico
  - discurso de ódio
  - discurso intolerante
  - discurso extremista



# Discursos simbolicamente prejudiciais

- Certos discursos possuem o potencial de causar danos significativos, inclusive críticos, e podem ser considerados tóxicos
  - podendo ser um discurso momentâneo ou persistente.
  - afetar indivíduos ou a sociedade como um todo.
  - causando danos temporários ou permanentes.



# Linguagem Ofensiva

- Tem a intenção de magoar, insultar ou provocar os sentimentos das pessoas, sem necessariamente ter um objetivo discriminatório.



“Linguagem, ou expressão indicativa de um tipo de conduta linguística, deliberadamente **intencional**, que **ataca ou diminui através da segregação ou da explicitação de preconceitos negativos** sobre pessoas ou grupos com base em **características específicas**, como aparência física, religião, descendência, origem nacional ou étnica, orientação sexual, identidade de gênero ou outra, reduzindo o seu valor e dignidade perante a sociedade, ameaçando e promovendo a sua insegurança e, nos casos mais extremos, incentivando à violência e ao extermínio. Uma linguagem que pode ocorrer com diferentes estilos linguísticos, **mesmo em formas sutis ou quando o humor é usado.**”

—Santana (2023)



# Discurso de Ódio

Discursos de ódio são manifestações que avaliam negativamente um grupo vulnerável ou um indivíduo enquanto membro de um grupo vulnerável, a fim de estabelecerem que ele é menos digno de direitos, oportunidades ou recursos do que outros grupos ou indivíduos membros de outros grupos, e, conseqüentemente, **legitimar a prática de discriminação ou violência**. Aquele que profere o discurso de ódio é aqui denominado o orador, aqueles a quem o discurso se dirige são a audiência e aqueles que são negativamente avaliados pelo discurso de ódio são o alvo. O grupo vulnerável é aquele que está mais propenso a sofrer violência ou discriminação em comparação com outros grupos sociais - *Fundação Getúlio Vargas (FGV)*; *Confederação Israelita do Brasil (CONIB)*.

Relatório de recomendações para o DISCURSO DE ÓDIO E AO  
EXTREMISMO NO BRASIL



# Discurso de Ódio

Qualquer tipo de comunicação falada ou escrita ou comportamento que ataque ou use **linguagem pejorativa ou discriminatória** com referência a uma pessoa ou grupo com base em quem eles são, em outras palavras, com base em sua religião, etnia, nacionalidade, raça, cor, descendência, gênero ou outro fator de identidade” - Organização das Nações Unidas (ONU).

Relatório de recomendações para o DISCURSO DE ÓDIO E AO EXTREMISMO NO BRASIL



# Exemplos

Exemplo 1: Um discurso político que descreve um grupo étnico como "inferior" em relação a outros grupos, com o objetivo de justificar políticas discriminatórias contra esse grupo.

Exemplo 2: Um comentário feito por uma pessoa em uma rede social usando linguagem racista para atacar um indivíduo com base em sua raça, sem a intenção explícita de incitar discriminação ou violência contra esse indivíduo.



# Impactos

- O impacto de toxinas discursivas é de natureza social
  - palavras ofensivas, insultos, discriminação, discurso de ódio, difamação, ameaças ou qualquer forma de linguagem que busque macular, menosprezar ou ferir a dignidade e a integridade de indivíduos pertencentes ao grupo alvo.
  - comentários tóxicos são a principal forma de ódio e assédio online.





**[Grupo religioso] são vírus. Eles estão deixando este país doente.**



Credooooo! A Miss Piaui tem cara de empregadinha, cara comum, não tem perfil de miss, não era pra ta ai. Sorry.  
[#MissBrasil](#) [#MissBrasil2017](#)

20/08/17 00:25



**Tem muito verme  
[nacionalidade/raça/etnia] no nosso  
país e eles têm que ir embora!**





# Características

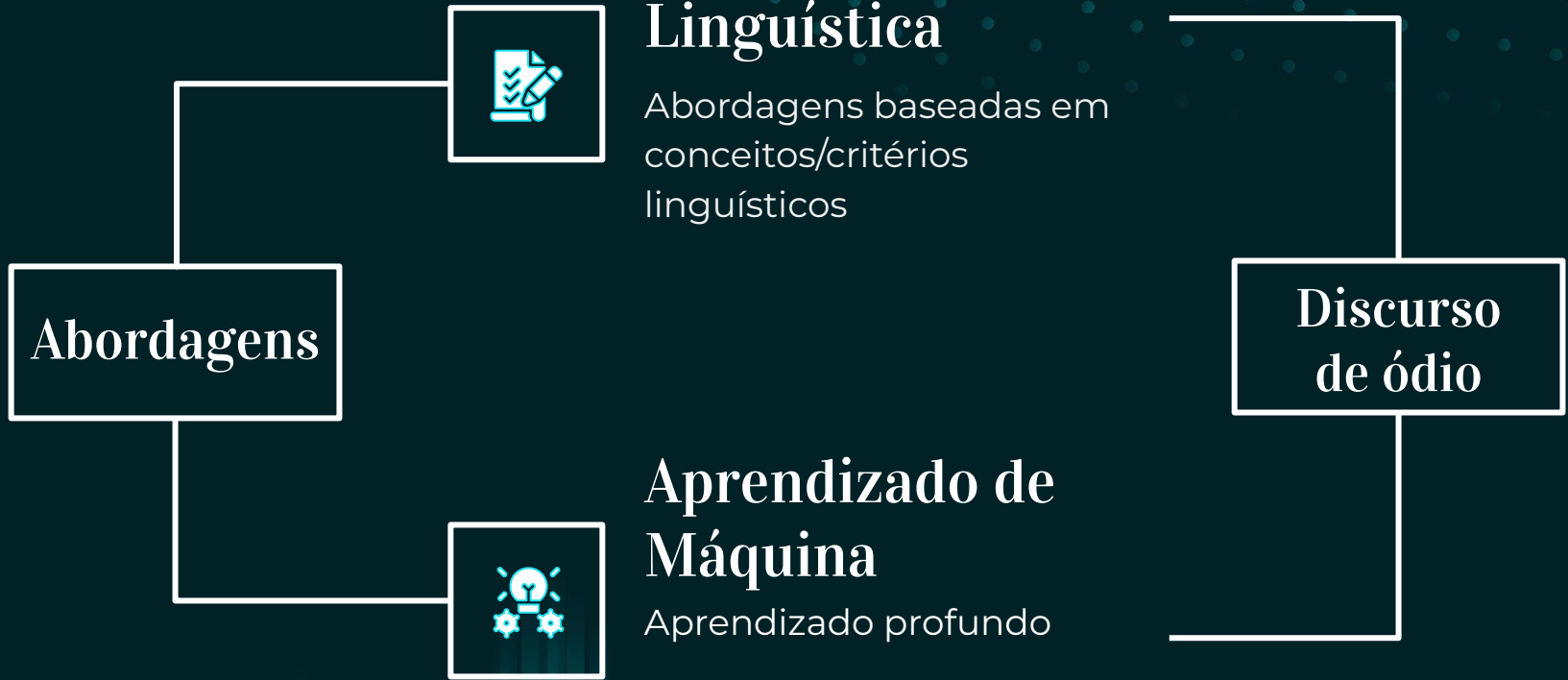
1. **Discurso de Sanção:** discursos que pretendem punir sujeitos considerados maus cumpridores de determinados contratos sociais.
2. **Ódio Passional e Aversão ao Diferente:** prevalecem as paixões do ódio e do medo em relação ao que é considerado diferente.
3. **Temas e Figuras de Oposição:** Discursos que desenvolvem temas e figuras a partir da oposição entre igualdade ou identidade e diferença.
4. **Falácia com intenção de propagar ódio:** situações onde se faz uso de argumentos que fogem ao contexto, passando a atacar a pessoa, e não as ideias apontadas por ela, a quem o discurso se dirige, até distorcendo o argumento utilizado por esta.



02

## ABORDAGENS

Como detectar esse tipo de discurso e o que vem sendo feito para isso



**Abordagens**



**Linguística**

Abordagens baseadas em  
conceitos/critérios  
linguísticos



**Aprendizado de  
Máquina**

Aprendizado profundo

**Discurso  
de ódio**

---

# Abordagens baseadas em conceitos/critérios linguísticos

- Léxicos
- Recursos lexicais e linguísticos que têm como objetivo representar e estruturar o conhecimento semântico e lexical da língua
  - VerbNet
  - WordNet
  - FrameNet



---

# Abordagens baseadas em aprendizado de máquina (aprendizado profundo)

- Redes Neurais Convolucionais (CNNs)
  - detecção de padrões em dados de texto
- Redes Neurais Recorrentes (RNNs) e Long Short-Term Memory (LSTM)
  - dependências sequenciais em textos - entender o contexto em que certas palavras ou frases aparecem.
- Transformers
  - relações contextuais complexas em textos.



03

# DESAFIOS

O que nos impede de ir além

# Subjetividade e Contexto

- Discursos intolerantes têm nuances subjetivas e dependem do contexto cultural e social em que as palavras são usadas, levando em conta nuances culturais e sociais.

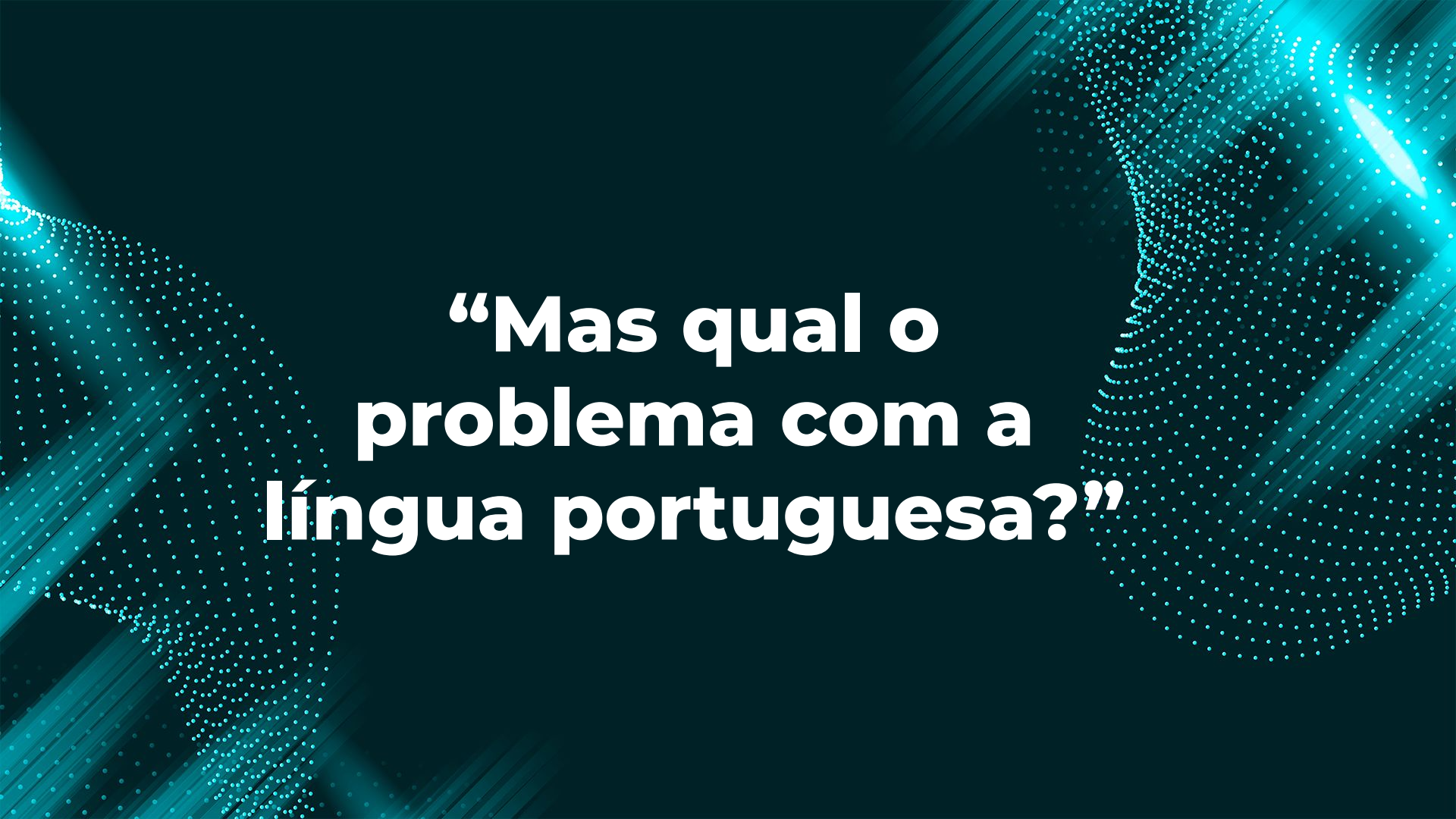


# Dinamismo da linguagem

- A língua é um sistema vivo e mutável que evolui apesar da nossa vontade de regrá-lo, que não obedece a normas impostas e a todo momento é moldado pelos falantes da língua
  - Pode evoluir ao longo do tempo, incorporando novos termos, gírias ou formas de expressão.
- Reflete quem somos, de onde viemos, o que somos, com quem nos relacionamos, os lugares onde circulamos, nossa classe social, nosso gênero, entre tantas outras coisas que nos representam

- A. Vanin (2024)





**“Mas qual o  
problema com a  
língua portuguesa?”**

# Dados

- Hate Speech Data
  - <https://hatespeechdata.com>
- International Network for Hate Studies
  - <https://internationalhatestudies.com>



# Equilíbrio entre Sensibilidade e Especificidade

- Exige um equilíbrio cuidadoso entre sensibilidade (a capacidade de detectar discurso de ódio real) e especificidade (evitar falsos positivos).
- Crucial para garantir que o sistema não rotule erroneamente expressões legítimas.



---

# Ética e Liberdade de Expressão

- Há desafios éticos significativos associados à detecção de discurso de ódio, especialmente em relação à liberdade de expressão.
- Algoritmos precisam ser projetados para evitar a supressão indevida da expressão legítima.





# Desafios

---

- **Treinamento e Dados Anotados**
  - A disponibilidade de dados rotulados para treinamento é um desafio.
  - Obtenção de conjuntos de dados representativos e equilibrados é crucial para garantir que o modelo seja eficaz em várias situações.



---

## Desafios

- Evolução rápida do discurso online
- A adaptação dos agressores a técnicas de detecção
- Interpretação de nuances culturais e contextuais na linguagem



# Fronteiras de Conhecimento

- **Aprendizado Profundo**
  - Abordagens mais avançadas baseadas em modelos de aprendizado profundo para melhorar o desempenho na detecção de nuances no discurso.
  - A atenção é dada a arquiteturas de modelos mais eficientes e escaláveis.
- **PLN Multimodal**
  - Integração de informações de diferentes modalidades, para melhorar a detecção, especialmente em redes sociais.

# Fronteiras de Conhecimento

- **Avaliação de Viés e Equidade**
  - Ênfase na compreensão e mitigação de vieses em modelos
- **Transferência de Aprendizado**
  - Técnicas de transferência de aprendizado para lidar com a falta de dados rotulados em idiomas ou domínios específicos.



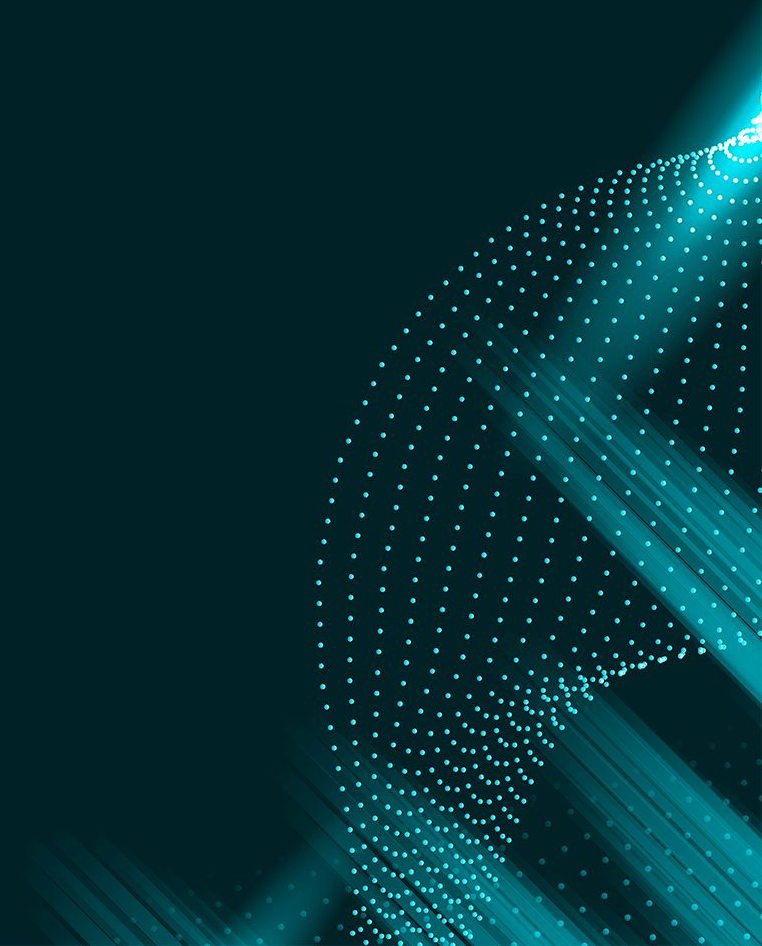



# Fronteiras de Conhecimento

- **Geração de Linguagem**
  - Desenvolvimento de modelos para não apenas detectar, mas também para gerar respostas ou contranarrativas construtivas para combater o discurso de ódio.
- **Adaptação Contínua**
  - Adaptação contínua dos modelos à evolução do discurso online, considerando que novos termos e padrões emergem com o tempo.

# Fronteiras de Conhecimento

- Identificação do alvo do discurso





# **Discurso de ódio não é liberdade de expressão**

---

discurso de ódio é considerado uma forma de  
violência verbal que pode incitar o ódio,  
prejudicar comunidades e violar os direitos  
humanos.

# Obrigada!

---

Brenda Salenave

[bssalenave@inf.ufpel.edu.br](mailto:bssalenave@inf.ufpel.edu.br)

