SOCIAL MEDIA TRENDS, 3RD OF OCTOBER

Health of the Media Conversation: Using AI to Detect and Understand Harmful Speech

Eric Karstens

Funding Consultant

European Journalism Centre

Sophia Karakeva

Communications and Marketing Executive

Datascouting

Stavros Vologiannidis

Professor for the Technological Educational Institute of Central Macedonia and Founder

Datascouting

50 FIBEP WORLD MEDIA INTELLIGENCE CONGRESS

COPENHAGEN
OCTOBER 1-3 2018
MARRIOTT HOTEL

European Journalism Centre

DATASCOUTING
Actionable Information

@EricKarstens | @soka1605 | @svologia | @ejcnet | @Datascouting | #FIBEP | #WMIC18

SOCIAL MEDIA TRENDS, 3RD OF OCTOBER

Health of the Media Conversation: Using AI to Detect and Understand Harmful Speech

Journalism and hate speech

Eric Karstens

Funding Consultant

European Journalism Centre

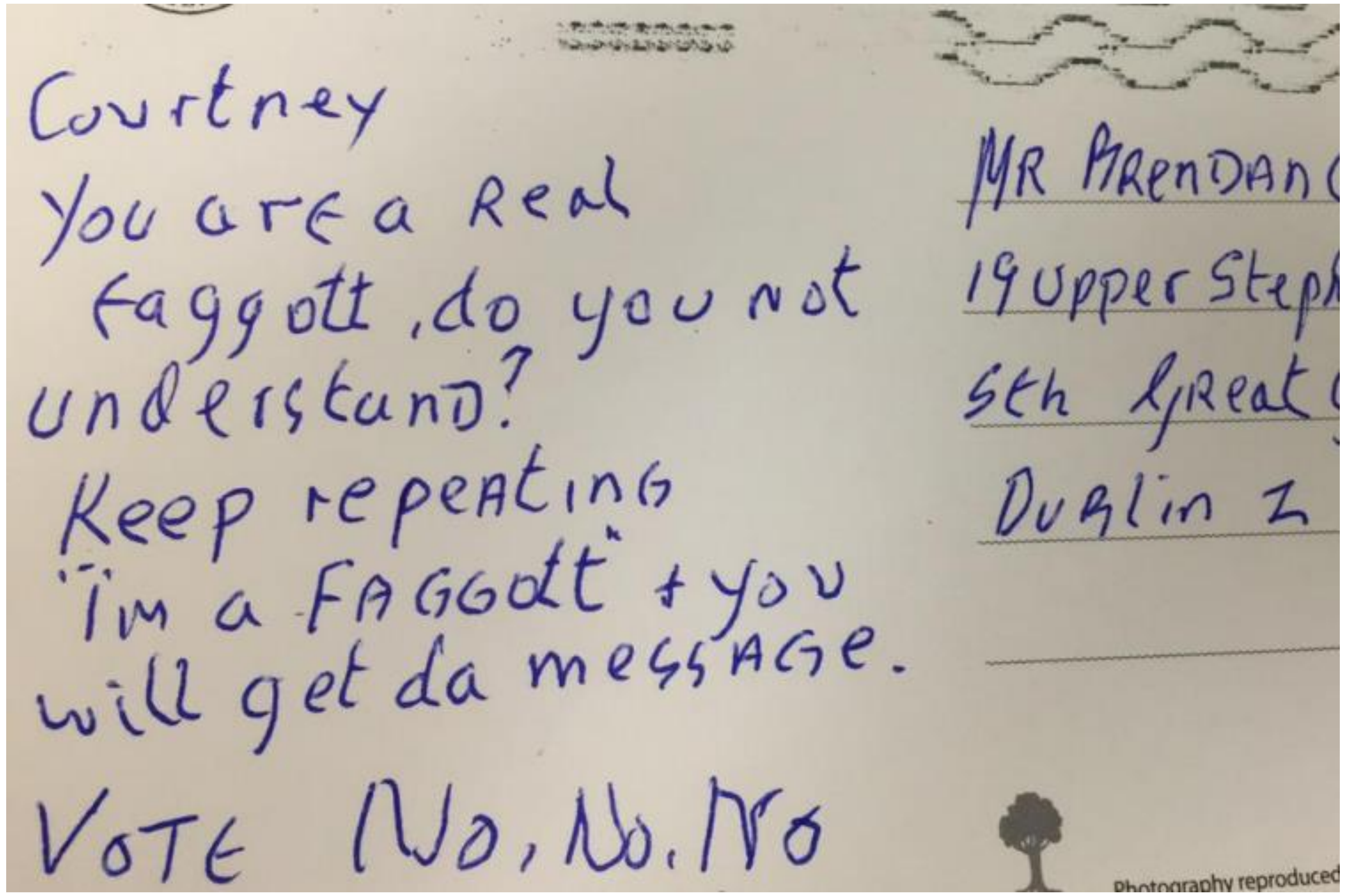
50 FIBEP WORLD MEDIA INTELLIGENCE CONGRESS
COPENHAGEN
OCTOBER 1-3 2018
MARRIOTT HOTEL

European Journalism Centre

DATASCOUTING
Actionable Information

@EricKarstens | @ejcnet | Eric Karstens | European Journalism Centre

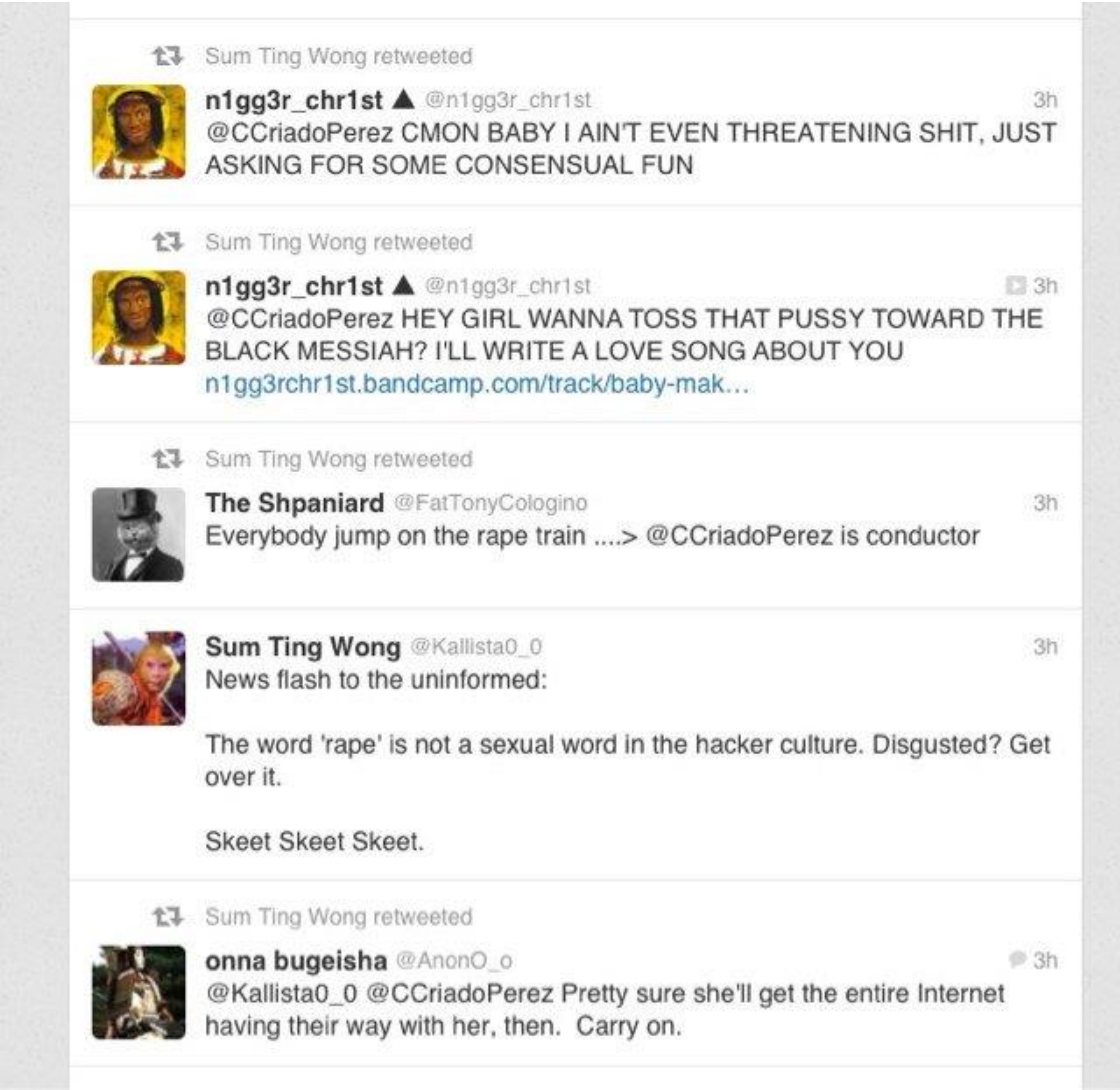
HATE MAIL THEN AND NOW



Courtney
You are a Real
faggott, do you not
understanD?
Keep repeating
'I'm a FAGGOtt' + you
will get da message.
VoTE No, No. No

MR BRenDAN C
19 Upper Steph
5th Great
Dublin 2

Photography reproduced



Sum Ting Wong retweeted

n1gg3r_chr1st ▲ @n1gg3r_chr1st 3h
@CCriadoPerez CMON BABY I AIN'T EVEN THREATENING SHIT, JUST ASKING FOR SOME CONSENSUAL FUN

Sum Ting Wong retweeted

n1gg3r_chr1st ▲ @n1gg3r_chr1st 3h
@CCriadoPerez HEY GIRL WANNA TOSS THAT PUSSY TOWARD THE BLACK MESSIAH? I'LL WRITE A LOVE SONG ABOUT YOU n1gg3rchr1st.bandcamp.com/track/baby-mak...

Sum Ting Wong retweeted

The Shpaniard @FatTonyCologino 3h
Everybody jump on the rape train ....> @CCriadoPerez is conductor

Sum Ting Wong @Kallista0_0 3h
News flash to the uninformed:

The word 'rape' is not a sexual word in the hacker culture. Disgusted? Get over it.

Skeet Skeet Skeet.

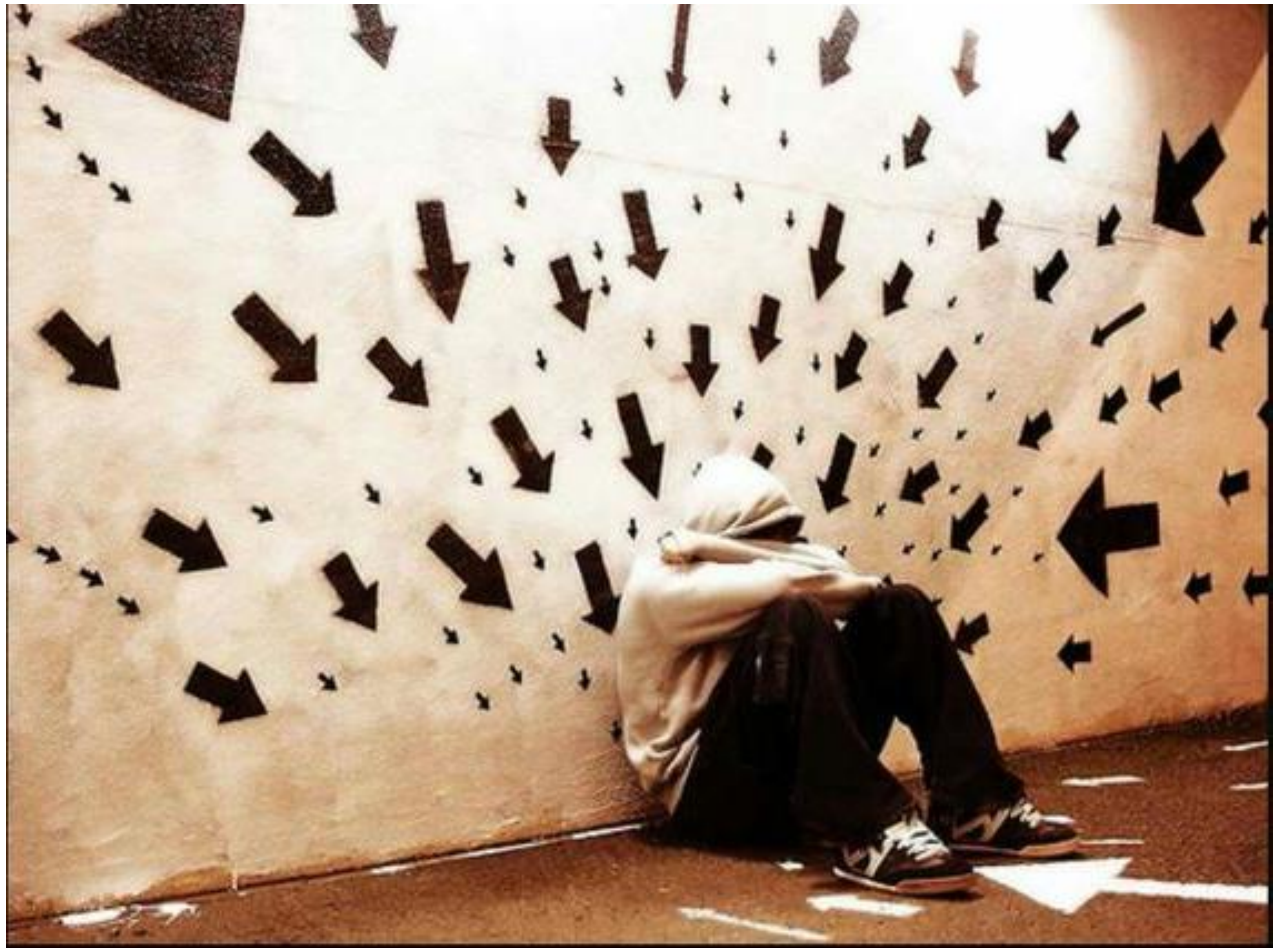
Sum Ting Wong retweeted

onna bugeisha @AnonO_o 3h
@Kallista0_0 @CCriadoPerez Pretty sure she'll get the entire Internet having their way with her, then. Carry on.

HATE SPEECH AGAINST JOURNALISTS

IMPACTS OF HATE SPEECH

RECLAIMING AGENCY



SOCIAL MEDIA TRENDS, 3RD OF OCTOBER

Health of the media conversation: using AI to detect and understand and harmful speech

Mapping and monitoring hate speech

Sophia Karakeva

Communications and Marketing Executive

Datascouting

What is hate speech?

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

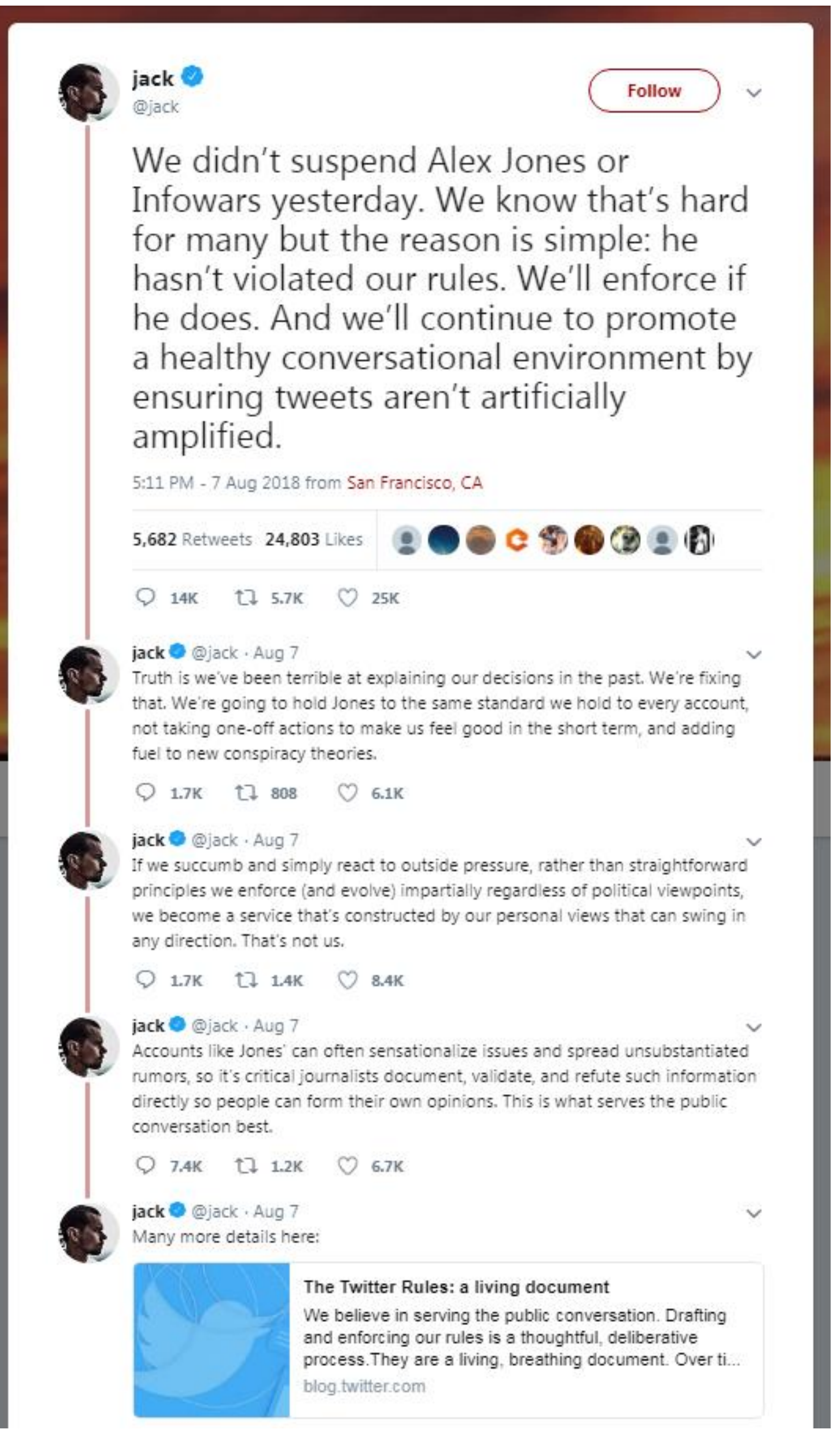
> #HateSpeech is just a word for pussies who can't stand up for themselves so they need legislation on language

> #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



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.

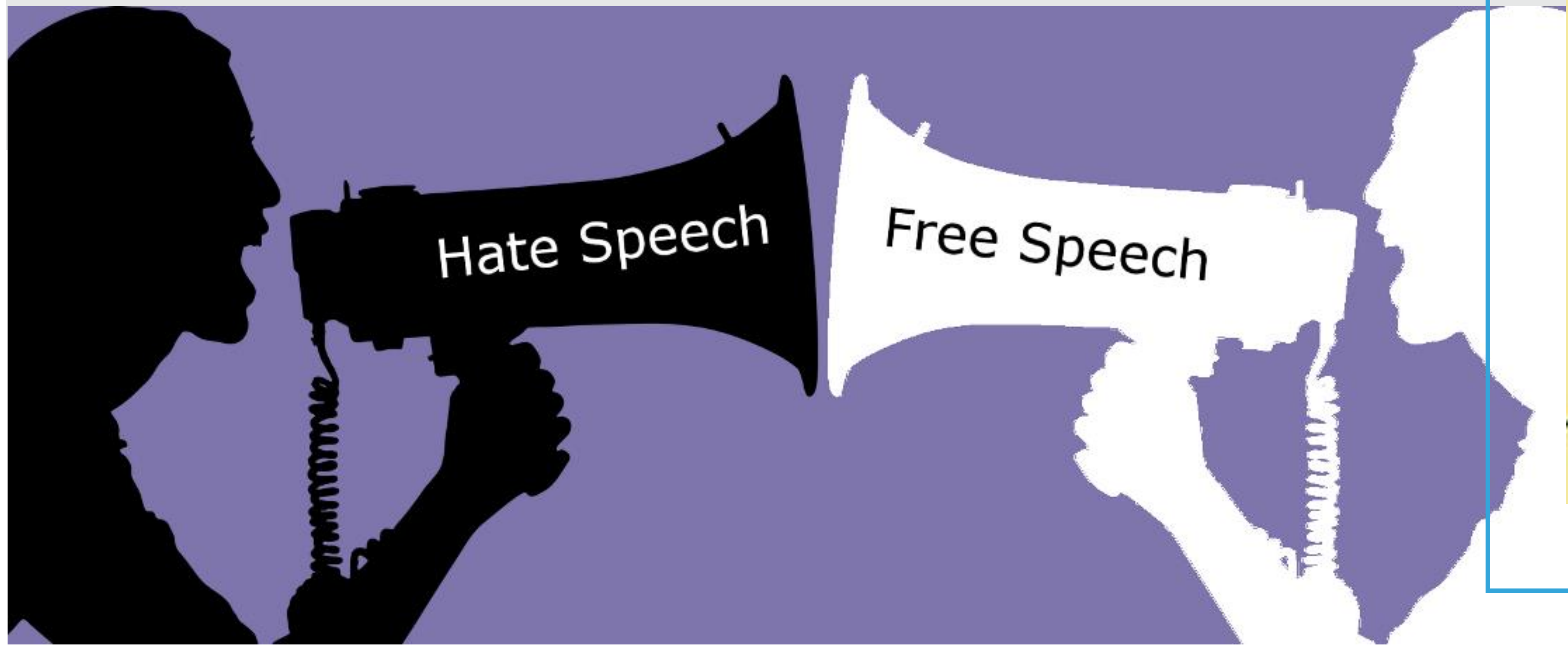
7.4K 1.2K 6.7K

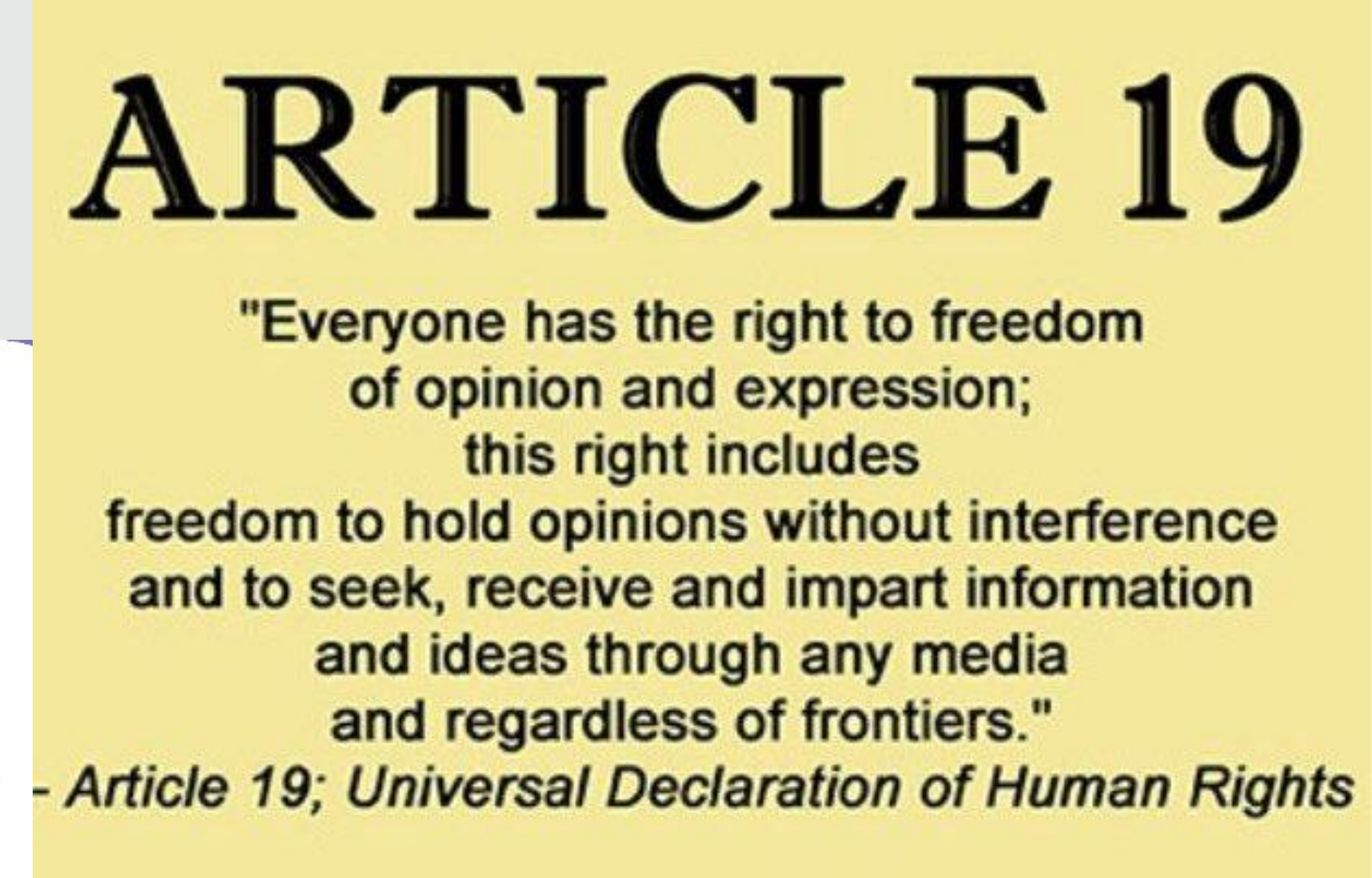
jack @jack · Aug 7
Many more details here:

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

The legal and discursive characteristics of hate speech

- Legal liability
- Danger of over-regulation

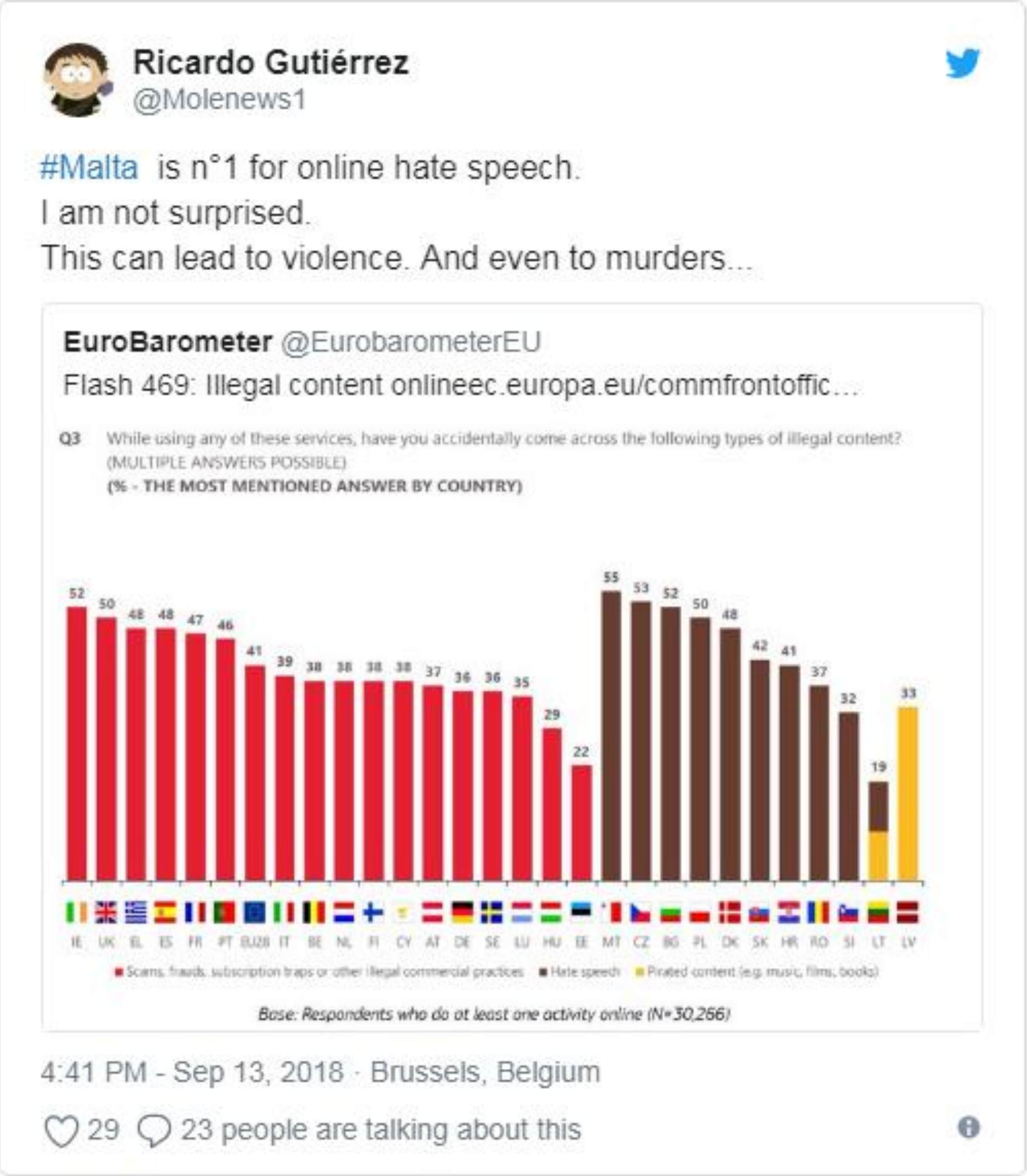




ARTICLE 19

"Everyone has the right to freedom
of opinion and expression;
this right includes
freedom to hold opinions without interference
and to seek, receive and impart information
and ideas through any media
and regardless of frontiers."
- *Article 19; Universal Declaration of Human Rights*

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…

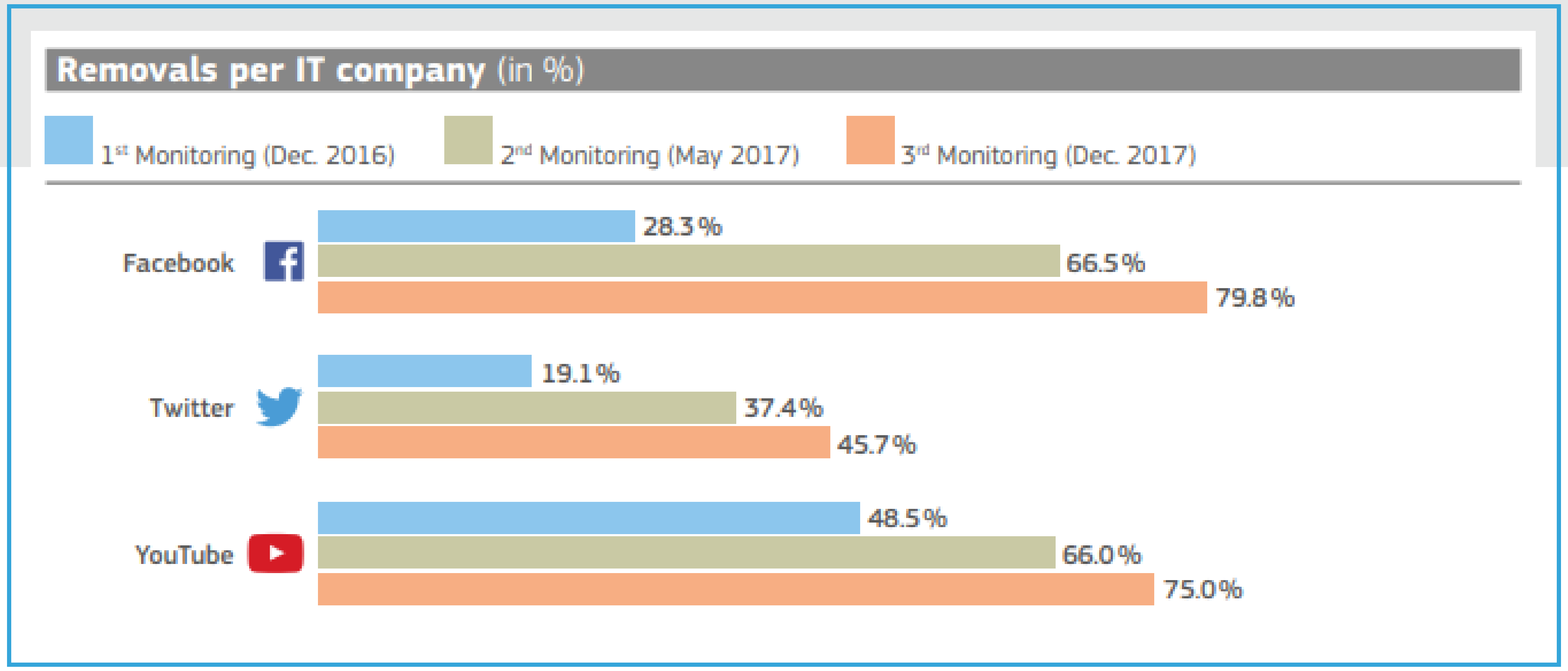
Q3 While using any of these services, have you accidentally come across the following types of illegal content?
(MULTIPLE ANSWERS POSSIBLE)
(% - THE MOST MENTIONED ANSWER BY COUNTRY)

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

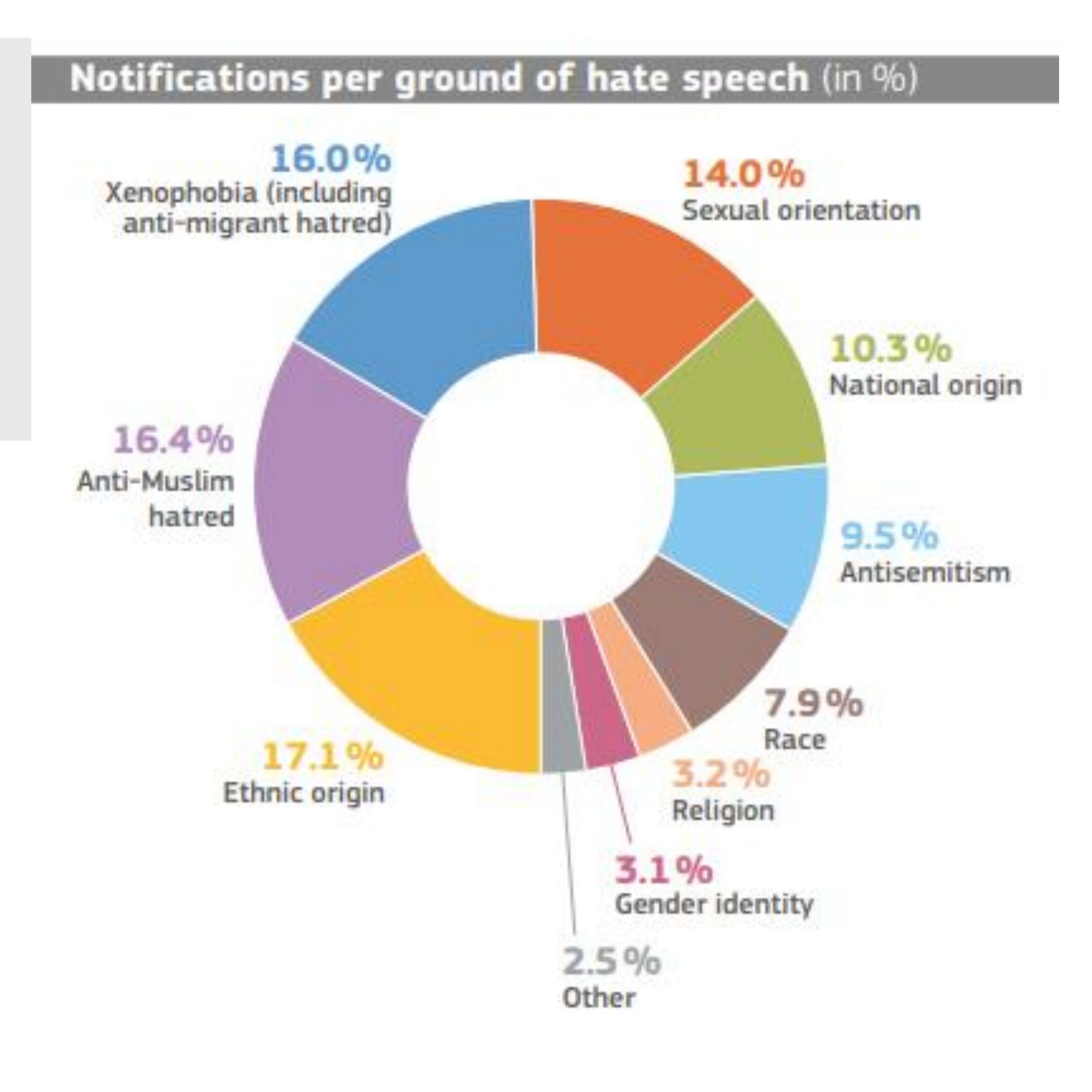


Source:
The Flash Eurobarometer, September 2018
The 3rd monitoring of EU's Code of Conduct, January 2018
2017 Pew Research Center survey about online harassment

Who is most targeted?

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

> “*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


Notifications per ground of hate speech (in %)

Source: 3rd monitoring exercise of the European Union’s Code of Conduct, released in January 2018

Hate me, hate me not

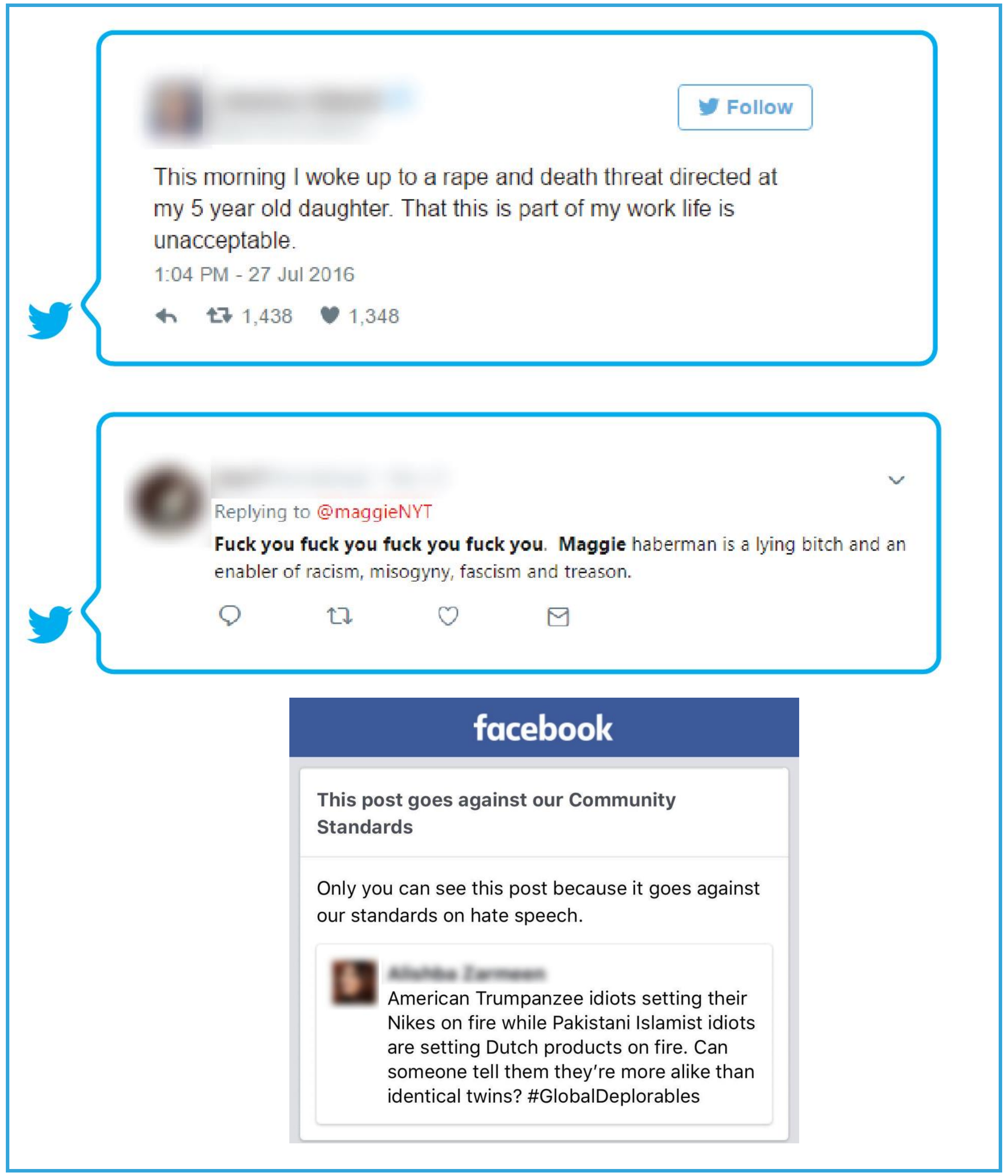
> ***"It's like public lynching… It has made be frightened for my physical safety when I am out in the streets."***
> Turkish journalist Amberin Zaman

- Hate me not: not allowed on our platform
- Hate me: often the most engaged content

- Hate me not: strong measures to take it down
- Hate me: not everything has the same value

- Hate me: highly reactive, high error rate
- Hate me not: automation and AI tools

- Hate me: "new normal"
- Hate me not: education, campaigns, regulation



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

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.

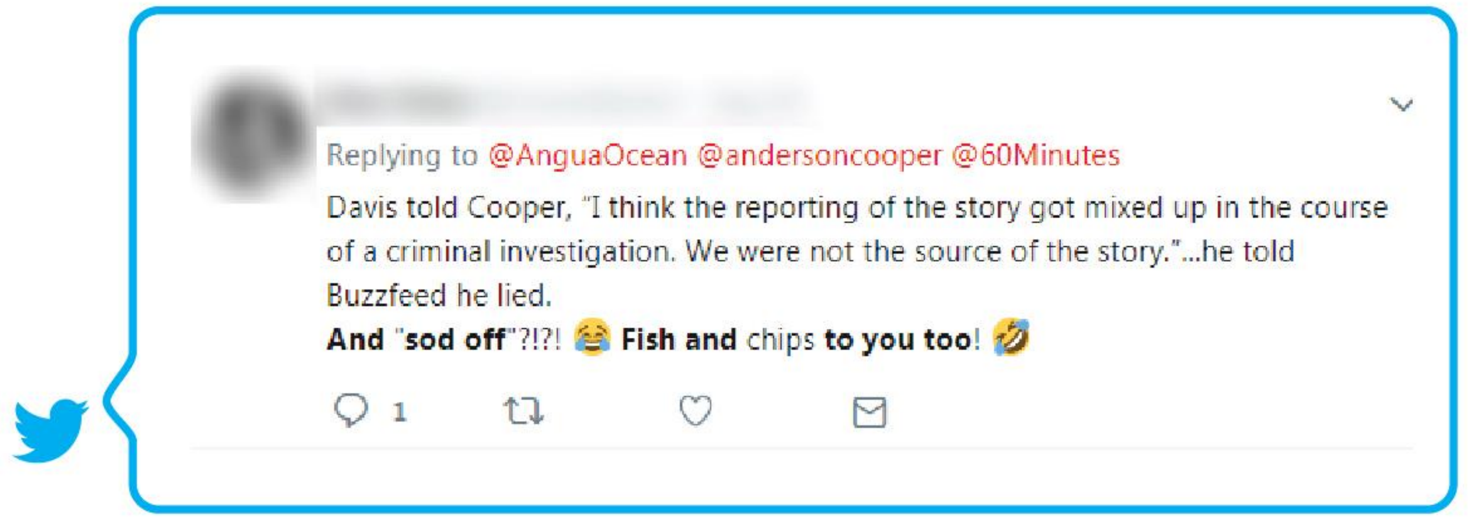
This post goes against our Community Standards

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

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

Typology and coding

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

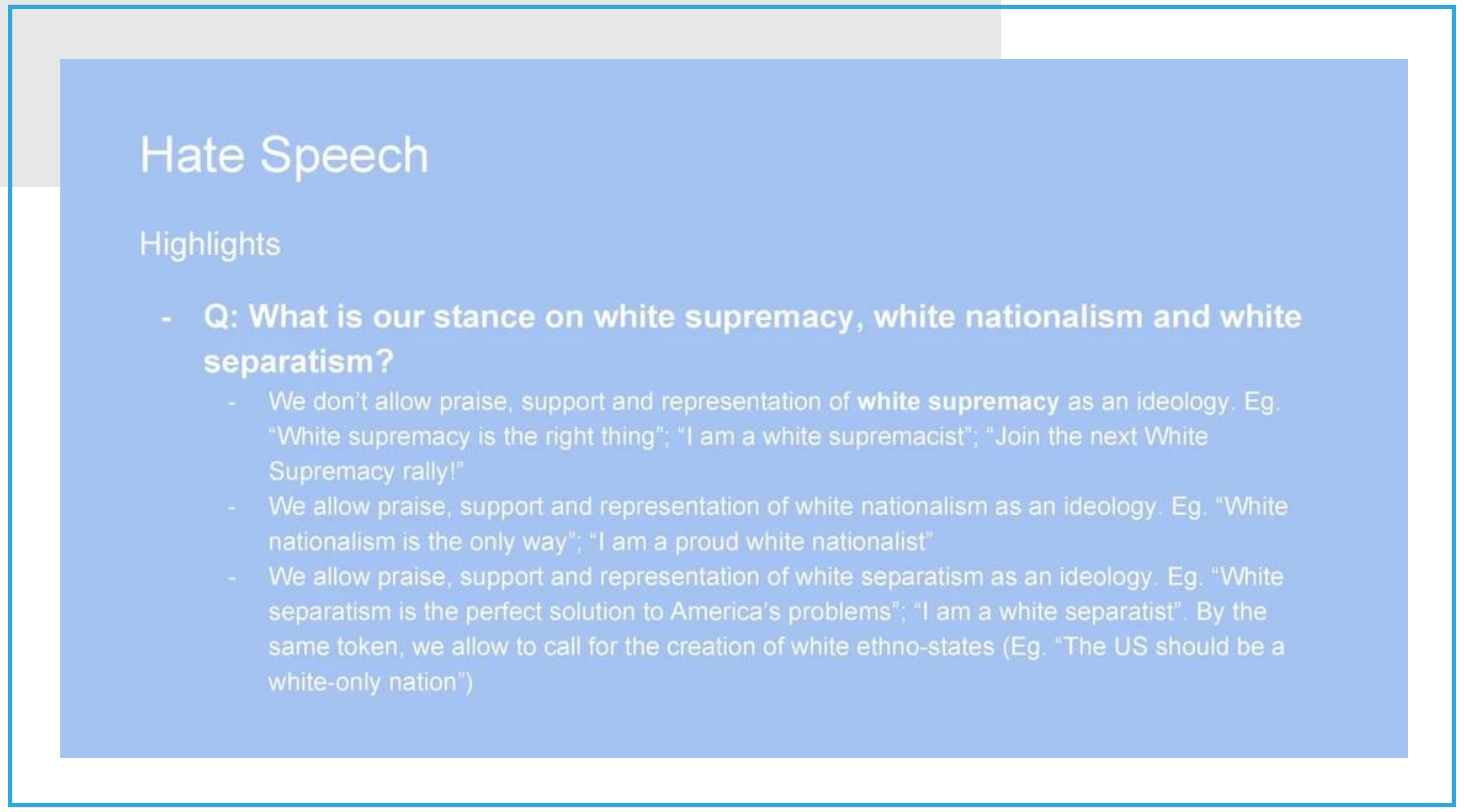


[redacted]

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.

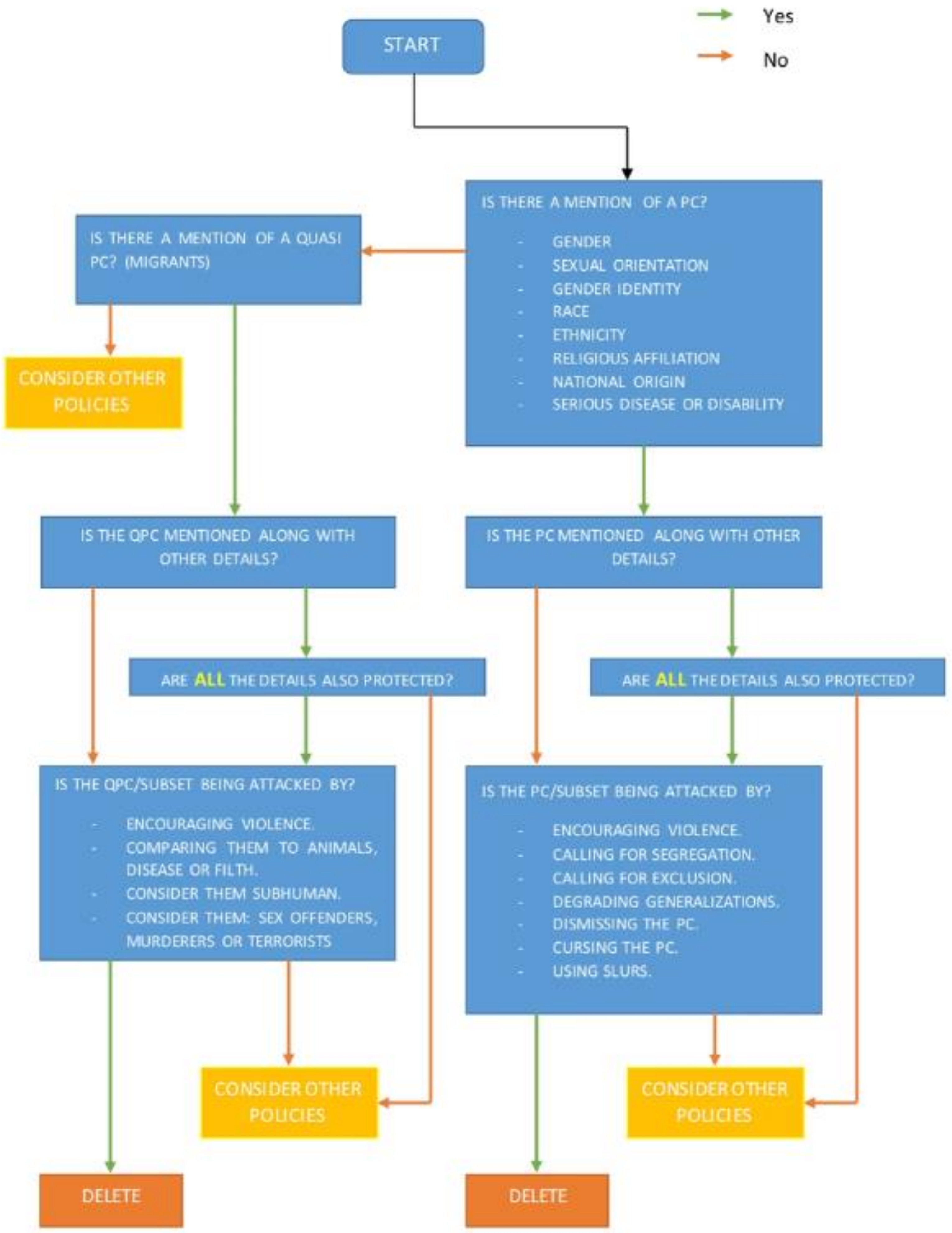
And "**sod off**"?!?! 😂 **Fish and** chips **to you too**! 🤣



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")

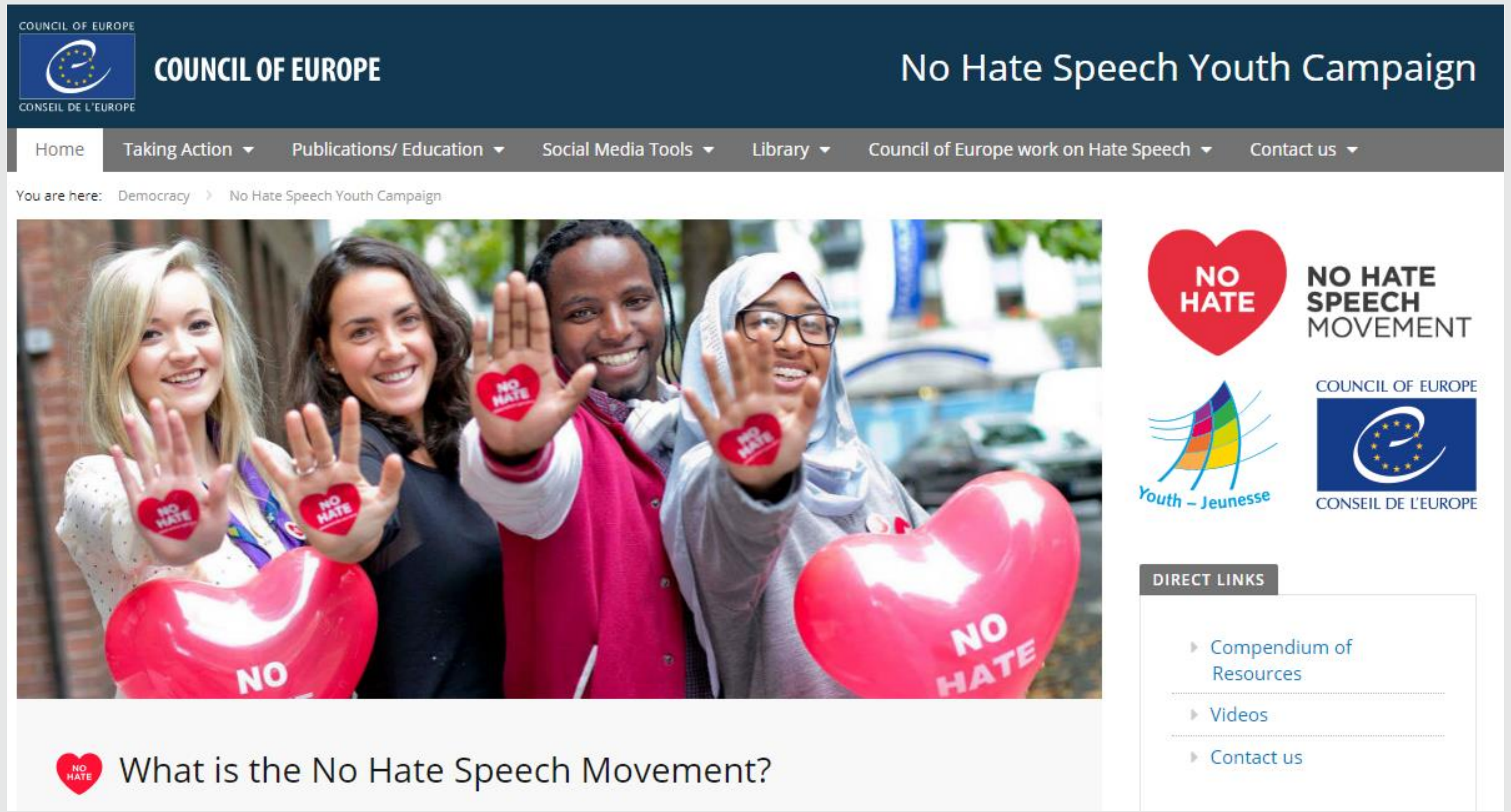


Yes
No
START
IS THERE A MENTION OF A PC?
- GENDER
- SEXUAL ORIENTATION
- GENDER IDENTITY
- RACE
- ETHNICITY
- RELIGIOUS AFFILIATION
- NATIONAL ORIGIN
- SERIOUS DISEASE OR DISABILITY
IS THERE A MENTION OF A QUASI PC? (MIGRANTS)
CONSIDER OTHER POLICIES
IS THE QPC MENTIONED ALONG WITH OTHER DETAILS?
IS 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.

What’s being done to fight it?

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



What is the No Hate Speech Movement?

SOCIAL MEDIA TRENDS, 3RD OF OCTOBER

HEALTH OF THE MEDIA CONVERSATION: USING AI TO DETECT AND UNDERSTAND AND HARMFUL SPEECH

AUTOMATICALLY TAGGING HATE SPEECH

Stavros Vologiannidis

ASSISTANT PROFESSOR
FOUNDER OF DATASCOUTING

50 FIBEP
WORLD MEDIA
INTELLIGENCE
CONGRESS

COPENHAGEN
OCTOBER 1-3 2018
MARRIOTT HOTEL

DATASCOUTING
Actionable Information

@SVOLOGIA | @DATASCOUTING | #FIBEP | #WMIC18

HATE SPEECH AND TECHNOLOGY

- Hate speech is nothing new
- But the problem is much more evident in the age of **social media**
- Several attempts have been made to automatically identify hate speech...
- **Challenge: separating hate speech from offensive speech**
    - Hate speech changes forms, uses different expressions
    - Keyword based identification techniques do not work well

MACHINE LEARNING APPROACH

- Use **Natural Language Processing** and **Machine Learning** techniques in order to identify hate speech
- **Main stages** in Text Analytics with Machine Learning using **Neural Networks**
  - Transform a **document** (social media article) to **numbers** (features)
  - Create a **Machine Learning model** based on your architecture of your choice and **train** it as follows:
  - Model **looks** at documents and calculates and answer
  - Model **compares** answer against the correct one (annotated by human experts)
  - Model **tweaks** itself so that if it finds the same document again it would be more likely to calculate the correct answer
  - **Repeat**

DOCUMENT -> NUMBERS

- **How to transform a document to numbers?**
  - A) Convert a **word to numbers** (vector of a few hundreds of numbers)
  - B) Convert a **document** = lots of words **to math**
- **A) Word -> numbers: Word embedding algorithms**
- We need to retain some **semantic properties** of the word such as similarity
  - Words that share common contexts are close as mathematical objects
  - King - Man + Woman = Queen
- We have several options of such algorithms and **pretrained** models (Wikipedia,...)
  - Word2vec (Google), FastText (Facebook), ...
- **B) Document -> numbers: math operations on words, can include NLP**

MODEL ACHITECTURE

- **Model architecture: Convolutional Neural Networks on top of character level word embeddings**
- **Input** : Document
- **Output**: Hate or not (or additional classes such as personal attack, etc)
- **Challenge:** we need **lots of annotated (labeled) documents to train** (a nd test) algorithms

ACTIVE LEARNING APPROACH

- Active Learning is a methodology that can greatly **reduce the amount of annotated data** required to train a machine learning model.
- It does this by **prioritizing** the labeling work for the experts .
  - The model **looks** at unlabeled data (cheap)
  - **Identifies which data** it is **most confused about** and **requests labels** from experts for just those (human is involved - expensive).
  - It **trains** on that small amount of labeled data

ACTIVE LEARNING - BETTER DATA > MORE DATA

- **Minimize costs/time** by reducing expert resources
  - Get the same or better classifier accuracy usually with **3 to 10 times less effort**
- **Better data** is **more useful** than **more data...**
- The best active learning based model is the one with **great accuracy** and the **minimum number of requests towards experts**

CURRENT STATE OF HATE SPEECH DETECTION

- **Social media companies** work extensively on hate speech detection
- Only **small datasets** are available in the research literature, even for the english language
- Different **hate speech definitions**, intensify the above problem
- Several researchers and research groups work on the subject, with a current **success rate of approximately 80%**

OUR VISION

- **Hate speech**
  - Create a scalable active learning based system for identifying hate speech with direct feedback from journalists
  - Support multiple languages (English, French, German, Greek, Spani sh)
- **DataScouting Media Intelligence Products**
  - Include technology from the project to DataScouting's media intelligence solutions

Thank you

Eric Karstens

Funding Consultant

European journalism centre

karstens@ejc.net

Sophia Karakeva

Marketing and Communications Executive

Datascouting

soka@datascouting.com

Stavros Vologiannidis

Professor for the Technological Educational Institute of Central Macedonia and Founder

Datascouting

svol@datascouting.com

European Journalism Centre

DATASCOUTING
Actionable Information