Hate Speech, Incitement, and Harassment

CS 152 — Trust and Safety

Alex Stamos

Stanford Cyber Policy Center

2026

Content Warning:

This lecture will contain frank discussion of harassment and hate speech targeted at people based on race, religion, gender or ethnicity. The content of this lecture may be shocking or triggering for some people. Our intent in showing it is not to offend, upset or sensationalize this material. Our learning objective is to understand the tactics for and impact of victimization on platforms and how platforms can respond.

What Will We Learn Today?

What Will We Learn Today?

- How hate speech is defined by platforms and around the world
- Why hate speech matters: connection to violence and genocide
- How quickly hate speech can morph as groups adapt their language
- Case studies: Myanmar, Meta Oversight Board decisions
- Possibilities and limits of AI for hate speech prevention
- Platform policies, interventions, and global enforcement challenges

Definitions

Platform Definitions of Hate Speech

Meta: "[D]irect attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation."

TikTok: "Hate speech and hateful behaviour attack, threaten, dehumanise or degrade an individual or group based on their characteristics."

UN: "[A]ny kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are."

Key Elements of Hate Speech Definitions

- **Attack** — violent, dehumanizing, harmful stereotypes, contempt
- **Protected characteristics** — race, ethnicity, religion, gender, etc.
- **Intent vs. Impact** — does intent matter? What about "jokes"? **The basic challenges in dealing with hate speech:**

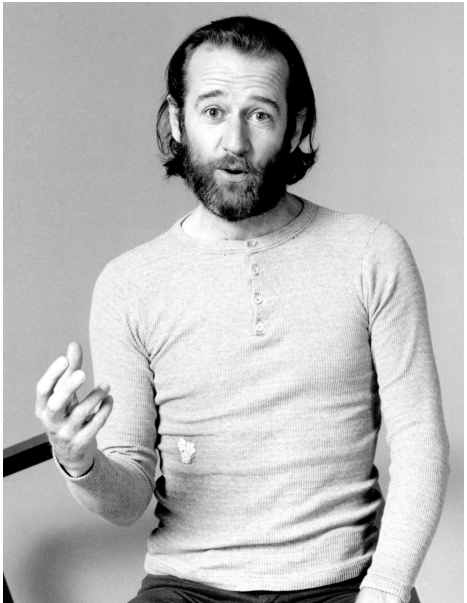
1) Hate speech can vary in impact from feelings being hurt to fomenting genocide
2) The global legal situation varies from the same speech being harshly prohibited to strongly legally protected
3) The relationship between the speaker and target is often important context
4) Different people can have very different reactions

Key Elements of Hate Speech Definitions

- **Attack** — violent, dehumanizing, harmful stereotypes, contempt
- **Protected characteristics** — race, ethnicity, religion, gender, etc.
- **Intent vs. Impact** — does intent matter? What about "jokes"? **The basic challenges in dealing with hate speech:**

1) Hate speech can vary in impact from feelings being hurt to fomenting genocide
2) The global legal situation varies from the same speech being harshly prohibited to strongly legally protected
3) The relationship between the speaker and target is often important context
4) Different people can have very different reactions

Do we just ban bad words?



What is a bad word?

Let’s see if you did the reading…

What is the most common three word phrase on Facebook ending in “…bitch”?

Happy Birthday

Legal Framework

LEGAL PARAMETERS OF HATE SPEECH

UNITED STATES

- No legal definition of hate speech
- Most hateful language is protected under the First Amendment, which guarantees the right for private citizens to be free from government interference in speech
- Under the First Amendment, hate speech is not protected when it directly incites imminent criminal activity or consists of specific threats of violence targeted against a person or group

Most social media companies are American and were founded to **privilege freedom of speech** and related principles

LEGAL PARAMETERS OF HATE SPEECH

GERMANY

Illegal to publicly incite hatred against parts of the population or to call for violent or arbitrary measures against them or to insult, maliciously slur or defame them in a manner violating their human dignity.

In 2017, criminalized hate speech on social media sites, with large fines for platforms failing to remove illegal content.

NORWAY

Defines hate speech as "threatening or insulting anyone, or inciting hatred or persecution of or contempt for anyone because of… a) skin color or national or ethnic origin, b) religion or life stance, or c) lifestyle or orientation."

FRANCE

Penal code and press laws prohibit communication that is defamatory or insulting, or that incites discrimination, hatred, or violence against a person or group based on specific criteria.

SOUTH AFRICA

Detailed and comprehensive laws against hate speech, specifying groups and attributes that are absent from other countries' laws such as pregnancy, marital status, conscience, language, color, and "any other group where discrimination… causes or perpetuates systemic disadvantage; (ii) undermines human dignity; or (iii) adversely affects the equal enjoyment of a person's rights and freedoms"

LAWS AGAINST HATE SPEECH

Many nations across the world have legal limitations on hate speech. Many of these laws came about in the wake of World War II, designed to curb incitement to racial, ethnic, and religious hatred after the Holocaust.

LEGAL PARAMETERS OF HATE SPEECH

INTERNATIONAL CONVENTION ON THE ELIMINATION OF ALL FORMS OF RACIAL DISCRIMINATION (CERD)

Article 4(a) requires governments to outlaw the spread of ideas based on racial superiority or hatred, along with incitement to racial discrimination and acts of violence or incitement to violence against any race or ethnic group.

International Law

Prohibitions on hateful speech in international law are contained in two main instruments: the CERD and the ICCPR.

INTERNATIONAL COVENANT ON CIVIL AND POLITICAL RIGHTS (ICCPR)

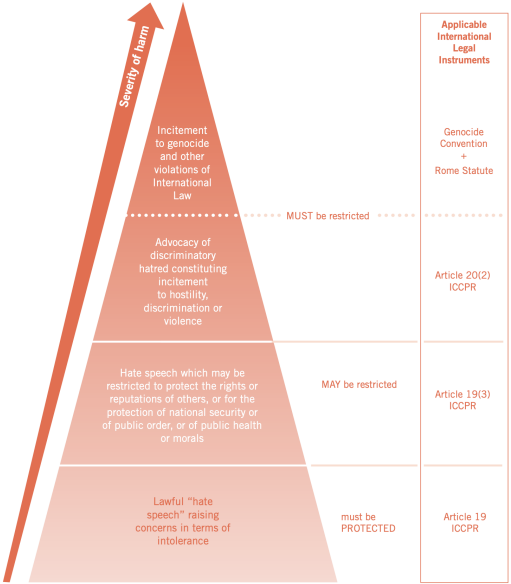
"Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

LEGAL PARAMETERS OF HATE SPEECH

HATE SPEECH AND FREEDOM OF SPEECH

- Freedom of expression has limitations under all legal codes and international human rights protections
  - Even the First Amendment does not protect certain types of unprotected speech, like incitement to violence
- Silencing voices can be counterproductive: it may make perpetrators more likely to resort to violence if they have no peaceful way of expressing and resolving their grievances
- Broad or vague definitions of hate speech and related crimes can jeopardize freedom of speech, because vagueness allows for subjective application.
- Laws against hate speech or hateful speech are often misused to punish and silence journalists, dissenters, and minorities, recently in countries as varied as Hungary, India, Rwanda, Kazakhstan, and Bahrain

Definitions Under International Law



United States

- No legal definition of hate speech
- Most hateful language is protected under the First Amendment, which guarantees the right for private citizens to be free from government interference in speech
- Under the First Amendment, hate speech is not protected when it directly incites imminent criminal activity or consists of specific threats of violence targeted against a person or group

Most social media companies are American and were founded to privilege freedom of speech and related principles

Nigeria

National Commission for the Prohibition of Hate Speech bill

4. Hate speech

(1) A person who uses, publishes, presents, produces, plays, provided, distributes and/or directs the performance of any material, written and or visual which is threatening, abusive or insulting or involves the use of threatening, abusive or insulting words or behavior commits an offence if such person intends thereby to stir up ethnic hatred, or having regard to all the circumstances, ethnic hatred is likely to be stirred up against any person or person from such an ethnic group in Nigeria.

Broad, vague, and subject to politicized enforcement

South Africa

10 Prohibition of hate speech

(1) Subject to the proviso in section 12, no person may publish, propagate, advocate or communicate words based on one or more of the prohibited grounds, against any person, that could reasonably be construed to demonstrate a clear intention to-

(a) be hurtful;

(b) be harmful or to incite harm;

(c) promote or propagate hatred.

South Africa (continued)

'prohibited grounds' are-

(a) race, gender, sex, pregnancy, marital status, ethnic or social origin, colour, sexual orientation, age, disability, religion, conscience, belief, culture, language and birth; or

(b) any other ground where discrimination based on that other ground-

(i) causes or perpetuates systemic disadvantage;

(ii) undermines human dignity; or

(iii) adversely affects the equal enjoyment of a person's rights and freedoms in a serious manner that is comparable to discrimination on a ground in paragraph ***(a)*** ;

Germany

Volksverhetzung = “incitement to hatred”

(1) Anyone who, in a manner likely to disturb public peace,

1. incites hatred, calls for violent or arbitrary measures against a national, racial, religious or ethnic group, against a section of the population or against an individual because of his or her membership of a aforementioned group or section of the population
2. attacks the human dignity of others by insulting, maliciously slandering or slandering a aforementioned group, part of the population or an individual because of his or her membership of a aforementioned group or part of the population,

shall be punished with imprisonment from three months to five years.

EU: Digital Services Act

- Requires big platforms to respond to government requests to remove illegal content (including hate speech, if illegal)
- Requires big platforms to allow users to easily flag content as hate speech
- Transparency reporting requirements
- First fine under DSA: X fined EUR 120M (December 2025)

Hate Speech Can Precede Atrocities

- **The Holocaust**: The Nazis first banned independent media and pushed out anti-Semitic hate speech on state-controlled media
- **The Cambodian Genocide**: The Khmer Rouge first pushed narratives labeling intellectuals and various ethnic and religious minorities “enemies of the people”
- **The Rwandan Genocide**: Decades of dehumanizing hate speech against the Tutsi preceded the genocide

The Eight Stages of Genocide

1. Classification
2. Symbolization
3. Dehumanization
4. Organization
5. Polarization
6. Preparation
7. Extermination
8. Denial

From “The Eight Stages of Genocide”, G. Stanton, Genocide Watch

Case Study: Myanmar

3. **DEHUMANIZATION:** One group denies the humanity of the other group. Members of it are equated with animals, vermin, insects or diseases. Dehumanization overcomes the normal human revulsion against murder. At this stage, hate propaganda in print and on hate radios is used to vilify the victim group. In combating this dehumanization, incitement to genocide should not be confused with protected speech. Genocidal societies lack constitutional protection for countervailing speech, and should be treated differently than in democracies. Hate radio stations should be shut down, and hate propaganda banned. Hate crimes and atrocities should be promptly punished.

New Light of Myanmar

11th Waxing of Tabaung 1374 ME — Thursday, 21 February, 2013

No 'Rohingya' race in Myanmar, says Deputy Minister

17th day sixth regular session of first Pyithu Hluttaw session in progress—news

8 ANALYSIS — Global NEW LIGHT of MYANMAR — 1 November 2016

Opinion

Write for us

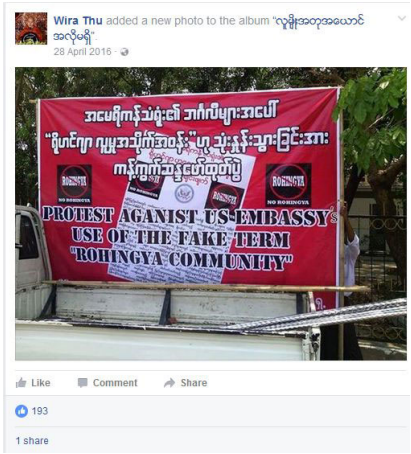
The Thorn Needs Removing If It Pierces!

Khin Maung Oo

Singing Anyeint and Dancing Anyeint Part (2)

Maha Sukhhumasa Jotikadhaja Sithu, Dr. Khin Maung Nyunt

Hate Speech Can Easily Become Incitement To Violence



Roots of Facebook's Failure in Myanmar

1. Unconstrained hyper-growth and I18N
2. Lack of content policy knowledge
3. Government-sponsored genocide, no legal protections and mass media control
4. No employees on the ground due to government feedback loop
5. No AI/ML capability in the local language
6. Extremely limited content moderation teams in relevant languages
7. Content moderators drawn from the ethnic "winners"

Facebook's Subsequent Reactions Have Not Improved the Reality

VIOLENCE, CONFLICTS, AND PROTECTION OF CIVILIANS

There is **widespread use of extrajudicial executions** by the military often following arrests carried out in villages and towns after raids.

In the past year, there were **at least 24 incidents where 5 or more people were detained and then killed** in a single incident.

At least **920 people have died while in military custody** since 1 February 2021.

Individuals were also **killed in airstrikes, by artillery fire and during arson attacks on villages**.

Fig. 1 Tactic (airstrikes, artillery, burnings, detentions) by number of incidents by year since the coup

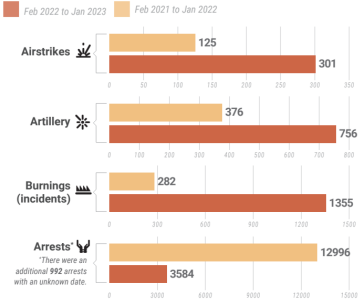
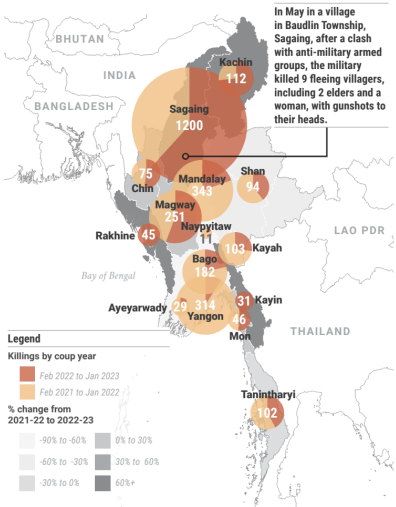
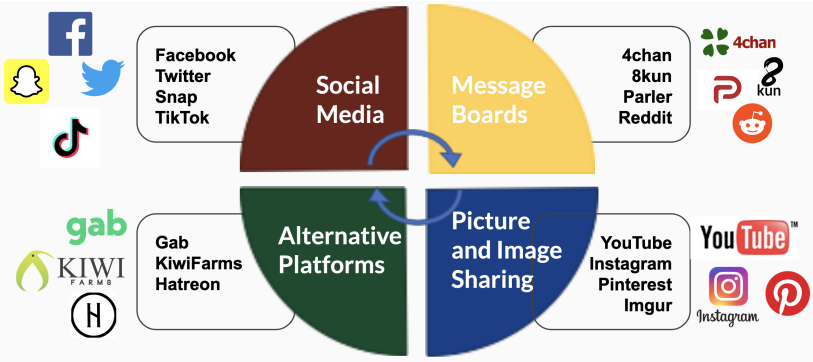


Fig. 2 Killings by region by year since the coup



All figures are based on available information. The boundaries and names shown and designations

Organized Hate Speech Is Often Multi-platform



Context and Symbols

Context Matters

Active Co-opting of Popular Symbols

Is It Real or Ironic?

File: IMG_2263.jpg (75 KB, 1073x531)

INTRODUCING: OPERATION O-KKK Anonymous (ID: GqQuAU8x)
02/27/17(Mon)03:32:18 No.114482325 ▶ >>114482375 >>114482572 >>114483103
>>114483235 >>114483753 >>114483969 >>114484471 >>114485569 >>114485688
>>114485751 >>114486015 >>114486043 >>114486249 >>114486520 >>114488048
>>114488197 >>114489614 >>114489933 >>114490057 >>114490153 >>114491182
>>114492623 >>114492860 >>114492995 >>114493101 >>114493153 >>114493173
>>114493296 >>114493406 >>114493456 >>114493524 >>114493725 >>114493753
>>114493814 >>114494238 >>114494627

We must flood twitter and other social media websites with spam, claiming that the OK hand sign is a symbol of white supremacy. Make fake accounts with basic white girl names and type shit like: OMG that's so truuuuu

Use as many emojis as you please. It would also be good for us to associate the OK sign being a symbol of white supremacy because Mel Gibson used it.

Use the hashtag "#PowerHandPrivilege" in all of your tweets and whatnot related to this.

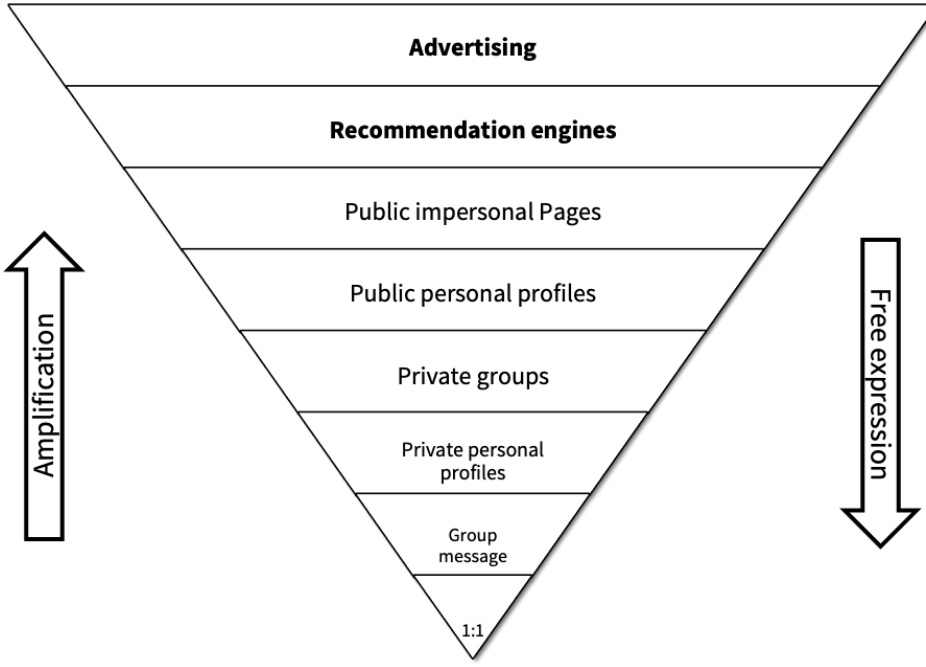
Bonus points if your profile pic is something related to supporting feminism.

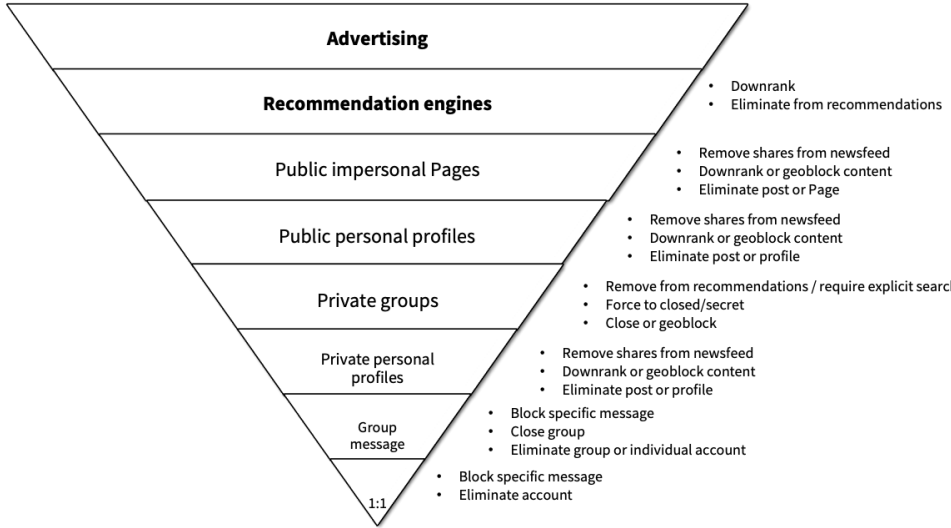
Leftists have dug so deep down into their lunacy. We must force to dig more, until the rest of society ain't going anywhere near that shit.

Responses and Mitigations

Policy Responses

1. Set clear policies broad enough to catch speech that sets the stage for violence
2. Allow for complicated contexts, including criticism and counter-speech
3. Stay aware of shifting meanings in the cultures in which you operate
4. Rapid response from moderators, policy teams and engineers is key


Advertising
Recommendation engines
Public impersonal Pages
Public personal profiles
Private groups
Private personal profiles
Group message
1:1
Amplification
Free expression


Advertising
Recommendation engines
Public impersonal Pages
Public personal profiles
Private groups
Private personal profiles
Group message
1:1
• Downrank
• Eliminate from recommendations
• Remove shares from newsfeed
• Downrank or geoblock content
• Eliminate post or Page
• Remove shares from newsfeed
• Downrank or geoblock content
• Eliminate post or profile
• Remove from recommendations / require explicit search
• Force to closed/secret
• Close or geoblock
• Remove shares from newsfeed
• Downrank or geoblock content
• Eliminate post or profile
• Block specific message
• Close group
• Eliminate group or individual account
• Block specific message
• Eliminate account

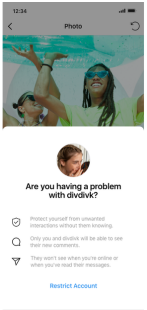
Product/Technical Responses

1. Reporting flows
2. Proactive monitoring
3. Productive friction
4. User-controls
5. Shadow-blocking

What Are Platforms Doing to Prevent Hate Speech?

Platforms take a variety of actions to mitigate the spread or creation of hate speech:

- Blocking content outright
- Putting warnings or disclaimers on content
- Blocking users
- Putting users in “time out”
- Warning users before they post likely violating content



Limits of AI

Vox recode

The algorithms that detect hate speech online are biased against black people

A new study shows that leading AI models are 1.5 times more likely to flag tweets written by African Americans as "offensive" compared to other tweets.

By Shirin Ghaffary | Aug 15, 2019, 11:00am EDT

AI has limited understanding — training data must be appropriately labeled with context

- Platforms can only see the hate they understand yet miss other types
- Algorithmic error cases can be biased towards certain races, ethnicities
- Those building training data need to be experts in specific hate groups
- Is labeling up to date as language evolves?

Facebook's Hate Speech Policy on Slurs

"We also prohibit the usage of slurs that are used to attack people on the basis of their protected characteristics. However, we recognize that people sometimes share content that includes slurs or someone else's hate speech to condemn it or raise awareness. In other cases, speech, including slurs, that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content."

Final Thought: Is There Academic Merit to Preserving This Info?

Hate speech detection with comment embeddings [PDF] acm.org
N Djuric, J Zhou, R Morris, M Grbovic… - Proceedings of the 24th …, 2015 - dl.acm.org Find it@Stanford
We address the problem of **hate speech** detection in online user comments. **Hate speech**, defined as an" abusive **speech** targeting specific group characteristics, such as ethnicity, religion, or gender", is an important problem plaguing websites that allow users to leave …
☆ 99 Cited by 252 Related articles All 9 versions

Automated **hate speech** detection and the problem of offensive language [PDF] aaai.org
T Davidson, D Warmsley, M Macy, I Weber - Eleventh international aaai …, 2017 - aaai.org
A key challenge for automatic **hate-speech** detection on social media is the separation of **hate speech** from other instances of offensive language. Lexical detection methods tend to have low precision because they classify all messages containing particular terms as **hate** …
☆ 99 Cited by 408 Related articles All 10 versions »

[BOOK] **Hate speech**: The history of an American controversy [PDF] semanticscholar.org
S Walker - 1994 - books.google.com
The First Amendment protects even the most offensive forms of expression: racial slurs, hateful religious propaganda, and cross-burning. No other county in the world offers the same kind of protection to offensive **speech**. How did this free **speech** tradition develop …
☆ 99 Cited by 364 Related articles All 3 versions »

Detecting **hate speech** on the world wide web [PDF] aclweb.org
W Warner, J Hirschberg - Proceedings of the second workshop on …, 2012 - dl.acm.org
We present an approach to detecting **hate speech** in online text, where **hate speech** is defined as abusive **speech** targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation. While **hate speech** against any group may exhibit …
☆ 99 Cited by 250 Related articles All 13 versions

Deep learning for **hate speech** detection in tweets [PDF] acm.org
P Badjatiya, S Gupta, M Gupta, V Varma - Proceedings of the 26th …, 2017 - dl.acm.org Find it@Stanford
Hate speech detection on Twitter is critical for applications like controversial event extraction, building AI chatterbots, content recommendation, and sentiment analysis. We define this task as being able to classify a tweet as racist, sexist or neither. The complexity of …
☆ 99 Cited by 240 Related articles All 10 versions

[BOOK] Countering online **hate speech**
I Gagliardone, D Gal, T Alves, G Martinez - 2015 - books.google.com
The opportunities afforded by the Internet greatly overshadow the challenges. While not forgetting this, we can nevertheless still address some of the problems that arise. **Hate speech** online is one such problem. But what exactly is **hate speech** online, and how can we …
☆ 99 Cited by 164 Related articles »

How can researchers stop the spread of something they can't see?

What can/should the limitations be on research materials?

What is the role of companies in this dynamic?

Word of the Day

gerbil

Exercise

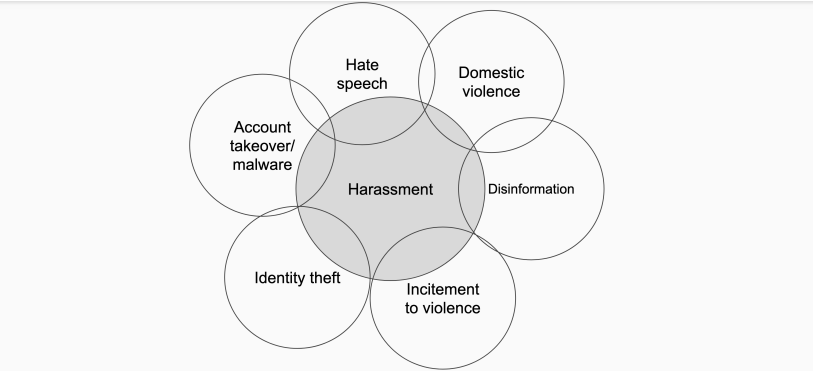
Exercise

Take a minute to read the entire worksheet, then form a small group and answer the questions.

You have 15 minutes to do the whole worksheet. Do not get bogged down with step 1. Make some decisions then move on.

Harassment

Harassment Is Complicated and Adjacent to Many Other Abuses



The State of Online Harassment (Pew Center 2021)

- Surveyed 10,093 American adults in September, 2020
- Roughly four-in-ten have experienced online harassment, half due to political reasons and half experiencing more severe behaviors

Trolling Tactics

Public

- Swatting
- Doxxing
- Astroturfing
- Extortion
- Impersonation
- Public Non-Consensual Intimate Imagery (NCII)
- Sealioning

Communal

- Dogpiling
- Group-to-individual harmful or threatening speech
- Outreach to outside communities (school, church, employer, family)

Trolling Continued

Private

- Non-Consensual Intimate Imagery (NCII)
- Individual-to-individual harmful or threatening speech
- Email/text/platform based messaging
- Cyberstalking
- Catphishing
- Sextortion

Tactics Deep-Dive: Sealioning

- Users appear facially to be engaging in dialogue.
- However, the speaker intends to turn the tables on the target, inundating them with questions.
- When the target withdraws, the speaker claims that they were the one who was attempting to be open and civil.

Tactics Deep-Dive: Dogpiling

- Reliant on cybermobs.
- Often targeted based on personal characteristics (e.g. race, ethnicity, religion, gender, sexual orientation, sexual identity) or by expressing unpopular opinions.
- Target is then overwhelmed with negative content by an online mob.
- Goal is to make the person so tired of dealing with abuse that they either recant, withdraw, or are driven from online spaces completely.
- Sometimes the dogpiling is so extreme that a person can no longer participate in online forums.

Tactics Deep-Dive: Swatting

- Hoax call placed to law enforcement claiming a life-threatening event that the caller claims is taking place in a target's home or business.
  - Hostage situation
  - Murder/suicide
  - Home intrusion
- Police and emergency response teams show up to a target's home primed for violence
- Swattings are common within gamer communities and often target celebrities and public figures.
- Uninvolved man shot & killed as a result of his address being used in a gaming dispute & swatting in December 2017.

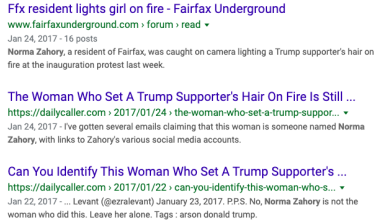
Examples of Harassment

Foundational Harassment Case: Gamergate

- Triggered by real-life boyfriend but picked up by strangers
- Coordinated dogpiling through many attack channels against a small number of female game designers
- Included doxxing, account hacking, revenge porn, and creation of vast amount of disinformation to confuse neutral observers
- Possibly amplified by foreign disinformation actors

Impersonation

Impersonation of a potential romantic partner is known as “catfishing” – but assuming another’s identity online can result in havoc in their professional and personal lives.



Impersonation

DC CLOTHESLINE

AIRING OUT AMERICA'S DIRTY LAUNDRY

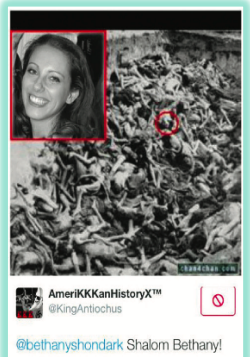
DEPLORABLES NETWORK | NEWS | POLITICS | GUNS/2A | VIDEOS

Woman who set fire to Trump supporter's hair is identified

January 26, 2017 | Dr. Eowyn | Uncategorized | 0

Man in brown jacket tries to extinguish flames from victim's smoking hair

Targeting Journalists to Silence Voices

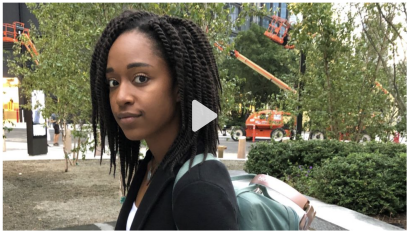


Targeting Journalists to Silence Voices

Targeting Journalists to Silence Voices

He tweeted hate at her. She sued. Then she met him

By Sara Sidner and Mallory Simon, CNN
Updated 6:50 PM EDT, Sat September 21, 2019



UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

TAYLOR DUMPSON,

Plaintiff,

v.

BRIAN ANDREW ADE,

EVAN JAMES MCCARTY,

ANDREW ANGLIN, in his personal capacity and d/b/a DAILY STORMER,

and

MOONBASE HOLDINGS, LLC, d/b/a ANDREW ANGLIN and/or DAILY STORMER,

Defendants.

FIRST AMENDED COMPLAINT

Case No. 1:18-cv-01011 RMC

A Combination of Numerous Tactics: Amanda Todd

- 15 year-old Amanda took her own life in 2012 after years of targeted harassment.
- Case combined cyberstalking, sextortion, doxxing, child sexual abuse, and in-person harassment.
- Abuser was extradicted from Netherlands, finally convicted in 2022.
- Lead to new NCII and cyberstalking laws in Canada, major changes at tech companies.

Targeting the Moderators

POLITICS

Kellyanne Conway unleashes Trumpers on Twitter integrity czar Yoel Roth — over new fact-checking policy and his own tweets

BY RON KAMPEAS MAY 27, 2020 12:49 PM

Responses and Mitigations

Policy Responses

1. Set clear policies that escalate along with escalating abuse
2. Recognize that lots of sub-violating content is abusive in volume
3. Protection of public figures - especially young ones
4. Partnerships and Education

Set Clear Policies for Awful but Lawful Content

Do not threaten, harass, or bully

reddit

7 months ago · Updated

We do not tolerate the harassment, threatening, or bullying of people on our site; nor do we tolerate communities dedicated to this behavior.

Community Standards

Do not:

Repeatedly contact someone in a manner that is:

- Unwanted or
- Sexually harassing or
- Directed at a large number of individuals with no prior solicitation

Abuse/harassment: You may not engage in the targeted harassment of someone, or incite other people to do so. This includes wishing or hoping that someone experiences physical harm. Learn more.

Dogpiling or Accountability?

Justine Sacco — Tweets racist joke to 170 followers, gets on plane, fired by landing

Dogpiling or Accountability?

Emmanuel Cafferty — SDG&E lineman, stranger posts photo to Twitter of his fingers in an “OK” sign, is doxxed and fired

Dogpiling or Accountability?

Amy Cooper — Calls cops on innocent black birdwatcher. Fired, doxxed, lost dog, charged by NY DA, plead out

Dogpiling or Accountability?

Sarah Jeong — Tech journalist, gets job at NYT. Old tweets unearthed, massive abuse thrown her way

Potential Partners

1. Targets
2. Victim advocates
3. Organizations vulnerable to manipulation
4. Law enforcement
5. Lawmakers

Targets

- Harassment is pervasive!
- Targets report being “not listened to” when they go to platforms.
- Many technical solutions place the onus on **targets themselves** to proactively report harassment and bullying.
- Whack-a-mole for victims, especially with cross-platform harassment.

Victim Advocates

HEARTMOB

Welcome to 7 Cups

7 Cups connects you to caring listeners for free emotional support

CENTER FOR ANTI-VIOLENCE EDUCATION

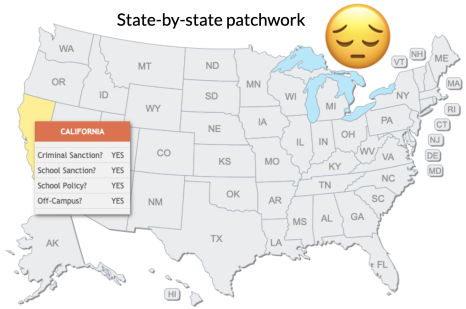
CYBER CIVIL RIGHTS INITIATIVE

THE CYBERSMILE FOUNDATION

Law Enforcement & Bullying Laws Across America

CYBERBULLYING RESEARCH CENTER

- LE often first POC
- Non-sophisticated departments will often tell targets to “turn off their screens” or will not have the sophistication/resources to investigate cyberharassment.



Despite the gravity of their predicaments, cyberharassment victims are often told that nothing can or should be done about online abuse. . . . If victims seek legal help, they are accused of endangering the internet as a forum of public discourse. . . . These views are wrongheaded and counterproductive.

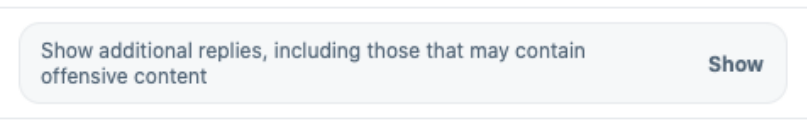
– Danielle Keats Citron, *Hate Crimes in Cyberspace*

Product/Technical Responses

1. Reporting flows
2. Proactive monitoring and measurement
3. Productive friction
4. User-controls
5. Shadow-blocking
6. Third-party tools

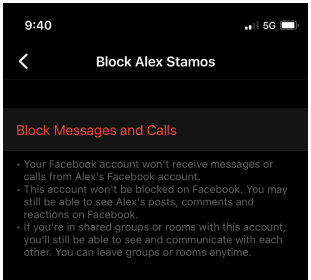
Public Reporting

- Automated flagging and reporting of messages that meet a harmfulness threshold
  - Messages may be immediately removed or flagged for moderator review
- User reports of messages

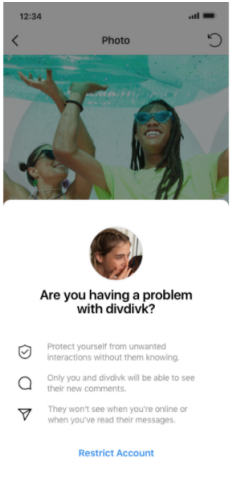


Private Reporting

- In order to report harmful content, users must at least:
  - Have awareness on the how and why of reporting
  - Believe that reports are being processed in a meaningful way that promotes their safety while maintaining privacy
- Reporting can be tied to:
  - Accounts
  - Conversations
  - Messages
- Specificity => effectiveness of operational teams



Quiet Blocking



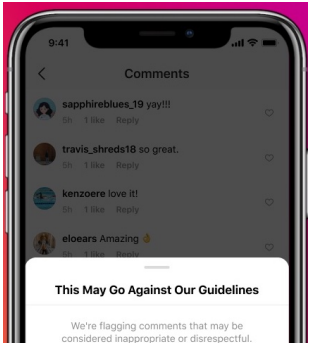
Platforms: Challenges in Proactive Monitoring

- Recognizing sarcasm, joyful playing and culturally specific use of certain words/phrases can be hard for ML
- Important signals:
  - Overall user karma
    - How many people have blocked?
    - How often contacting strangers?
  - Relationship between sender and receiver

Platforms: Productive Friction

How can you build features into platforms to make it “more work” or “less fun” for harassers?

- Instagram: Restrict feature, where users block abusers and the abusers aren’t notified
- Instagram: Rethink feature, asking users to pause before posting material meeting AI patterns for bullying



Third-Party Tools

BLOCK PARTY

How it works About us Blog LOG IN SIGN UP

Take back control of your online experience

Use Block Party to filter out unwanted @mentions from Twitter, and continue to use Twitter as normal.

Welcome, Alex Stamos

Third-Party Tools

DJSmith
@davidjacksmith

Of course @alexstamos BLOCKS ME...

Because he'd like to block from existence like ALEX THANOS anyone who disagrees with his manic totalitarianism.

Gulags next folks. And I'm only half-kidding.

Of course I knew he would. He's an extremist Stalinist see. So I Samizdated him 1st

Bobby "Axe" Axelrod
@RobAxelrod

@alexstamos But... you still are a child molester. Let's not divert our attention away from the fact. 😉

11:34 AM · Jan 17, 2021

Jeff Haley
@Jeffrhal97113834
2 Following • 1 Followers
Bio: Nobody of any consequence whatsoever.
Joined January 2021

View profile on Block Party

Add to watchlist

Account blocked on Twitter. Unblock

Alex Stamos @alexstamos · Jan 20, 2021
I strongly support net neutrality. I don't support AT&T/Verizon giving DANN a monthly rev share per subscriber. There is a difference between a neutral ISP and carrying a channel as part of a financial deal.

Shockingly, the networks I talked about are also lying about this.

hackajar @hackajar
There is a real Liberty problem in the tech community when @alexstamos decided to abandon support for #NetNeutrality live on CNN 🙄

Jeff Haley
@Jeffrhal97113834

Oh, Mr Trustworthy Tech is trying to close down free

Privacy/Response Controls

Alex Stamos
@alexstamos

Not that accuracy matters to Mr. Greenwald, but this is not what I said. On the segment, which he clearly did not watch, I specifically talked about how there are two different issues that need to be addressed: organized violent groups and the lying to the wider mob.

> **Glenn Greenwald** @ggreenwald · 7h
> Here's a former Facebook executive on CNN *explicitly urging that the same tactics that were used in cooperation with the US security state to remove ISIS from the internet now be used against "conservative influencers," to deprogram everyone to believe the CNN/liberal consensus:
>
> > **Daily Caller** @DailyCaller · 7h
> > Former Facebook insider Alex Stamos tells @brianstelter: "We have to turn down the capability of these Conservative influencers to reach these huge audiences... There are people on YouTube for example that have a larger audience than daytime CNN."
>
> 
>
>
> 497 3.9K 8.5K

9:27 PM · Jan 17, 2021 · Twitter Web App

72 Retweets **17** Quote Tweets **318** Likes