SAFENET: CHALLENGES AND STRATEGIES IN COMBATTING ONLINE HATE SPEECH

Peter Trolle Luca Rossi Christian Hardmeier

IT University of Copenhagen

SafeNet

- 21 partners in 18 European countries
- Aim to report 16,000 cases of hate speech on social media and monitor responsiveness
- Engage with stakeholders to expedite removal

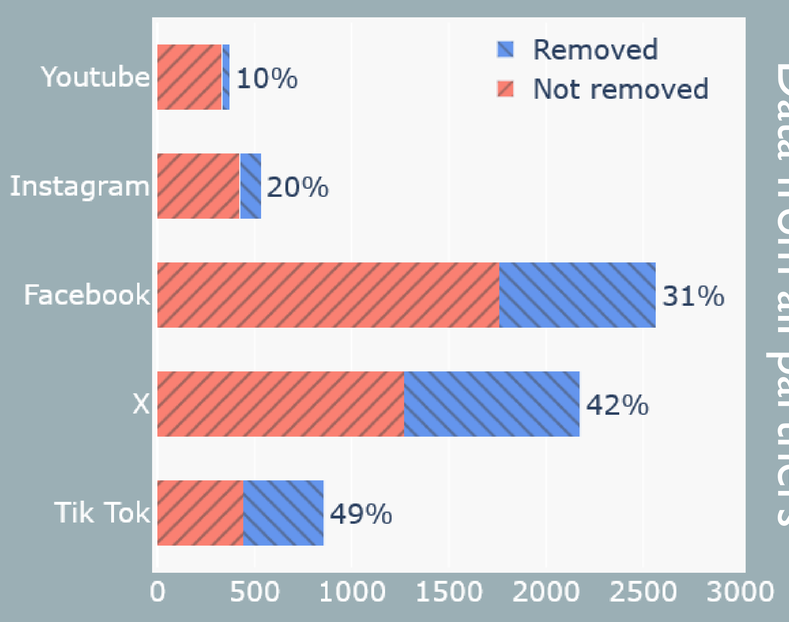
Issues

- Low removal rates
- Applicability of our definition of hate speech
- Selection bias of different types of hate speech

Results so far:

6507 reported cases
4285 replies
2282 removals
Removal rate: 35%

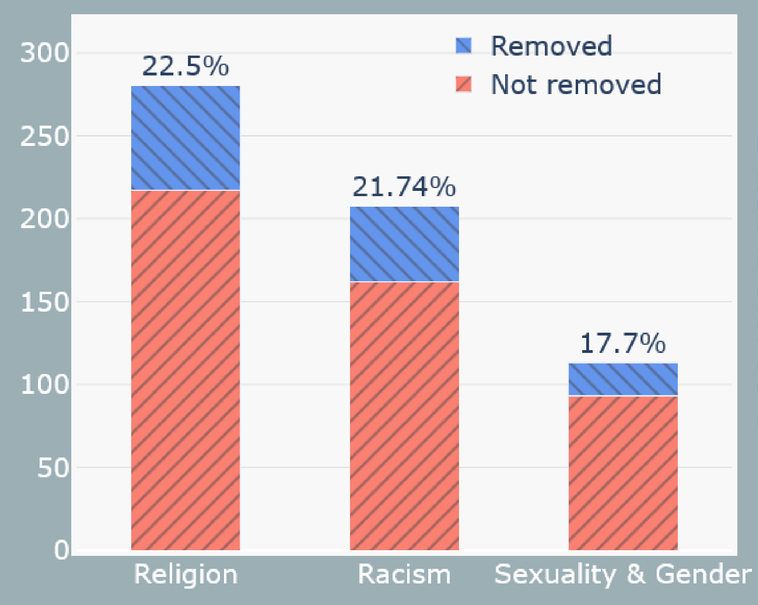
Removal Rate by Social Media



Data from all partners

Lower rates due to
- Removal policies?
- Lower severity?
- Something else?

Removal Rate by Hate Type



Data from ITU

Lower rates due to
- Reporting tolerance?
- Removal policies?

Hate speech definition & applicability

Very long

Complicated wordings

Vague terms

Official definition

According to the Committee of Ministers of the Council of Europe, hate speech is understood as « all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race", colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation ».

Members of the SafeNet consortium also define hate speech as « intentional or unintentional public discriminatory and/or defamatory statements; intentional incitement to hatred and/or violence and/or segregation based on a person's or a group's real or perceived race, ethnicity, language, nationality, skin colour, religious beliefs or lack thereof, gender, gender identity, sex, sexual orientation, political beliefs, social status, property, birth, age, mental health, disability, disease. Hate speech also includes intentional or unintentional public discriminatory and/or defamatory statements; intentional incitement to hatred and/or violence and/or segregation of a person or persons based on their real or perceived belonging to a certain group or community, or lack thereof. Hate speech also includes intentional and public apologist arguments, negationism, revisionism and denial of genocides especially of the holocaust and other crimes [illegible]; or crimes committed by certain oppressive political regimes ».

Extraction of key points

Shorter

Easier to read

Vague terms

Expressions that:
incite/promote/spread/justify violence/hatred/segregation/discrimination against a **person** or **group of persons**, or that **denigrates** them, based on their real or perceived **race/ethnicity/language/nationality/skin colour/ religious beliefs/gender/gender identity/sex/ sexual orientation/age/disability**, or **lack thereof**.
It also includes intentional or unintentional **discriminatory/defamatory statements**, **apologist arguments**, **negationism**, **revisionism** and **denial of genocides**.

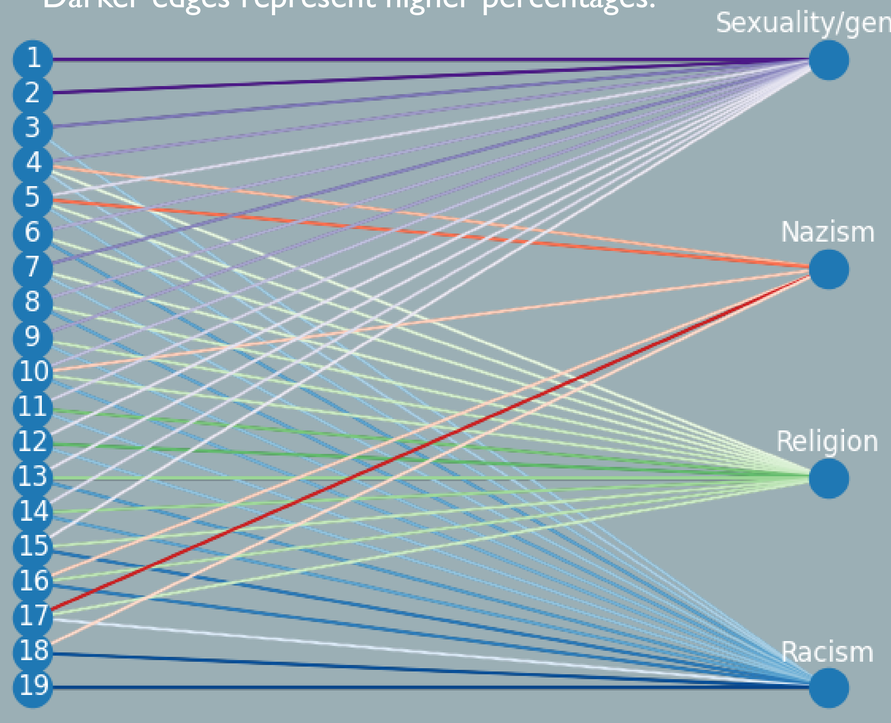
Training is necessary to ensure consistent application of the definition
- How to do it efficiently across many partners and cultures?

Selection bias

- Specialisation bias
- Cultural/geographical bias
- Sampling bias

How to ensure balanced coverage?

Top 95% of cases from each organisation included.
Darker edges represent higher percentages.



< 5% of reports for **4/19** partners

< 5% of reports for **13/19** partners

< 5% of reports for **5/19** partners

< 5% of reports for **2/19** partners

IT UNIVERSITY OF COPENHAGEN

SafeNet safer net for all

GA #: 101084457

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.