

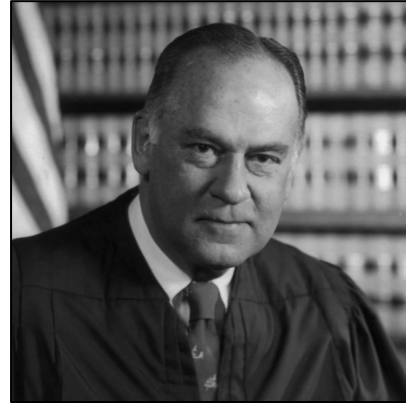
Monitoring hate speech

Challenges and strategies



May 2019 Geneva

What is hate speech?



“I know it when I see it.”

Justice Potter Stewart
Jacobellis v. Ohio, 1964



Is hate speech...

- Criticism of a specific country or group?
- Insults or jokes that are based on a specific group identify?
- Holocaust denial / revisionism?
- Intra-group / reappropriated language?
- Threats against a specific population?
- Generalizations about the attitudes, motives or predilections of a specific group?

All of these boundaries are difficult to quantify.



Is hate speech best defined as a potential predictor of conflict?



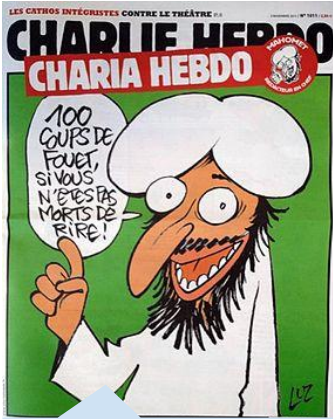
“Any advocacy of national, racial or religious hatred that **constitutes incitement** to discrimination, hostility or violence shall be prohibited by law.”



“[Hate speech is] broad discourse that is extremely negative and **constitutes a threat** to social peace.”

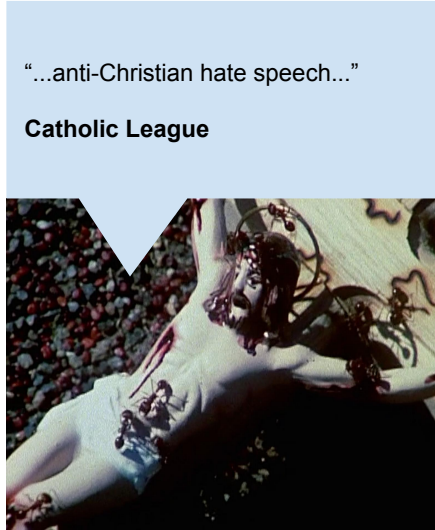


Is hate speech, to some extent, in the eye of the beholder?



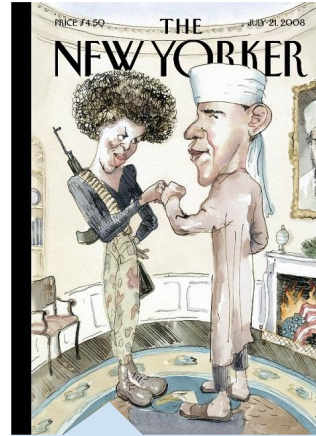
"We cannot credibly dismiss that many do find Charlie Hebdo racist and Islamophobic..."

Al Jazeera



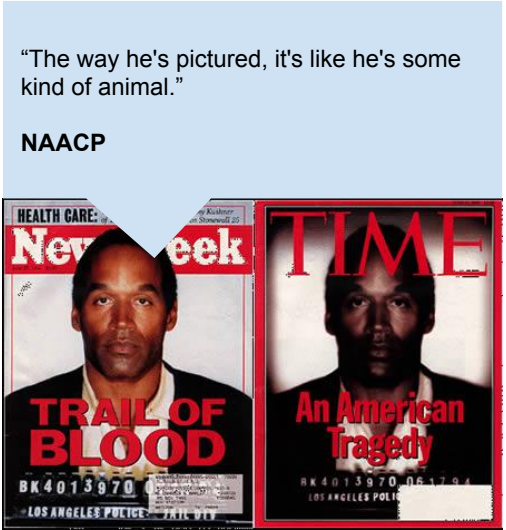
"...anti-Christian hate speech..."

Catholic League



"...reinforces a critical piece of misinformation... and make[s] it... difficult for a person of color to be elected president."

ShadowProof



"The way he's pictured, it's like he's some kind of animal."

NAACP



One reason hate speech is difficult to define is that it exists at the intersection of several related types of expression

- **Discriminatory language** which disparages people based on a shared identify
- **Generalization** which, regardless of intent, stereotypes people based on a shared identify
- **Dangerous speech** which incites or presages violence (hate crime)
- **Symbolic (non-verbal) expression**, e.g. swastikas, repurposed emojis and memes

These different aspects also pose a significant challenge for automation.



Hatebase defines hate speech as:

Any expression, regardless of offensiveness, which broadly characterizes a specific group of people based on **malignant**, **qualitative**, and/or **subjective** attributes -- particularly if those attributes pertain to:

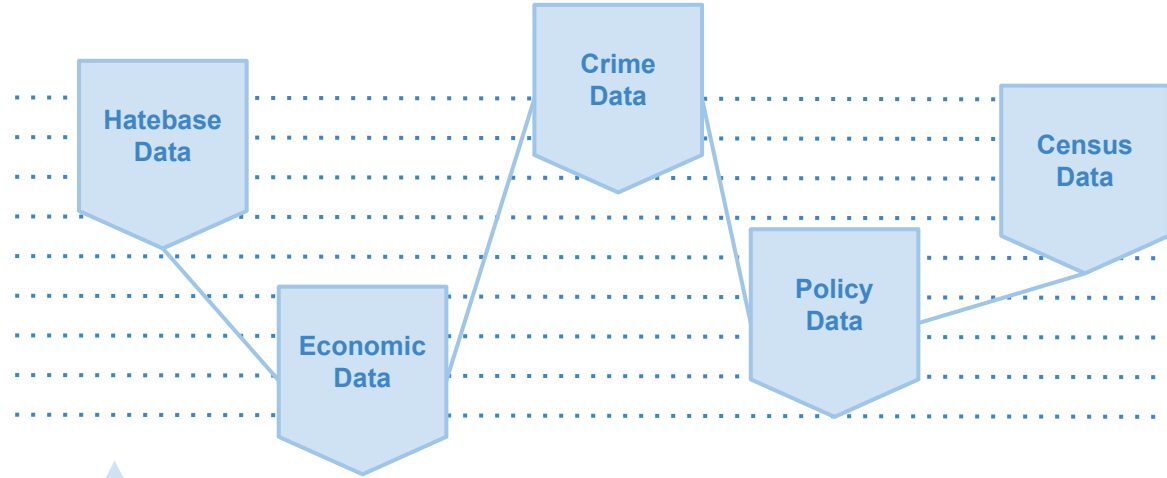
- **ethnicity**
- **nationality**
- **religion**
- **sexuality**
- **disability**
- **class**

Excluding offensiveness from a definition of hate speech allows for a less opinionated perspective of what is and isn't hateful.



A broad definition of hate speech allows for a wide variety of actionable use cases

- Monitoring tensions across areas of concern
- Triaging distribution of human, material, and financial resources
- Performing long-term analysis on underlying causes and apply predictive results to future planning efforts



Combining data from numerous datasets can help reveal important relationships between government, citizens and external actors



Can hate speech be used to predict violence?



‘The graves of the Tutsi are only half full — we must complete the task’

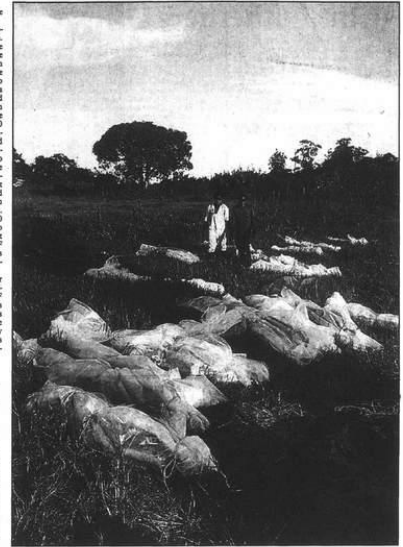
Richard Dowden, Africa Editor, reports on the rising Rwanda



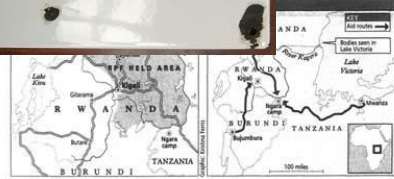
THE KILLING goes on. It could get worse. For seven weeks now, the massacres have continued.

the population — want revenge.”
 Revenge must precede.
 There is little or nothing to be done to stop it. Eight African countries — Zimbabwe, Tanzania, Ghana, Nigeria, Namibia, Senegal, Zambia and Congo — have agreed to contribute troops to the UN mission in Rwanda but they have not got the necessary finance and logistics to move their troops. Yesterday there was supposed to be a 48-hour ceasefire to allow the United Nations special representative, Iqbal Raza, to visit both sides in Kigali to discuss the deployment of the UN force. The guns fell silent at 8am but at 9.30 an artillery barrage began and Mr Raza had to turn back.
 Until yesterday the rebel of the Rwanda Patriotic Front, who are largely Tutsi, had been restrained and did not let their military struggle degenerate into revenge. But even if they maintain their discipline, their advance into the capital, expected in the coming days, will cause more pain. Hundreds of thousands of Hutus are expected to flee south and the Hutu militia, retraining with Hutu and with nothing to lose, will seek out more targets among

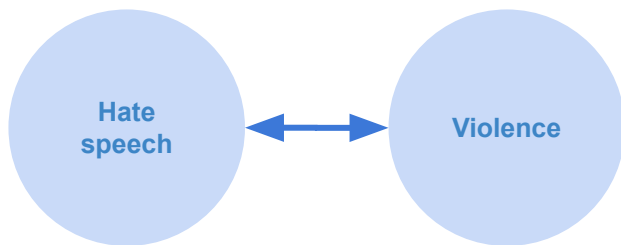
the Tutsi refugees who are unable to escape.
 This is perhaps the ghastly counterpoint to the heart of Rwanda. The RPF is mainly made up of the children of Tutsi who were driven out by a Hutu uprising in the early 1960s. Traditionally the Tutsi were an aristocracy who dominated the Hutus in Rwanda and Burundi. The two groups had a symbiotic relationship, though the Hutu always outnumbered the Tutsi by about ten to one. In 1990 the exiles formed the RPF and invaded. They were well organised and disciplined and had a “non-ethnic” ideology. They wanted to come home and fight for a share of power — not a complete takeover. However, led by government propaganda, many Hutus believed the RPF was a Tutsi army which would restore the Tutsi overlords. In Ngora camp there are 250,000 Hutus — and some Tutsi — who will see the RPF carries out the massacre. So even if the RPF maintains its discipline and takes over the country, it may not be accepted by Hutus.
 But there are signs that their discipline is slipping. Reliable reports yesterday suggest that RPF soldiers have murdered Hutus. As more and more RPF fighters realise their own families have been wiped out, how long will they maintain their self-control? In the spirit of revenge spreads, the future is unimaginable.
 Leading article, page 15



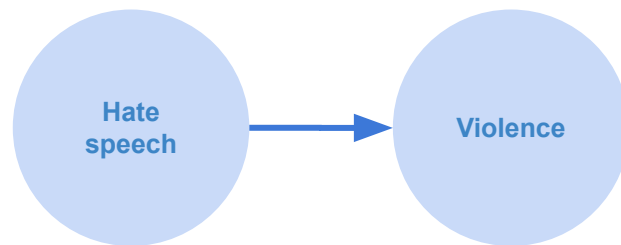
Bodies taken from Lake Victoria after being washed down the river Kagera. Mariella Ferrer/Saba/Kat



When we analyze the relationship between hate speech and violence, we're looking for **correlation** and **causation**



Correlation



Causation

Both correlation and causation have been extensively studied in a variety of real-world contexts.

Correlating spikes in use of “jew” on /pol

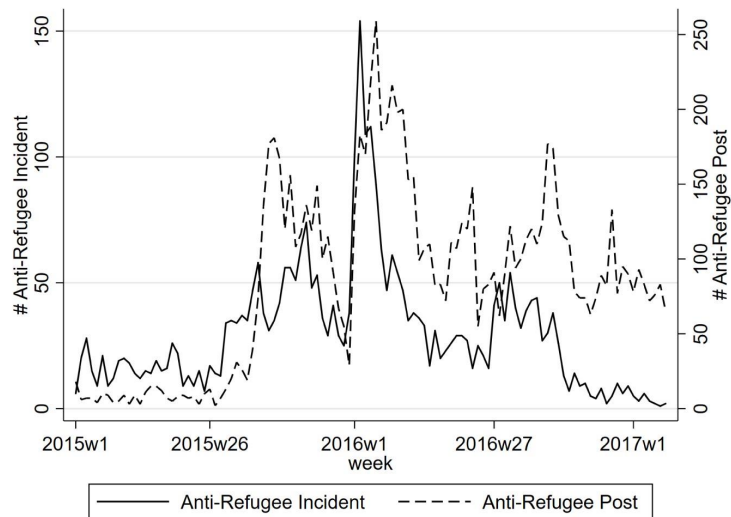
Rank	Date	Events
1	2016-12-25	2016-12-23: Samantha Power, US ambassador to the UN abstains from voting in a 140 Security Council vote to condemn Israel’s construction of settlements into the Palestinian territories 2016-12-19: ISIS truck attack in Berlin Germany
2	2017-01-17	2017-01-17: Presidential inauguration of Donald Trump 2017-01-17: Benjamin Netanyahu attacks the latest peace-conference by calling it “useless”
3	2017-04-02	2017-04-05: President Trump removes Steve Bannon from his position on the National Security Council 2017-04-06: President Trump orders a strike on the Shayrat Air Base in Homs, Syria, using 59 Tomahawk cruise missiles
4	2017-11-26	2017-11-29: According to a New York Times report, it is revealed that Jared Kushner has been interviewed by Robert Mueller’s team in November
5	2016-10-08	2016-10-09: Second presidential debate 2016-10-09: A shooting takes place in Jerusalem that kills a police officerser and two innocent people, wounding several others

Source: “A Quantitative Approach to Understanding Online Antisemitism” by Joel Finkelstein (Princeton University), Savvas Zannettou (Cyprus University of Technology), Barry Bradlyn (University of Illinois at Urbana-Champaign) and Jeremy Blackburn (University of Alabama at Birmingham)



Correlation between anti-refugee hate speech and violence

Anti-Refugee Posts and Incidents over time



This chart shows correlations between anti-refugee posts on the Facebook page of the right-wing group “Alternative fur Deutschland” (Alternative for Germany) and anti-refugee incidents reported in Germany from 2015 to 2017.

Source: “Fanning the Flames of Hate: Social Media and Hate Crime” by Karsten Muller (University of Warwick) and Carlo Schwarz (University of Warwick)



Historically, hate speech has emboldened violent perpetrators...

Psychological Science

aps | ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Us Versus Them: Social Identity Shapes Neural Responses to Intergroup Competition and Harm

Mina Cikara, Matthew M. Botvinick, Susan T. Fiske

First Published January 26, 2011 | Research Article
<https://doi.org/10.1177/0956797610397667>

Article information ▾

Abstract

Intergroup competition makes social identity salient, which hardships. The failures of an in-group member are painful, pleasure—a feeling that may motivate harming rivals. The neural responses to rival groups' failures correlate with like rivals. Avid fans of the Red Sox and Yankees teams view magnetic resonance imaging. Subjectively negative outgroup (rival team) activated anterior cingulate cortex and insula, v team or failure of the rival team, even against a third team effect, associated with subjective pleasure, also correlated fan of the rival team (controlling for general aggression). C effort within a social psychological research paradigm, which is

The screenshot shows the article 'Fueling the Fire: Violent Metaphors, Trait Aggression, and Support for Political Violence' by Nathan P. Kalmoe. It includes a sidebar with metrics (1,067 Views, 19 CrossRef citations, 125 Altmetric) and a list of actions (Full Article, Figures & data, References, Supplemental, Citations, Metrics, Reprints & Permissions). The abstract text is partially visible.

Journal
Political Communication >
Volume 31, 2014 - Issue 4

Articles

Fueling the Fire: Violent Metaphors, Trait Aggression, and Support for Political Violence

Nathan P. Kalmoe ✉
Pages 545-563 | Published online: 16 Oct 2014
[Download citation](#) <https://doi.org/10.1080/10584609.2013.852642>

Full Article | Figures & data | References | Supplemental | Citations | Metrics | Reprints & Permissions

Abstract

The recent concurrence of violent political rhetoric and violence against political targets in the U.S. and abroad has raised public concern about the effects of language on citizens. Building from theoretical foundations in aggression research, I fielded two nationally representative survey experiments and a third local experiment preceding the 2010 midterm elections in Florida.

“...certain types of hate speech can serve as both a warning sign and a catalyst of genocide and mass atrocities.”

US Holocaust Memorial Museum

“There's been a significant, sustained increase in anti-Semitic activity since the start of 2016, and... the numbers have accelerated over the past five months.”

Anti-Defamation League



...either directly or indirectly

Psychosomatic Medicine. 66(3):343-348, MAY 2004

PMID: 15184693

Issn Print: 0033-3174

Publication Date: 2004/05/01

Immigrant Suicide Rates as a Function of Ethnophaulisms: Hate Speech Predicts Death

Brian Mullen; Joshua Smyth;

[- Author Information](#)

From the Department of Psychology, Syracuse University, Syracuse, NY.

Abstract

Objective:

The purpose of this study was to determine whether suicide rates among ethnic immigrant groups were predicted by the ethnophaulisms, or the hate speech, used to refer to those ethnic immigrant groups.

Methods:

Data were obtained for 10 European ethnic immigrant groups during the 1950s. These 10 European ethnic immigrant groups accounted for approximately 40% of all immigration into the United States during this time period. Both the suicide rates for these ethnic immigrant groups in the United States and suicide rates for those ethnic immigrant groups in their countries of origin were derived. The complexity and valence of ethnophaulisms used to refer to these ethnic immigrant groups were derived from the historical record of hate speech in the United States.

Results:

Consistent with previous research, immigrant suicide rates were strongly correlated with origin suicide rates. As expected, the suicide rates for ethnic immigrant groups in the United States were significantly predicted by the negativity of the ethnophaulisms used to refer to those ethnic immigrant groups. This pattern was obtained even after taking into account the suicide rates for those ethnic immigrant groups in their countries of origin, and even after taking into account the size of these ethnic immigrant groups.



Teaching humans to detect written hate speech is difficult.
Teaching machines is exponentially more difficult.

- Small sample sizes
- Lack of continuity
- Group identity of speaker, recipient and/or subject
- Intent / sentiment
- Location of conversation
- Loan words, patois, mixed languages
- Misspellings
- Homonyms
- Obfuscation

What many of these challenges have in common is **context**.



Why is context important?



124a T. F. Kuboye Rd, Lagos, Nigeria

Photo by Joshua Oluwagbemiga

In Nigeria, when a Hausa talks to another Hausa:

“aboka”: **friend**

But when another ethnicity talks about a Hausa:

“aboka”: **uneducated**



Context is challenging because language is challenging

Not only can units of vocabulary have a hateful and non-hateful context, but language can be structured to communicate hateful context using sarcasm, double entendre, innuendo, euphemism, metaphor and other forms of rhetorical obfuscation



Data is a tool, not a solution

Large, geographically diverse datasets are prone to various types of artifact:



Granularity artifacts

Occur when data is analyzed at too granular a magnification, rather than at a level where crests and troughs can even out



Volume artifacts

Occur when a dataset is large enough to be assumed reflective of reality, even though the activity it models is much larger





Geolocation artifacts

Occur when data is filtered for specific locations, ignoring data that hasn't been geotagged



Evolutionary artifacts

Occur when the technology for acquiring data improves and new data is compared against data acquired from older technology





Technological artifacts

Occur when data is acquired across regions of varying technological infrastructure and adoption



Cultural artifacts

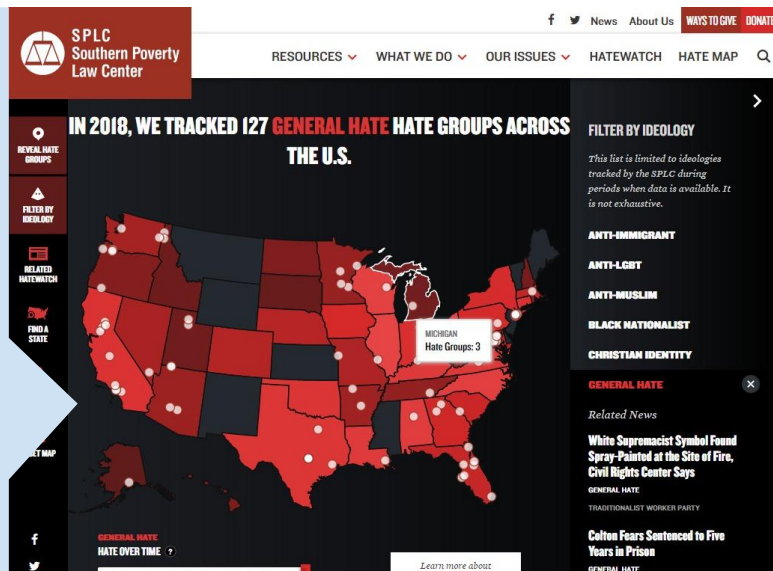
Occur when data is impacted by community attitudes toward identity and/or discrimination



Most hate speech monitoring technologies involve a mix of manual and automated processes

“[SPLC’s] hate map, which depicts the groups’ approximate locations, is the result of a year of monitoring by analysts and researchers and is typically published every January or February. It represents activity by hate groups during the previous year.”

Southern Poverty Law Clinic



<https://www.splcenter.org/hate-map>



“PeaceTech Lab’s series of hate speech Lexicons identify and explain inflammatory language on social media while offering alternative words and phrases that can be used to combat the spread of hate speech in conflict-affected countries”

PeaceTech Lab

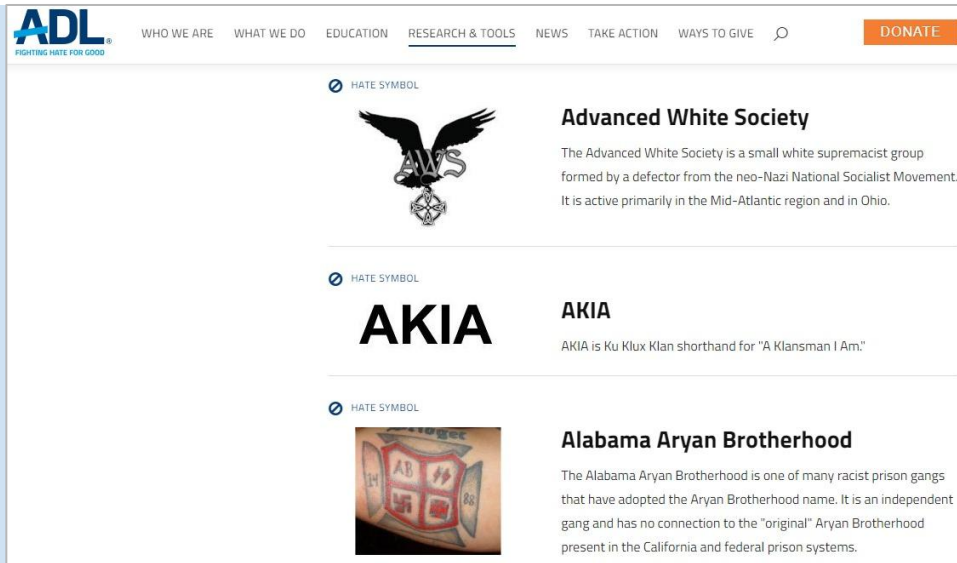


<https://www.peacetechnology.org/the-peacetechnology-toolbox>



“This database provides an overview of many of the symbols most frequently used by a variety of white supremacist groups and movements, as well as some other types of hate groups.”

Anti-Defamation League



The screenshot shows the ADL (Anti-Defamation League) website's "Hate Symbols Database". The navigation bar includes "WHO WE ARE", "WHAT WE DO", "EDUCATION", "RESEARCH & TOOLS", "NEWS", "TAKE ACTION", "WAYS TO GIVE", and a "DONATE" button. The database lists three hate symbols:

- Advanced White Society**: A black eagle with wings spread, perched on a white cross. Description: "The Advanced White Society is a small white supremacist group formed by a defector from the neo-Nazi National Socialist Movement. It is active primarily in the Mid-Atlantic region and in Ohio."
- AKIA**: The letters "AKIA" in a bold, black, sans-serif font. Description: "AKIA is Ku Klux Klan shorthand for 'A Klansman I Am.'"
- Alabama Aryan Brotherhood**: A photograph of a prison cell door with a red and white sign. Description: "The Alabama Aryan Brotherhood is one of many racist prison gangs that have adopted the Aryan Brotherhood name. It is an independent gang and has no connection to the 'original' Aryan Brotherhood present in the California and federal prison systems."


<https://www.adl.org/hatesymbolsdatabase>



“The Wiesenthal Center has tracked, over the past year, the continued emergence of Alt.Tech – a new generation of social media platforms that serves the Alt-Right – as well as the emergence on popular gaming platforms of bigotry, anti-Semitism and the glorification of radical Islamic terror.”

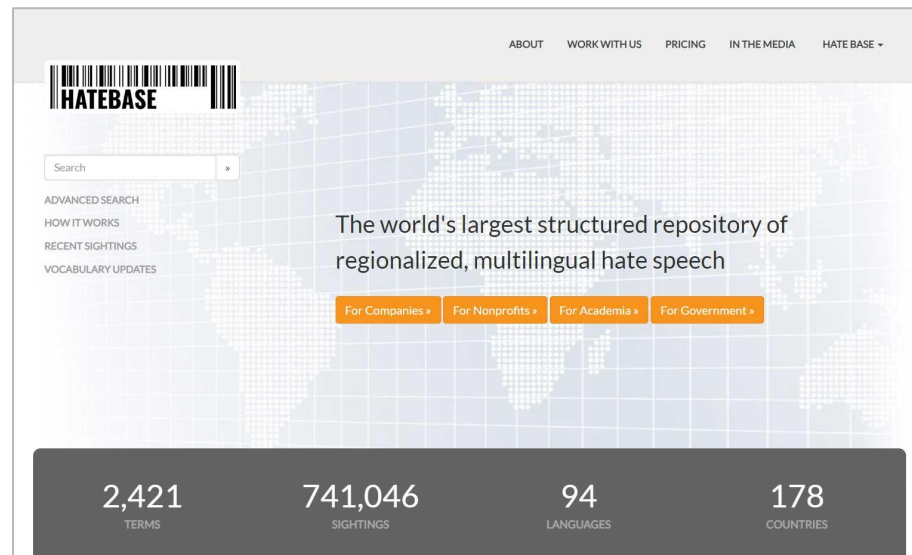
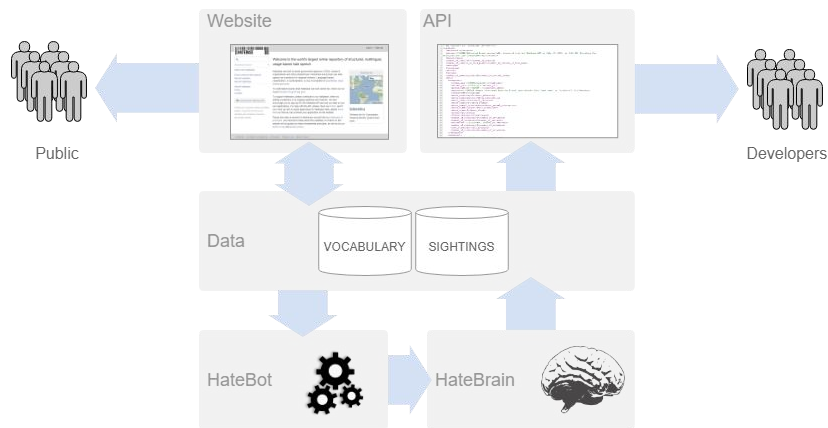
Simon Wiesenthal Center

The screenshot displays the 'digitalterrorism.hate' website interface. At the top, it features the logo 'simon wiesenthal center digitalterrorism.hate 2019+' and navigation tabs for 'REPORT CARD', 'REPORT ACTIVITY', 'STRATEGIES 2.0', 'TERRORISM', and 'GEOGRAPHIC'. The main content area shows a document titled 'Mosque Hunting' with the following text: 'Tom Warriner is a British neo-Nazi whose vk.com avatar pictures him in camouflage with a Swastika arm band and a baseball bat. In July he re-posted a picture of a specialized shotgun and a box of 12 gauge shells, adding the statement "Mosque Hunting!" Warriner has over 1,500'. To the right, a VK profile for 'Tom Warriner' is visible, showing a profile picture of a person in camouflage, a 'Friends' list, and a 'Photos' section. Below the document, there are search and site app icons, and a world map with markers for 'server location' and 'site owner'. At the bottom, a footer reads 'simon wiesenthal center snider social action institute' and 'To report a hate or terror website, blog, newsgroup or video please email: report@wiesenthal.com'.

<http://digitalhate.net> 



Hatebase is a technology platform for monitoring and analyzing **multilingual** and **regionalized** hate speech



<https://hatebase.org>



Our data

Hatebase ingests approximately 10,000 unique datapoints every 24 hours.

Vocabulary	2,691
-------------------	-------

Sightings	818,075
------------------	---------

Languages	95
------------------	----

Countries	178
------------------	-----

Users	1,294
--------------	-------

API Users	599
------------------	-----

Nonprofits	32
-------------------	----

Universities	193
---------------------	-----



Where our data comes from

Our **sightings** (incidents) dataset is generated from a variety of public data sources (e.g. social networks, online comments, forums).

Much of our **vocabulary** dataset comes from NGOs and other partners in linguistically diverse areas of the world.

We have several other internal datasets which we've built from public sources to help analyze hate speech (e.g. inflammatory language).



Hatebase is used by various **public and private entities**



Governments



Tech Companies



Academia



Law Enforcement



Publishers



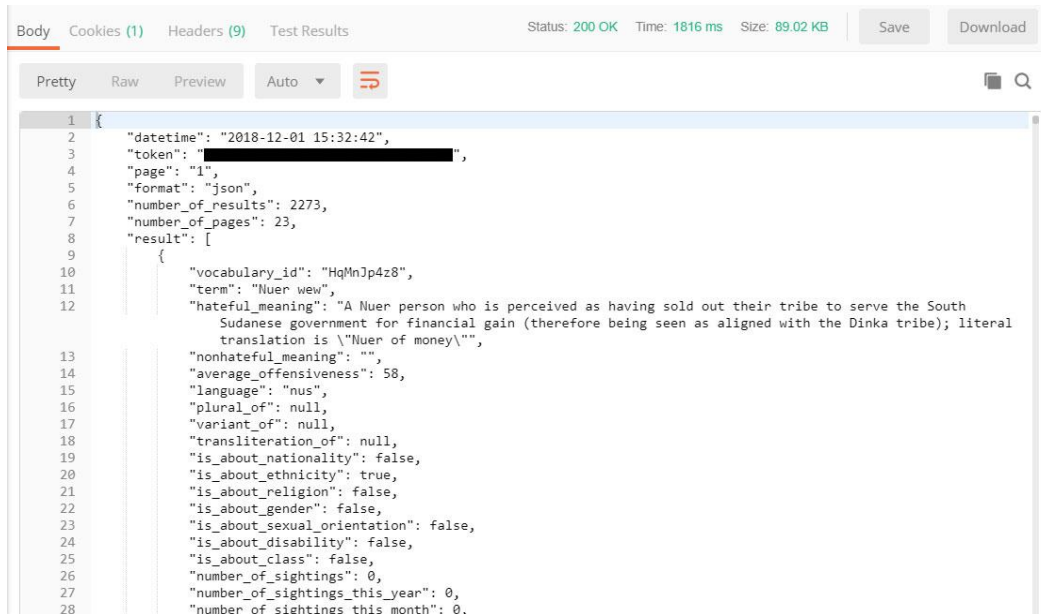
Nonprofits



Most users interact with Hatebase through our API

The current version of our API currently provides several endpoints for:

- Downloading vocabulary
- Downloading sightings
- Submitting content for analysis



```
Body Cookies (1) Headers (9) Test Results Status: 200 OK Time: 1816 ms Size: 89.02 KB Save Download
Pretty Raw Preview Auto
1 {
2   "datetime": "2018-12-01 15:32:42",
3   "token": "████████████████████",
4   "page": "1",
5   "format": "json",
6   "number_of_results": 2273,
7   "number_of_pages": 23,
8   "result": [
9     {
10      "vocabulary_id": "HqMnJp4z8",
11      "term": "Nuer wew",
12      "hateful_meaning": "A Nuer person who is perceived as having sold out their tribe to serve the South
13        Sudanese government for financial gain (therefore being seen as aligned with the Dinka tribe); literal
14        translation is \"Nuer of money\"",
15      "nonhateful_meaning": "",
16      "average_offensiveness": 58,
17      "language": "nus",
18      "plural_of": null,
19      "variant_of": null,
20      "transliteration_of": null,
21      "is_about_nationality": false,
22      "is_about_ethnicity": true,
23      "is_about_religion": false,
24      "is_about_gender": false,
25      "is_about_sexual_orientation": false,
26      "is_about_disability": false,
27      "is_about_class": false,
28      "number_of_sightings": 0,
29      "number_of_sightings_this_year": 0,
30      "number_of_sightings_this_month": 0,
```



Hatebase is supporting research at several universities and research labs



Cornell University



UNIVERSITY OF
OXFORD



HARVARD
UNIVERSITY



Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



Yale University



UCLA | SCHOOL OF LAW



universität
wien



UNIVERSITY OF
TORONTO

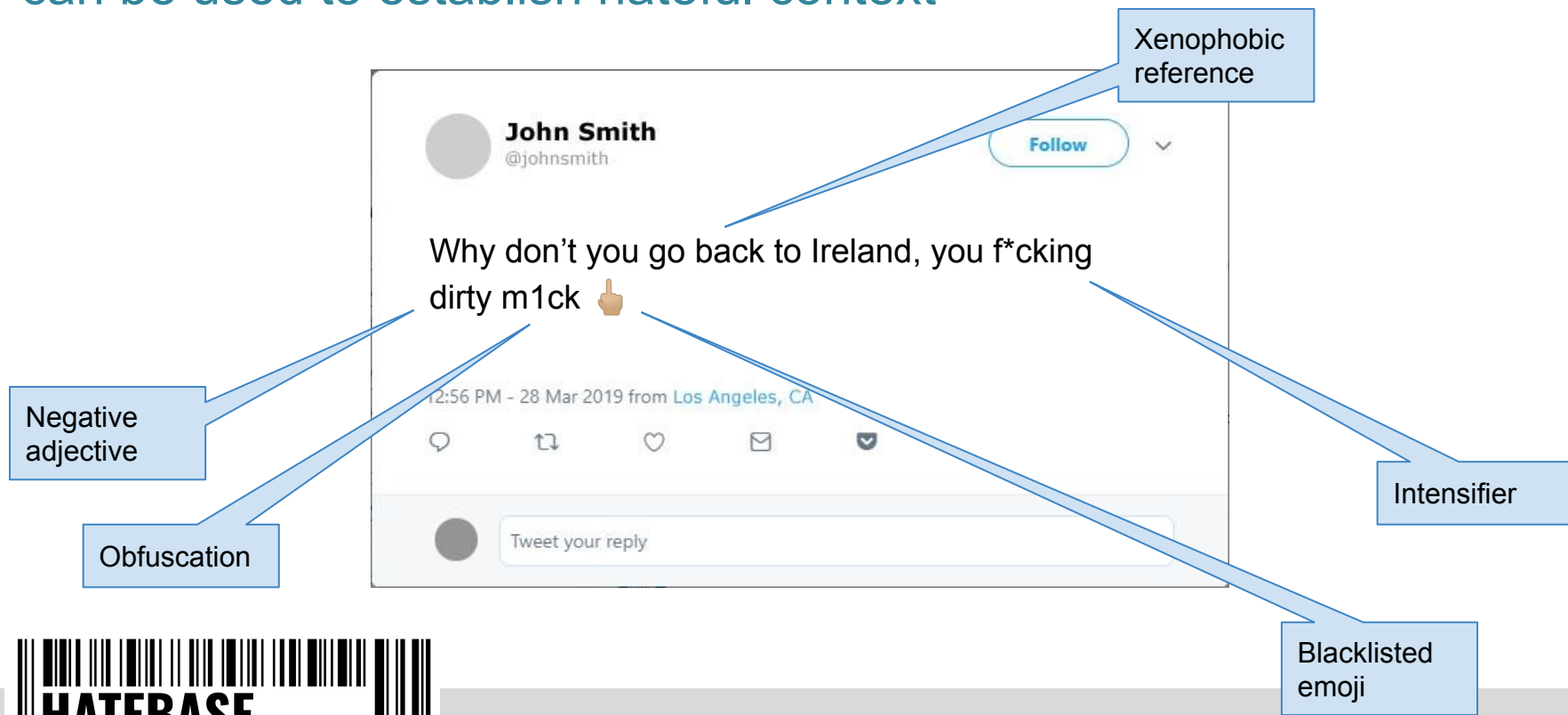


Hatebase is built around a natural language processing (NLP) engine called **HateBrain**

- Recognizes hate speech terms, even if obfuscated (e.g. leetspeak)
- Eliminates homonyms using rudimentary language detection
- Recognizes clinical (non-hateful) contexts
- Assesses the probability of hateful context using helper language which we call “pilotfish”



Hatebase's **pilotfish** are helper terms and grammatical cues which can be used to establish hateful context



Hate speech vs. free speech

Azhar Ahmed sentenced over Facebook soldier deaths slur

9 October 2012

f t e Share

A man who posted an offensive Facebook message following the deaths of six British soldiers has been given a community order.

Azhar Ahmed, 20, of Fir Avenue, Ravensthorpe, West Yorkshire, was found guilty in September of sending a grossly offensive communication.

He said he did not think the message, which said "all soldiers should go to hell!", was offensive.



Ahmed posted the message on just days after the soldiers' death

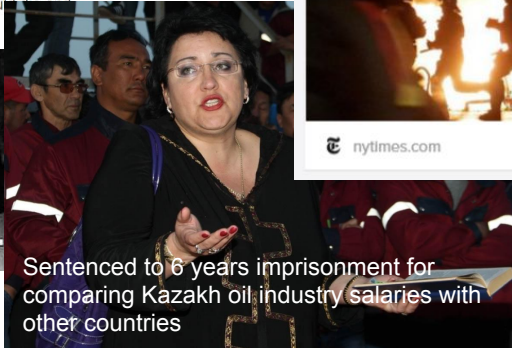
Germany, in a First, Shuts Down Left-Wing Extremist Website



nytimes.com

Venezuela's new "anti-hate" law seeks to silence media

November 9, 2017 3:55 PM ET



Sentenced to 6 years imprisonment for comparing Kazakh oil industry salaries with other countries

Countries with laws against holocaust denial



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 and 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)



Government, law enforcement and peacekeepers have a right / responsibility to monitor hate speech as an early indicator of violence



We strongly support **constructive, self-sustaining, actionable approaches** to hate speech reduction

- **Research and analysis** to understand the root causes of hate speech, as well as the complex relationship between hate speech and violence
- **Informed resource allocation** to help focus timely attention on mitigating the impact of hate speech in specific fragile regions
- **Counter-messaging of** hateful disinformation and misinformation

These use cases ultimately inform the **design and architecture** of the Hatebase technology platform.



Hatebase recommendations



Establish a working definition of hate speech based on actionable, long-term monitoring goals



Be alert to data artifacts, particularly in small sample sizes (and inquire about sample sizes)



Promote a culture of sharing data and methods, and discourage findings which aren't replicable



Collaborate with governments and non-government entities to monitor regional hate speech and analyze trends both contemporaneously and historically



Delineate hate speech from free speech, and unambiguously reject the misuse of hate speech as a means of suppressing dissent



Hatebase.org

Timothy Quinn, Principal

in [linkedin.com/in/timothyquinn](https://www.linkedin.com/in/timothyquinn)

 tim@hatebase.org

