



The
University
Of
Sheffield.



Diving Deep into the Murky World of Social Media

Dr. Diana Maynard

Dept. of Computer Science
University of Sheffield



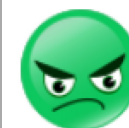
Who are we?



GATE team <http://gate.ac.uk>

- Research team of approx. 15 based in the NLP group of the Computer Science Department
- 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,

GATE Hate



Analyses text to find abusive phrases and attempts to determine who the abuse is aimed at. As well as finding abusive phrases it tags standard named entities, UK political topics/entities and, when processing Tweets, hashtags and user mentions.

📄 **1,200 free requests / day**
Larger batches **£0.80 / CPU hour**

Journalist Safety Analyser



Annotates descriptions of violations against journalists such as killings, threats etc. Identifies key information about the event and people involved.

📄 **1,200 free requests / day**
Larger batches **£0.80 / CPU hour**

Source Credibility



Annotates URLs to highlight the credibility of the source at which they ultimately point.

📄 **1,200 free requests / day**
Batch processing not available

Tweet Stance Classification



Classifies a reply to a tweet based on it's stance (support, deny, query, or comment) to the original

📄 **1,200 free requests / day**
Larger batches **£0.80 / CPU hour**



Social Media

['sō-shəl 'mē-dē-ə]

A computer-based technology that facilitates the sharing of ideas, thoughts, and information through virtual networks and communities.

Social media is rapidly changing our lives

There are more than 4.5 billion social media users around the world.

Even nomadic herders in Mongolia



But isn't all this staring at screens bad for us?

Mental health and wellbeing:

- “**Little association**” between technology use and mental health problems (study by Oxford University published in May 2021).
- Children say social media allows them to **do the things they want to do** and keeps them **entertained and feeling happy** (research from the Children's Commissioner)
- Social media can make young people feel **less lonely** (research by Talk Talk)

Literacy & creativity:

- People actually read more now than ever before, even if it's on a screen rather than a book or newspaper
- Audiovisual platforms like Youtube, TikTok and Instagram focus on creative arts (music, dance, photography)

Social Media and Disaster Relief

- Hurricane Sandy in the US:
 - 1.1 million tweets in the first day; over 20 million in total
 - > 800K photos with #Sandy hashtag on Instagram
- Haze in Singapore: > 23 million
- Nepal earthquake in 2015: more than half a million posts



Social media comes to the rescue

Sign In | Register  0

SCIENTIFIC AMERICAN™

Search ScientificAmerican.com

How Social Media Is Changing Disaster Response

Congress is grappling with the benefits and risks of using Facebook, Twitter and other social media during emergencies

By Dina Fine Maron | June 7, 2013

CNN News Regions Video TV Features

World Sport Technology Entertainment Style Travel

Heading off disaster, one tweet at a time

By Jim Spellman, CNN

September 22, 2010 — Updated 1713 GMT (0113 HKT) | Filed under: [Social Media](#)

ForbesBrandVoice Connecting marketers to the Forbes audience. [What is this?](#)

BUSINESS 10/29/2013 @ 8:17AM | 2,894 views

Social Media's Role In Disaster Response Improves Over Organizational Resiliency

theguardian
Winner of the Pulitzer prize 2014

home UK world politics sport football opinion culture business life  all

[Global development professionals network](#)

Social media, crisis mapping and the new frontier in disaster response

During a crisis the large amounts of data produced can be overwhelming to analyse. Microtasking could help solve that problem by harnessing the power of the crowd

 **ALJAZEERA**

NEWS ▾ PROGRAMMES ▾ OPINION INVESTIGATION

Topics: [Greece](#) [Boko Haram](#) [Egypt](#) [Cuba](#) [ISIL](#)

OPINION

When disaster strikes, count on tweets and data

Social media is increasingly becoming an important tool for emergency management and response

 **República**
República Dominicana
July 12, 2015

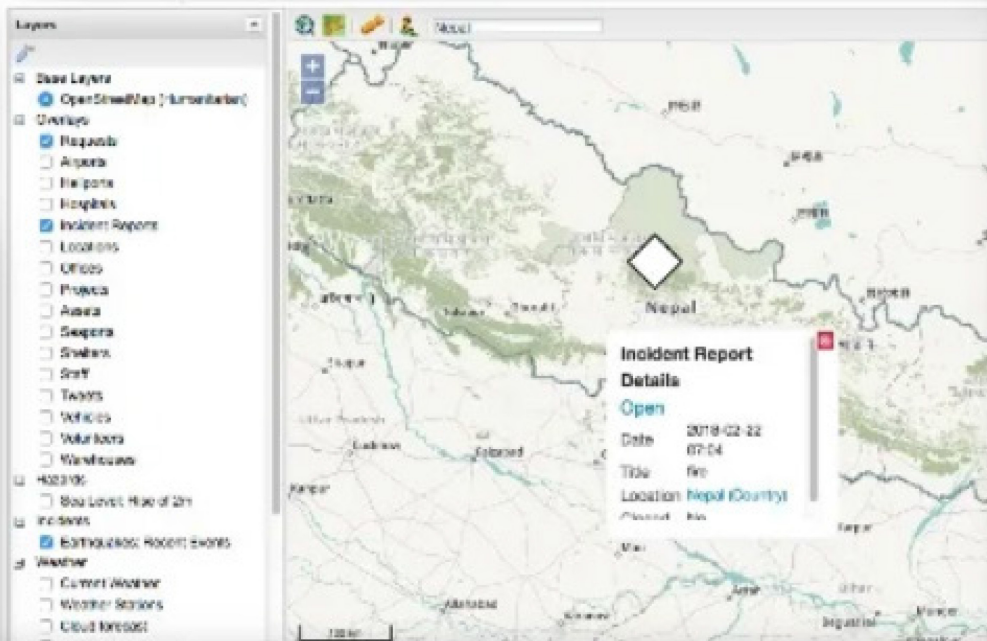
Earthquake rescue: Twitter for crisis communication

12 Jul 2015 | 20:16pm | | Dr Rajib Subba | 0 Comments

Aid workers in Nepal discussing strategy



CROWDSOURCING EMERGENCY RESPONSE



Google Person Finder

Kerala flooding

English

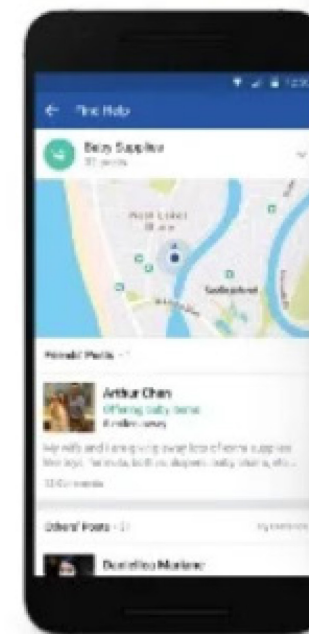
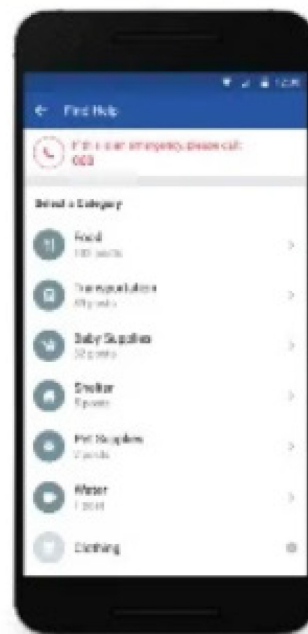
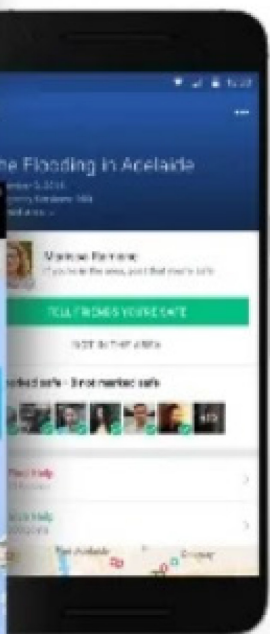
I'm looking for someone

I have information about someone

Currently tracking about 22300 records.

You can use this tool to help people find each other in the aftermath of the August 2018 floods in Kerala and nearby regions.

Please note: To request help or to find other rescue related information, please see keralarescue.in. For rescue phone numbers and additional information on the flooding, see Google Search.



Person Finder SMS

Find a Person:
Send an SMS to Person Finder by texting +91-

To search, text Search [name]. For example, to search for John, text Search John.

Add a person, to say you're okay:
Send an SMS to Person Finder by texting +91-

Text I am [full name]. For example, if you are John Doe and you are OK, text I am John Doe.

Tools to help disaster victims get aid quickly

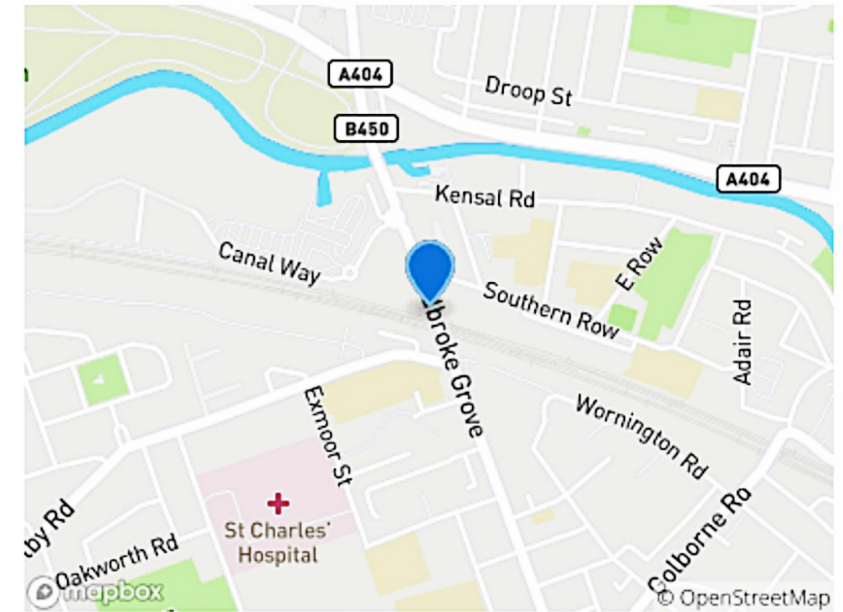
Problem:

- Many NGOs are not local to the disaster area and may not have a good grasp of the geography
- Place names are ambiguous

Solution:

- Find mentions of locations in the text, match them to a knowledge base, and plot them on a map

Location



Annotation from YODIE

A huge fire has engulfed a tower block in West London. The fire broke out shortly before 1am on Wednesday at Grenfell Tower in Latimer Road near Notting Hill. The tower is at least 24 storeys high and contains 120 apartments. 200 firefighters have been tackling the blaze with 40 engines. A number of people have been treated for a range of injuries according to the fire brigade. There have been multiple reports of people trapped in the blaze. These have not been confirmed by police or the fire brigade. Streets around the tower have been sealed off and residents in their houses evacuated. London

30 injured in tube train crash



Under review



Kenny

2 days ago

via Web



- How important and urgent is the message?
- What actions need to be taken?

Auto categories



related, derailment, affected_individuals

More than 30 people were taken to hospital after a crowded Tube train crashed into a tunnel wall, then hit the station platform, when it derailed in central London. Shocked passengers who were on board the Central Line train spoke of flying glass, and panic as they discovered some carriage doors were jammed shut. No-one was seriously injured. Fleets of ambulances surrounded the exit to the station as a major emergency services operation swung into action following the incident, which occurred at around 1.50pm on Saturday. A London Underground spokeswoman said it was believed all 800 passengers had been taken off the train and casualties had been taken to three nearby hospitals, but the Fire Brigade said crews were still searching the scene.

CREES Google Sheets Add-on

Adding semantic information (from BabelNet) improves cross-lingual crisis event classification

Notes	CREES		
Text of request	RELATED	EVENT	INFO
	non-related	none	sympathy_and_support
Young nurse needing rescue!	related	floods	donations_and_volunteering
	non-related	none	sympathy_and_support
Together we will rebuilt page - Celia Torres	related	floods	infrastructure_and_utilities
Only have bank address from Twitter	non-related	floods	other_useful_information
#disabled lady is #stranded in #LeagueCity #dickinsontexas #needrescue #help #hurricaneharveyanimalrescue	related	floods	affected_individuals
	non-related	none	sympathy_and_support
	non-related	none	sympathy_and_support
Been trying numbers for hrs - busy signal. Elderly couple desperately #NeedWaterRescue:10202 Willowgrove, 77035. Near S Post Oak & W Belfort	related	floods	other_useful_information
	non-related	none	sympathy_and_support
Can't get through on the coast guard number reported 2:19 pm CT	non-related	floods	affected_individuals

Understanding Social Media Behaviour: Intervention Strategies for Social Media Environmental Campaigns

Based on Robinson's 5 Doors Theory of Behaviour Change



Behaviour Analysis

- Users in different behavioural stages communicate differently
- We can map these to linguistic features to track behavioural change on social media

You are an organisation, not an individual!!



Pajarito @lindopajarito . 2h

Our building needs 40% of all energy consumed in Switzerland!

Desirability: Negative sentiment (expressing personal frustration-anger/sadness)



DJPajarito @DJPajaritoGenial . 12h

I'm so proud when I remember to save energy and I know however small it's helping

Buzz: Positive sentiment (happiness/joy). I/we + present tense



HotelPajarito @HotelPajarito . 18h

Join us today today to switch of a light for EH

Invitation: Positive sentiment (happy) + use of vocatives



Social Media Campaign Recommendations

- **Provide messages with very concrete suggestions on climate change actions**
 - Most users are in the desirability stage: they want to change but they don't know how
- **Identify really engaged individuals and community leaders and involve them closely**
 - Few users in the invitation stage; most are organisations
 - Exception is COP21 (a movement more oriented to act and change policy)
 - The right person issuing the invitation is vital to its effectiveness
- **Dedicate effort to engaging in discussions and providing direct feedback**
 - Communication in these campaigns generally functions as one-way broadcasting from the organisations to the public
 - Frequent and focused feedback can help build self-efficacy and nudge users in the direction of change

But now the
flipside....

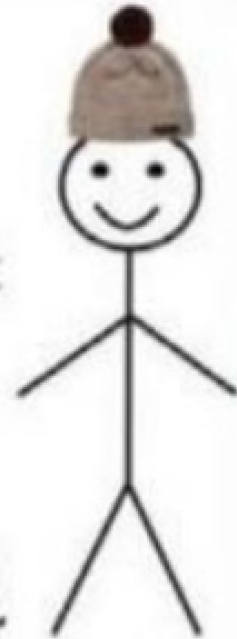
This is Bill.

Bill has **lots of opinions** about different things. Bill knows that other people have their own opinions about things too.

Bill doesn't think that everyone else is wrong just because their opinions aren't the same as his.

Bill is smart.

Be like Bill.



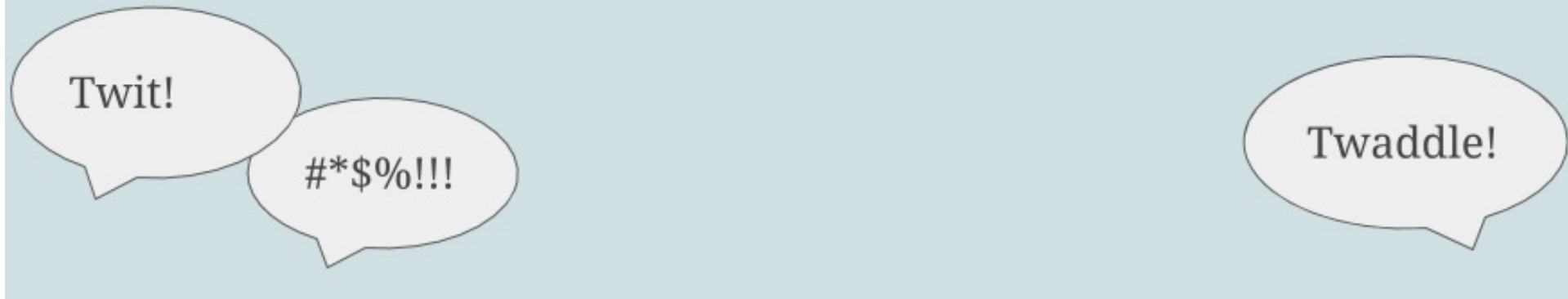
This Is Daphne Caruana Galizia.

Galizia was a journalist who helped lead the Panama Papers exposé, which showed that many rich people, internationally, were keeping their fortune in illegal offshore accounts. None of them were punished for it.

After the story broke, assassin planted a bomb in her car, killing her.

REMEMBER HER.





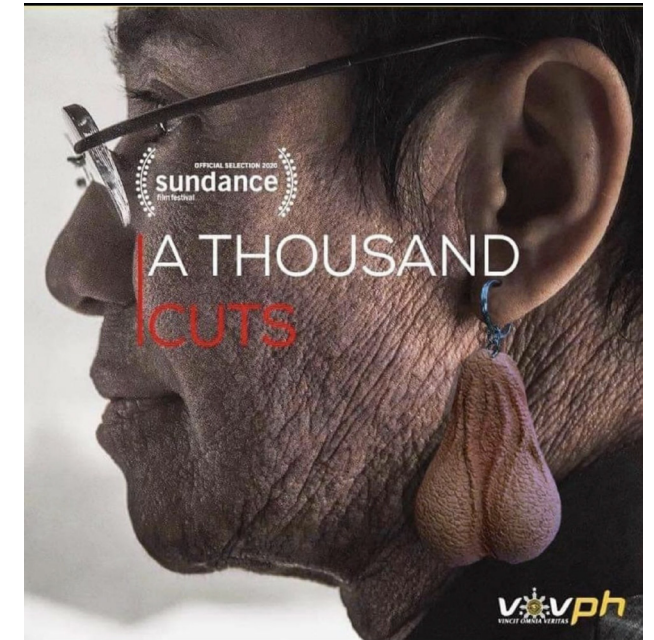
Twits, tw@ts and twaddle: analysis of hate speech towards public figures



Online abuse

- Puts people off debating online
- Puts people off becoming politicians, journalists etc.
- Seems to be getting worse
- Might be particularly bad for particular groups (females, ethnic minorities, LGBT etc)

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



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

"death threats"

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

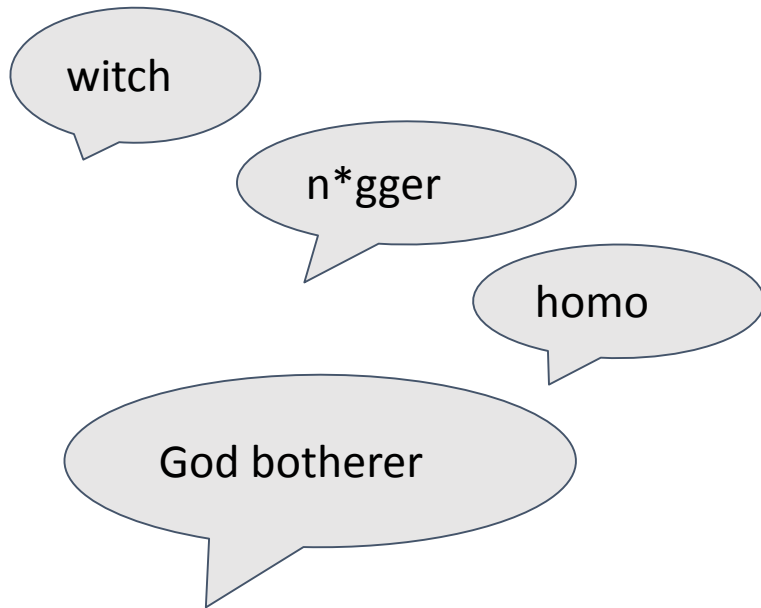
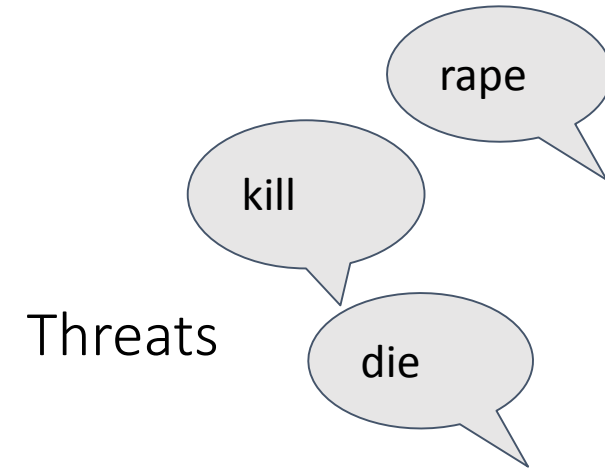
" They were also calling for her to be sexually assaulted, killed and even "raped repeatedly to death "

Analysis of online abuse

- Who is being abused?
- Who is abusing them?
- What is the abuse about?
- What kind of people send abuse?
- Is it getting worse?
- How do people respond to abuse? (victims/bystanders)
- How can we prevent/mitigate it?
- How can we prevent it escalating (eg to offline abuse)?
- How do abusers avoid detection and how can we mitigate this?

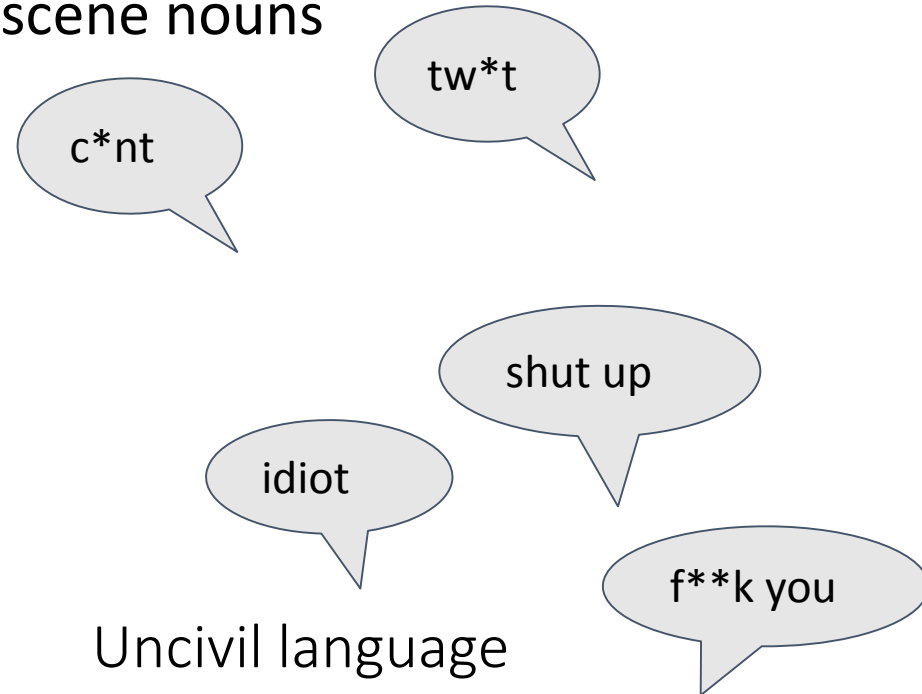
Finding abusive terms

There are a lot of offensive words and phrases that we can try to spot



Racist and bigoted language

Obscene nouns



Uncivil language

Identifying the right idiot

- Just mentioning the word “idiot” isn’t precise enough
 - “I’m an idiot” – self-abusive
 - “You idiot!” – abusive towards addressee
 - “What kind of idiot would do that?” – ambiguous
 - “They’re idiots” – abusive towards others
 - “Donald Trump is an idiot” – directed towards a specific person

Abusive terms are often found in hashtags

@theresa_may If you legalise #foxhunting I will fire an arrow through your face.
#promises #killthewitch

12:32 AM - 10 May 2017

@RanaAyyub Dont call urself Journalist, u insult the #Journalist #Community.. u r just a #Hypocrite #Fake #FakeNewsMedia #communal #antihindu #antinationa..

Separating out hashtags can be tricky

#powergenitalia

#lesbocages

#molestationnursery

#teacherstalking

#therapist

#expertsexchange

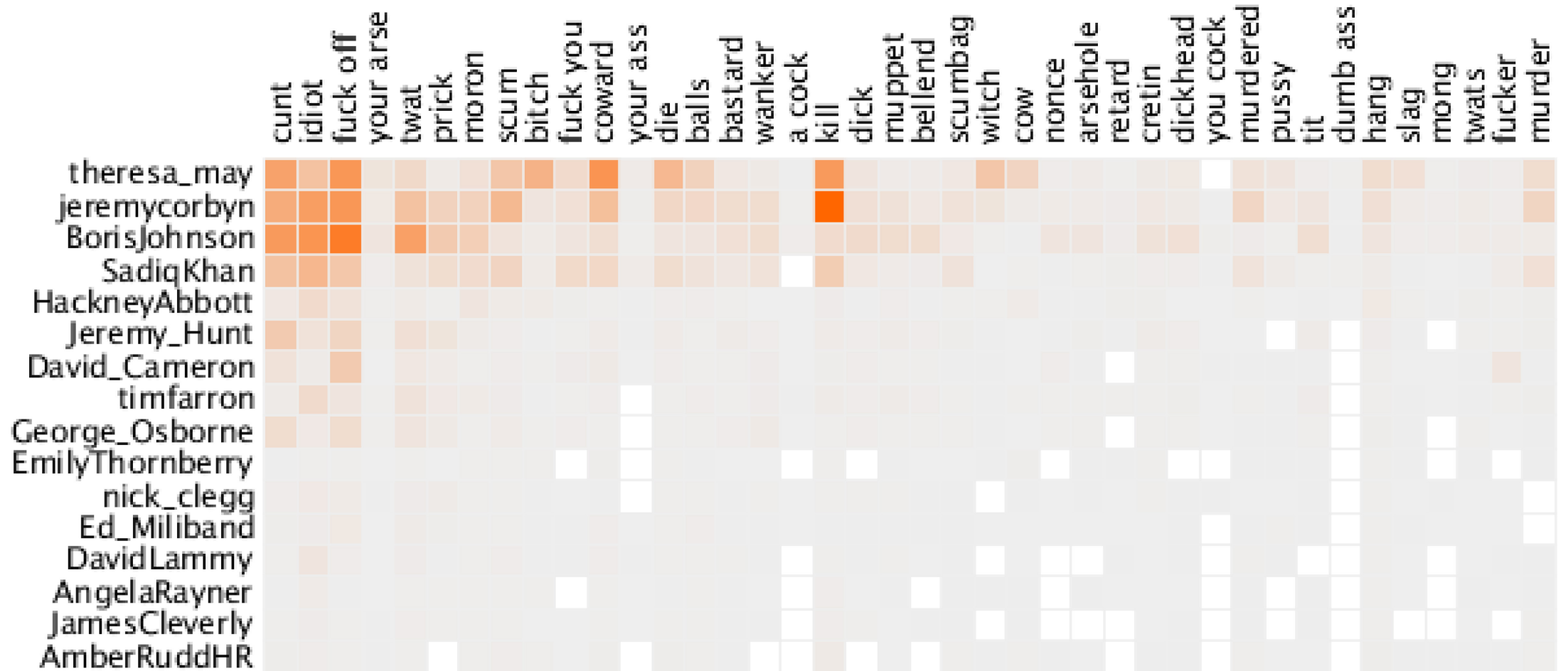


And we also have to be careful about language
– what about #slagroom?

Methodology for Analysing Abuse

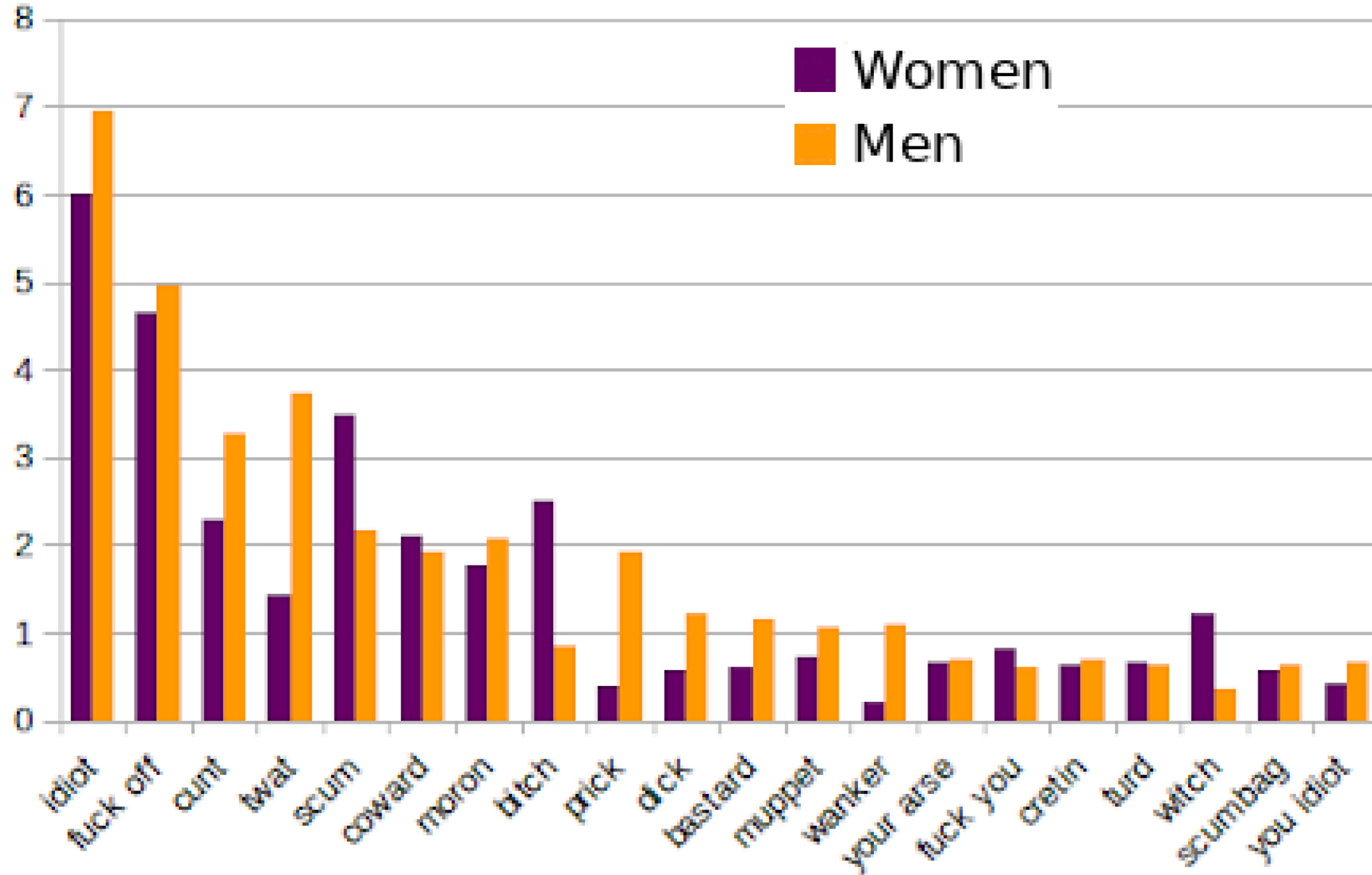
- Collect all tweets to, from and about our targets (e.g. politicians, journalists, footballers, etc.)
- Annotate all the interesting information (who, what, when, where) with our social media toolkit
- Run an abuse classifier
- Index all the information and apply sophisticated search
- Build a dashboard to visualize the results of queries
- Mixed methods approach to combine with contextual knowledge and a manual deep dive into interesting cases

Analysing the data



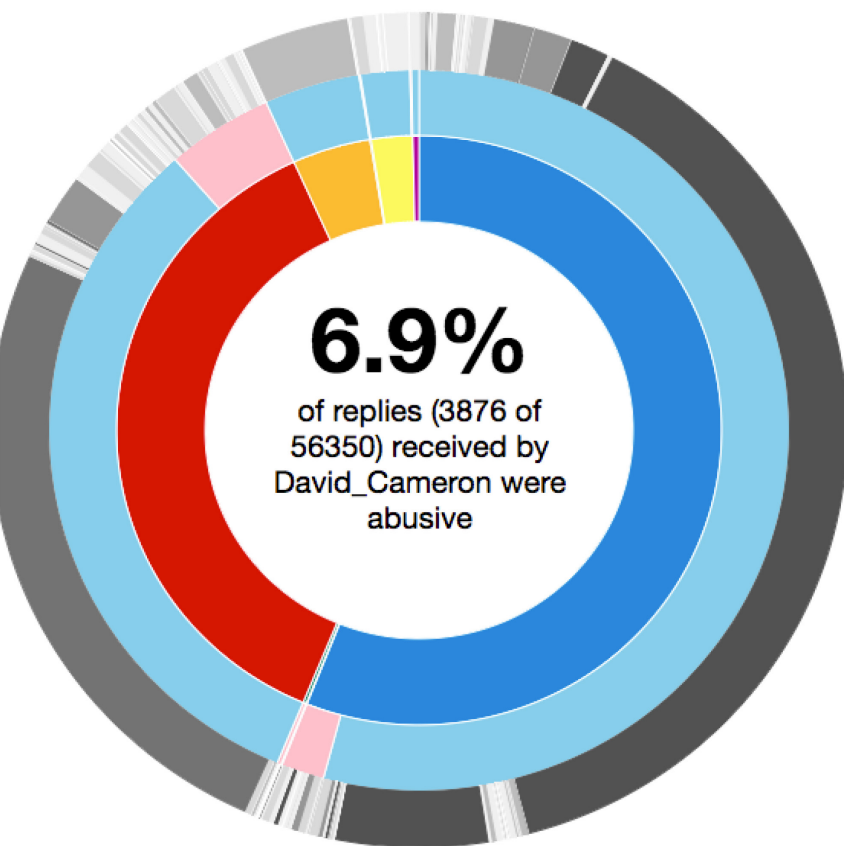
➤ Abuse differs in style when directed at men and women

Top abusive words directed at MPs

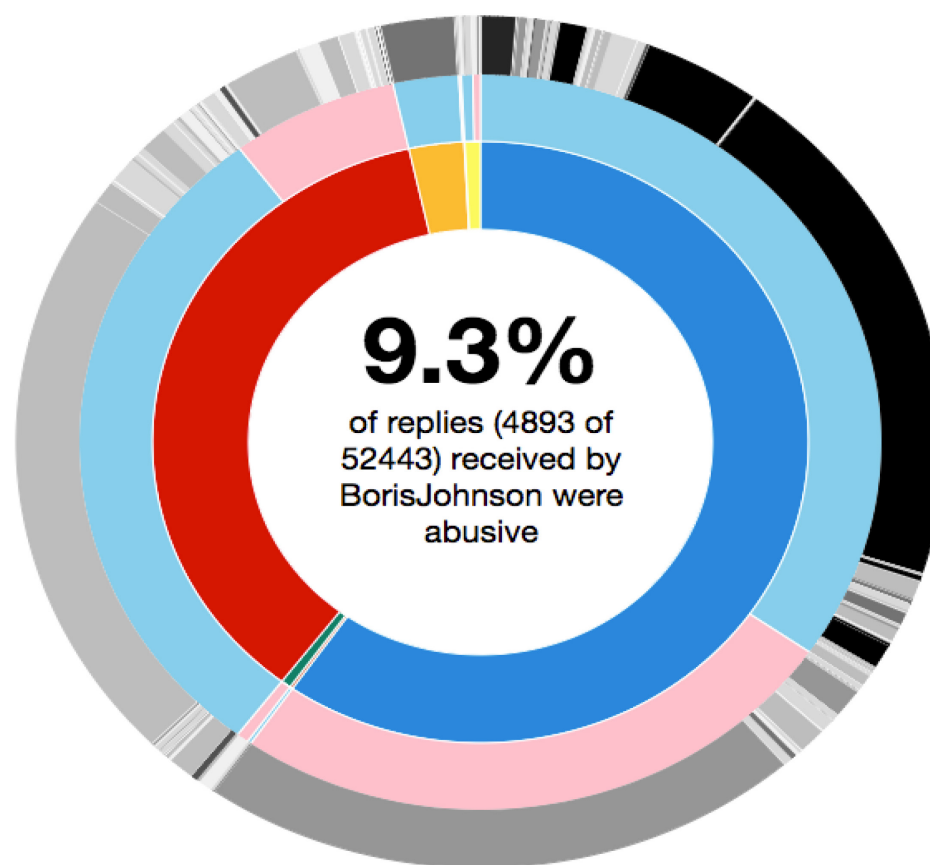


Quantifying online abuse directed at UK MPs

2015



2017

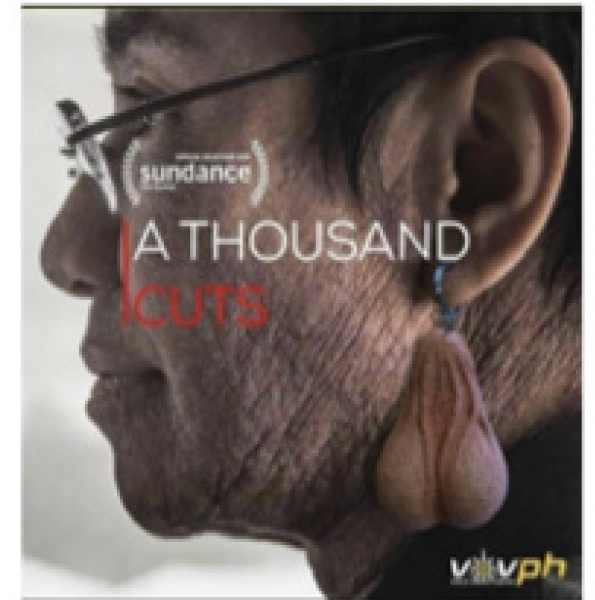


- There was more abuse in 2017
- Men got more abuse than women
- Conservatives got more than Labour

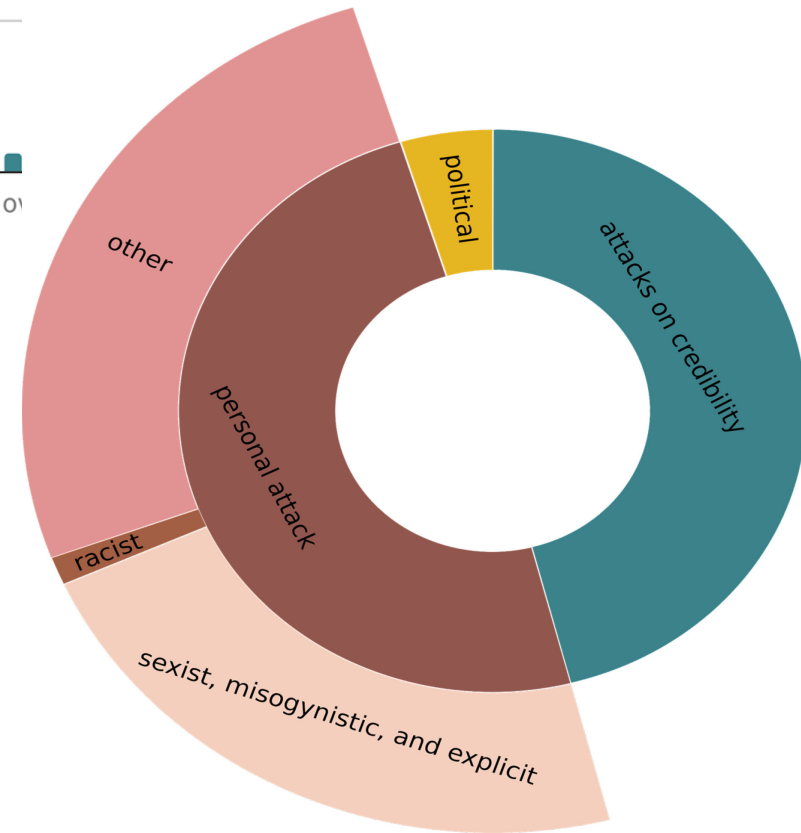
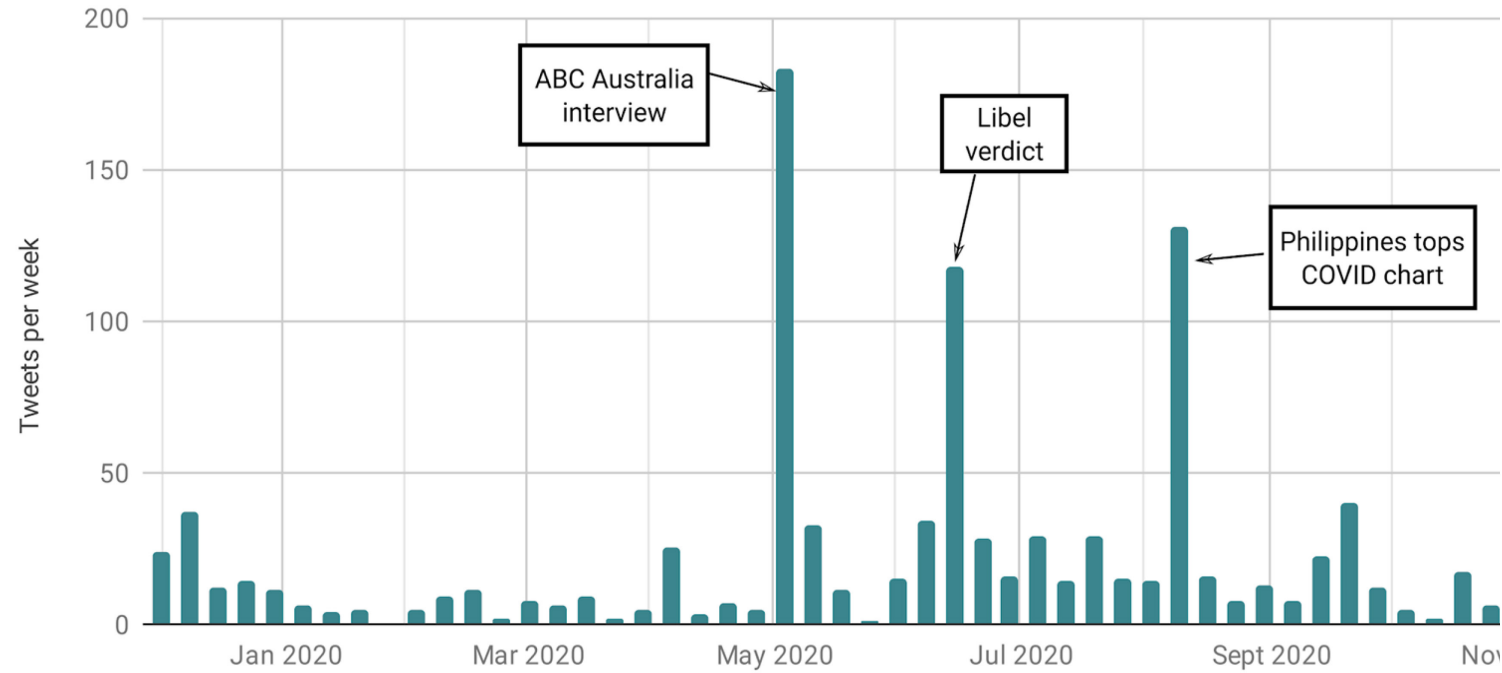
In the eye of the storm: 5 years of online abuse

- Using NLP, we analysed almost half a million Facebook and Twitter posts
- The most comprehensive assessment of online violence against a prominent woman journalist
- Almost 60% of the attacks were designed to undermine her professional credibility, frequently deploying disinformation tactics
- Over 40% of the attacks were personal
- Evidence of coordinated and orchestrated attacks
- Offline consequences: the enabling environment for her persecution, prosecution and conviction

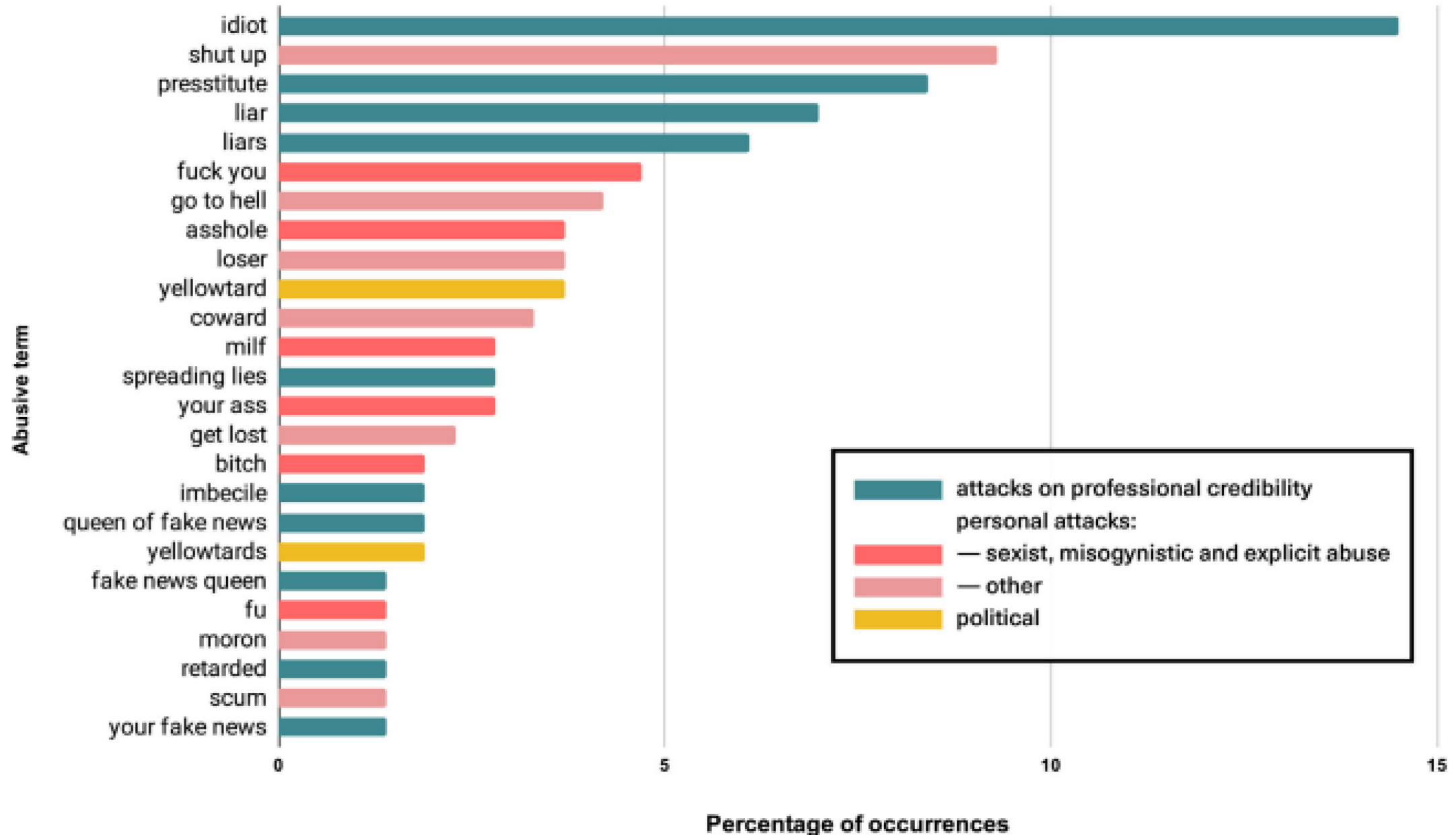
presstitute
#dicksucker
liars go to hell
scrotum face
#yestoshutdownrapper











Characterising Abuse Against Women Journalists

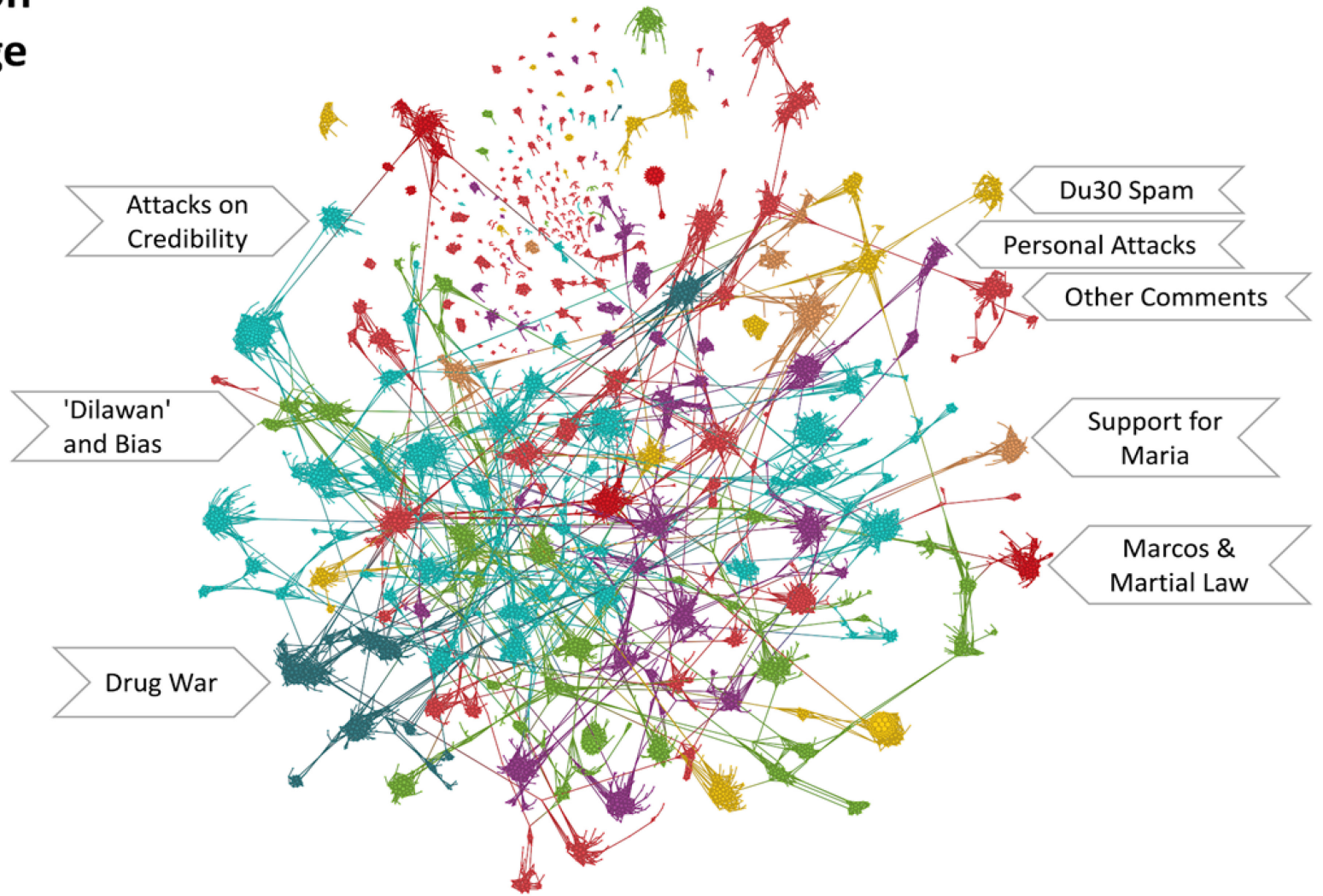


Top abusive terms in Facebook comments to Maria Ressa



Topic cluster of comments on Maria Ressa's Facebook page

Narratives		
	Attacks on Credibility	25%
	Other Comments	22%
	'Dilawan' and Bias	15%
	Personal Attacks	14%
	Du30 Spam	8.1%
	Drug War	5.6%
	Marcos and Martial Law	4.2%
	Support for Maria	4.0%



Typical methods

- Key significant attacks appear to be orchestrated (with the use of fake and bot accounts), and on occasion this has led Facebook to remove networks of accounts identified as participating in what they call ‘coordinated inauthentic behavior.’ However, the company’s response to the attacks on Ressa has been wildly inconsistent and “woefully inadequate,” in her words.

Hashtags designed to encourage swarms of attackers and fuel ‘patriotic trolling’ are frequently used, and sometimes include threats within them e.g., #ArrestMariaRessa.

- Memes and manipulated images are deployed to increase engagement with the attacks on Ressa and avoid automated abuse detection tools.
- Doxxing (publishing private or personal identifying information) is used to motivate Ressa’s online attackers to attack her offline as well.

Abuse and Disinformation

#presstitute

#FakeNewsQueen

- Many people are distrustful of journalists generally.
- Journalists often get accused of lying or being bad at their job in order to discredit them

“Carole Codswallop”

- Journalists who report on disinformation get a lot of abuse



Marianna Spring ✓
@mariannaspring

lying bitch



After reporting first-hand on online conspiracy theories about covid promoted at protests, I have been bombarded with horrific abuse and threats.

I am very happy the BBC has a specialist disinformation reporter investigating their impact. No-one should receive this vitriol.

A few thousand? More like a million you daft cow.

One day your children will resent everything about you, the choices you made and the choices that affected their future and their friends. Hell awaits you.

Marianna I for one believe you should have the 'vaccine' As soon as possible

21:48



@mariannaspring Your an arsepiece



@mariannaspring And yet you punt them daily ya fucking mutant

(Nice rack btw, put that on your news!)

You are a simple piece of sh*t who works for a bigger sh*t! A simple licky-licky slave! Ugly in and out disgusted!

You're a miserable cunt, get laid.

You're part of the propaganda machine Marianna, you're a traitor to your country.

These traitors seem to not take to account we all have firsthand experience of who was on what side of history - the traitors they are.

We will never forget.

9:02 AM · Apr 25, 2021





Congratulations MARIA RESSA. Ikaw na ang reyna ng Fake News.

[Translate Tweet](#)



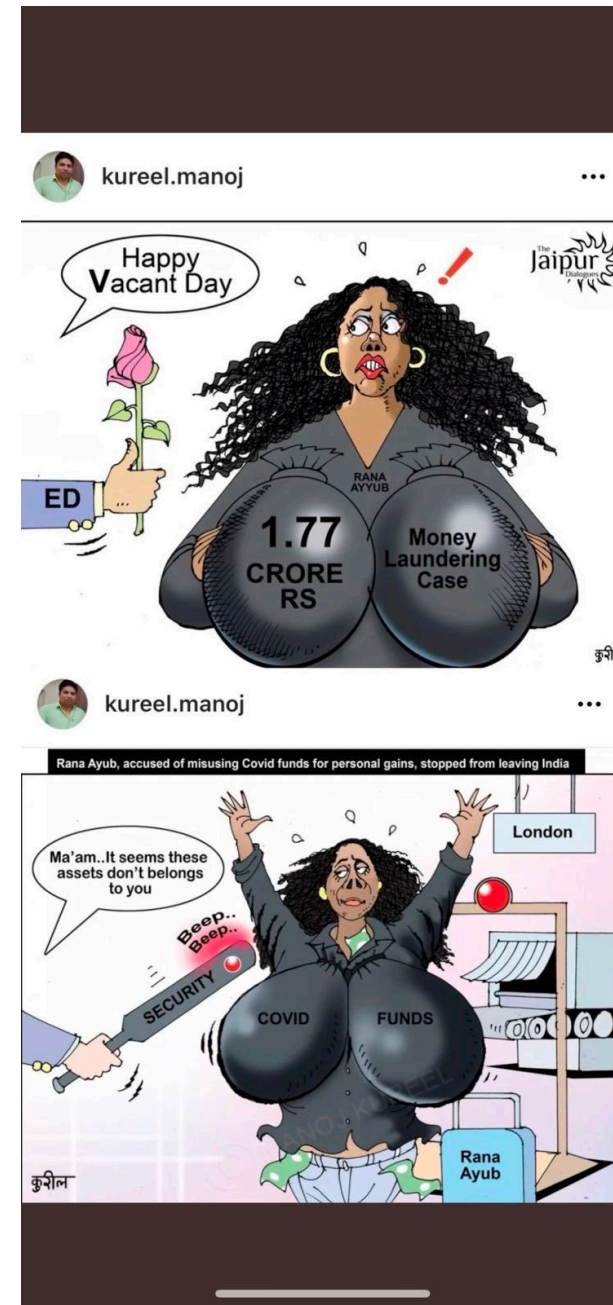
2:08 PM · Jun 15, 2020 · Twitter for iPhone

102 Retweets 63 Quote Tweets 451 Likes

But many problems!

- Under the radar abuse (especially manipulated images)
- Subtle gaslighting (targeted at credibility and reputation, effect only seen over time)
- Contextual references
- Correct identification of abuse target

@carolecadwalla @Nigel_Farage Dear Arron Banks I am also sorry. Sorry you are a slippery cunt.



Understanding the Escalation of Violence



Illustration credit: Franziska Barzyck; UNESCO

20%

of women journalists surveyed in a recent UNESCO-ICFJ survey said they had been attacked or abused **offline** in connection with online violence they had experienced.

Now, we're launching a research project to develop an early warning system to help detect, predict, and ultimately prevent violence against women journalists.

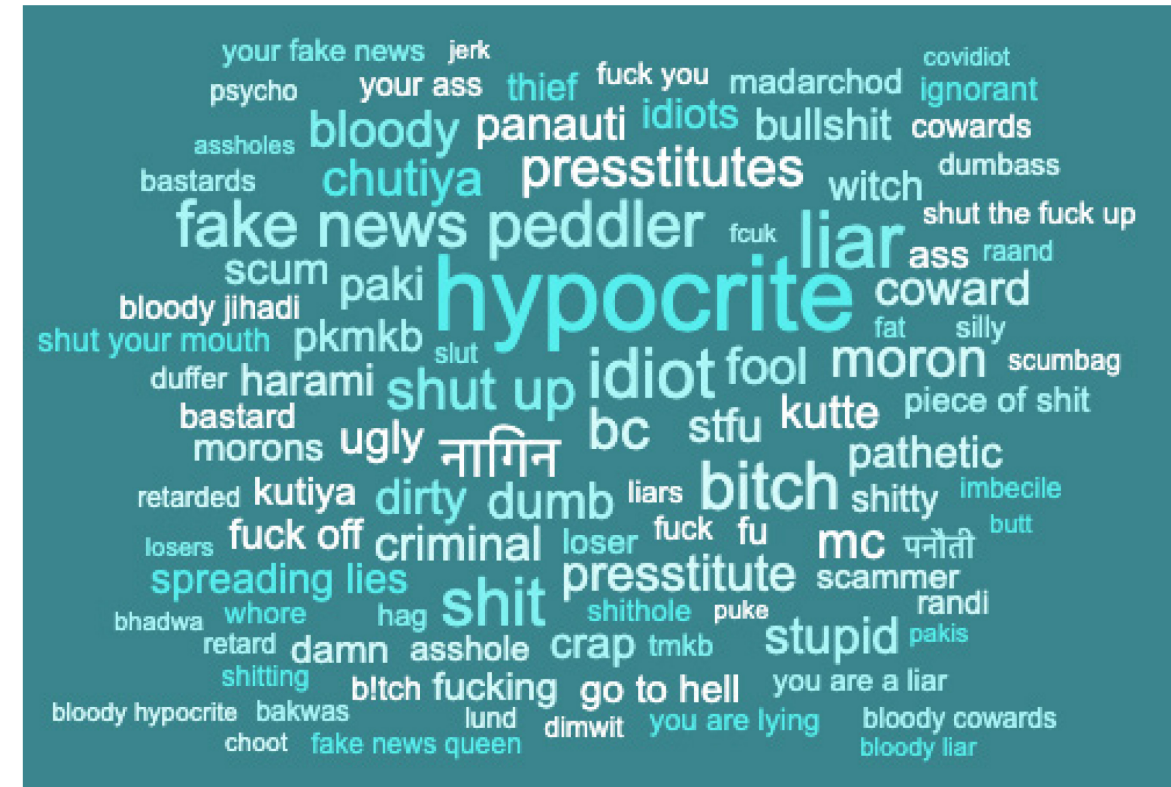
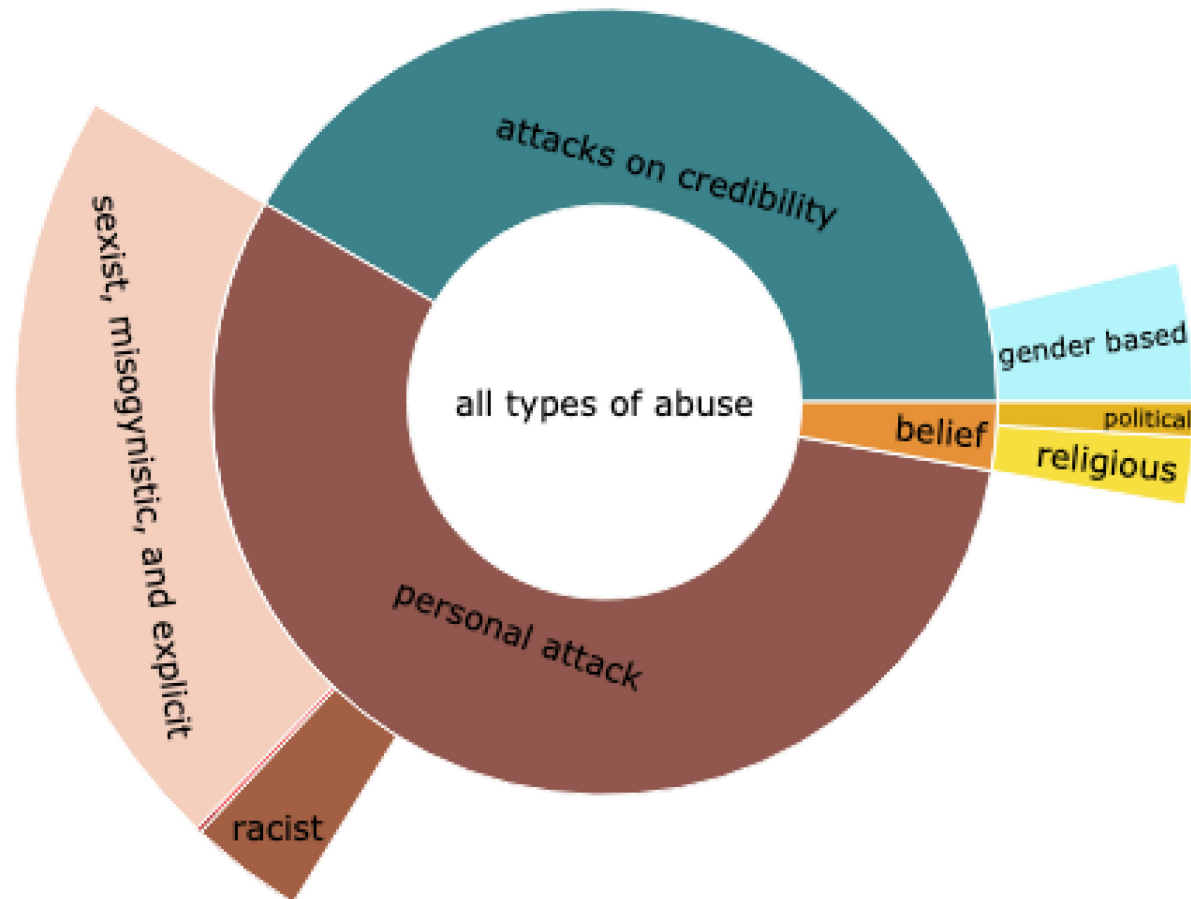
ICFJ



The
University
Of
Sheffield.

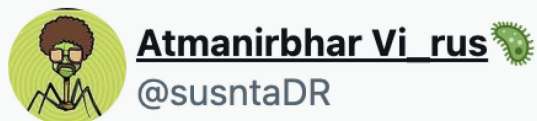
Abuse against Rana Ayyub

- Analysed 13 million tweets sent over the last 2 years
- >44,000 blatantly abusive tweets



Racist

ANALYSE TWEET



Atmanirbhar Vi_rus

@susntaDR



Replying to @RanaAyyub

Shameless paki living here with managed
Endian passport

8:01 AM · Jun 29, 2020



(May 20, 2020, 2102) Retweets

>40% of Rana's tweets had at least
1 abusive reply

Sexual

ANALYSE TWEET



पी के

@bullbule_



Replying to @RanaAyyub

Mocking indian products.....ok ok I got
it.....even ur D!LDO is made in china.....

Must be printed on it..

In the Name of Jihaad.....

Keep it up....up

7:54 AM · Jun 29, 2020



Unique Tweets

Retweets

Jan 2020 Apr 2020 Jul 2020 Oct 2020 Jan 2021 Apr 2021 Jul 2021 Oct 2021 Jan 2022

What Triggered the Abuse?

Tweet URL	All Replies	Abusive Replies	% Abuse	First Abusive Reply
https://twitter.com/ghadaoueiss/status/1453169069314789376	686	24	3.50%	1h 28m 17s
https://twitter.com/ghadaoueiss/status/1459620996089004035	516	13	2.52%	14m 14s
https://twitter.com/ghadaoueiss/status/1465239554181447686	368	12	3.26%	30m 36s

TRANSLATE TWEET



Ghada Oueiss غادة عويس
@ghadaoueiss · Follow

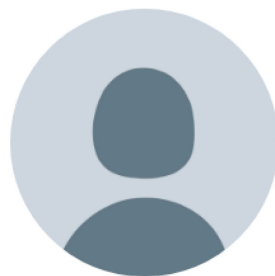


ويخلط جورج قرداحي بين الشعب اليمني وحزب او حركة
يمنية سياسية دينية مسلحة.
لكنه حزب يناسب اصطفااته فانجاز له وهاجم أعداءه
بسذاجة قل نظيرها جعلته اليوم والحكومة التي ينتمي لها
في حرج لا يحسد عليه.
ليته اكتفى ببرامج المسابقات وابتعد عن السياسة وأعفى
نفسه وحكومته المأزومة من الإحراج

2:17 AM · Oct 27, 2021



3.8K Reply Copy link



Ghada Oueiss غادة عويس
@ghadaoueiss

884 Following

1.05M Followers

Reputation

TRANSLATE TWEET

ANALYSE TWEET



عبد الخالق منصور
@RE3yGSrsm1W75gu · Follow



Replying to @ghadaoueiss

منافقه وناعقه.
جورج رجل حر ابدى موقف شخصي له من حرب
عبيته ظالمه فما الضير في ذلك وانتم شعاركم الرأي
والرأي الاخر

4 AM · Oct 27, 2021



Reply



Read more

Responses

Filter by

TYPE OF ABUSE

Reputation ☒

Sexist ☒

Sexual ☒

General ☒

No Abuse ☐

CONTAINS

Hashtag ☐

30 Tweets

Sexual ☒

TRANSLATE TWEET

ANALYSE TWEET



الشاعر الجلال
@samd2030 · Follow



Replying to @ghadaoueiss

حتى أنتي يا حمالة الحطب حشرتي نفسك في
مؤخرة بن سلمان وبن زايد

3:46 AM · Oct 27, 2021



Reply

Copy link

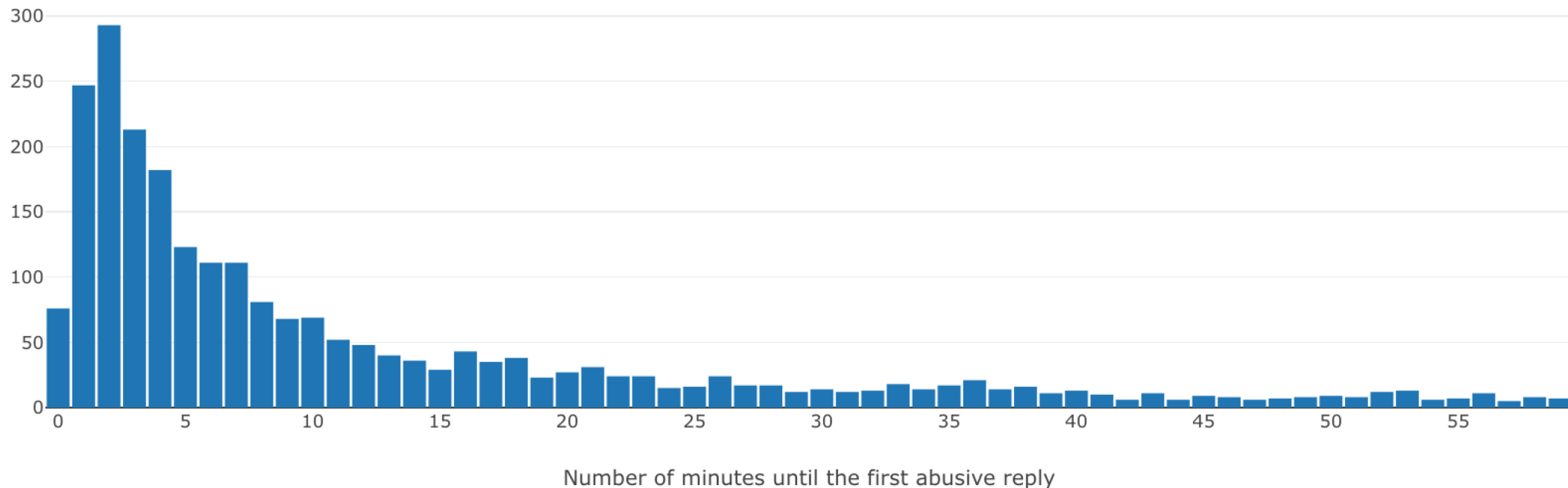
Read more on Twitter

Hashtags

#جورج_قرداحي
#البحر
#سوريا
#كوهين
#برلمان_شعب
#درعا
#جورج
#الدنوع
#ترامب
#جورج_قرداحي
#المملكة_السعودية_أرضنا_المقدسة
#متابعين
#قرداحي

How quickly do abusive replies get sent?

- For Rana Ayyub, the first abusive reply often comes almost instantly
- Some occur within 2 seconds
- The majority occur within 3 minutes
- Strongly suggests these are **co-ordinated attacks**



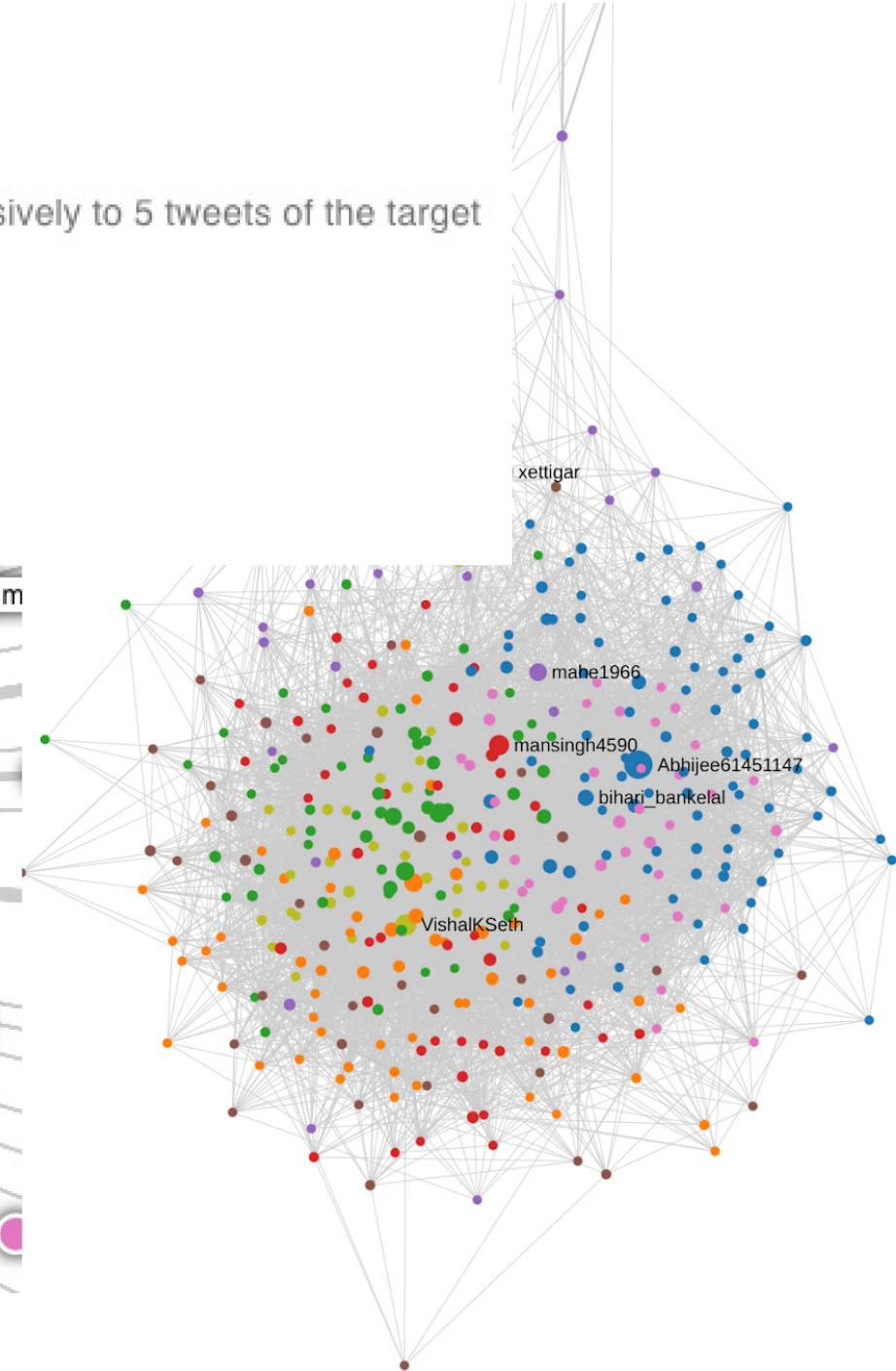
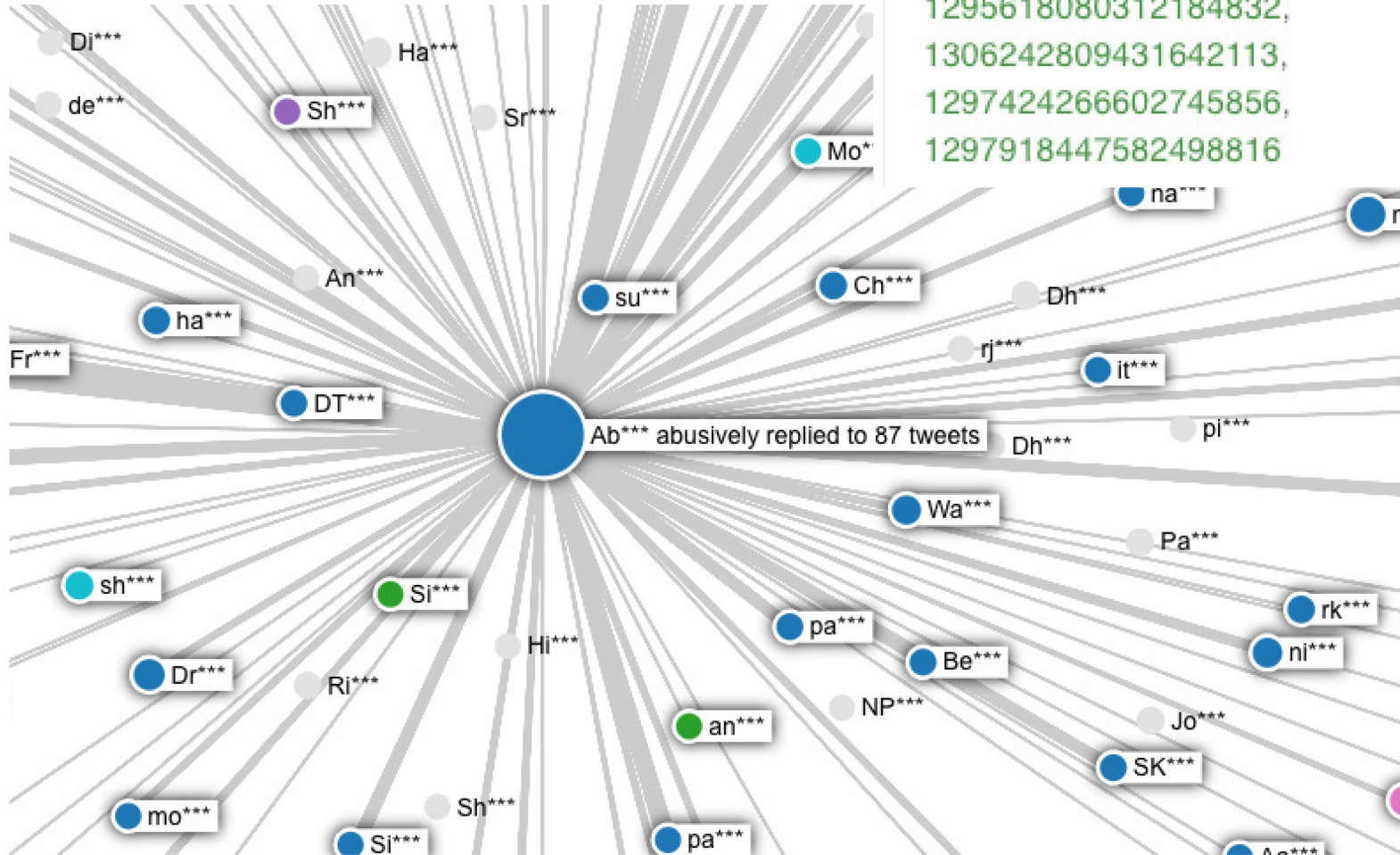
Co-reply network

Edge Details

ve*** and Ab*** co-replied abusively to 5 tweets of the target

Abused Tweets:

1296078111780429824,
1295618080312184832,
1306242809431642113,
1297424266602745856,
1297918447582498816



What more can be done?

- Methods of attack are growing more sophisticated and evolving with technology.
- They are also increasingly networked and fuelled by political actors.
- Need for responses to online violence to grow equally in technological sophistication and collaborative coordination.
- Most women journalists do not report or make public the online attacks they experience
- People are still reluctant to take online violence seriously.
- Failure of the internet communications companies - whose services facilitate much of the abuse - to take effective action

Potential Indicators of Escalation to Offline Harm

- Doxxing can lead to physical stalking & violence
- Death and rape threats
- Evidence of orchestrated attacks and disinformation e.g., large scale & instantaneous pile-ons
- Targeted attacks or threats against family members
- Hashtags and trending narratives associated with abuse
- Potential for significant long-term psychological harm, e.g. gaslighting, high volume abuse over a long time period

Challenges Ahead

- How do we maintain the balance with freedom of speech and online safety / integrity?
- Accuracy of and bias in ML models for abuse detection
- How can we detect the real subtleties of speech?
- How can we work with social media platforms to ensure best practices?
- Many ethical issues around the use of personal information (e.g. predicting individual characteristics)
- Ensuring our own mental and physical safety

More info

- GATE NLP toolkit <http://gate.ac.uk>
- Our Social Media Analysis work <https://gate-socmedia.group.shef.ac.uk/>
- Our studies of online abuse against journalists:
 - [The Chilling: Global Trends in Online Violence Against Women Journalists](#)
 - [Maria Ressa: Fighting an Onslaught of Online Violence](#)
- EU projects studying misinformation
 - WeVerify <http://weverify.eu>
 - Vera.AI @veraai_eu
 - EDMO Ireland Hub <https://t.co/vMxExC5ASH> @Ireland_EDMO

Work supported by Unesco, the UK FCDO, and the European Union/EU under the Information and Communication Technologies (ICT) theme of the 7th Framework and H2020 Programmes for R&D WeVerify (825297). vera.ai (101070093)

