Brazil#WithoutHate

Self-Explaining and Moral-Aware AI for Hate Speech Detection

Francielle Vargas

HARVARD UNIVERSITY

APPLIED SOCIAL MEDIA LAB

BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

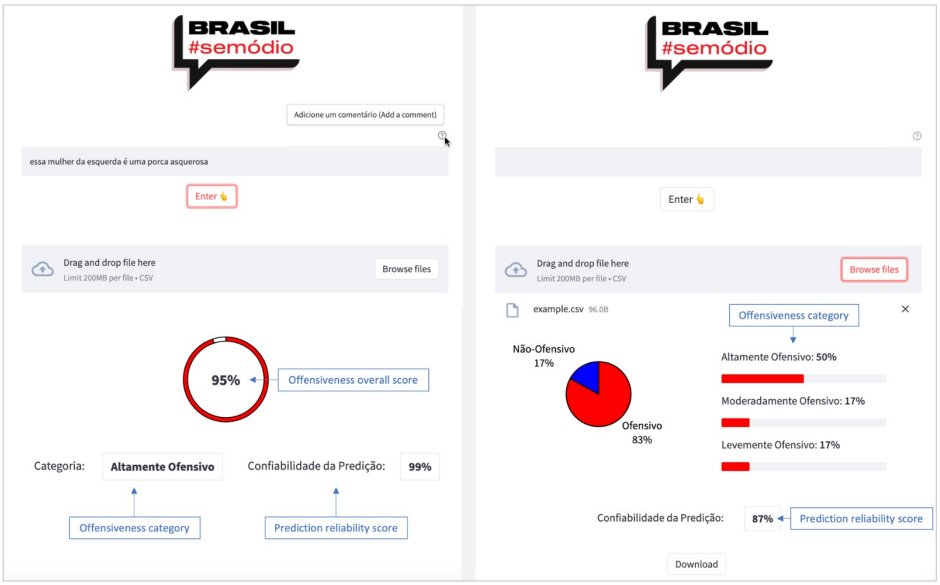


Figure 1. **BrazilWithoutHate**: A Transparent and Morally-Aware AI System for Detecting Hate Speech in Brazilian Portuguese.

Hate Speech Detection with Moral Rationales

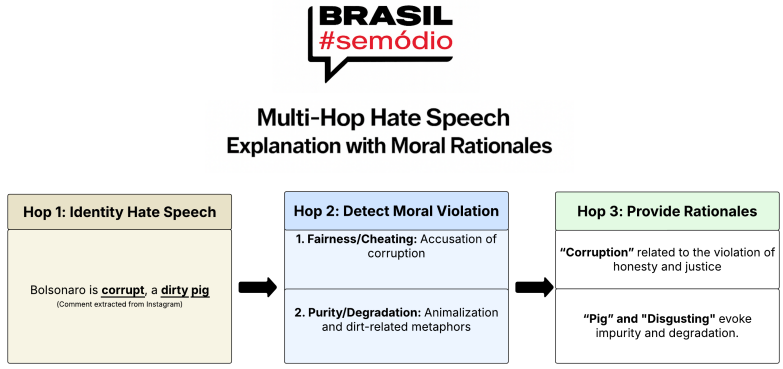


Figure 2. **Brazil#WithoutHate**: Mult-hop Hate Speech Explanation with Moral Rationales.

HateBR Dataset

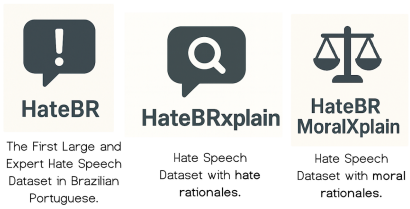


Figure 3. HateBR dataset and its versions.

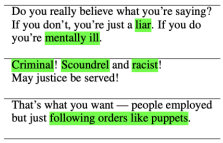
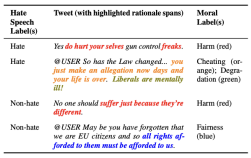


Figure 4. Annotation of hate speech rationales.



Hate Speech Label(s)	Tweet (with highlighted rationale spans)	Moral Label(s)
Hate	*Yes do hurt your selves gun control freaks.*	Harm (red)
Hate	*@USER So has the Law changed... you just make an allegation now days and your life is over. Liberals are mentally ill!*	Cheating (orange); Degradation (green)
Non-hate	*No one should suffer just because they're different.*	Harm (red)
Non-hate	*@USER May be you have forgotten that we are EU citizens and so all rights afforded to them must be afforded to us.*	Fairness (blue)

Figure 5. Annotation of moral rationales.

Self-Explaining AI

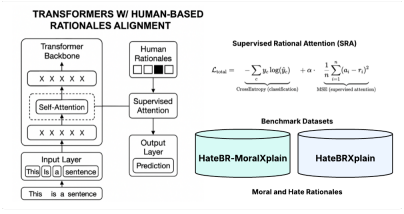


Figure 6. Supervised Rational Attention for Self-Explaining Hate Speech Detection.

SCAN ME

← Check out the **Brazil#WithoutHate** platform!

Google