IFRRO Athens World Congress

Connecting the Dots: The Future of Collective Management, 23rd of October, 2018

# Monitoring and Tagging Hate Speech in Social Media

Sophia Karakeva

Communications and Marketing Executive | Datascouting

Vice President | FIBEP

# Καλησπέρα

DATASCOUTING
actionable information

FIBEP

## Some interesting facts about our very “hateful” social world

- Tech platforms now **control** the majority of online conversation
- Tech platforms have undertaken a shift towards **censorship** and **moderation**
- The American tradition of **free speech** on the internet is **no longer viable**
- Users, governments, and tech firms are all **behaving badly**
- **Free speech** has become a social, economic and political **weapon**
- **Racists, misogynists and oppressors** are allowed a voice

*Source: The Good Censor | Cultural context report, March 2018*

## What is hate speech?

- No universal definition
- Freedom of expression vs hate speech
- Relying on hate speech policies

> “ *#HateSpeech is just a word for pussies who can't stand up for themselves so they need legislation on language*

> “ *#HateSpeech is free speech. Otherwise you're not allowed to hate nazis.*

> “ *Truth is now considered #HateSpeech. Thanks progressives.*

## What is the role of social media when it comes to hate speech?

- Code of conduct
- Still not there
- Cooperation but diversity

# Monitoring and Tagging Hate Speech in Social Media

## What is the role of social media when it comes to hate speech?

## Still not there

> “ *Sometimes our systems get it wrong … Our system sometimes make mistakes in understanding context and nuances…*
>
> **Johanna Wright**
> *YouTube VP of Product Management (2017)*

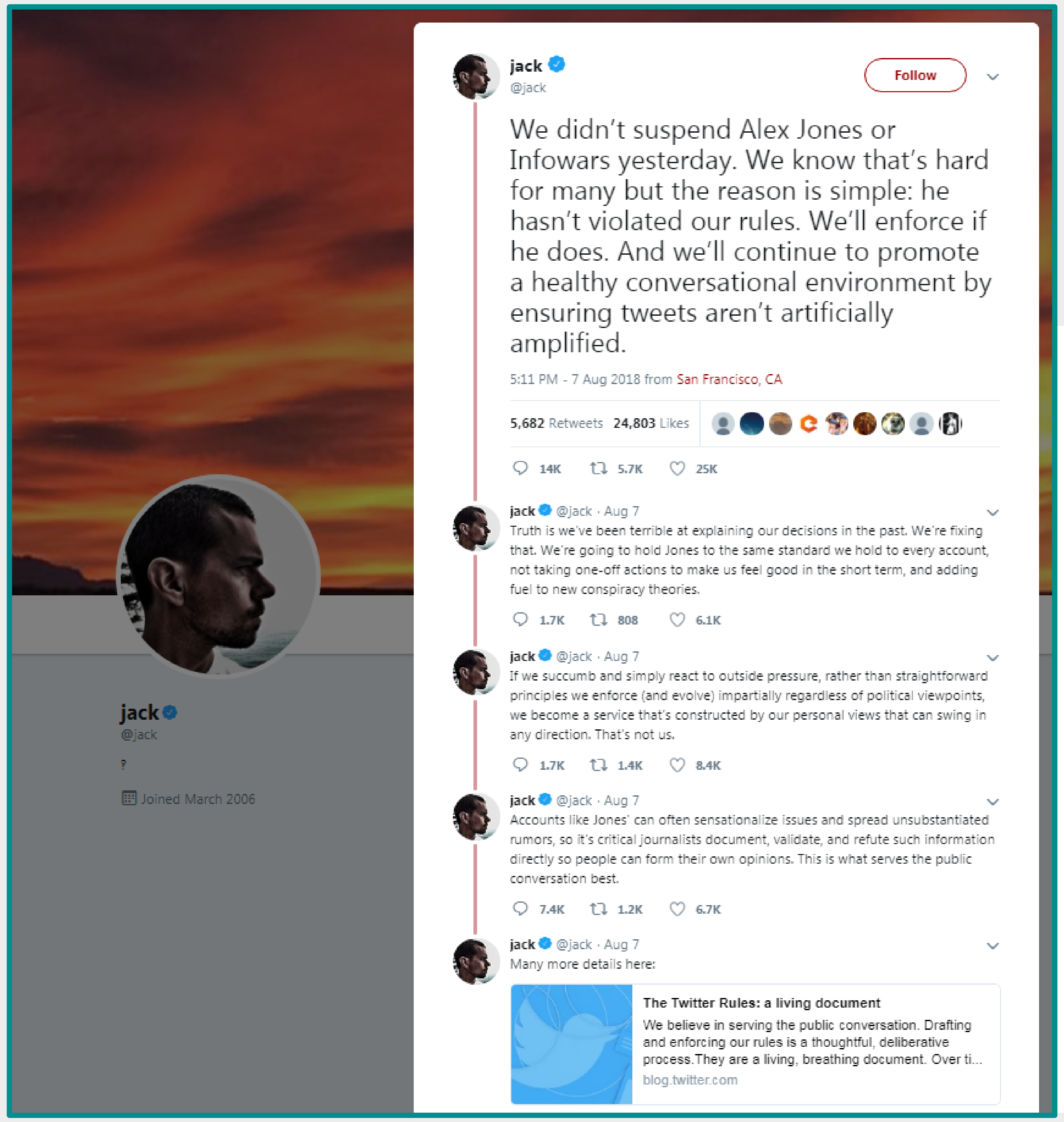
> “ *We see voices being silenced on Twitter every day. We’ve been working to counteract this for the past 2 years… We prioritized this in 2016. We updated our polices and increased the size of our teams. It wasn’t enough.*
>
> **Jack Dorsey**
> *Twitter CEO (2017)*

> “ *… It’s much easier to build an AI system that can detect a nipple that it is to determine what is linguistically hate speech.*
>
> **Mark Zuckerberg**
> *Facebook CEO (2018)*

# Monitoring and Tagging Hate Speech in Social Media

## What is the role of social media when it comes to hate speech?

### Collaboration



jack
@jack
Joined March 2006

jack @jack
Follow

We didn't suspend Alex Jones or Infowars yesterday. We know that's hard for many but the reason is simple: he hasn't violated our rules. We'll enforce if he does. And we'll continue to promote a healthy conversational environment by ensuring tweets aren't artificially amplified.

5:11 PM - 7 Aug 2018 from San Francisco, CA

5,682 Retweets 24,803 Likes

14K 5.7K 25K

jack @jack · Aug 7
Truth is we've been terrible at explaining our decisions in the past. We're fixing that. We're going to hold Jones to the same standard we hold to every account, not taking one-off actions to make us feel good in the short term, and adding fuel to new conspiracy theories.

1.7K 808 6.1K

jack @jack · Aug 7
If we succumb and simply react to outside pressure, rather than straightforward principles we enforce (and evolve) impartially regardless of political viewpoints, we become a service that's constructed by our personal views that can swing in any direction. That's not us.

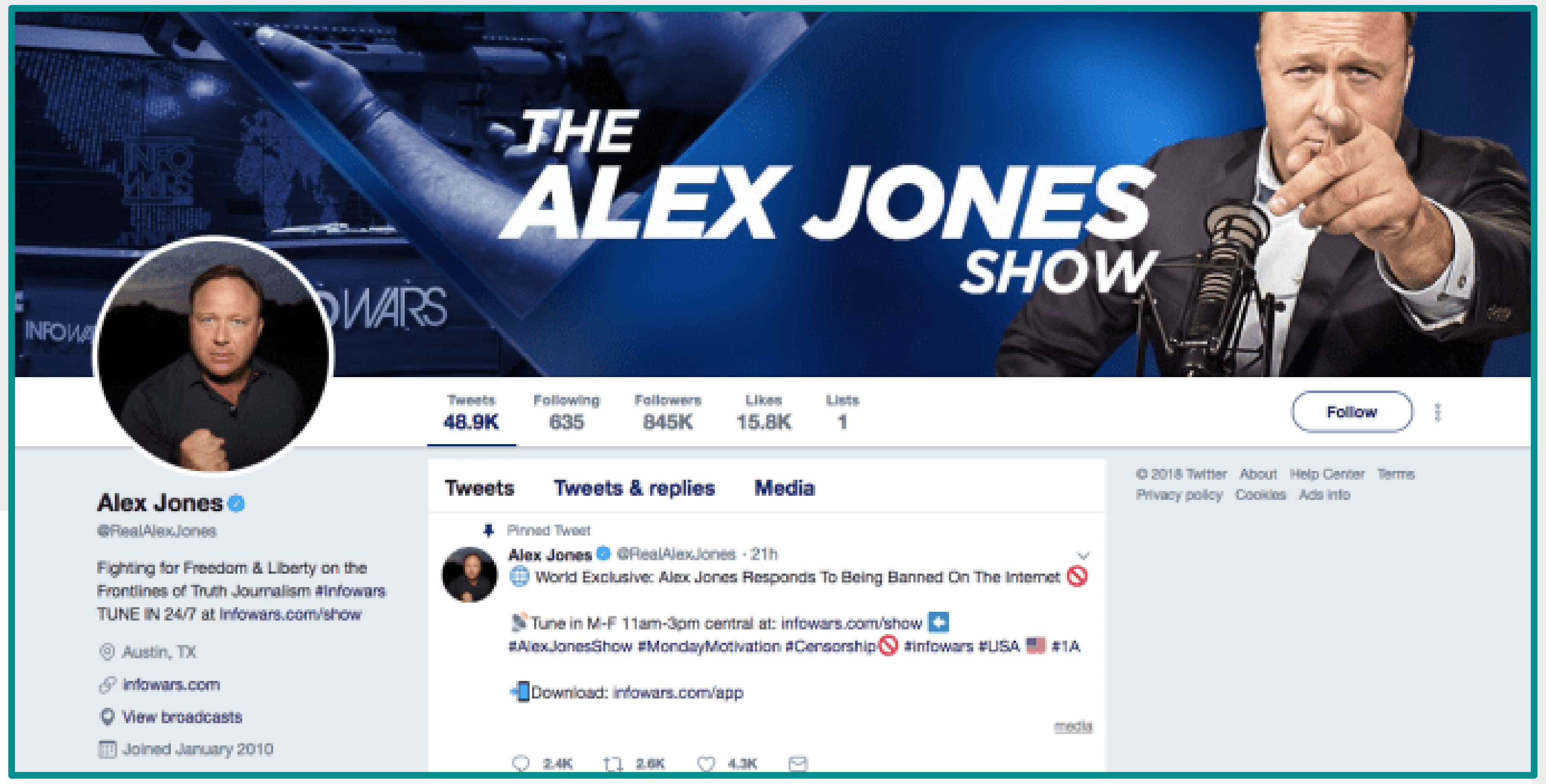
1.7K 1.4K 8.4K

jack @jack · Aug 7
Accounts like Jones' can often sensationalize issues and spread unsubstantiated rumors, so it's critical journalists document, validate, and refute such information directly so people can form their own opinions. This is what serves the public conversation best.

7.4K 1.2K 6.7K

jack @jack · Aug 7
Many more details here:

The Twitter Rules: a living document
We believe in serving the public conversation. Drafting and enforcing our rules is a thoughtful, deliberative process.They are a living, breathing document. Over ti...
blog.twitter.com



| Tweets | Following | Followers | Likes | Lists |
|---|---|---|---|---|
| 48.9K | 635 | 845K | 15.8K | 1 |

Follow

Alex Jones
@RealAlexJones

Fighting for Freedom & Liberty on the Frontlines of Truth Journalism #Infowars TUNE IN 24/7 at Infowars.com/show

Austin, TX
infowars.com
View broadcasts
Joined January 2010

Tweets Tweets & replies Media

Pinned Tweet
Alex Jones @RealAlexJones · 21h
🌐 World Exclusive: Alex Jones Responds To Being Banned On The Internet 🚫

Tune in M-F 11am-3pm central at: infowars.com/show
#AlexJonesShow #MondayMotivation #Censorship🚫 #infowars #USA 🇺🇸 #1A

Download: infowars.com/app

media

2.4K 2.6K 4.3K

© 2018 Twitter About Help Center Terms
Privacy policy Cookies Ads info

> Alex Jones against CNN Senior Media Correspondent Brian Stelter: ***"enemy . . . drunk on children's blood," "You will pay! You will fall!"***
>
> Alex Jones against people of Muslim faith: ***"And we are going to stop you again! Do you understand?!""***
>
> Alex Jones against drag queens: ***"We're going to destroy you."***

*Source: CNN*

## The legal and discursive characteristics of hate speech

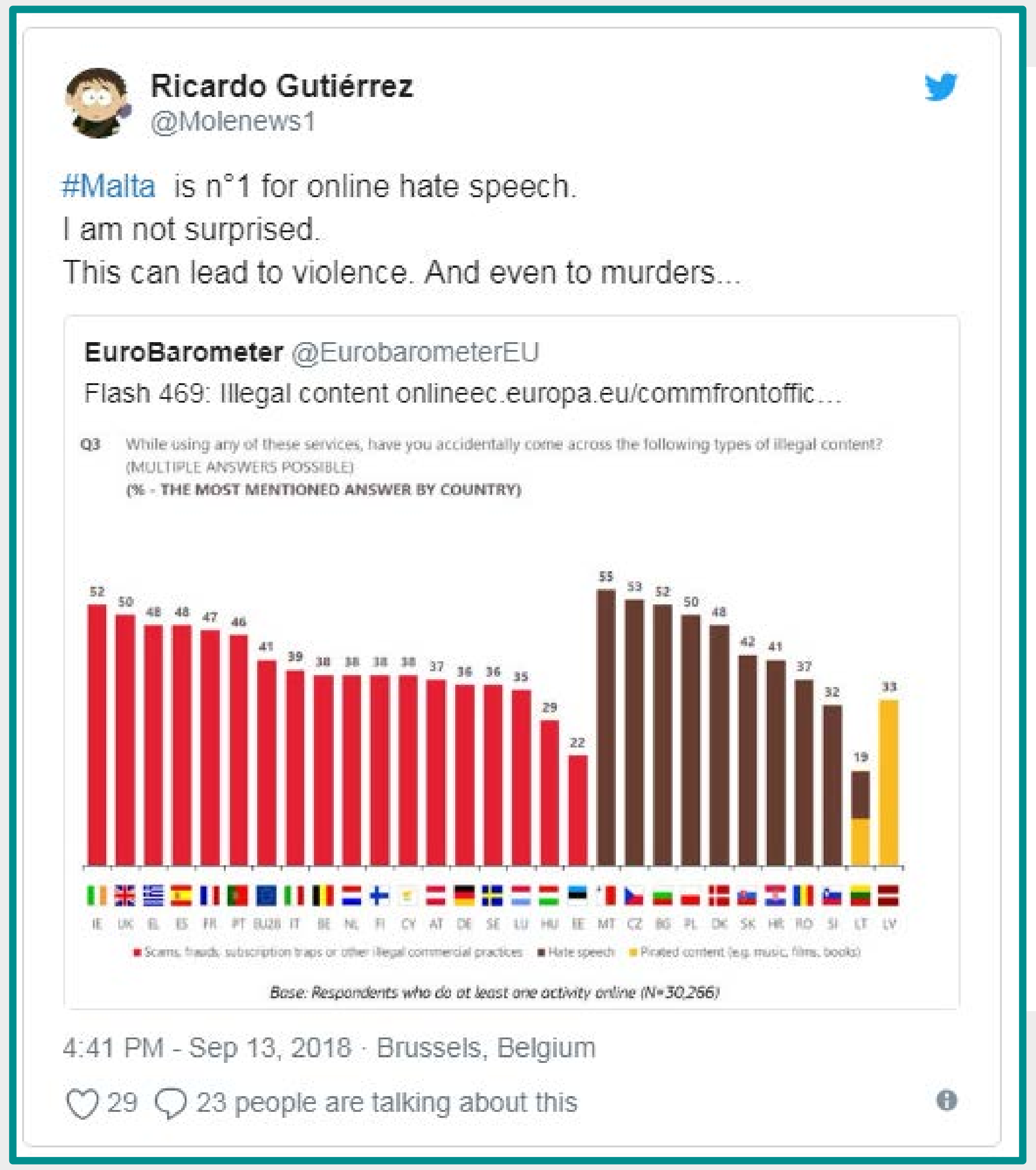
Create unmediated marketplaces of ideas **VS** Create well-ordered spaces for safety and civility

- Legal liability
- Danger of over-regulation
- Platforms vs editors / publishers

> ***Platforms have to deny that they're media companies in order to retain their immunity from liability. But at the same time, they're exercising more influence as media companies…***
>
> **Jeffrey Rosen**
> *Professor of Law at the George Washington University and legal affairs editor of the The New Republic CEO (2018)*

# Monitoring and Tagging Hate Speech in Social Media

## Where does hate speech spread the most?



Ricardo Gutiérrez
@Molenews1

#Malta is n°1 for online hate speech.
I am not surprised.
This can lead to violence. And even to murders...

EuroBarometer @EurobarometerEU
Flash 469: Illegal content onlineec.europa.eu/commfrontoffic...

Q3 While using any of these services, have you accidentally come across the following types of illegal content?
(MULTIPLE ANSWERS POSSIBLE)
**(% - THE MOST MENTIONED ANSWER BY COUNTRY)**

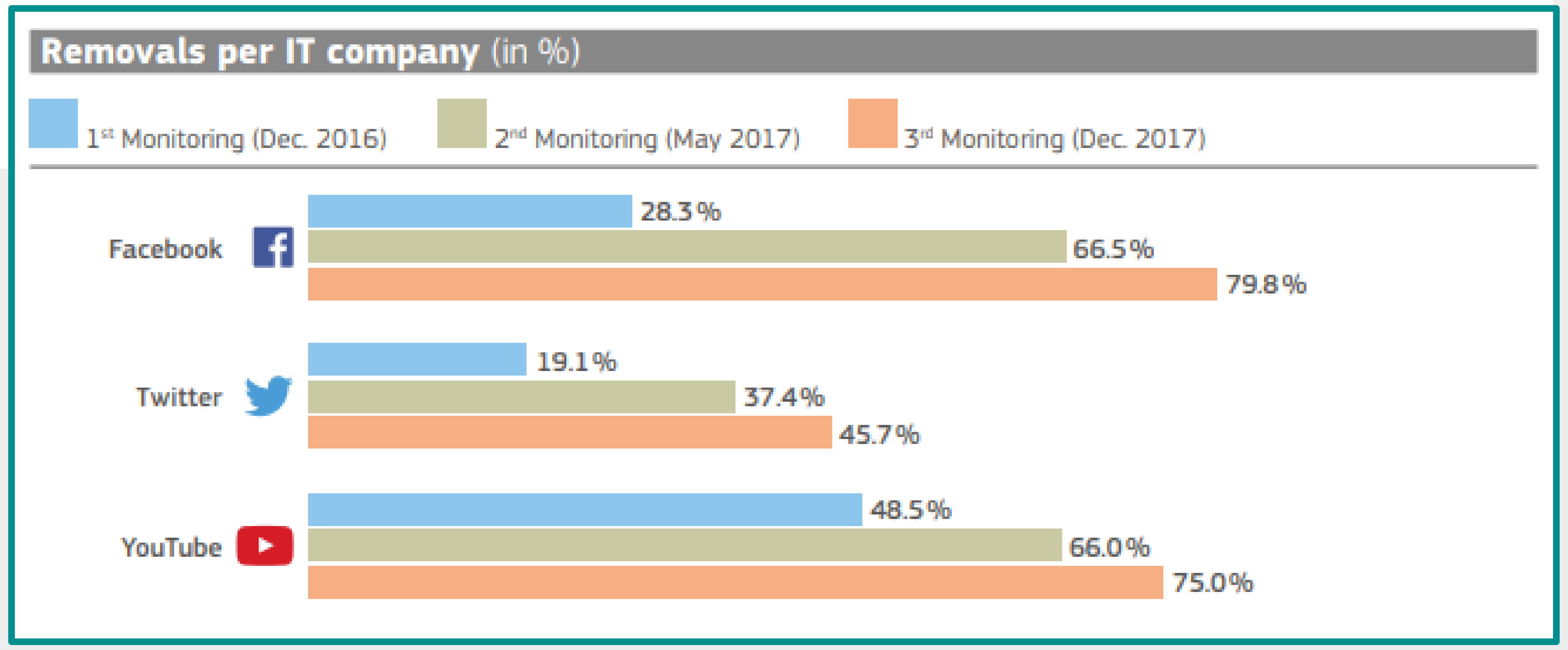
*Base: Respondents who do at least one activity online (N=30,266)*

4:41 PM - Sep 13, 2018 · Brussels, Belgium

♡ 29 💬 23 people are talking about this

- Social networks the most common online daily activity
- Malta has the highest online hate speech in the EU
- Facebook has the largest amount of notifications
- 70% of hate speech content was removed in 2017

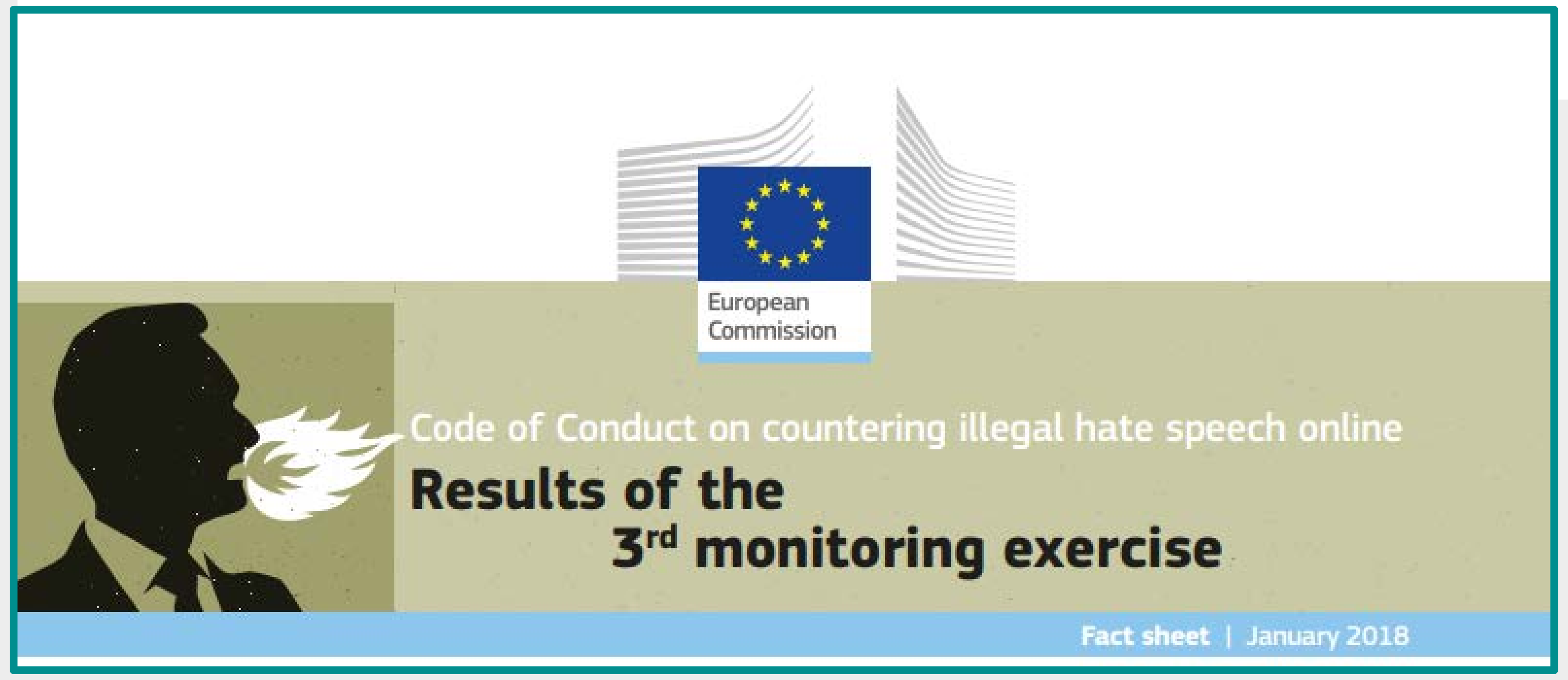
*Source:*
*The Flash Eurobarometer, September 2018*
*The 3rd monitoring of EU's Code of Conduct, January 2018*
*2017 Pew Research Center survey about online harassment*



# Monitoring and Tagging Hate Speech in Social Media

## Who is most targeted?

- Ethnic origin, anti-Muslim hatred, & xenophobia
- Gender matters to social media
- The developing world

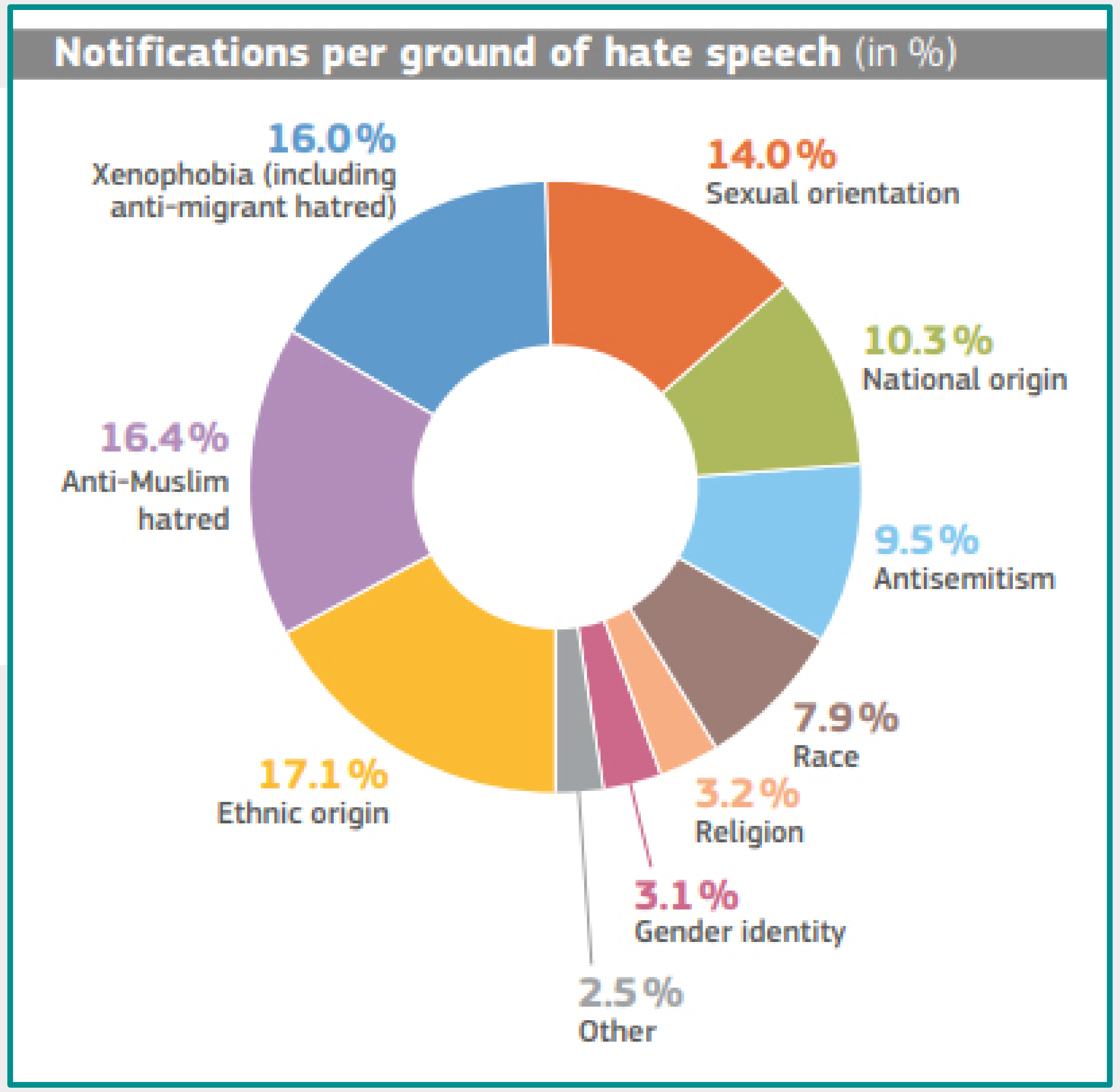


European Commission

Code of Conduct on countering illegal hate speech online

**Results of the 3rd monitoring exercise**

Fact sheet | January 2018



**Notifications per ground of hate speech** (in %)

# Monitoring and Tagging Hate Speech in Social Media

## Who is most targeted?

## Gender matters to social media

> “ I think Twitter is the worst of the social media platforms, just because of the quickened and masked flow [of abuse] that happens. The content feels pretty similar across the platforms but the sheer volume of it on Twitter is what’s different. ”

Jessica Valenti, US journalist and writer

> “It’s like public lynching. It has made me frightened for my physical safety when I am out in the streets”

**Amberin Zaman,** Turkish journalist

# Monitoring and Tagging Hate Speech in Social Media

## Who is most targeted?

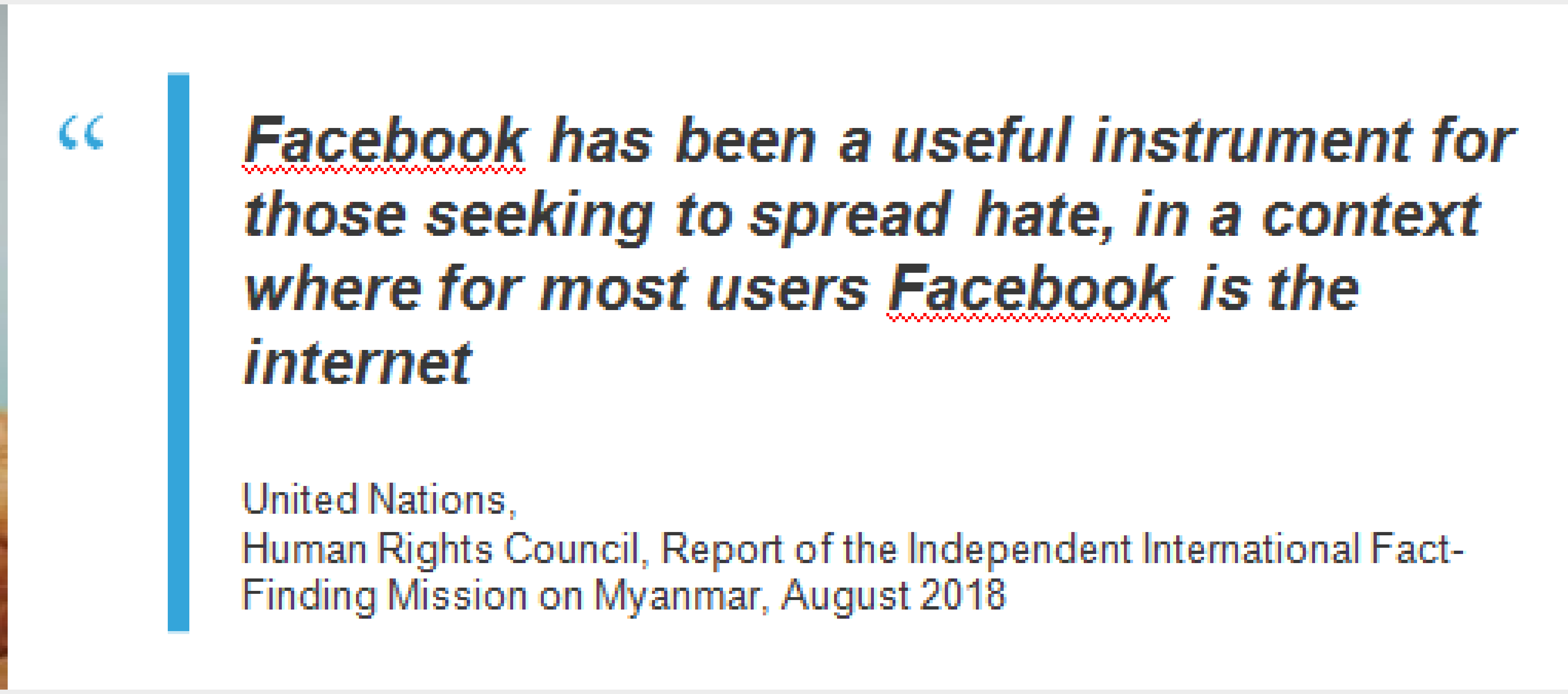
## Digital Colonization

BBC

The country where Facebook posts whipped up hate

The New York Times

Where Countries Are Tinderboxes and Facebook Is a Match



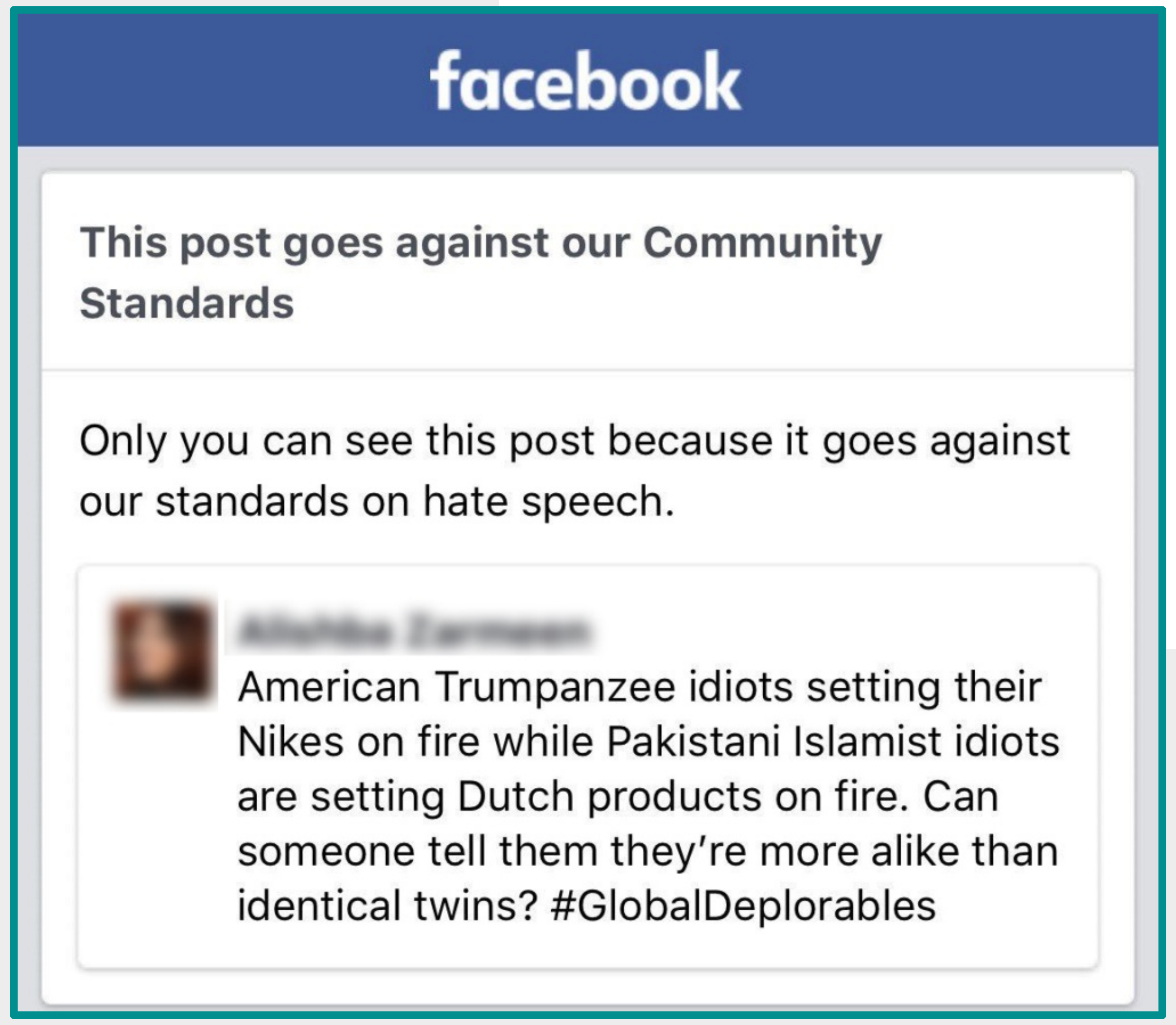
> “ ***Facebook has been a useful instrument for those seeking to spread hate, in a context where for most users Facebook is the internet***
>
> United Nations,
> Human Rights Council, Report of the Independent International Fact-Finding Mission on Myanmar, August 2018

*Source: John Oliver, Last Week Tonight Show, 23 September 2018*

## How are tech companies mismanaging hate speech?

Hate me, Hate me not

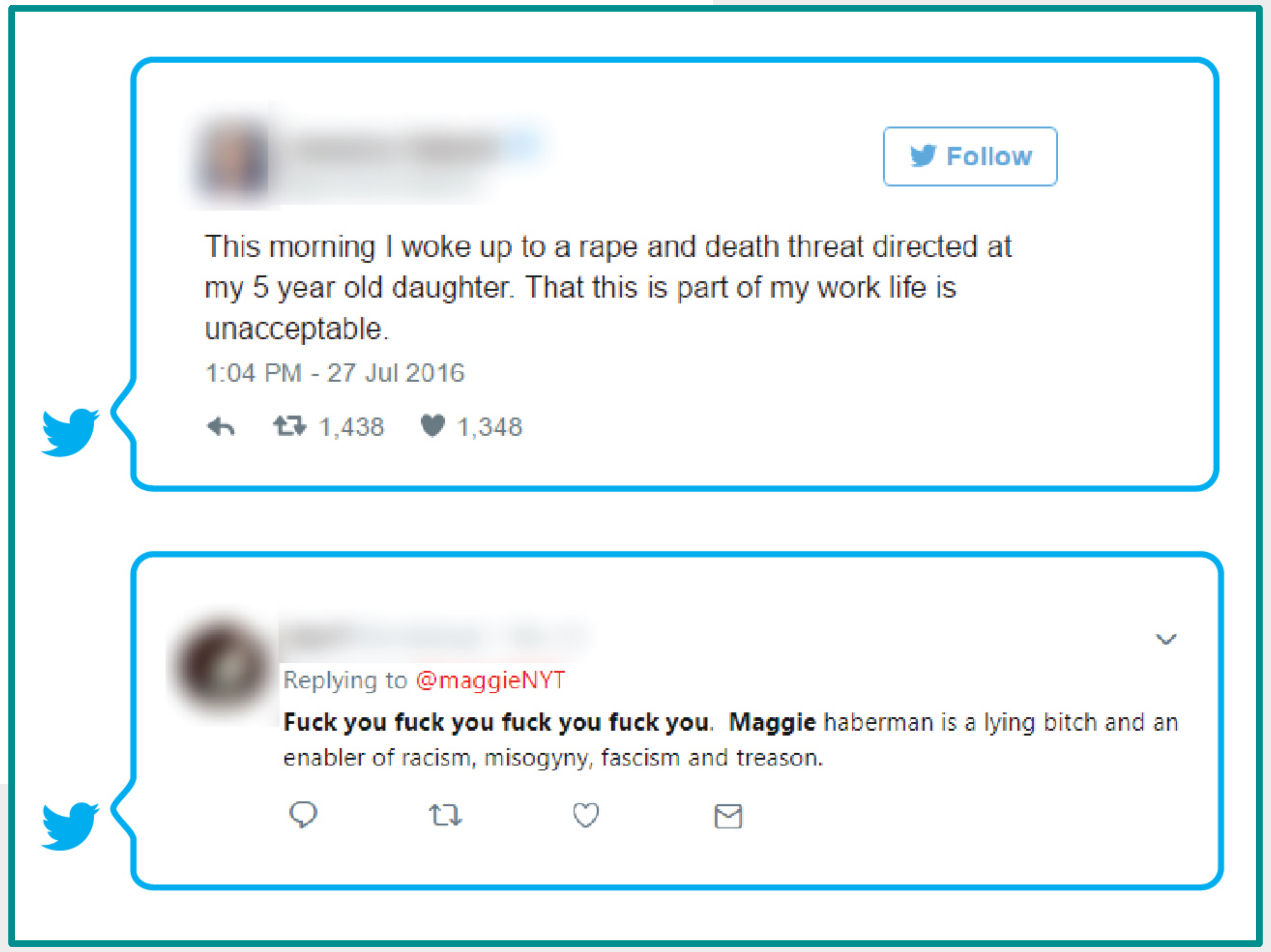
- Commercialized conversation
- Inconsistent interventions
- Lack of transparency
- Ineffective automation
- Underplaying the issue



**This post goes against our Community Standards**

Only you can see this post because it goes against our standards on hate speech.

[redacted]
American Trumpanzee idiots setting their Nikes on fire while Pakistani Islamist idiots are setting Dutch products on fire. Can someone tell them they're more alike than identical twins? #GlobalDeplorables



[redacted]

This morning I woke up to a rape and death threat directed at my 5 year old daughter. That this is part of my work life is unacceptable.

1:04 PM - 27 Jul 2016

1,438 1,348

[redacted]

Replying to @maggieNYT

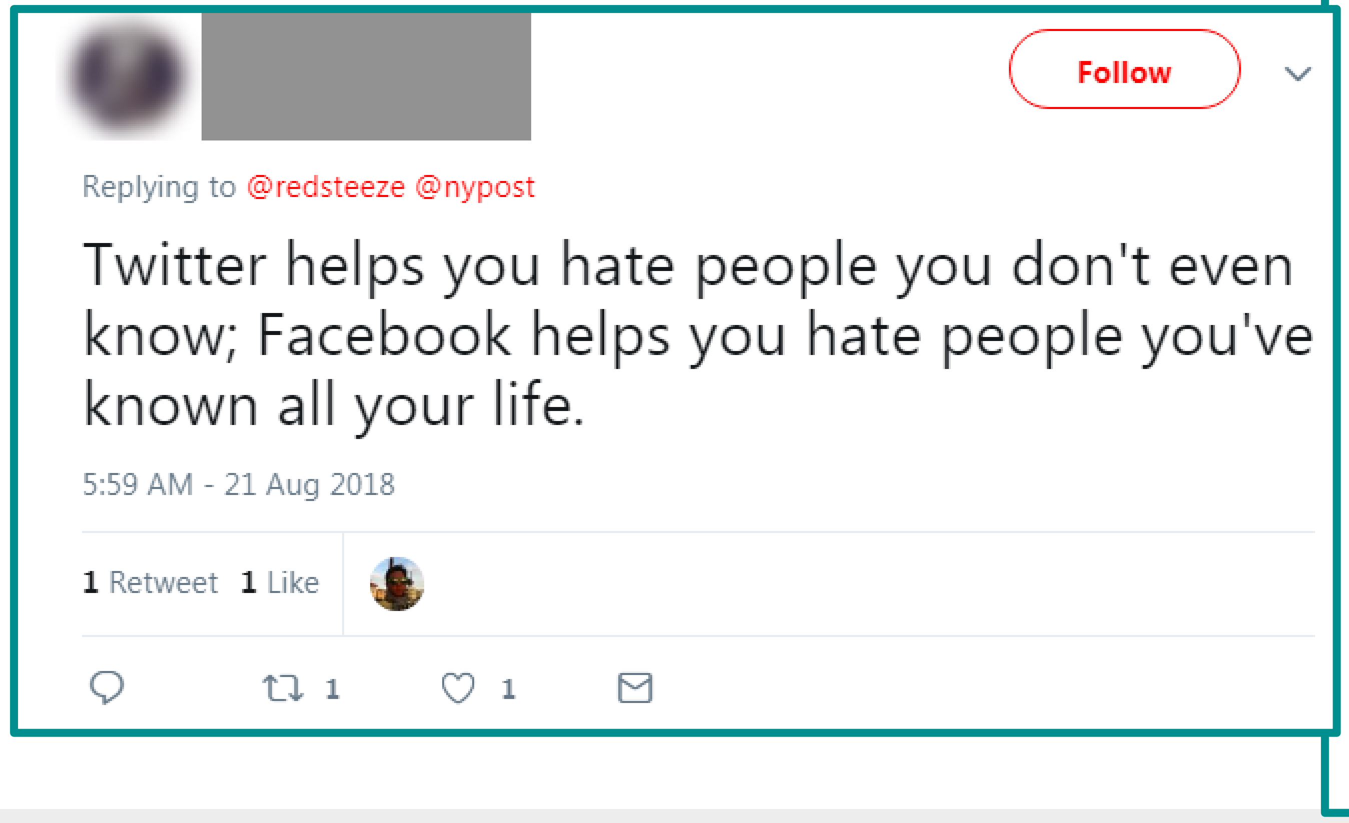
**Fuck you fuck you fuck you fuck you.** **Maggie** haberman is a lying bitch and an enabler of racism, misogyny, fascism and treason.

# Monitoring and Tagging Hate Speech in Social Media

## Hate me, hate me not

## Commercialized conversation

- **Hate me not**: hate speech is not allowed on our platforms
- **Hate me**: hate speech is often the most engaged content



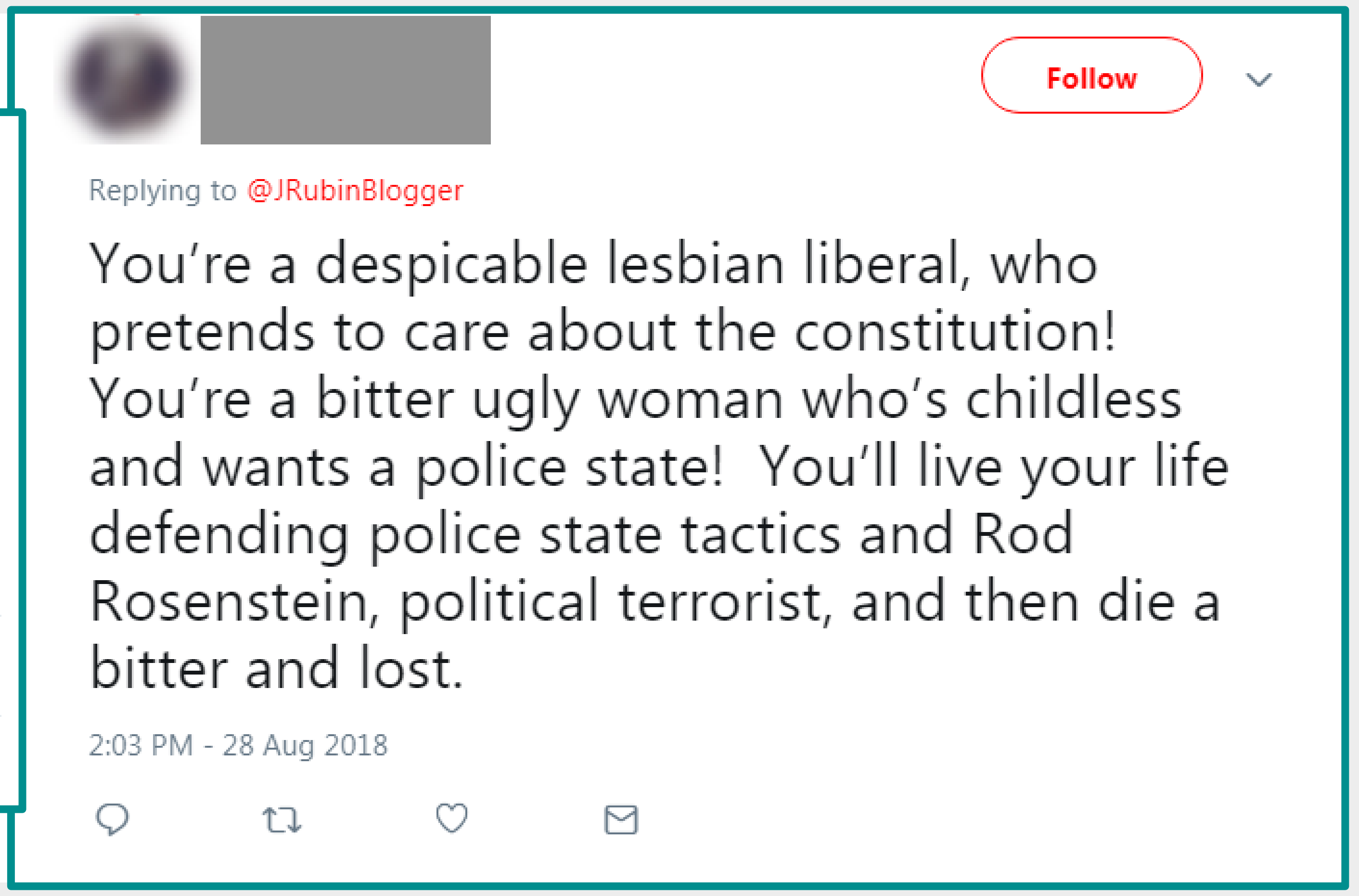
[redacted]

Replying to @redsteeze @nypost

Twitter helps you hate people you don't even know; Facebook helps you hate people you've known all your life.

5:59 AM - 21 Aug 2018

1 Retweet 1 Like



[redacted]

Replying to @JRubinBlogger

You're a despicable lesbian liberal, who pretends to care about the constitution! You're a bitter ugly woman who's childless and wants a police state! You'll live your life defending police state tactics and Rod Rosenstein, political terrorist, and then die a bitter and lost.

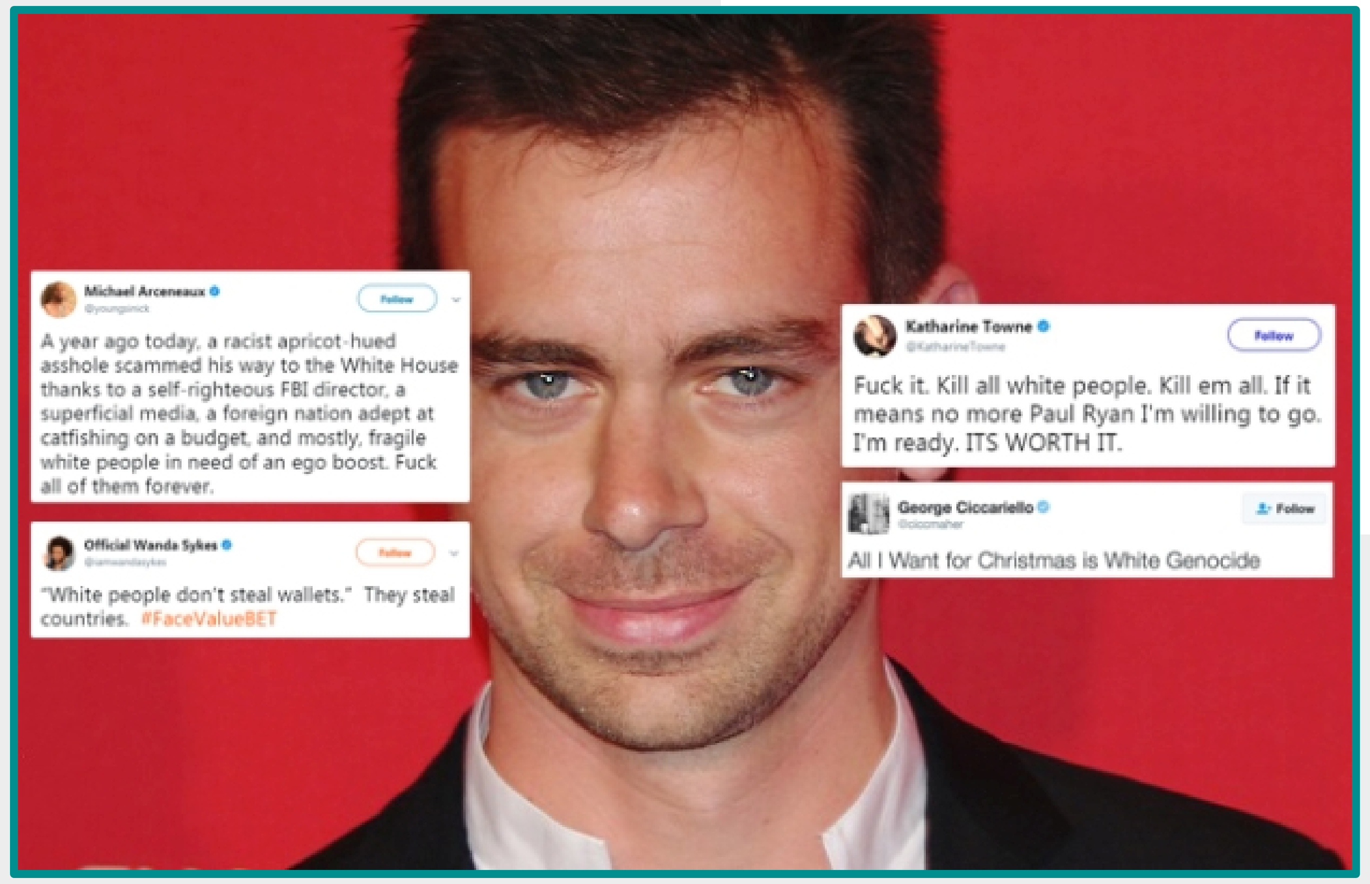
2:03 PM - 28 Aug 2018

# Monitoring and Tagging Hate Speech in Social Media

## Hate me, hate me not

## Inconsistent interventions & lack of transparency

- **Hate me not**: we take strong measures to take hate speech down
- **Hate me**: not everything has the same value

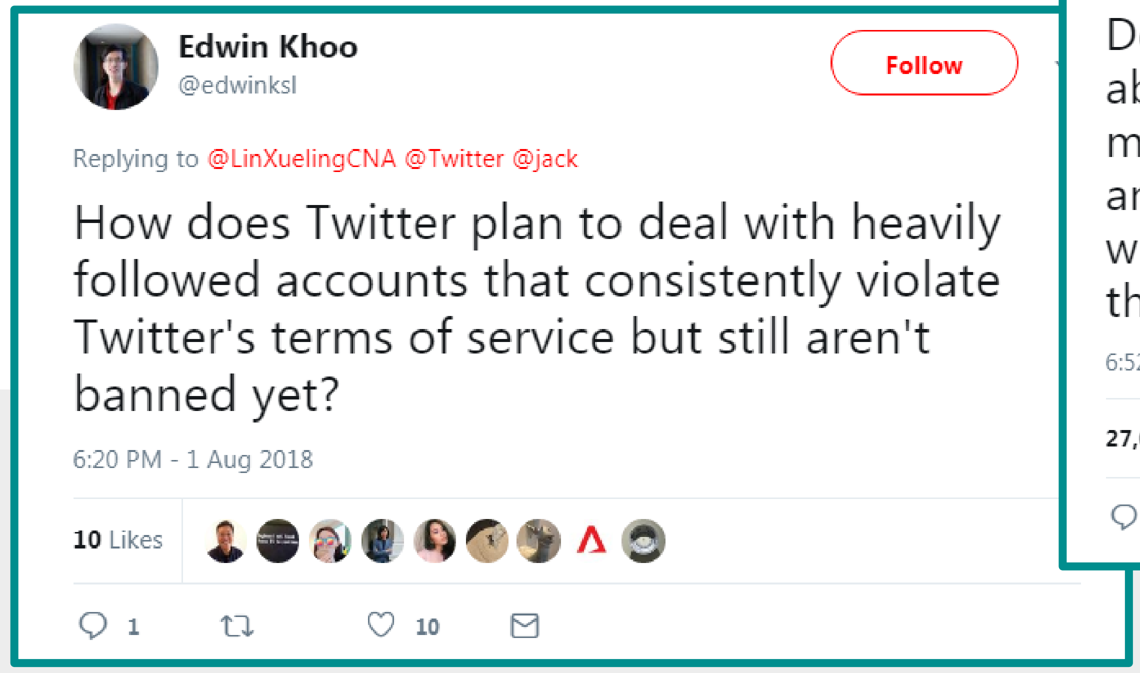


Michael Arceneaux
Follow
A year ago today, a racist apricot-hued asshole scammed his way to the White House thanks to a self-righteous FBI director, a superficial media, a foreign nation adept at catfishing on a budget, and mostly, fragile white people in need of an ego boost. Fuck all of them forever.

Official Wanda Sykes
Follow
"White people don't steal wallets." They steal countries. #FaceValueBET

Katharine Towne
@KatharineTowne
Follow
Fuck it. Kill all white people. Kill em all. If it means no more Paul Ryan I'm willing to go. I'm ready. ITS WORTH IT.

George Ciccariello
Follow
All I Want for Christmas is White Genocide



Edwin Khoo
@edwinksl
Follow

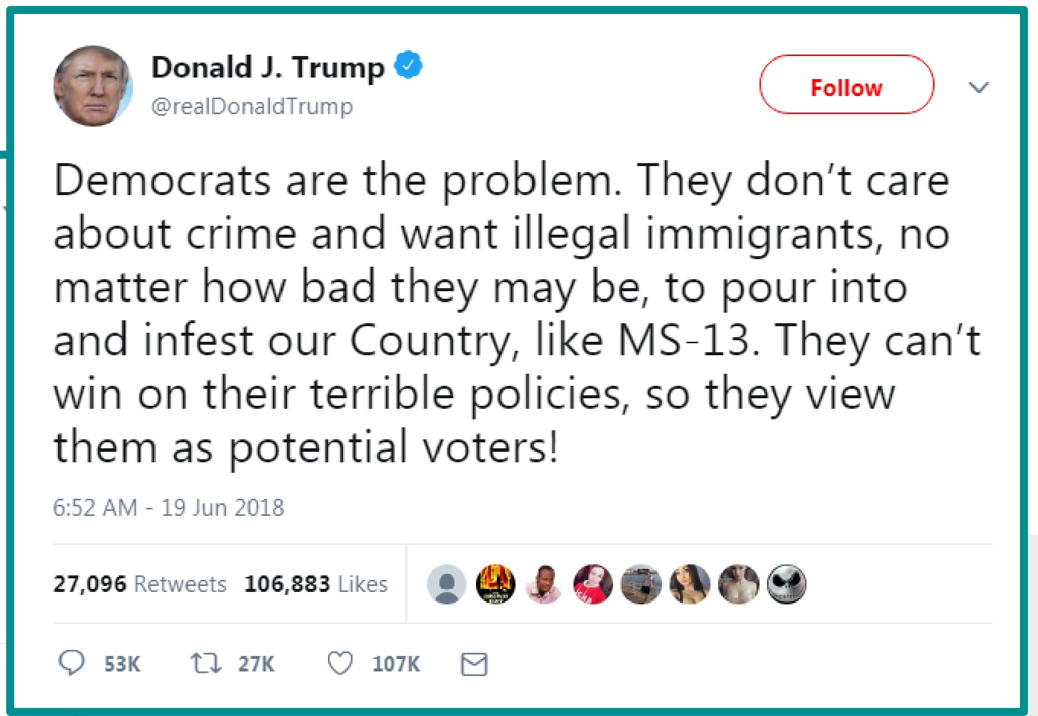
Replying to @LinXuelingCNA @Twitter @jack

How does Twitter plan to deal with heavily followed accounts that consistently violate Twitter's terms of service but still aren't banned yet?

6:20 PM - 1 Aug 2018

10 Likes

1 10



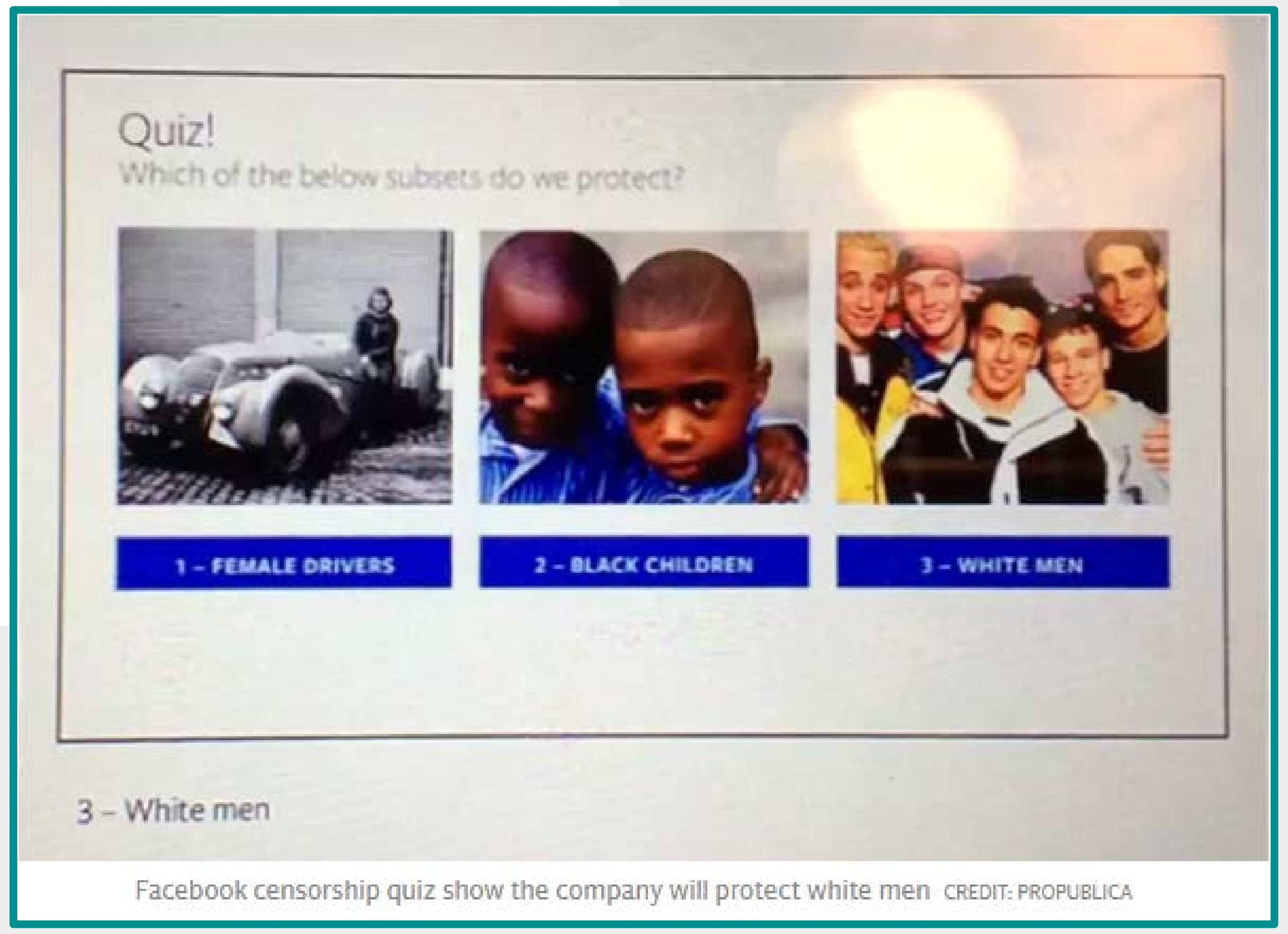
Donald J. Trump
@realDonaldTrump
Follow

Democrats are the problem. They don't care about crime and want illegal immigrants, no matter how bad they may be, to pour into and infest our Country, like MS-13. They can't win on their terrible policies, so they view them as potential voters!

6:52 AM - 19 Jun 2018

27,096 Retweets 106,883 Likes

53K 27K 107K



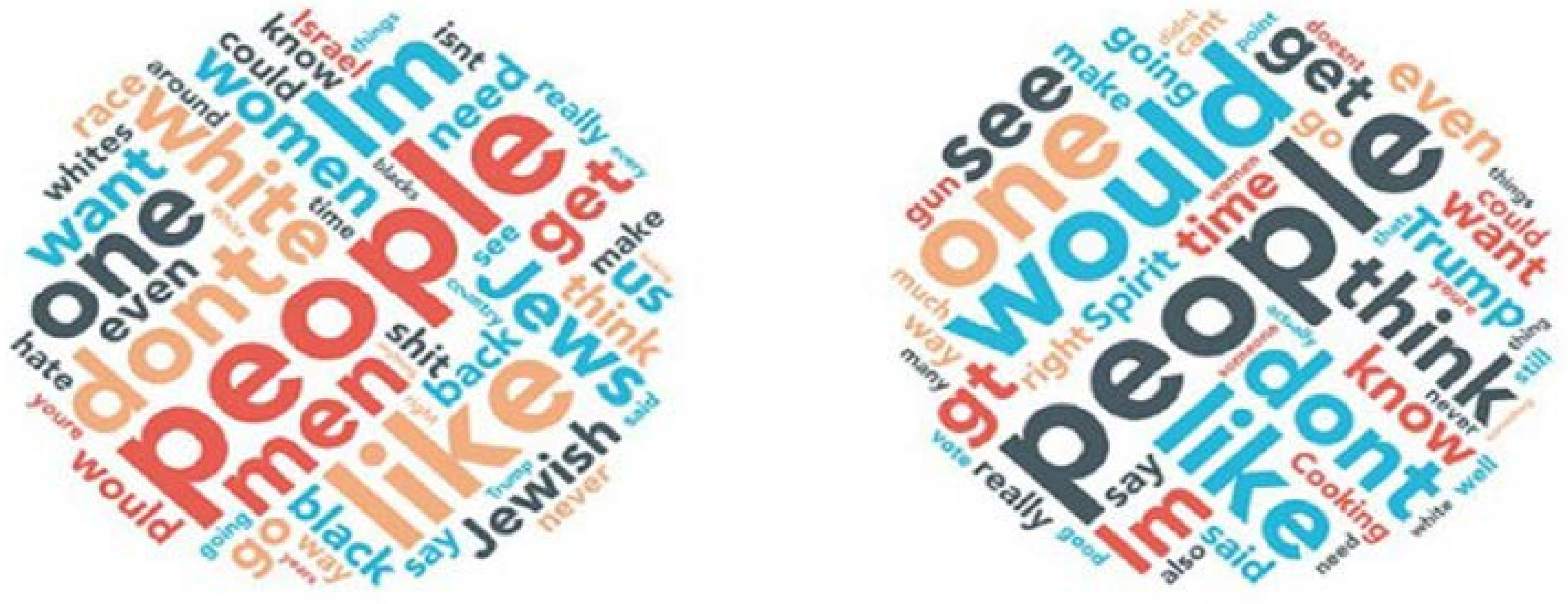
Facebook censorship quiz show the company will protect white men CREDIT: PROPUBLICA

# Monitoring and Tagging Hate Speech in Social Media

## Hate me, hate me not

### **Ineffective automation**

- **Hate me**: hate speech is highly reactive with high error rate
- **Hate me not**: automation and AI tools



Most common words found in Hate Comments

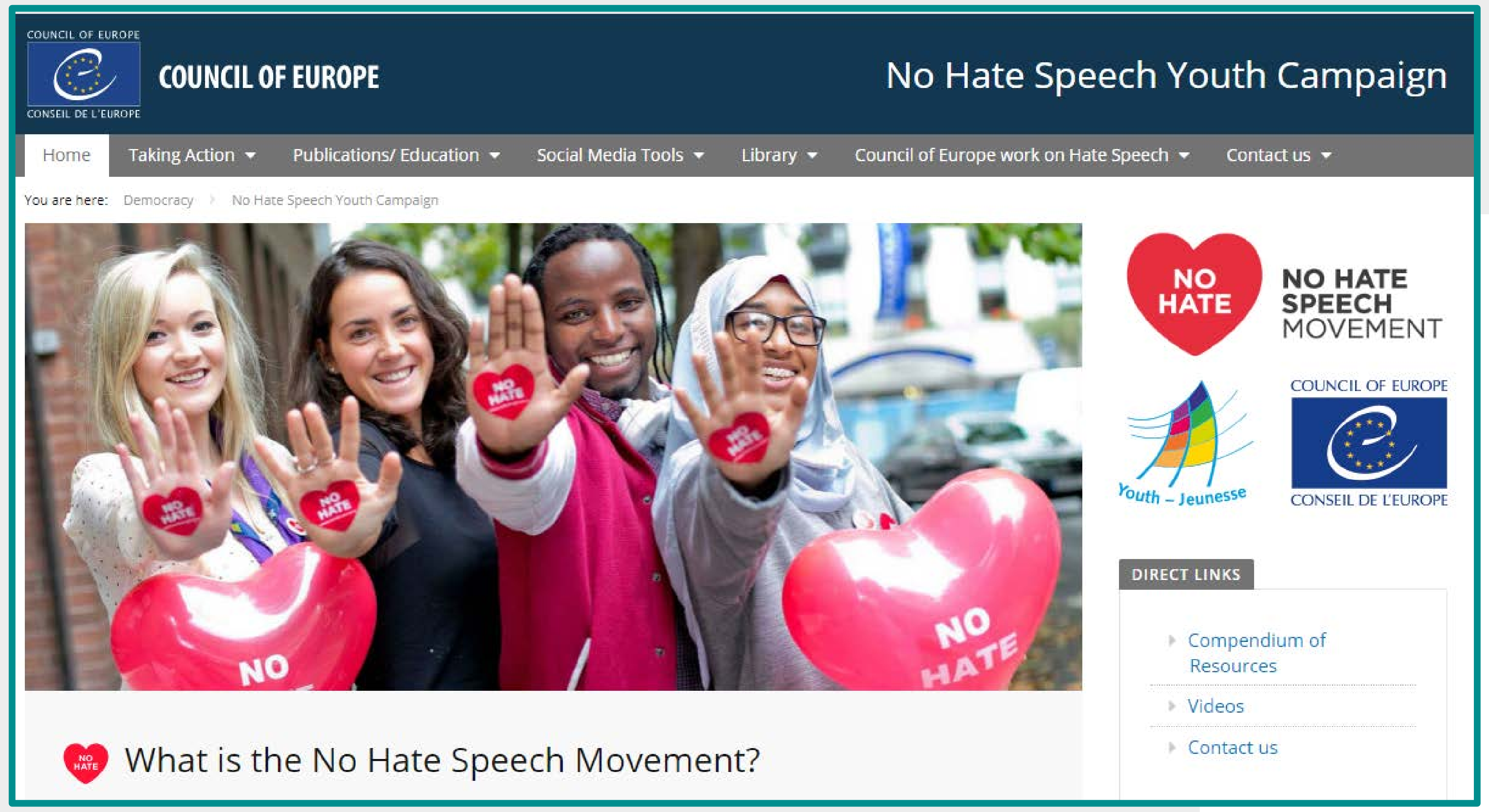
Most common words found in Non-hate Comments

This data representation shows common terms used in hateful comments on Reddit, Twitter, and other social media sites compared to non-hateful comments. (courtesy of the Anti-Defamation League)

# Monitoring and Tagging Hate Speech in Social Media

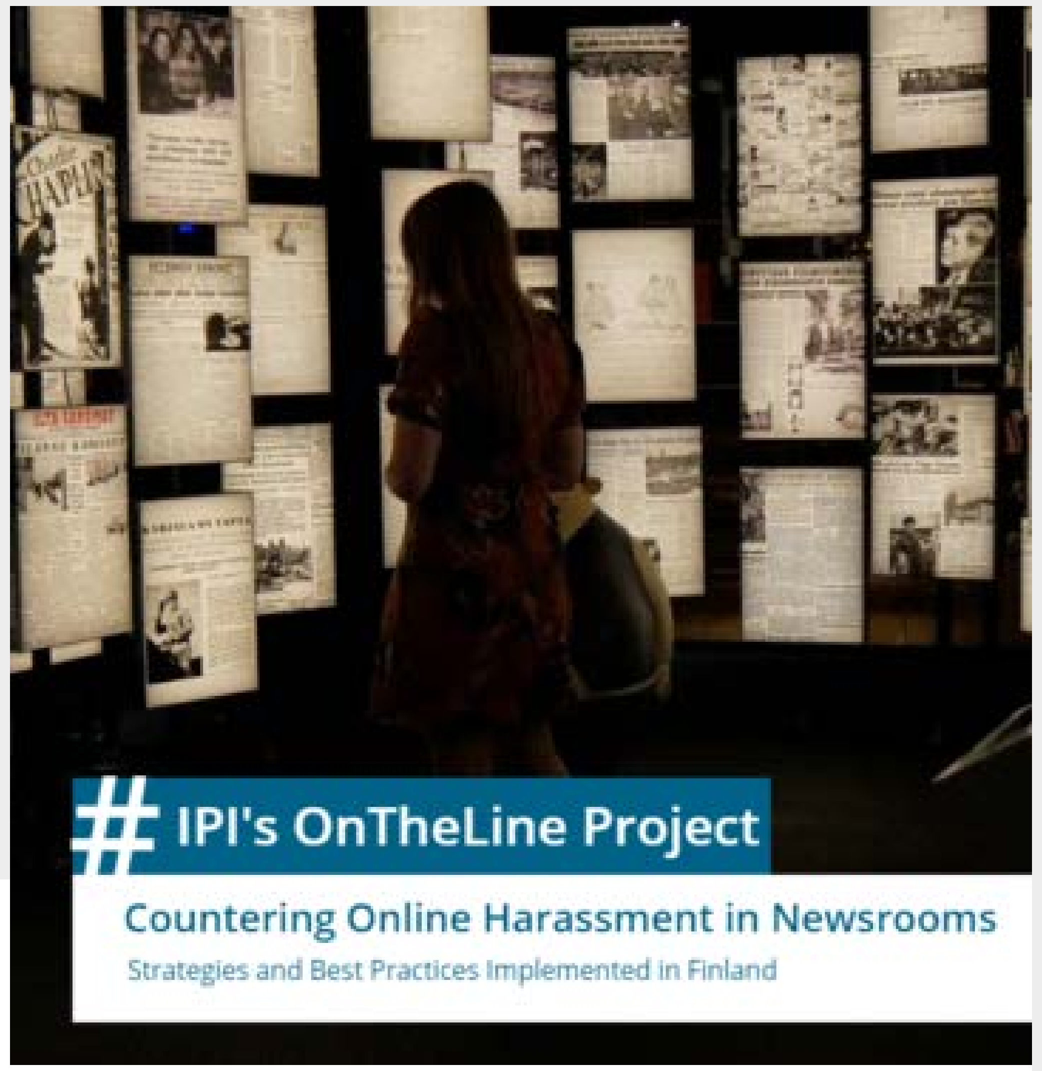
## Hate me, hate me not

## Underplaying hate speech


COUNCIL OF EUROPE

No Hate Speech Youth Campaign

What is the No Hate Speech Movement?


# IPI's OnTheLine Project

Countering Online Harassment in Newsrooms

Strategies and Best Practices Implemented in Finland

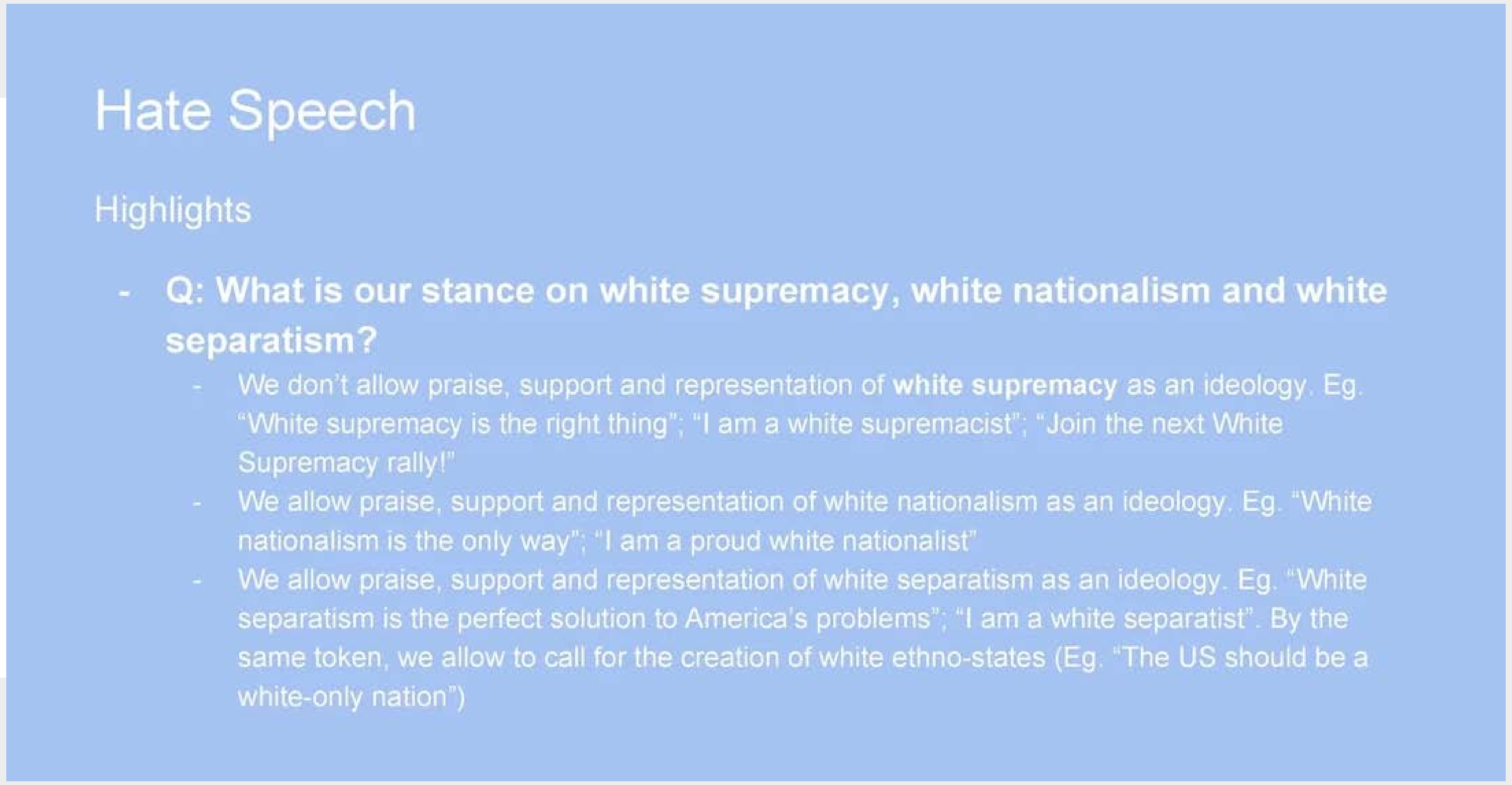
- **Hate me**: hate speech becomes the "new normal"
- **Hate me not**: education, awareness campaigns, media ethics, regulation

SELMA

HACKING HATE

## Typology and coding

- Scale vs human annotation
- Typology and coding
- Main challenges



### Hate Speech

Highlights

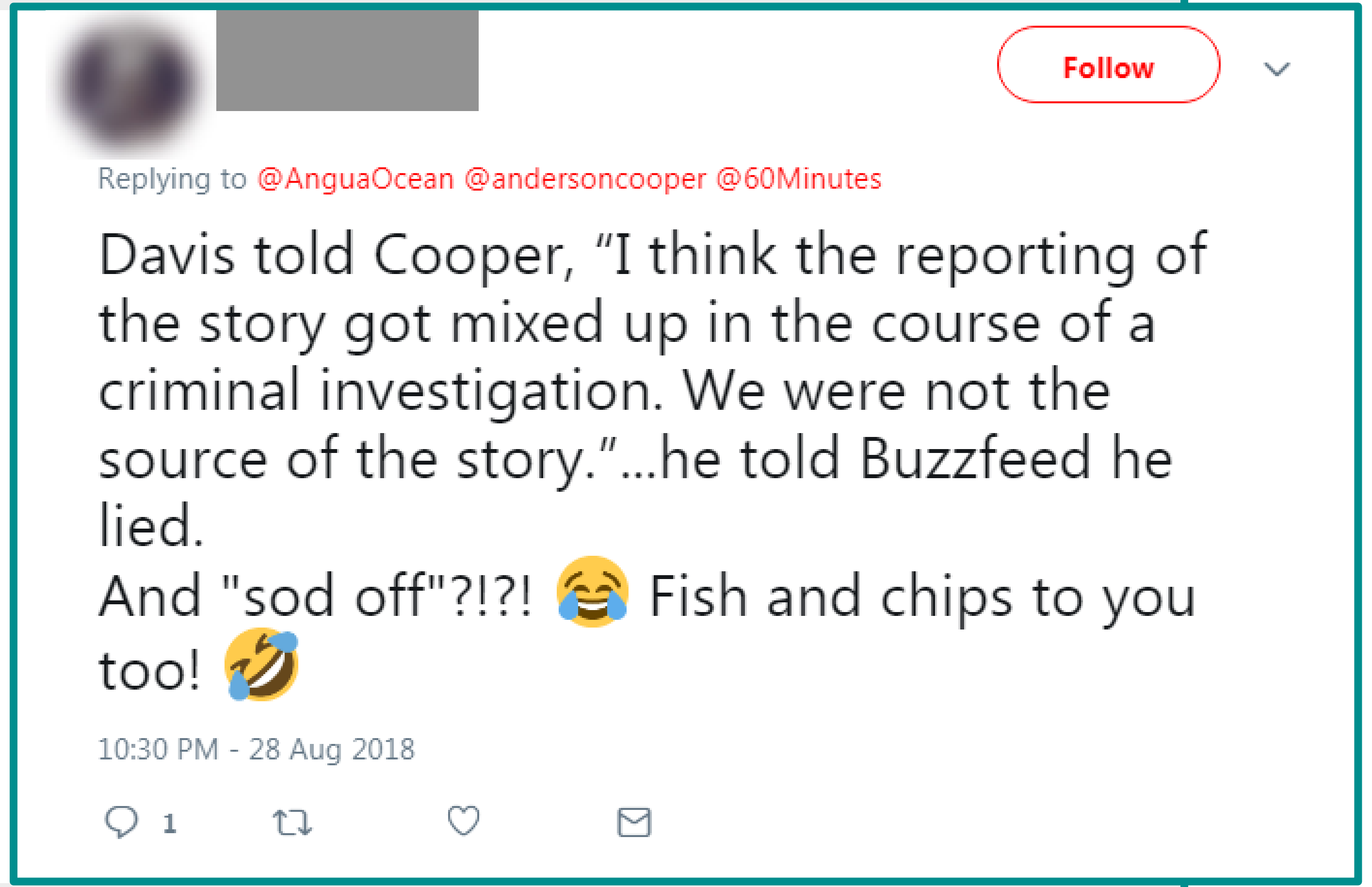
- **Q: What is our stance on white supremacy, white nationalism and white separatism?**
  - We don't allow praise, support and representation of **white supremacy** as an ideology. Eg. "White supremacy is the right thing"; "I am a white supremacist"; "Join the next White Supremacy rally!"
  - We allow praise, support and representation of white nationalism as an ideology. Eg. "White nationalism is the only way"; "I am a proud white nationalist"
  - We allow praise, support and representation of white separatism as an ideology. Eg. "White separatism is the perfect solution to America's problems"; "I am a white separatist". By the same token, we allow to call for the creation of white ethno-states (Eg. "The US should be a white-only nation")

# Monitoring and Tagging Hate Speech in Social Media

## Typology and coding

## Coding

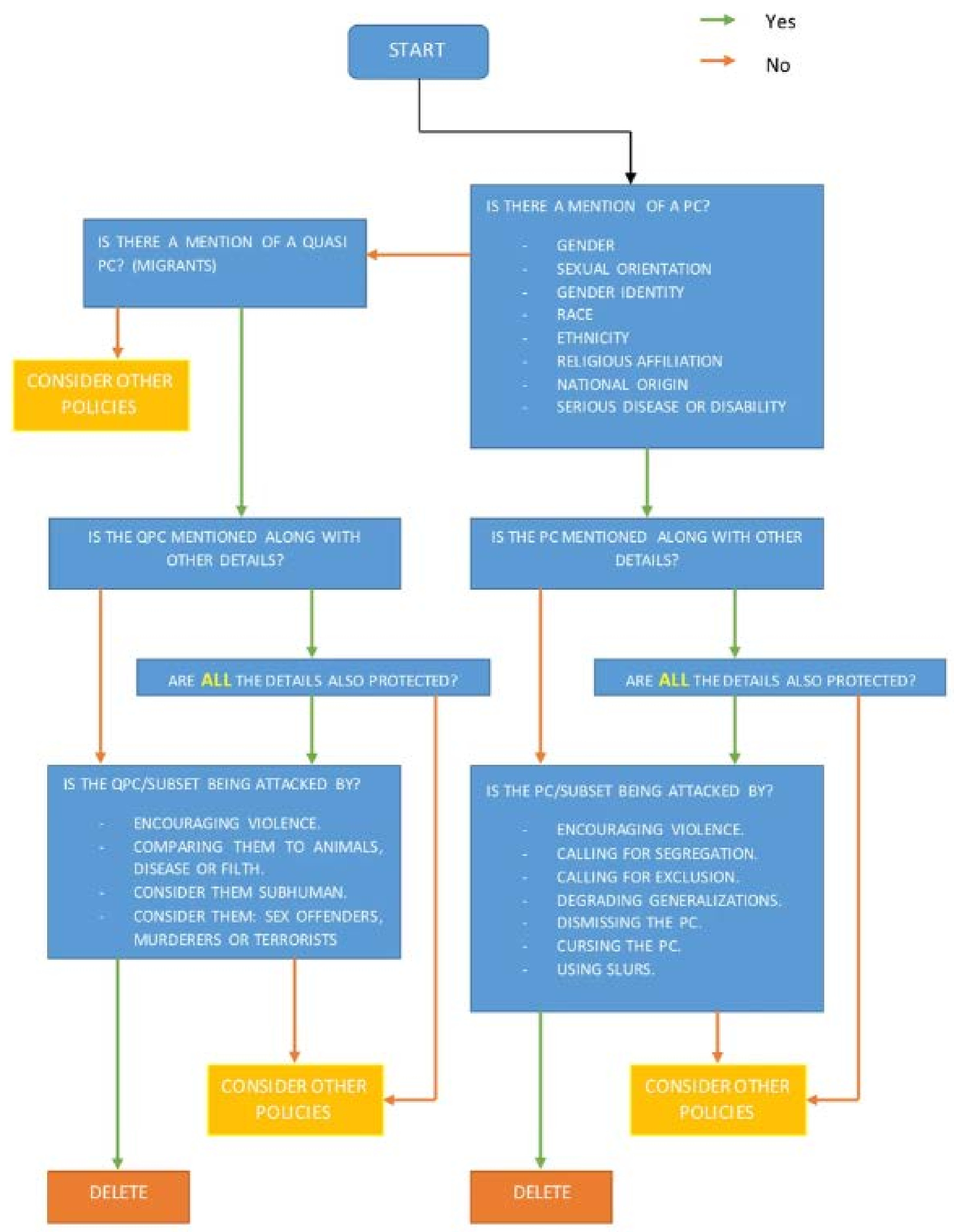
- A global set of policies to achieve consistency
- YES or NO **OR** multiple choice questions
- Test cases as words take on new meaning
- Offensive speech vs hateful territory



Replying to @AnguaOcean @andersoncooper @60Minutes

Davis told Cooper, “I think the reporting of the story got mixed up in the course of a criminal investigation. We were not the source of the story.”...he told Buzzfeed he lied.
And "sod off"?!?! 😂 Fish and chips to you too! 🤣

10:30 PM - 28 Aug 2018


START
Yes
No
IS THERE A MENTION OF A PC?
- GENDER
- SEXUAL ORIENTATION
- GENDER IDENTITY
- RACE
- ETHNICITY
- RELIGIOUS AFFILIATION
- NATIONAL ORIGIN
- SERIOUS DISEASE OR DISABILITY
IS THERE A MENTION OF A QUASI PC? (MIGRANTS)
CONSIDER OTHER POLICIES
IS THE QPC MENTIONED ALONG WITH OTHER DETAILS?
IS THE PC MENTIONED ALONG WITH OTHER DETAILS?
ARE ALL THE DETAILS ALSO PROTECTED?
ARE ALL THE DETAILS ALSO PROTECTED?
IS THE QPC/SUBSET BEING ATTACKED BY?
- ENCOURAGING VIOLENCE.
- COMPARING THEM TO ANIMALS, DISEASE OR FILTH.
- CONSIDER THEM SUBHUMAN.
- CONSIDER THEM: SEX OFFENDERS, MURDERERS OR TERRORISTS
IS THE PC/SUBSET BEING ATTACKED BY?
- ENCOURAGING VIOLENCE.
- CALLING FOR SEGREGATION.
- CALLING FOR EXCLUSION.
- DEGRADING GENERALIZATIONS.
- DISMISSING THE PC.
- CURSING THE PC.
- USING SLURS.
CONSIDER OTHER POLICIES
CONSIDER OTHER POLICIES
DELETE
DELETE

**Migrants:** people who cross an international border with intent to establish residency in a new country, regardless of whether their motivation is economic or political.

Typology and coding

## Main challenges

- Language diversity
- Ineffective automation
- Incubating fake news and misinformation
- Commercialized conversation

- Slow corrections – unfair censorship
- Global inconsistency
- Geopolitical situations
- Historical and cultural implications

## What’s being done to fight it?

- Expanded hate speech rules
- More human annotation
- Sophisticated technology
- Academic & scientific research

European Commission

Rights, Equality and Citizenship Programme

**DACHS**

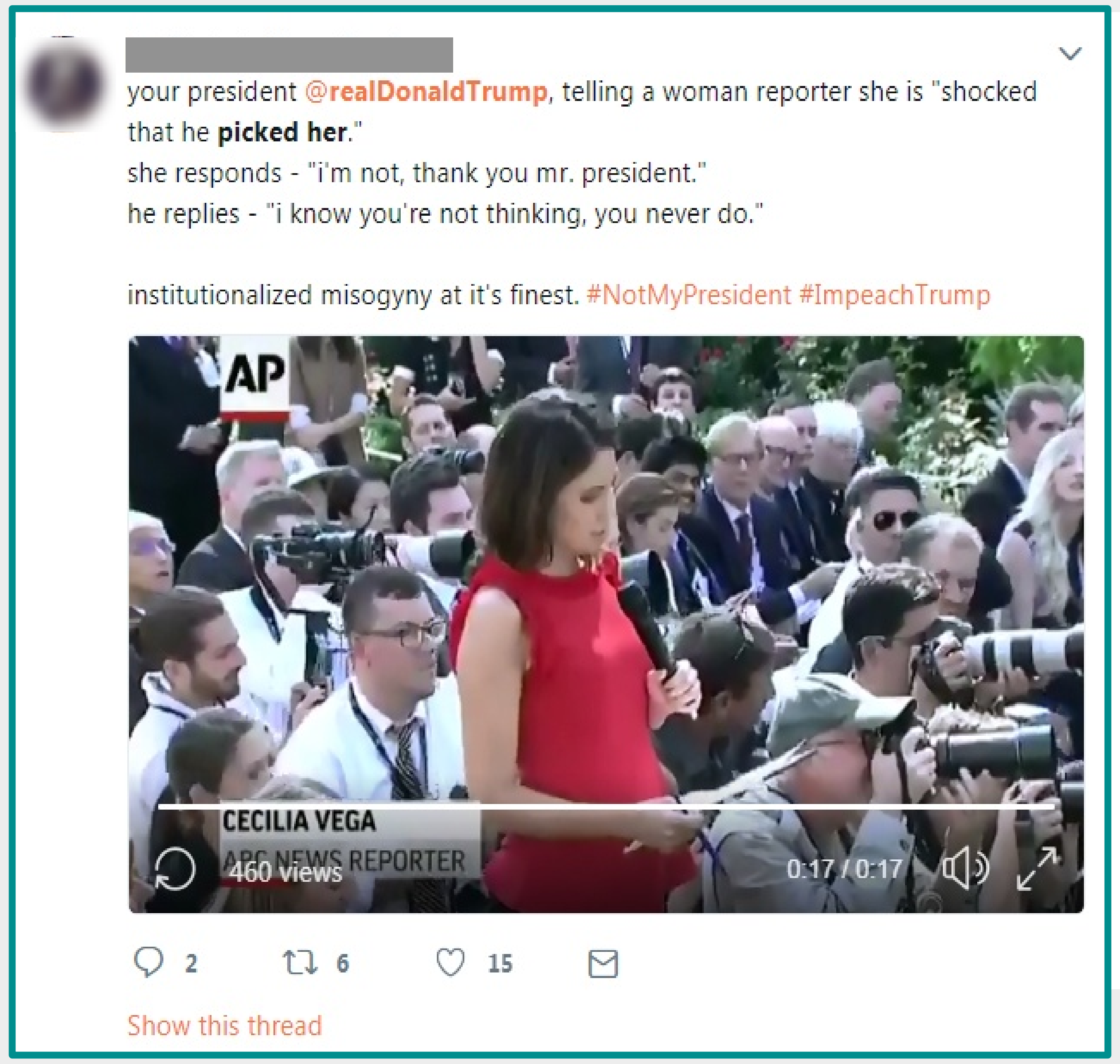
**A Data-Driven Approach to Countering Hate Speech**

DATASCOUTING actionable information

European Journalism Centre

# Monitoring and Tagging Hate Speech in Social Media

## Hate speech against journalists



[redacted]

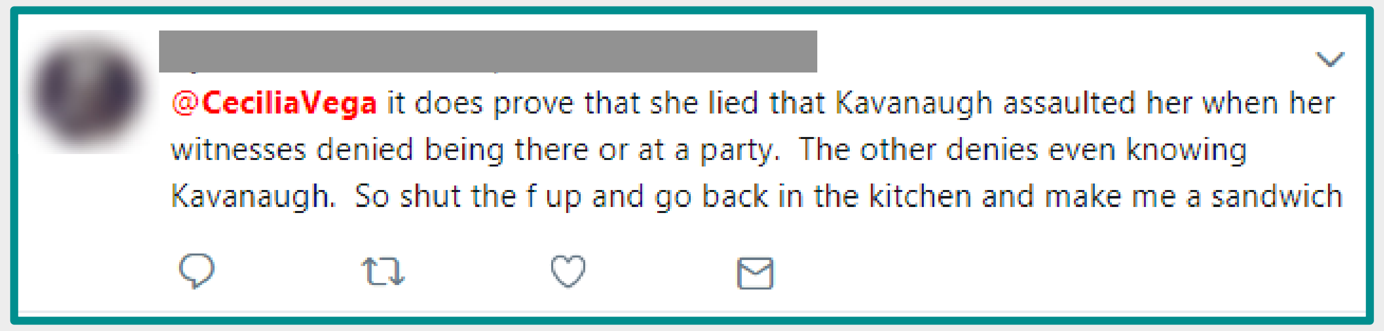
your president @realDonaldTrump, telling a woman reporter she is "shocked that he **picked her**."
she responds - "i'm not, thank you mr. president."
he replies - "i know you're not thinking, you never do."

institutionalized misogyny at it's finest. #NotMyPresident #ImpeachTrump

2 6 15

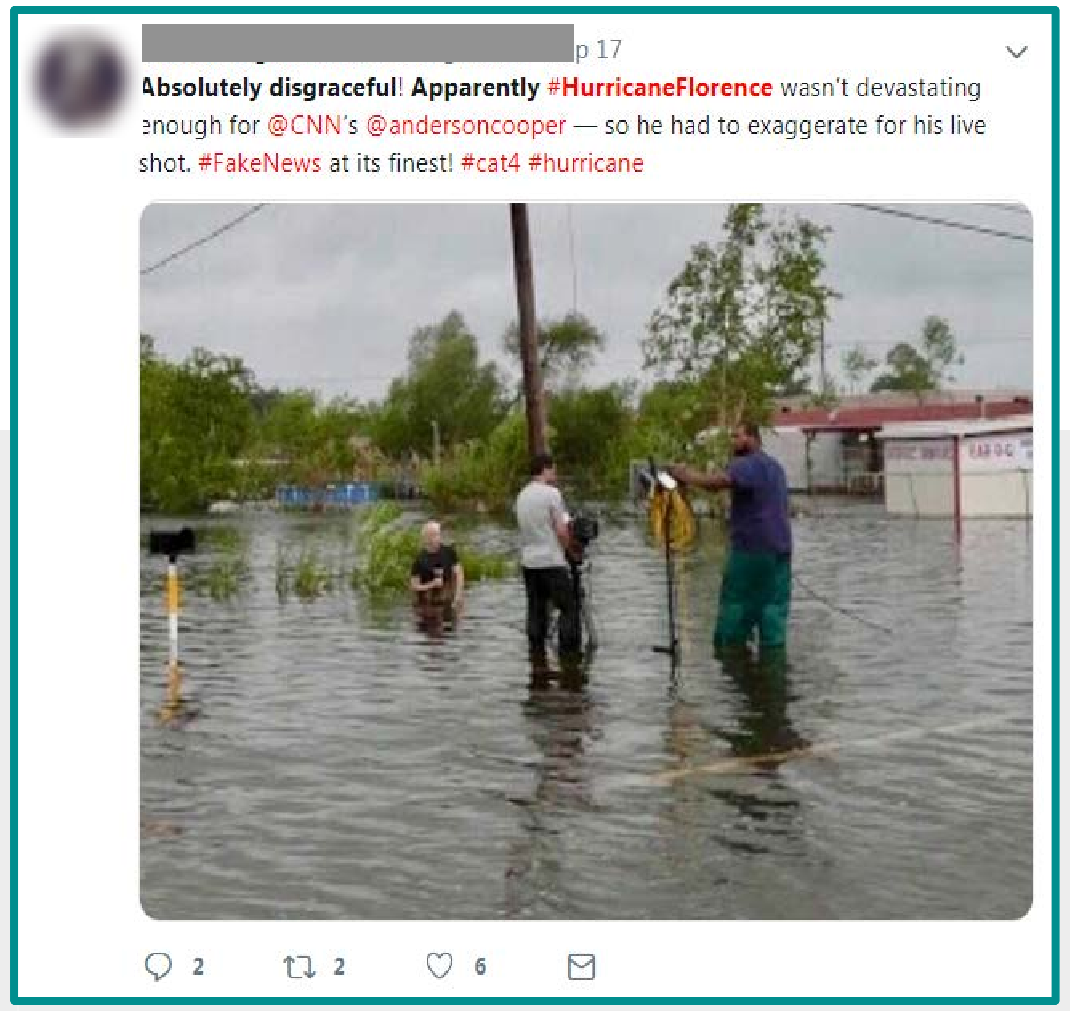
Show this thread

- Special targets of hate speech
- Public personas
- Report on somebody else
- Fake news runs faster than the truth



[redacted]

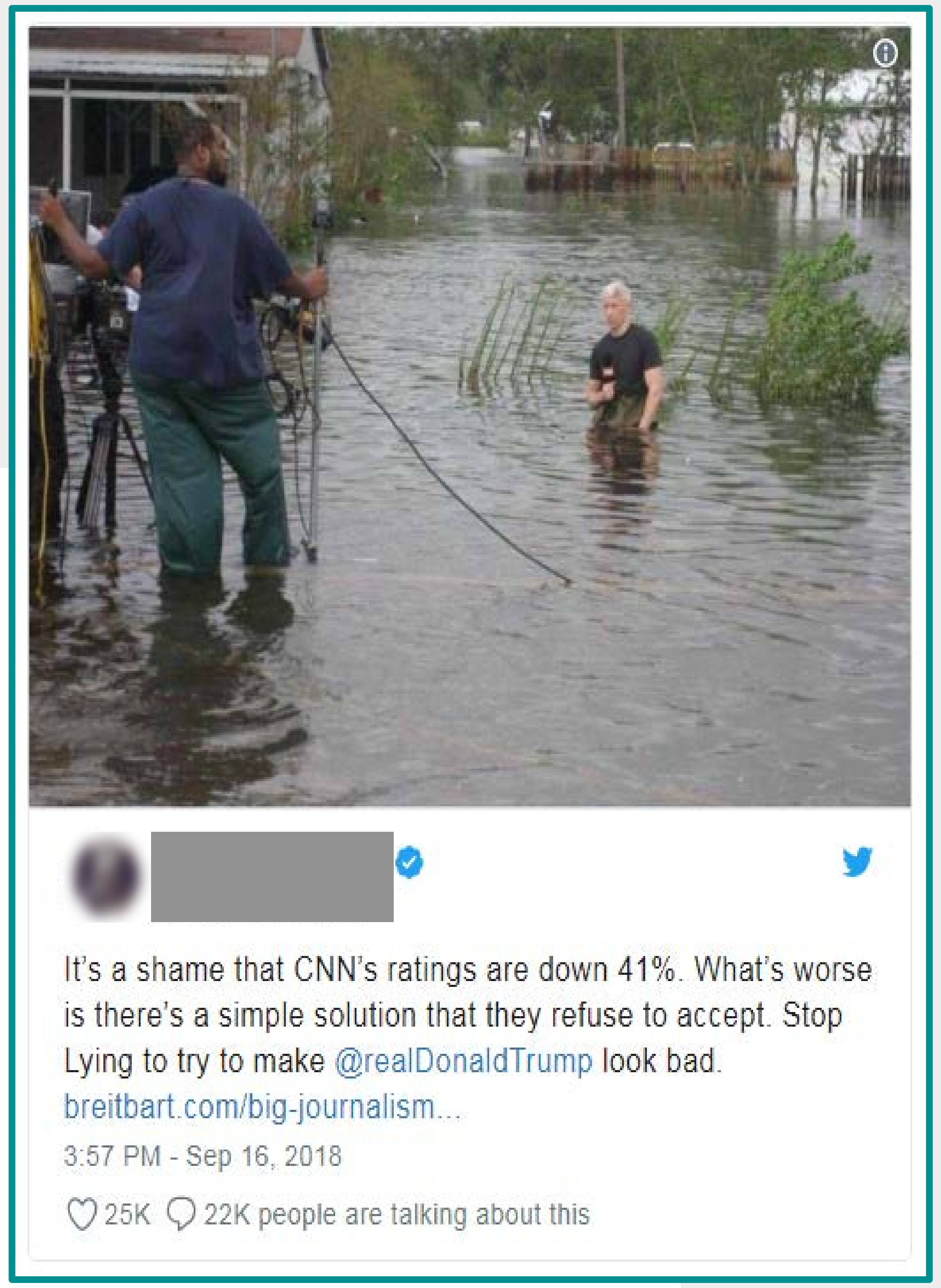
@CeciliaVega it does prove that she lied that Kavanaugh assaulted her when her witnesses denied being there or at a party. The other denies even knowing Kavanaugh. So shut the f up and go back in the kitchen and make me a sandwich



[redacted] p 17

**Absolutely disgraceful**! **Apparently** #HurricaneFlorence wasn't devastating enough for @CNN's @andersoncooper — so he had to exaggerate for his live shot. #FakeNews at its finest! #cat4 #hurricane

2 2 6

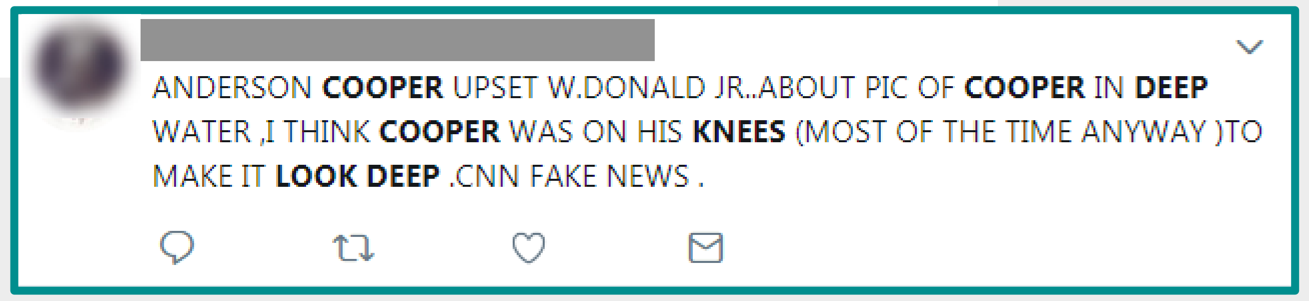


[redacted]

It's a shame that CNN's ratings are down 41%. What's worse is there's a simple solution that they refuse to accept. Stop Lying to try to make @realDonaldTrump look bad. breitbart.com/big-journalism...

3:57 PM - Sep 16, 2018

25K 22K people are talking about this



[redacted]

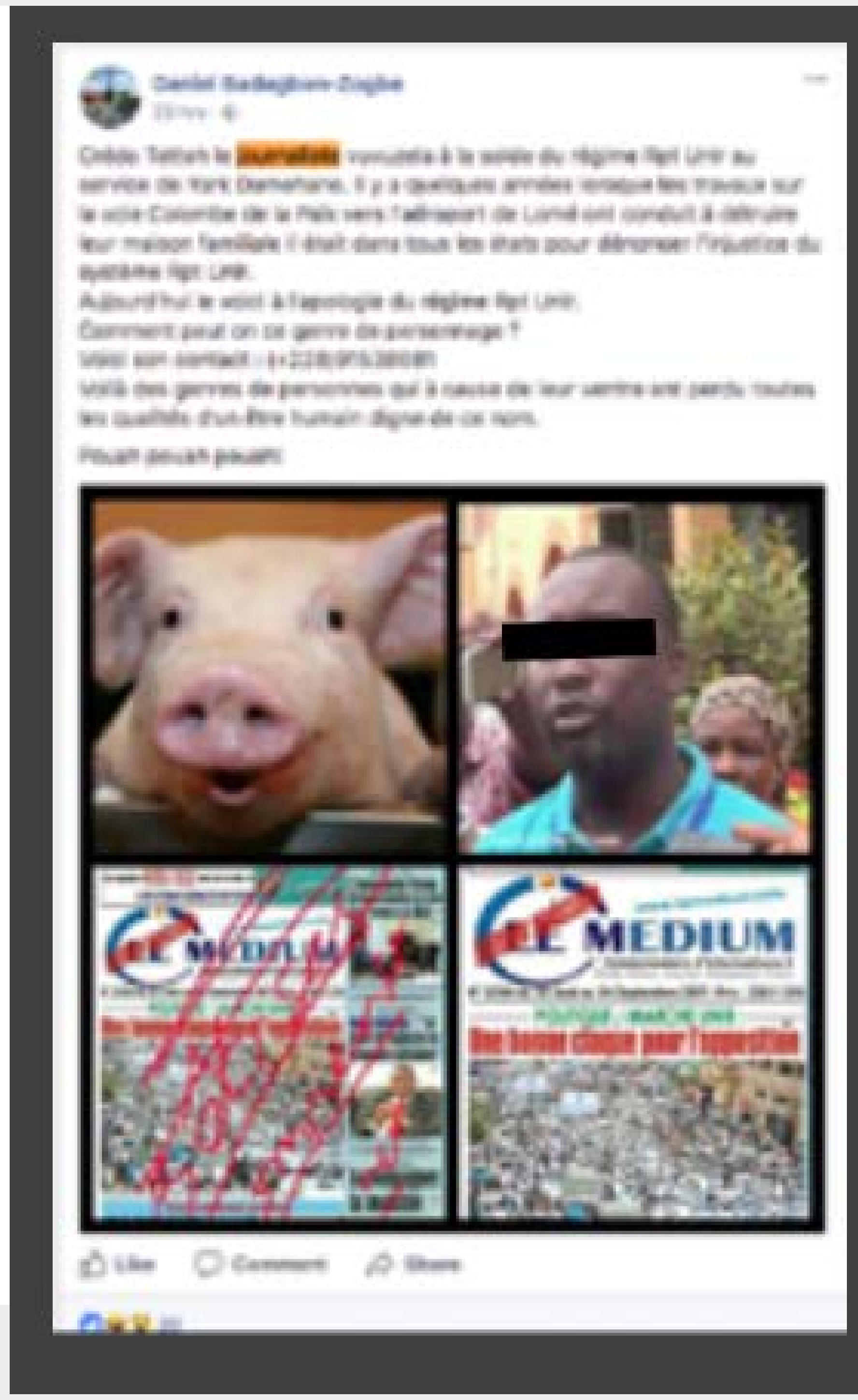
ANDERSON **COOPER** UPSET W.DONALD JR..ABOUT PIC OF **COOPER** IN **DEEP** WATER ,I THINK **COOPER** WAS ON HIS **KNEES** (MOST OF THE TIME ANYWAY )TO MAKE IT **LOOK DEEP** .CNN FAKE NEWS .

# Monitoring and Tagging Hate Speech in Social Media

## Impact of hate speech on journalists

- Undermines the role of journalism
- Impacts trust
- Incites criticism – instigates violence
- Intimidation and harassment
- Self-censorship, mental well-being
- Fear of repercussion

– jailed, disappeared, abducted, killed



[illegible]

### “Kill these journalists once and for all”

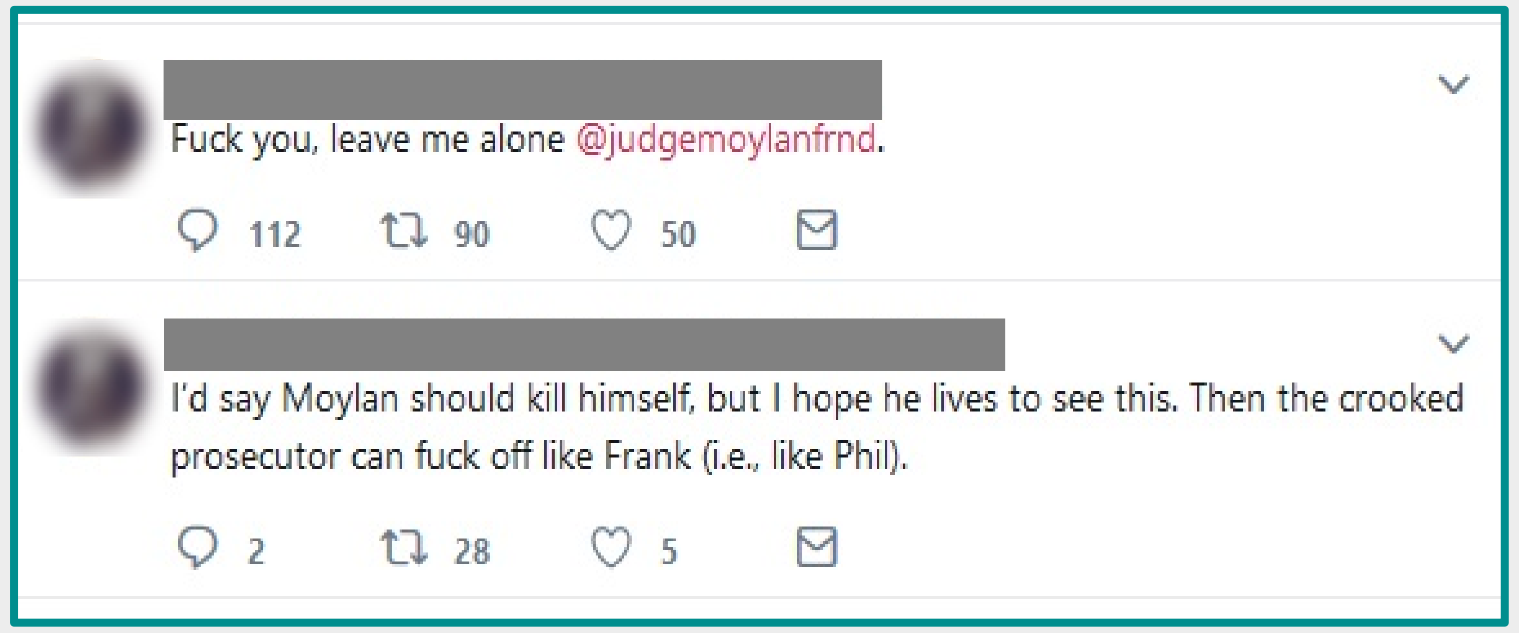
In the summer of 2017, a fearful message was spread on social networks in Togo: *“Kill all these journalists once and for all.”* It was accompanied by photos of four journalists, pasted onto images of pigs. The journalists’ personal data was disseminated and they were accused by their critics of supporting the Lomé regime.

# Monitoring and Tagging Hate Speech in Social Media

## From malicious threats to actual intent

The Baltimore Sun

From left up: John McNamara, Rob Hiaasen, Gerald Fischman
From left down: Rebecca Smith, Wendi Winters

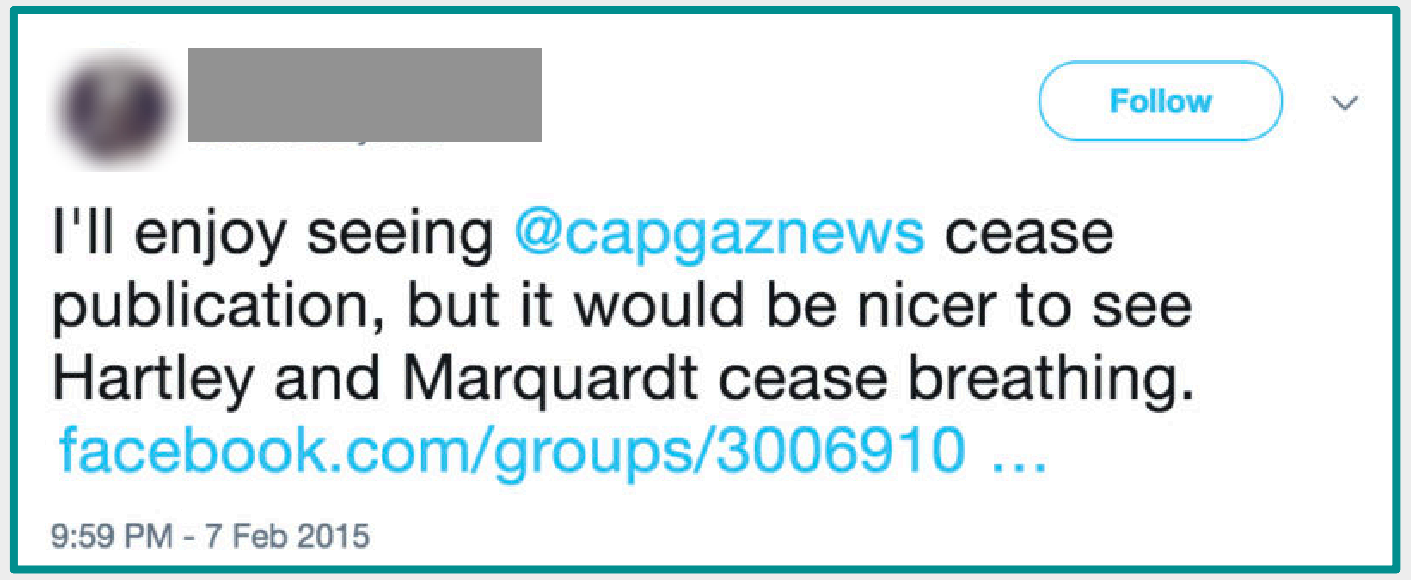


[redacted]

Fuck you, leave me alone @judgemoylanfrnd.

[redacted]

I'd say Moylan should kill himself, but I hope he lives to see this. Then the crooked prosecutor can fuck off like Frank (i.e., like Phil).



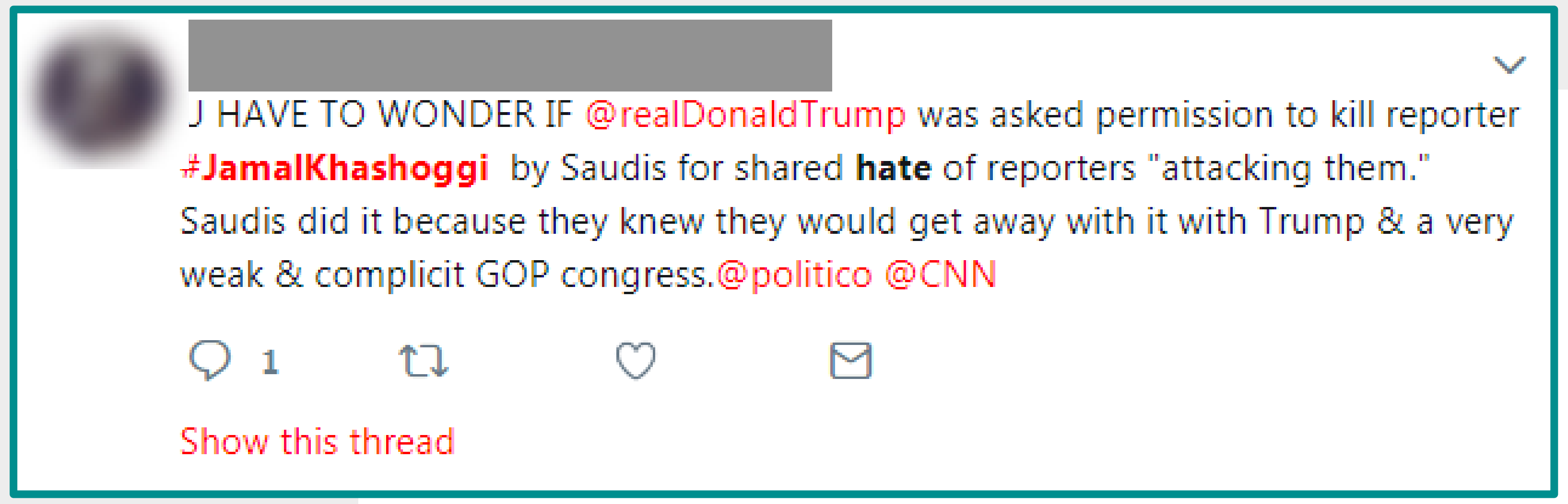
[redacted]

I'll enjoy seeing @capgaznews cease publication, but it would be nicer to see Hartley and Marquardt cease breathing. facebook.com/groups/3006910 …

9:59 PM - 7 Feb 2015

# Monitoring and Tagging Hate Speech in Social Media

## Reclaiming Agency

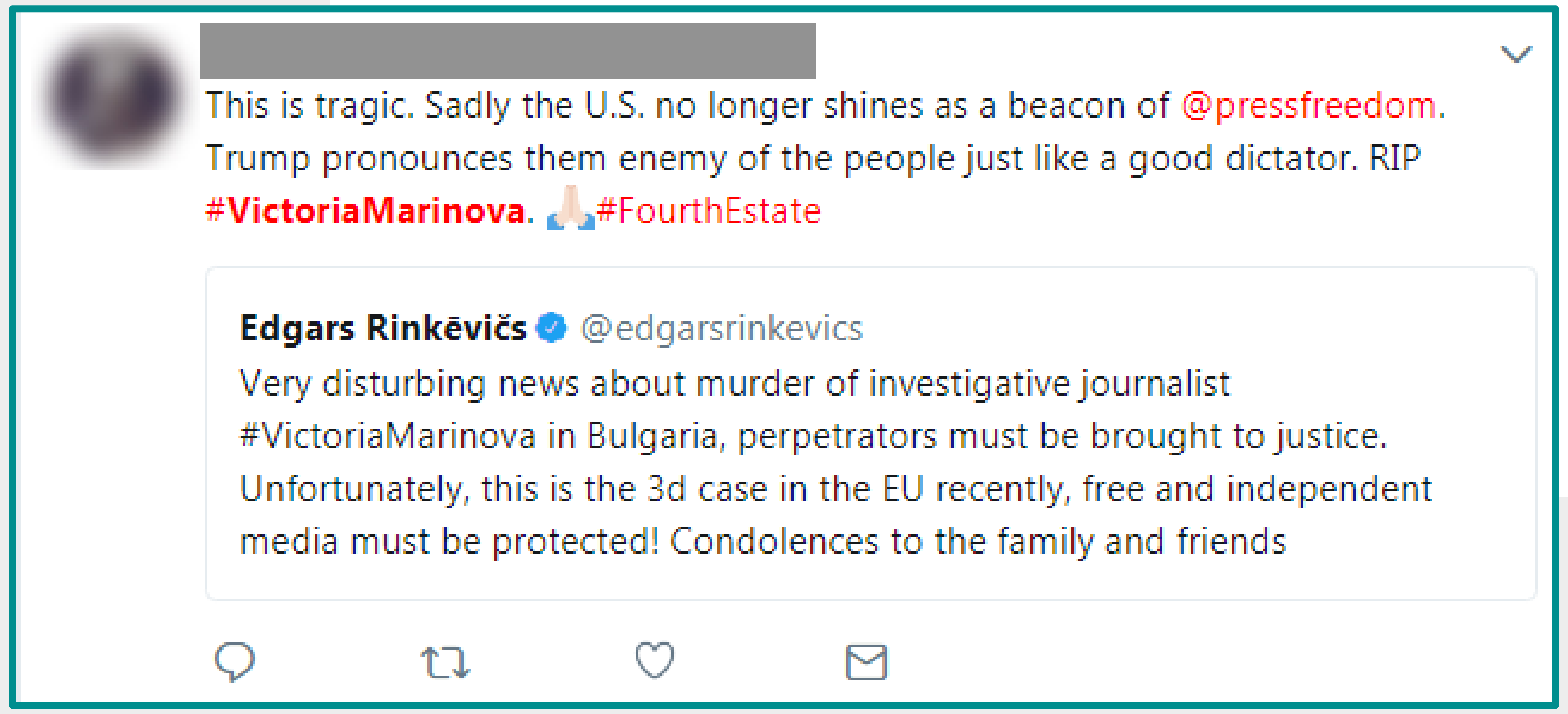


[redacted]

J HAVE TO WONDER IF @realDonaldTrump was asked permission to kill reporter #JamalKhashoggi by Saudis for shared **hate** of reporters "attacking them." Saudis did it because they knew they would get away with it with Trump & a very weak & complicit GOP congress.@politico @CNN

1

Show this thread



[redacted]

This is tragic. Sadly the U.S. no longer shines as a beacon of @pressfreedom. Trump pronounces them enemy of the people just like a good dictator. RIP #VictoriaMarinova. 🙏🏻#FourthEstate

> **Edgars Rinkēvičs** @edgarsrinkevics
> Very disturbing news about murder of investigative journalist #VictoriaMarinova in Bulgaria, perpetrators must be brought to justice. Unfortunately, this is the 3d case in the EU recently, free and independent media must be protected! Condolences to the family and friends

- Hate speech detection
- Early warning mechanism
- Counter measures

Bulgarian investigative journalist Viktoria Marinova and Saudi journalist Jamal Khashoggi. | Photo: Reuters

# Ευχαριστώ

Sophia Karakeva

Communications and Marketing Executive | Datascouting

Vice President | FIBEP

DATASCOUTING
actionable information

FIBEP