The University Of Sheffield.

GATE General Architecture for Text Engineering

Visualising Toxicity: Interactive Dashboards for Social Media Abuse Monitoring

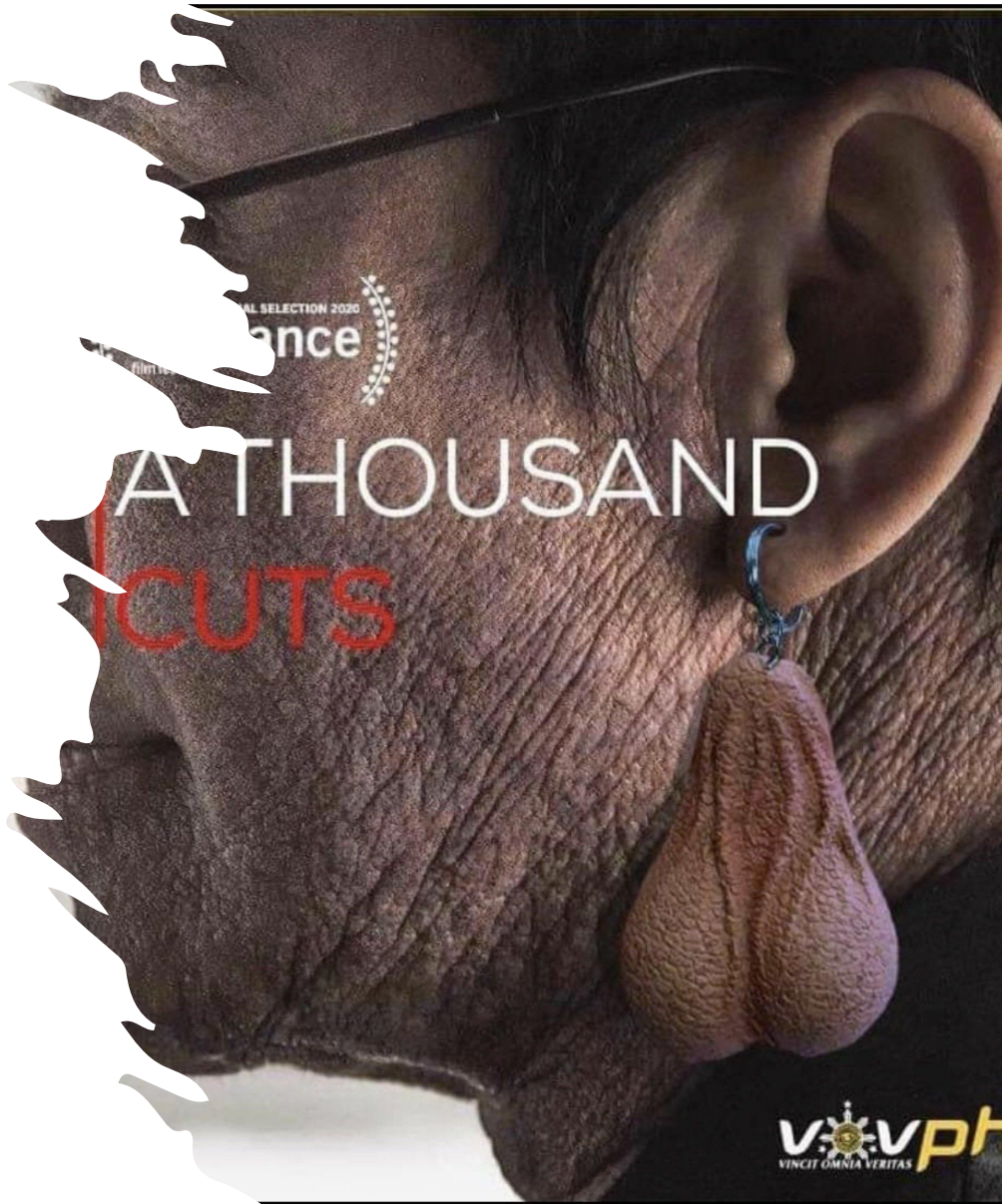
Dr. Diana Maynard

Dept. of Computer Science
University of Sheffield

CEOM Centre for Freedom of the Media

Maria Ressa scrotum-skinned, scrotum-looking, scrotum-minded, lives like a scrotum! You don't know math! Like her master, Leni [Robredo]!

Gender-based online violence against women journalists is one of the biggest contemporary threats to press freedom globally



What’s the problem?

- Methods of online abuse and disinformation are growing more sophisticated and evolving with technology.
- They are increasingly networked and fuelled by political actors.
- Need for responses to online violence to grow equally in technological sophistication and collaborative coordination.
- Most women do not report their online attacks or make them public
- People are reluctant to take it seriously.
- Failure of the internet communications companies - who facilitate much of the abuse - to take action

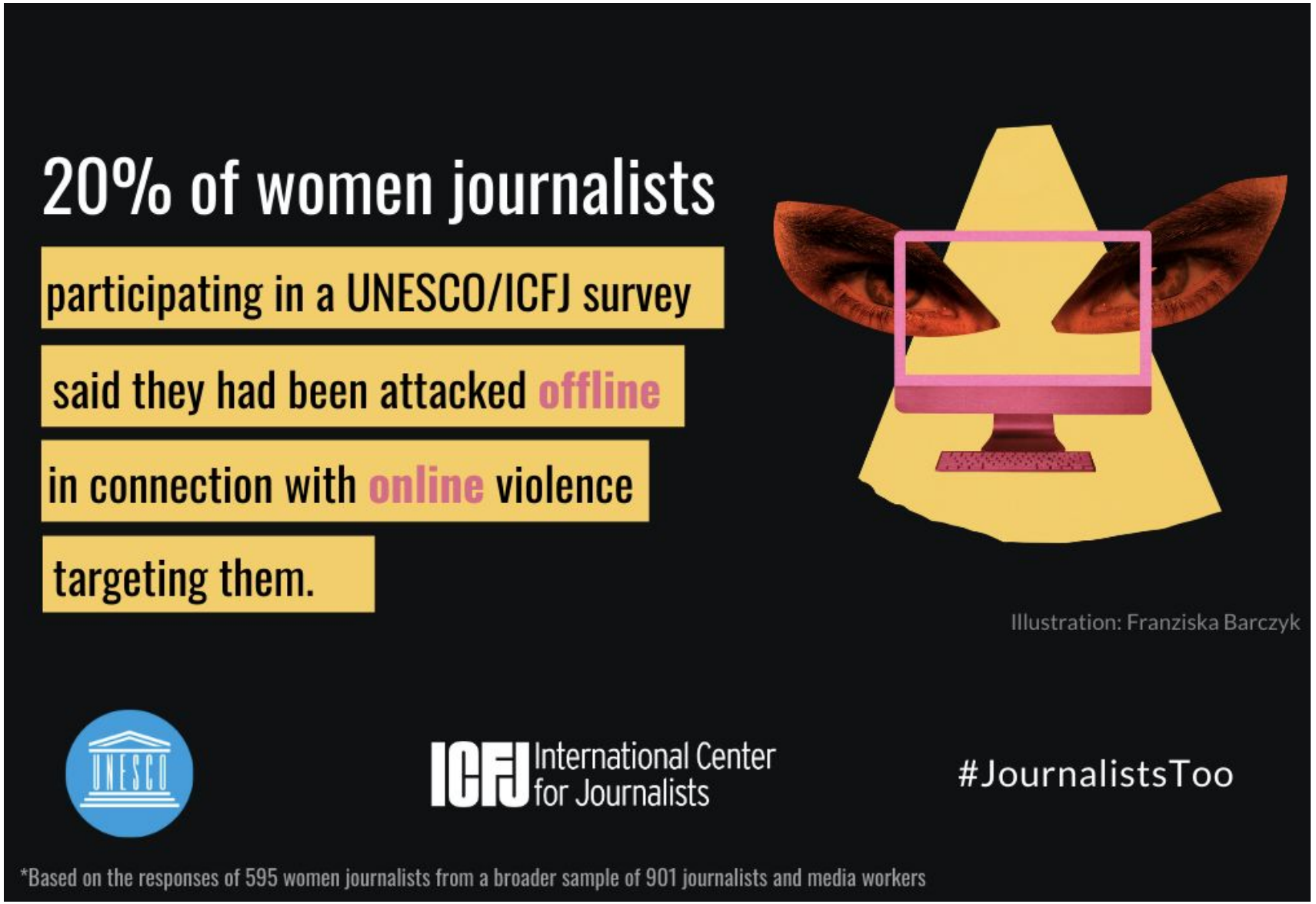
"death threats"

"Misogynist comments, sexual abuse … My children saw this"

"My staff try not to let me go out alone"

" They were also calling for her to be sexually assaulted, killed and even “raped repeatedly to death "

“But it’s just words, it’s not real abuse…”

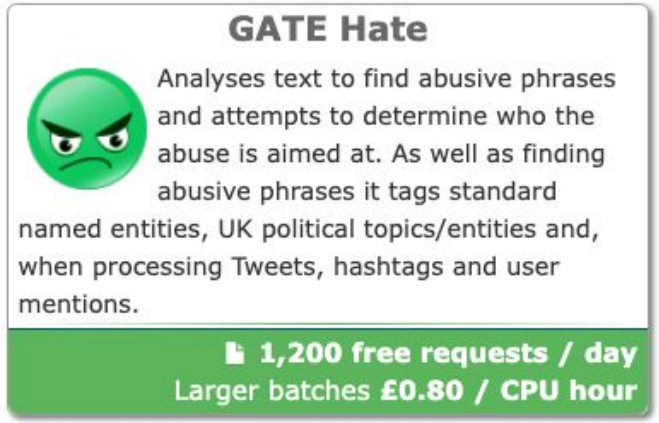


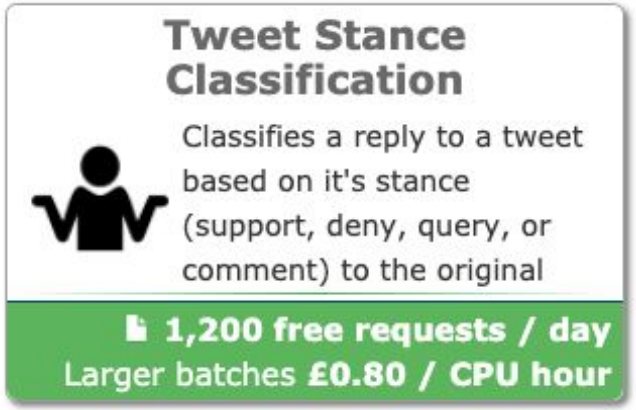
Who are we?

GATE General Architecture for Text Engineering

GATE team http://gate.ac.uk

- Research team based in the NLP group of the Computer Science Department at Sheffield University
- Developing tools for analysing language
- GATE toolkit developed since 2000
- Media and social media analysis, information extraction, abuse detection, misinformation, legal text mining, food and climate change research, medical and biomedical NLP, ….

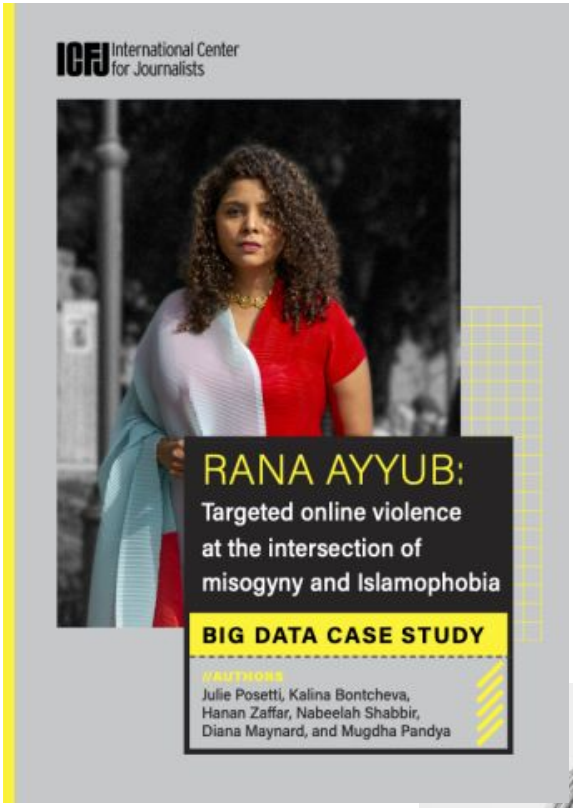




Musk: “I quite like trolls”

- Investigation into escalation of misogynistic hate speech since October 2022.
- Abuse against her tripled in that time
- Attacks increased as a result of social media protection and safety systems which they were failing to maintain
- Horrific threats and abuse in public domain, cross-platform, including graphic threats of rape and murder.

Big Data Case Studies

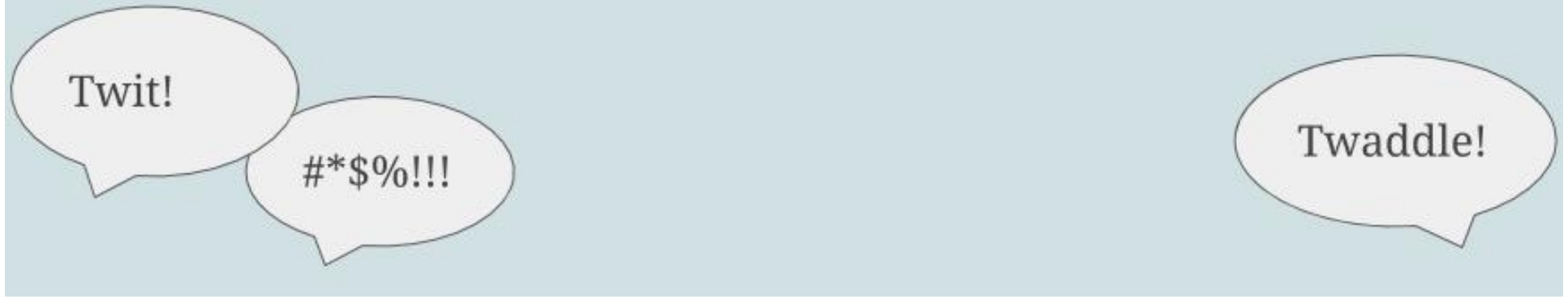


- Prominent Filipino-American journalist, co-founder of the Rappler digital news site and winner of the Nobel Peace Prize in 2021
- Has faced constant online threats and attacks since Rodrigo Duterte came to power in 2016, on account of her critical reporting
- Convicted of cyber-libel in 2020, her life is constantly at risk and she fought a number of other cases on trumped-up charges, for which she potentially faced decades in prison.
- Over 60% of the attacks were designed to undermine her professional credibility
- Typically also death and rape threats; doxxing; racist, sexist, and misogynistic abuse and memes.

- Lebanese journalist and an Al Jazeera Arabic principal presenter based in Qatar
- Ongoing target of brutal misogynistic online violence campaigns since the Arab Spring of 2011
- She didn't even have a social media account until 2014 (now >1 million Twitter followers and 2 million on Facebook)
- Now over one million followers on Twitter and two million on Facebook.
- Routinely threatened with rape and death, and smeared as a 'prostitute'.
- Also targeted because of her age, her employer's geopolitical vulnerability, and her Christian faith.

- Award-winning Washington Post columnist from India
- In 2010 her high-impact undercover investigation into India's Gujarat Riots, in which the Prime Minister Modi was implicated, helped to get Amit Shah arrested.
- An army of trolls evidently aligned with the ruling BJP party in India bombards her constantly with abuse on Facebook, Instagram and Twitter
- Misogynistic, disinformation-laced abuse, radiating also to her family, and shows signs of orchestration

- Multi award-winning British journalist whose investigative work helped to expose the Cambridge Analytica scandal
- Since then, she has become the target of a misogynistic disinformation-laced campaign of online abuse on Twitter
- This has become the enabler for the ongoing legal harassment by political actors
- Unlike the other case studies, a lot of the abuse is in the form of low-level but persistent gaslighting, questioning her authority and credentials as a journalist



Twits, tw@ts and twaddle:

analysis of hate speech towards public figures

Analysis of online abuse – what do we want to know?

- Who is being abused?
- Who is abusing them? How are the abusers connected?
- What is the abuse about?
- Why do people send abuse? What kind of people are they?
- Is it getting worse?
- How do people respond to abuse? (targets/bystanders)
- What are the effects of abuse?
- How can we prevent/mitigate it?
- How can we predict and prevent escalation?

So we built a dashboard to:

- Collect and analyse tweets in real time relevant to a particular person or group of people
  - Identify and characterise abuse directed at them)
  - Provide key statistics and graphs
- Provide a means for detailed **forensic analysis** combining automated tools and expert knowledge
  - users/hashtags/tweets/connections/abuse terms/abuse types/timeline spikes in volume/abuse spikes/events/topics/keywords
- Enable provision of comprehensive **evidence** of abuse
- Enable **alerts and warnings** of serious problems

Twitter Abuse Dashboards

This tool provides ways to explore collections of tweets and abuse directed towards different Twitter accounts. Select an account below to be redirected to the relevant dataset.

Rana Ayyub
India

Marianna Spring
United Kingdom

Marianna Spring
United Kingdom (Live Index)

Carmen Aristegui
Mexico

Carmen Aristegui
Mexico

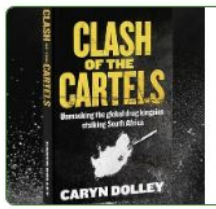
Ferial Haffajee
South Africa

Ferial Haffajee
South Africa (Backfill)

Pauli Van Wyk
South Africa

Marianne Thamm
South Africa

Rebecca Davis
South Africa



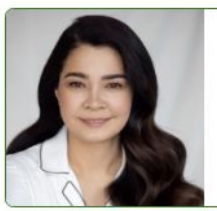
Caryn Dolley
South Africa

Ghada Oueiss
Lebanon

Caroline Muscat
Malta

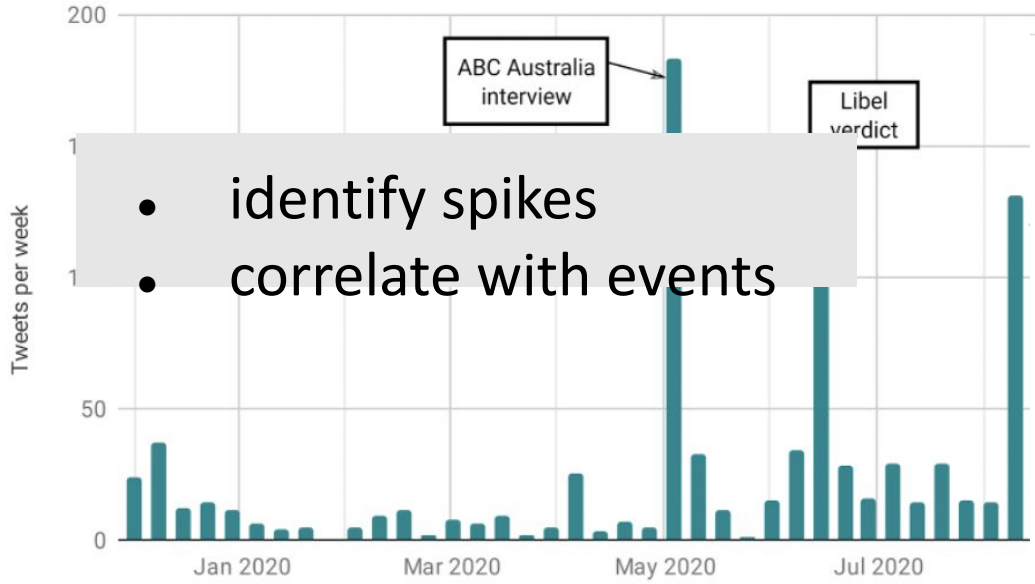
Katerina Sergatskova
Ukraine (Live Index)

Olga Rudenko
Ukraine (Live Index)



Sevgil Musaieva
Ukraine (Live Index)

Let’s Dive into the Data!



- identify spikes
- correlate with events

- examine abuser network
- identify coordinated behaviour

- investigate language, tropes
- attacks with psychological harm

- identify specific abuse types

Questions?