Stop raging against the machines: Building a citizen linguist lab that detects hate speech

By Christopher Tuckwood, Drew Boyd, and Raashi Saxena
It’s difficult to **understand** hate speech without talking about hate speech. And it’s difficult to talk about hate speech **without** looking at hate speech.

Although actual hate speech and profanity will be minimized in the talk, expect to see or hear some hateful language.
Hatebase was built to assist companies, government agencies, NGOs and research organizations in moderating online conversations and potentially use hate speech lexicons as an early warning predictor for regional violence.

https://miro.com/app/board/o9J_IRSQqTg=/
Any expression regardless of offensiveness**, which broadly categorises a specific group of people based on malignant, qualitative and or/subjective attributes particularly if those attributes pertain to

- Ethnicity
- Nationality
- Religion
- Class
- Disability
- Sexuality

**Excluding offensiveness from the definition of hate speech allows for a non/unbiased less opinionated perspective of hate speech
Hatebase does not support censorship or the criminalization of speech (with a few caveats)

Online communities have a right / legal responsibility to moderate user activity to ensure fair and respectful treatment of all users.

While hate speech as an expression of opinion is (and should be) protected, hate speech which carries the threat of violence isn’t (and shouldn’t be).

Allowing discriminatory content to proliferate silences marginalized voices, which is itself a form of censorship.
TRANSPARENCY
Awareness, monitoring, content moderation which is both transparent and proactive

EDUCATION
Education and counter-messaging in high friction environments

POLICY
Create informed govt and non-govt policies, leading to more effective triage of resources to impacted populations

RESEARCH
Independent research and analysis
98 languages
177 countries
350 universities
Regionalised Structural multilingual hate speech

- Languages: 98
- Terms: 3,786
- Countries: 177
LEVERAGING HATEBASE FOR DATA DRIVEN DECISION MAKING
The Third point of the manifesto was the labeling of Amharas as The Enemy of Tigray.

Translation: The main enemy of the Tigrayan people are the Amhara
HOW DOES IT WORK?
Questions posted on Miro Board, please cite with examples

01. Insults based on a specific group or identity

02. Holocaust denial?

03. Criticism of a specific country or group?

04. If you were to write a program which monitors hate speech on Twitter, would finding it in a random tweet be potentially meaningful? If not, it's probably not hate speech.
Questions posted on Miro Board, please cite with examples

05  Threats against a specific population?

06  Generalisations of the attitudes, motives, predilections of groups?

07  Intragroup / reappropriated language?

08  Threat against a specific population
How can YOU help?
Sign up for the Citizen Linguist Lab

1. Add multilingual vocabulary
2. Spend some time to add and edit hate speech vocabulary/terms
3. Providing offensiveness rating to existing vocabulary
Join the Citizen Linguist Lab

Hatebase's multilingual vocabulary is created and maintained by staff and volunteers with experiece in languages. Our Citizen Linguists serve as a critical function in the organization, providing valuable nuances of lexicographic usage in various linguistically complex areas of the globe.

You don't have to be a professional linguist to pitch in -- all you need is a willingness to commit some time editing vocabulary and other data through the Hatebase website. For frequent contributors, we'll make sure your name is listed as part of the advisory team on our About page.

Sign up now

Which languages are you fluent in?
Q&A