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?

“...anti-Christian hate speech...”
Catholic League

“We cannot credibly dismiss that many do find Charlie Hebdo racist and Islamophobic…”
Al Jezeera

“...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 identity
- **Generalization** which, regardless of intent, stereotypes people based on a shared identity
- **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

Nur für Arier! Juden unerwünscht!

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

Both correlation and causation have been extensively studied in a variety of real-world contexts.
## Correlating spikes in use of “jew” on /pol

<table>
<thead>
<tr>
<th>Rank</th>
<th>Date</th>
<th>Events</th>
</tr>
</thead>
</table>
| 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 officer 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

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 2107.

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...

"...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
Immigrant Suicide Rates as a Function of Ethnophobiaisms: Hate Speech Predicts Death

Brian MullengJoshua Smyth;

Abstract

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

Methods:
Data were obtained for 10 European ethnic immigrant groups during the 1990s. 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 ethnophobiaisms 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 ethnophobiaisms 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...
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?

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.
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.
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.

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
“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

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
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.

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>2,691</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sightings</td>
<td>818,075</td>
</tr>
<tr>
<td>Languages</td>
<td>95</td>
</tr>
<tr>
<td>Countries</td>
<td>178</td>
</tr>
<tr>
<td>Users</td>
<td>1,294</td>
</tr>
<tr>
<td>API Users</td>
<td>599</td>
</tr>
<tr>
<td>Nonprofits</td>
<td>32</td>
</tr>
<tr>
<td>Universities</td>
<td>193</td>
</tr>
</tbody>
</table>
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
- Law Enforcement
- Tech Companies
- Publishers
- Academia
- 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
Hatebase is supporting research at several universities and research labs.
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

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.

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

linkedin.com/in/timothyquinn
tim@hatebase.org