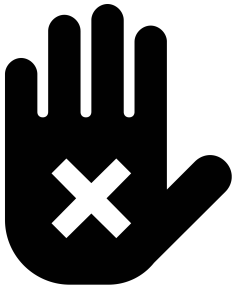
Stop raging against the machines: Building a citizen linguist lab that detects hate speech

By Christopher Tuckwood, Drew Boyd, and Raashi Saxena

DISCLAIMER



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

Introduction to SP & Share link here

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_lRSQqTg=/

HATEBASE

HATEBASE

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

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

HATEBASE

TRANSPARENCY

Awareness, monitoring, content moderation which is both transparent and proactive

POLICY

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

EDUCATION

Education and counter-messaging in high friction environments

RESEARCH

Independent research and analysis

HATEBASE

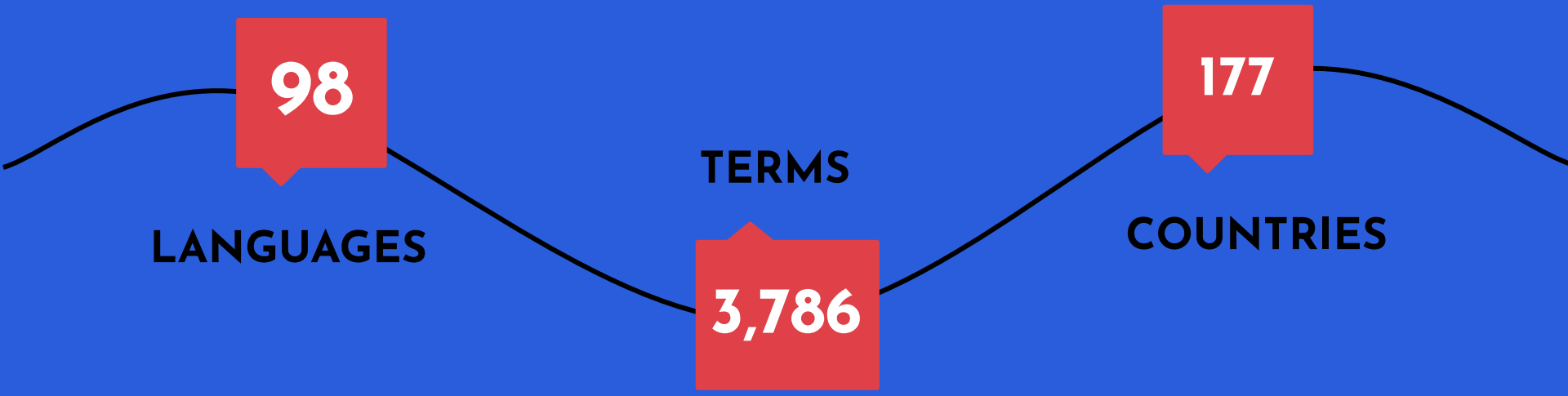
98 languages

177 countries

350 universities

HATEBASE

Regionalised Structural multilingual hate speech



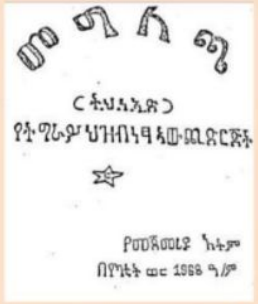
HATEBASE

LEVERAGING HATEBASE FOR DATA DRIVEN DECISION MAKING

HATEBASE

RECENT EXAMPLES OF HATE SPEECH

The Third point of the manifesto was the labeling of Amharas as The Enemy of Tigray.

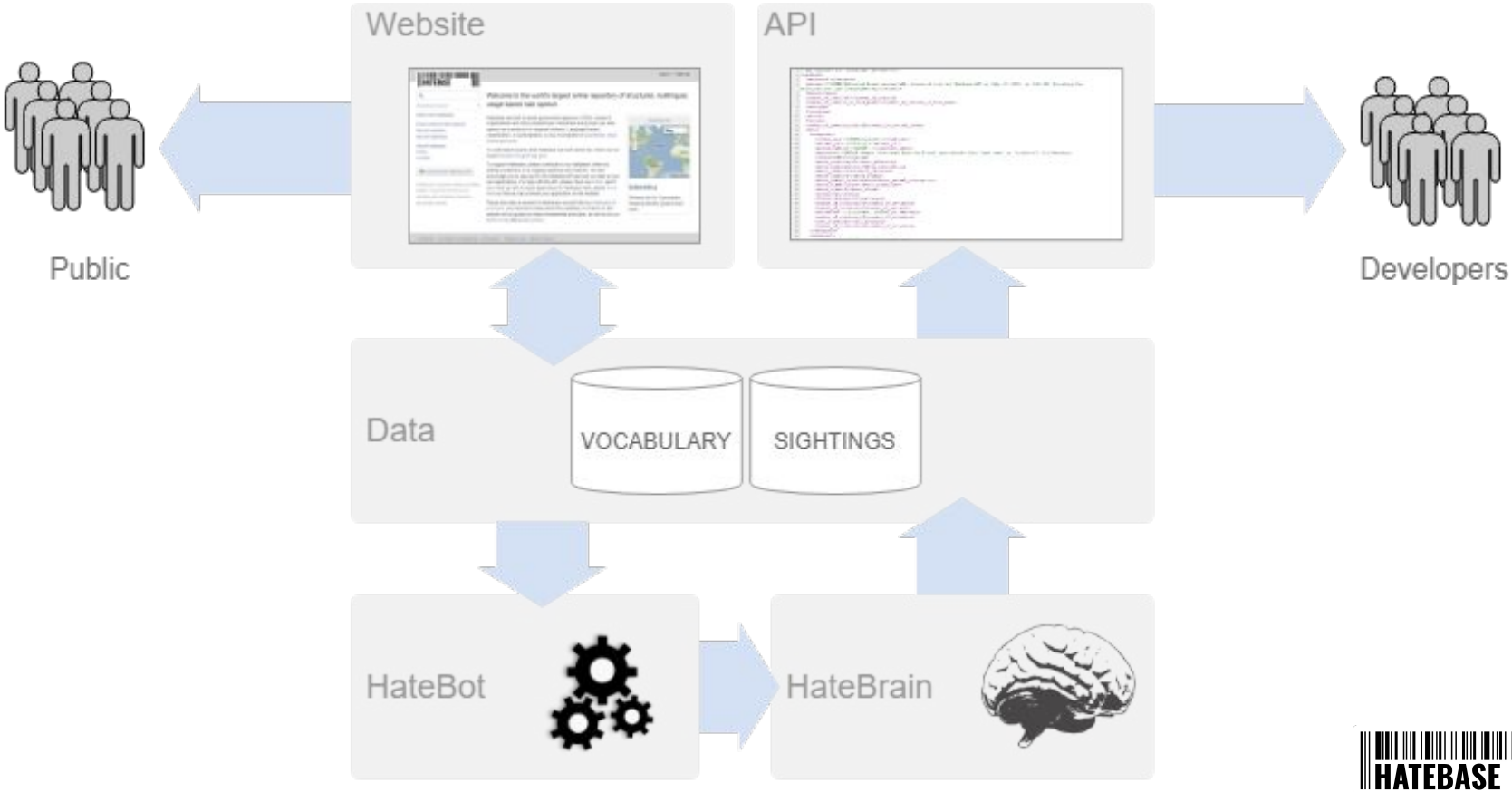


ሀ- ዲሳማጭና ሥራው።
የትግራይ ህዝብ ብሔራዊ ታሚል ፀረ-የአማራ ብሔራዊ ጭቆና፣ ፀረ-ኢምፐርያሊዝም እንደሁም ፀረ-ንኡስ ከበርቴያ ዊ. ጠንኛ ስውጥ ነው። ሰለዚህ የአብየታዊው ትግል ዓለማ ከበሳበታዊ ሥርዓትና ከኢምፐርያሊዝም ነፃ የሆነ የትግራይ ጨቆነሲኒያዊ ሪፓብሊክ መቋቋም ይሆናል።

Translation- The main enemy of the Tigrayan people are the Amhara

HATEBASE

HOW DOES IT WORK?



HATEBASE

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.

HATEBASE

Questions posted on Miro Board, please cite with examples

05 Threats against a specific population?

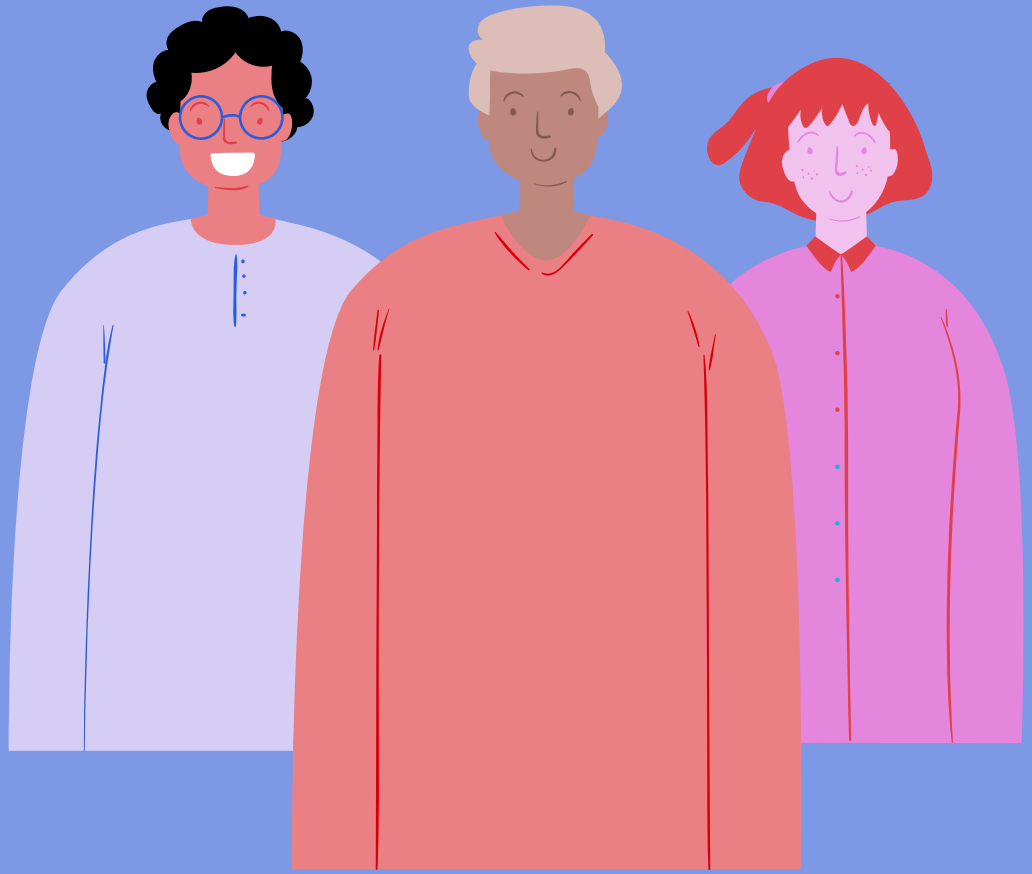
06 Generalisations of the attitudes, motives, predilections of groups?

07 Intragroup / reappropriated language?

08 Threat against a specific population

HATEBASE

How can YOU help?



HATEBASE

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

hatebase.org/citizen_linguist/Raashi

ABOUT WORK WITH US PRICING IN THE MEDIA

HATEBASE

Join the Citizen Linguist Lab

Hatebase's multilingual vocabulary is created and maintained by staff and volunteers with experie languages. Our Citizen Linguists serve as a critical function in the organization, providing valuable nuances of lexiconographic 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 som editing vocabulary and other data through the Hatebase website. For frequent contributors, we'll whether you're comfortable being listed as part of the advisory team on our About page.

Sign up now

Which languages are you fluent in?

Continue

HATEBASE

HATEBASE

All terms

All ethnicities

All religions

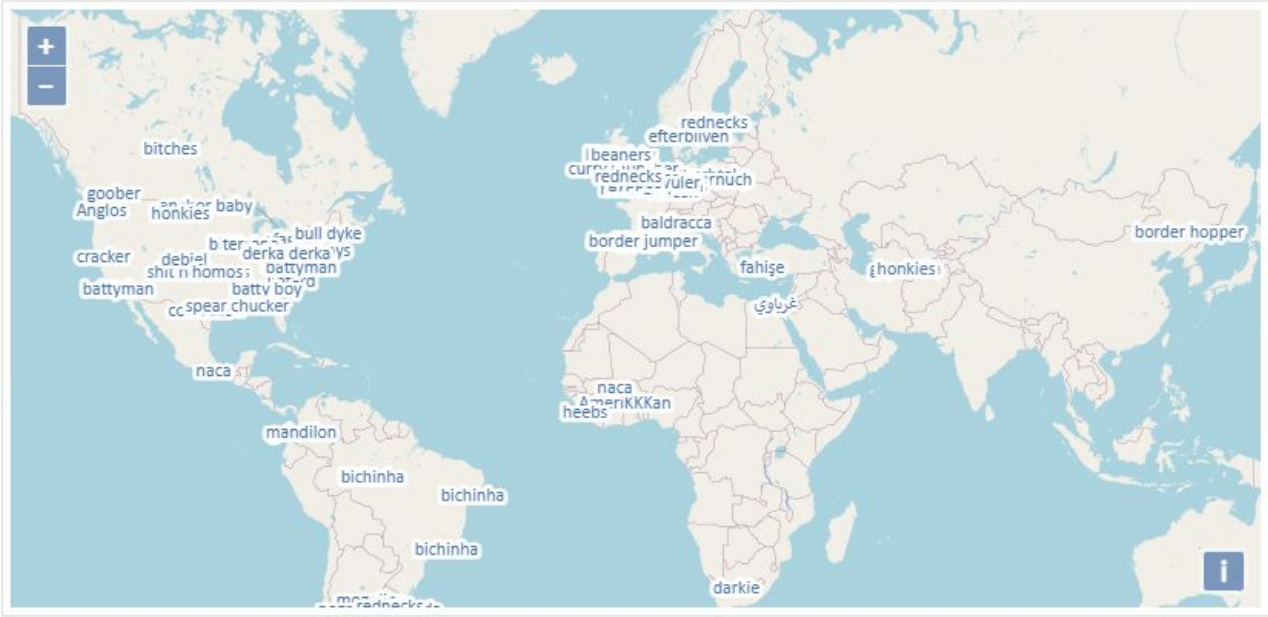
All languages

- [ ] Pertains to gender
- [ ] Pertains to sexual orientation
- [ ] Pertains to disability
- [ ] Pertains to class

Search Clear

Recent Sightings

MAP TABLE



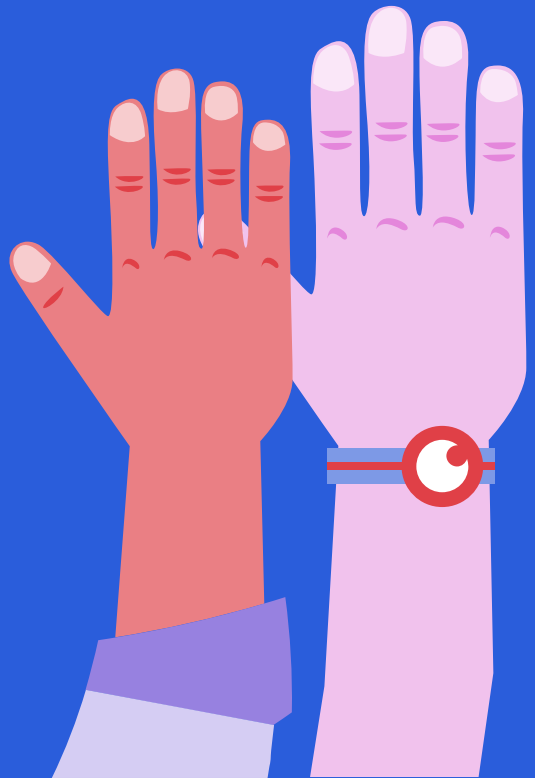
HATEBASE

PROBLEM

SOLUTION

HATEBASE

Q&A



01

02

HATEBASE