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](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, ....

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

TWEETSTANCE 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 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
Aid workers in Nepal discussing strategy
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

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

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.

• How important and urgent is the message?
• What actions need to be taken?
### CREES Google Sheets Add-on

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

<table>
<thead>
<tr>
<th>Notes</th>
<th>CREES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text of request</strong></td>
<td><strong>RELATED</strong></td>
</tr>
<tr>
<td>Young nurse needing rescue!</td>
<td>related</td>
</tr>
<tr>
<td>Together we will rebuild page - Celia Torres</td>
<td>related</td>
</tr>
<tr>
<td>Only have bank address from Twitter</td>
<td>non-related</td>
</tr>
<tr>
<td>#disabled lady is #stranded in #LeagueCity #dickinsontexas #needrescue #help #hurricaneharveyanimalrescue</td>
<td>related</td>
</tr>
<tr>
<td>Been trying numbers for hrs - busy signal. Elderly couple desperately #NeedWaterRescue:10202 Willowgrove, 77035. Near S Post Oak &amp; W Belfort</td>
<td>related</td>
</tr>
<tr>
<td>Can't get through on the coast guard number reported 2:19 pm CT</td>
<td>non-related</td>
</tr>
</tbody>
</table>
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

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)

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

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

"death threats"

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

Maria Ressa scrotum-skinned, scrotum-looking, ...
scrotum-minded, lives like a scrotum! You don’t know math! Like her master, Leni [Robredo]!"
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:
- n*gger
- homo
- God botherer
- witch

Obscene nouns:
- c*nt
- tw*t
- shut up
- idiot
- rape
- f**k you
- kill
- die

Uncivil language:
- shut up
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 Don't call urself Journalist, u insult the #Journalist #Community.. u r just a #Hypocrite #Fake #FakeNewsMedia #communal #antihindu #antinational..
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

- Women
- Men
Quantifying online abuse directed at UK MPs

- 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
Characterising Abuse Against Women Journalists
Top abusive terms in Facebook comments to Maria Ressa

- idiot
- shut up
- prostitute
- liar
- liars
- fuck you
- go to hell
- asshole
- loser
- yellowtard
- coward
- milf
- spreading lies
- your ass
- get lost
- bitch
- imbecile
- queen of fake news
- yellowtards
- fake news queen
- fu
- moron
- retarded
- scum
- your fake news

Legend:
- attacks on professional credibility
- personal attacks:
  - sexist, misogynistic and explicit abuse
  - other
  - political

Percentage of occurrences
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

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

20% of women journalists surveyed in a recent UNESC-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.
Abuse against Rana Ayyub

- Analysed 13 million tweets sent over the last 2 years
- >44,000 blatantly abusive tweets
>40% of Rana’s tweets had at least 1 abusive reply
<table>
<thead>
<tr>
<th>Tweet URL</th>
<th>All Replies</th>
<th>Abusive Replies</th>
<th>% Abuse</th>
<th>First Abusive Reply</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://twitter.com/ghadaoueiss/status/1453169069314789376">https://twitter.com/ghadaoueiss/status/1453169069314789376</a></td>
<td>686</td>
<td>24</td>
<td>3.50%</td>
<td>1h 28m 17s</td>
</tr>
<tr>
<td><a href="https://twitter.com/ghadaoueiss/status/1459620996089004035">https://twitter.com/ghadaoueiss/status/1459620996089004035</a></td>
<td>516</td>
<td>13</td>
<td>2.52%</td>
<td>14m 14s</td>
</tr>
<tr>
<td><a href="https://twitter.com/ghadaoueiss/status/1465239554181447686">https://twitter.com/ghadaoueiss/status/1465239554181447686</a></td>
<td>368</td>
<td>12</td>
<td>3.26%</td>
<td>30m 36s</td>
</tr>
</tbody>
</table>
@ghadaoueiss

2:17 AM · Oct 27, 2021

884 Following
1.05M Followers

Replies

Hashtags
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,
129561808031284832,
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)
The Chilling: Global trends in online violence against women journalists


The Chilling: Assessing Big Tech's Response to Online Violence Against Women Journalists

Maria Ressa: Fighting an Onslaught of Online Violence
A big data analysis