Maria Ressa: Fighting an Onslaught of Online Violence

A big data analysis

Julie Posetti, Diana Maynard, and Kalina Bontcheva

With Don Kevin Hapal and Dylan Salcedo

ICFJ International Center for Journalists

Content warning:

This report includes graphic content that illustrates the severity of online violence against women, including references to sexual violence and gendered profanities. This content is not included gratuitously. It is essential to enable the analysis of the types, methods and patterns of attacks against Maria Ressa.

Acknowledgments:

This case study is a product of a Participatory Action Research project partnership between ICFJ, the University of Sheffield, and digital news outlet Rappler.

The authors thank ICFJ's Nabeelah Shabbir and Nermine Aboulez, and the University of Sheffield's Mark Greenwood, Ian Roberts, and Genevieve Gorrell for their contribution to research underpinning this case study, along with Graphika researchers who also contributed data and analysis. They are also grateful for the production assistance of ICFJ staff, especially Heloise Hakimi le Grand, Sharon Moshavi and Bob Tinsley.

Published by ICFJ, Washington, D.C., March 2021.

The easiest part is dealing with the impact of online violence and disinformation on *me*. I just see the impact on the world, and I don't know why we're not panicking."

Maria Ressa, Rappler CEO and co-founder

This groundbreaking collaborative case study is the most comprehensive assessment of online violence against a prominent woman journalist to date. We conducted a forensic analysis of the torrent of social media attacks on internationally celebrated digital media pioneer Maria Ressa over a five-year period (2016-2021). Here, we detail the intensity and ferocity of this abuse, and demonstrate how it is designed not only to vilify a journalism icon, but to discredit journalism itself, and shatter public trust in facts. These attacks also created an enabling environment for Ressa's persecution and prosecution in the Philippines. Now, her life is at risk and she faces the prospect of decades in jail, proving that there is nothing virtual about online violence.

Death threats. Rape threats. Doxxing. Racist, sexist, and misogynistic abuse and memes. These are just some of the features of the digital attacks that Maria Ressa, the Filipino-American journalist who founded the Manila-based news site Rappler, has faced daily since Rodrigo Duterte came to power in 2016. The former CNN war correspondent and UNESCO World Press Freedom Prize jury chair says none of her experiences in the field prepared her for the massive and destructive campaign of gendered online violence that's been directed at her over the past five years. At one point, in response to an investigative series on State-linked disinformation, she recorded more than 90 hate messages an hour on Facebook.

Ressa lives at the core of a very 21st century storm. It is a furor of disinformation and attacks - one in which credible journalists are subjected to online violence with impunity; where facts wither and democracies teeter.

She is not only attacked for being a journalist. She is attacked for being a woman. For the color and texture of her skin. For her American citizenship. And for her sexuality. Ressa is an emblematic case study in the global scourge of online violence against women journalists, which operates at the intersection of viral disinformation, networked misogyny, platform capture, press freedom erosion, and contemporary populist politics.

The attacks against Ressa are fueled by Duterte, who has publicly condemned

her - while musing that journalists are not exempt from assassination. His government has also employed a number of the key actors who have targeted Ressa online. And the worst attacks against her appear to have been orchestrated.

Despite fledgling attempts to address hate speech within the social media ecosystem, the platforms are the vectors which facilitate these attacks, creating an enabling environment for the State-led legal harassment of Ressa. This 'lawfare,' as Ressa refers to it, led to her conviction in mid-2020 on a trumped-up criminal 'cyberlibel' charge, and it continues to escalate. Issued with 10 arrest warrants in less than two years, and detained twice in the space of six weeks, she has paid more in bail and bonds than Imelda Marcos - corrupt wife of the former Philippines dictator - ever did. Ressa is currently fighting nine separate cases, and if she is convicted on all charges, she could spend the rest of her life in a Philippines jail. For publishing public interest journalism.

As Caoilfhionn Gallagher QC, the co-lead of Ressa's international legal team, explains, Ressa is attacked for her journalistic work on multiple fronts[1]:

> **"She faces a barrage of baseless lawsuits that seek to criminalize her work and expose her to a century in prison. The authorities vilify her, and President Duterte has helped to amplify online attacks against her. State authorities thus both directly attack Maria, and also create an enabling environment that facilitates and fuels abuse from others. In turn, online abuse emboldens the authorities in their persecution of her. In my view, there is a symbiotic relationship between the abuse Maria experiences online and the progress of the legal harassment offline."**
>
> Caoilfhionn Gallagher QC

[1] Caoilfhionn Gallagher QC co-leads the international legal team with Amal Clooney, acting with Can Yeginsu and Claire Overman.

Here, as part of a Participatory Action Research project involving Rappler, we present detailed evidence of the abuse, threats and harassment gathered from hundreds of thousands of posts directed at Ressa on Twitter and Facebook between 2016 and 2021. This study combines big data analysis - using Natural Language Processing (NLP) techniques (merging linguistics, computer science and artificial intelligence) and network analysis - with deep dive interviews. It also draws on research undertaken by others, including specialist 'internet cartographers,' civil society organizations and academics.

Our analysis of the online violence the Rappler CEO experiences helps us understand how and why she is targeted, and how the threats spread. It also gives us insights into the role of the State in manufacturing consensus by fueling the behavior of the social media users who target her. "If I wanted to see what the government was going to do, I only needed to look at social media because the attacks to arrest me and shut down Rappler were seeded as meta-narratives in 2017," Ressa says. "And now here we are." The function of the social media companies in facilitating the abuse is also spotlighted by Ressa: "The only way it will stop is when the platforms are held to account, because they allow it... They have enabled these attacks; they should not be allowing this to happen."

FIGURE 1

Abstract representation of the 100 most prevalent abuse terms in the analyzed tweets. Term size reflects frequency of occurrence. Elaborated below.

These are the **12 key findings** of our novel case study examining online violence against Maria Ressa:

1. Almost 60% of the attacks on Ressa we extracted from Facebook and Twitter for analysis were designed to undermine her professional credibility and public trust in her journalism.

2. Credibility or reputation-based attacks frequently deployed disinformation tactics and abuse conflating Ressa and her journalism with "fake news" (e.g., "Queen of Fake News"; "LIAR"; "#Presstitute").

3. Over 40% of the attacks in the combined datasets targeted Ressa at the personal level - often viscerally.

4. 14% of all abuse and 34% in the category of 'personal' attacks against Ressa could be classified as misogynistic, sexist and explicit abuse.

5. The use of abusive memes and manipulated images, which 'fly under the radar' of detection, is commonplace.

6. There is evidence that some of the attacks on Ressa are coordinated or orchestrated - a hallmark of State-led disinformation campaigns.

7. Much of the abuse and threats are fueled by President Duterte's public statements demonizing Ressa and Rappler as criminals, and pro-Duterte bloggers/social media influencers.

8. Lightning rods for attacks include: Rappler's investigative journalism; Ressa's reporting and commentary on State-linked disinformation; Ressa's high-profile media appearances; her industry accolades; and her court appearances.

9. Facebook is the main vector for the online violence Ressa faces. It is also the most used social media site in the Philippines - a country which spends more time online than any other.

10. Both Facebook and Twitter have promised to address the attacks on Ressa, but Facebook has failed dismally to effectively stem the tide of hate against her. Ressa says she feels "significantly safer" on Twitter.

11. For every one comment supportive of Ressa on her Facebook page, there were about 14 comments attacking her.

12. There is direct evidence that the online violence targeting Ressa has offline consequences. It has created the enabling environment for her persecution, prosecution and conviction. It also subjects her to very real physical danger.

A climate of impunity

The Philippines remains one of the most dangerous countries in the world to practice journalism, and targeted online violence attacks against journalists like Maria Ressa need to be examined in that context. In 2009, the country was the site of the deadliest attack on journalists ever recorded by the Committee to Protect Journalists (CPJ) - the Maguindanao massacre which killed 32 journalists and media workers in an "orgy of political violence." But when Rodrigo Duterte was asked by a reporter in 2016 what his incoming government would do to address the high rate of journalists continuing to be murdered with impunity in the Philippines, he answered: "Just because you're a journalist, you are not exempted from assassination, if you're a son of a bitch." A year later, he had Rappler and Maria Ressa in his sights in his 2017 State of the Nation Address, while his associates were seeding meta-narratives on social media, painting Ressa as a criminal, and calling for her arrest. They were also calling for her to be sexually assaulted, killed and even "raped repeatedly to death."

The impacts of online violence can be personally and professionally devastating for the individual journalists targeted. They also radiate - affecting their families, sources, colleagues and audiences. Increasingly, online violence is also spilling offline. An ICFJ-UNESCO survey of nearly 1,000 journalists in late 2020 found that 20% of women who responded had experienced offline attacks that they believed had originated online.

In a country like the Philippines, which also stands accused of State-sanctioned extrajudicial killings in the context of the current 'drug war,' the potential is high for online violence against women journalists to reap deadly results. As Rappler's Executive Editor Glenda Gloria noted: "We never doubted that those online threats would translate to physical threats. That's why we doubled not just the security of Maria, but of the newsroom, because a lot of the online threats against activists turned into reality," she says. "There was this female activist who was first blasted online and shot while on her way home. It's real. Especially against women."

But the impacts don't end there. They also include the potential for a 'disinfodemic,' the erosion of democracy, the chilling of independent journalism, restrictions on societies' access to information, and the furthering of gender inequality in and through the news media.

Timeline of attacks

JUNE 2016

Incoming President Rodrigo Duterte tells journalists they are not exempt from assassination.

OCTOBER 2016

Rappler publishes Ressa's seminal investigation into State-linked online disinformation 'Propaganda War: Weaponizing the Internet.' Pro-Duterte blogger Mocha Uson (who would go on to become the president's Communications Assistant Secretary) seeds the term "presstitutes" in an online campaign designed to silence Duterte's critics.

DECEMBER 2016

Rappler reports on the extrajudicial killings associated with Duterte's drug war and the online propaganda campaign to popularize pro-violence narratives.

MAY 2017

Pro-Duterte blogger 'Thinking Pinoy' seeds the hashtag #ArrestMariaRessa while accompanying Duterte on a State visit to Russia. #ShutdownRappler, #BringHerToTheSenate and #UnfollowRappler were offshoots of this campaign.

JUNE 3, 2017

Ressa draws on her history of investigating terrorism networks and reports on a suspected terrorist attack at 'Resorts World' in Manila. The second most intense set of attacks on Ressa visible in our Facebook datasets follows (See figure 3.)

JULY 10, 2017

Ressa speaks for the first time about being brutally attacked online in a book chapter published by the United Nations Educational, Scientific and Cultural Organization (UNESCO).

JULY 27, 2017

Duterte targets Rappler in his State of the Nation Address, amplifying the narrative seeded online two months earlier by 'Thinking Pinoy' that Rappler is foreign-owned (this was a precursor for the legal harassment to follow). Just over a week later they received their first subpoena. Online abuse suggests the dual Filipino-American citizen is a CIA operative and a terrorist. These accusations are even more threatening under the country's new anti-terror law (passed in 2020).

DECEMBER 2017

Duterte accuses Rappler of being funded by the CIA and online attacks referring to Ressa as a 'terrorist' and a 'foreign agent' surge.

DECEMBER 2017

The subject of a corruption story published by Rappler files a 'cyberlibel' complaint to the National Bureau of Investigations (NBI), naming Ressa, despite the fact she didn't write or edit the piece.

JANUARY 2018

The Duterte government revokes Rappler's license to operate (a decision which Rappler has appealed), calling the outlet "fake news." This becomes the dominant narrative of online attacks targeting Ressa.

FEBRUARY 2018

The NBI rejects the cyberlibel complaint.

MARCH 2018

The NBI reverses its decision and recommends Rappler be charged for 'cyberlibel' by the Department of Justice.

MARCH 2018

The Bureau of Internal Revenue files tax evasion charges against Ressa and Rappler, becoming the 5th government agency to go after them since Duterte claimed Rappler was funded by the CIA and referred to it as a "fake news outlet."

DECEMBER 2018

Rappler reports on Duterte's confession that he sexually assaulted a maid.

JANUARY 2019

Department of Justice recommends criminal charges be laid against Ressa and a former colleague in the 'cyberlibel' case.

FEBRUARY 2019

Ressa arrested in connection with the 'cyberlibel' charge in the Rappler newsroom. She is detained overnight.

MARCH 2019

Ressa arrested for a second time after flying into Manila, this time for alleged breaches of foreign ownership rules applying to news media, as suggested by Rodrigo Duterte in his 2017 SONA speech.

JULY-DECEMBER 2019

Cyberlibel trial takes place in Manila.

JUNE 2020

Ressa faces a second cyberlibel complaint brought by the same source who initiated the first action.

JUNE 2020

Ressa and her former colleague Rey Santos Jr. are convicted of criminal cyberlibel charges (the original charge) by a Manila court and sentenced to up to six years in prison, triggering an international outcry from human rights organizations, journalists and the UN Special Rapporteur on Freedom of Expression. The meta-narrative of 'journalist equals criminal' planted online in 2017 becomes reality. (The case is currently under appeal).

DECEMBER 2020

9th arrest warrant issued for Ressa as she is charged by the Department of Justice with the second criminal cyberlibel offense.

JANUARY 2021

Third cyberlibel charge laid against Ressa and Rappler by the Department of Justice after a complaint from another subject of a story about corruption is published by Rappler.

MARCH 2021

Ressa is cross-examined in a Manila court on tax-related charges brought by the State, while battling eight other continuing cases.

Shark tanks full of data

In this case study, we examine nearly 400,000 tweets directed at Maria Ressa during a 13-month period from December 2019 to February 2021, along with more than 57,000 posts and comments published on Facebook between 2016 and 2021.

The tweets were gathered and analyzed using Natural Language Processing (NLP) techniques by computer scientists at the University of Sheffield (which provided formal ethics clearance for the work). From the large dataset, we extracted a sample of 1,128 tweets demonstrating *highly explicit abuse* - predominantly expressed in English - for detailed analysis. In parallel, data analysts from the digital research firm Graphika conducted a network analysis on a subset of tweets from a spike of abuse in June 2020 that coincided with Ressa's conviction on the first cyberlibel charge. The aim of this analytical process was to map the types and methods of attack, along with the trajectory of the abuse, and the interconnectedness of those attacking Ressa.

The Facebook data represents over 9,400 public comments in response to Ressa's posts that were gathered from her professional Facebook page with her explicit permission, and other public Facebook data extracted from a massive database that Rappler maintains called 'Sharktank.' According to Ressa, 'Sharktank' maps the information ecosystem of the Philippines on Facebook (which is synonymous with the internet in the country, reaching 96% penetration in 2021). As of January 2021, the database had captured 471,364,939 public posts and 444,788,994 public comments made by 4,176,326 users, 68,000 public pages and 26,000 public groups on Facebook. From that database, over 47,000 posts mentioning Ressa were extracted for analysis. The same NLP techniques were applied to the Facebook data with the assistance of Rappler's research team.

The NLP analysis is based on a 'high accuracy detection' model and it is largely restricted to English-language posts, or those which blend English with Tagalog (the other national language of the Philippines). Consequently, the samples of online abuse extracted using this method are considered to be severely underreported, capturing only around 50% of all English-language abusive messages present in the target's social media stream, according to previous studies by the computer scientists attached to this project.

It is important to note that this analysis excludes the most brutal online violence Ressa has experienced, which she says came via Facebook Messenger. Such content is not only harder to detect, it's also more difficult to report. Similarly, public abuse is often subtle and potentially designed to 'slip under the radar'. A classic example of this is a death threat sent to Ressa on Twitter on February 12, 2021. This is subtle partly because it refers to text in an image which is not easily processed by automated tools, and partly because the message itself contains no abusive terminology, though the underlying meaning is clear. This message was still visible on Twitter at the time of writing - two weeks after being reported to the platform by the lead author.

FIGURE 2

A death threat sent to Maria Ressa via Twitter on February 12, 2021. The threat was still visible on the platform when this report was finalized.

Methods, themes and tropes deployed by the attackers

Using the methods we described above, we analyzed 399,131 tweets and approximately 56,400 Facebook comments or posts directed at Maria Ressa. While the vast bulk of online violence that Ressa experiences occurs via Facebook, the most common themes and methods of attack were relatively consistent across both platforms. We triangulated this data with online research and in-depth interviews with Ressa and her colleagues.

This is what we found when we examined the combined datasets.

1. Dominant themes and tropes deployed by Ressa's online attackers:

- 60% of the abuse Ressa received can be classified as damaging to her professional reputation or credibility. It includes disinformation designed to discredit her and erode trust in her journalism, and features calls for her to be charged, tried, raped, imprisoned and even killed for her work. This abuse frequently involves false claims that she is a purveyor of "fake news" and includes the pernicious hashtag #presstitute.

- Over 40% of the attacks against Ressa were designed to hurt her at the personal level

- 14% of all abuse and 34% of that in the category of 'personal' attacks against Ressa could be classified as misogynistic, sexist and explicit abuse. This includes abuse targeting Ressa's physical appearance (emphasizing her skin condition) and manipulated photographs depicting her head associated with male genitalia.

- Racist abuse and memes that include depicting her as a monkey (representing 1% of all abuse Ressa receives and 3% in the 'personal attacks' category).

- Homophobic slurs designed to question Ressa's sexuality and increase her vulnerability were determined to represent <1% of all abuse but 2% of personal attacks.

- Threats of physical violence, including death threats embedded in images, and threats of sexual violence (e.g., being "publicly raped to death") were associated with the worst attacks.

2. *Typical methods of attacks:*

- 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.

3. *Triggers for attacks:*

The attacks spike in association with:

- Rappler's investigative journalism focused on the bloody and increasingly dictatorial Duterte administration;
- Ressa's reporting and commentary on disinformation and Duterte;
- Ressa's high-profile media appearances;
- Ressa's international awards;
- Ressa's court appearances.

"Every time a complaint reaches the court, every time a statement is made in support of Maria, there's a troll army that really is commanded to respond."

Rappler's Executive Editor Glenda Gloria

Spikes and triggers

The largest surge of attacks we identified against Ressa on her Facebook page occurred in October 2016, when Rappler published a 3-part investigative series into State-linked disinformation networks - two of which were written by Ressa herself. The Facebook data covers a five-year timespan and clearly demonstrates other attack spikes connected to Rappler's critical coverage of Duterte's 'drug war', and the extrajudicial killings associated with it, along with international media attention (e.g., associated with the multi-award-winning film A Thousand Cuts, which depicts her struggle), her high-profile industry awards, and her arrests, detentions, and trials are also associated with increased attack spikes in the Facebook data.

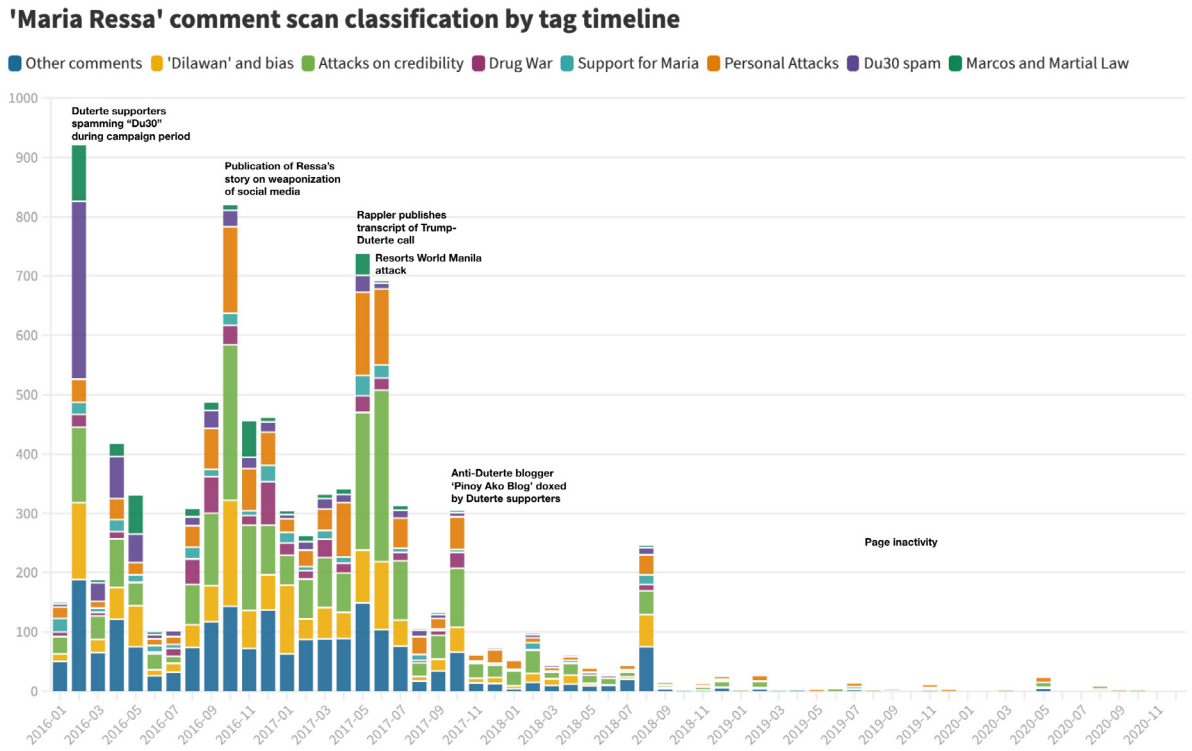

FIGURE 3

A timeline of comments scraped from Maria Ressa's professional Facebook page highlighting peaks associated with key events. The categories 'Attacks on credibility', 'Personal attacks', and 'Dilawan and bias' represent identified abuse. These categories are further analysed below.

An examination of the Twitter data detailing frequency of abuse (see figure 4. below) reveals three major spikes, each of which have more than 50 tweets per week identifiable as *highly explicit abuse* (largely in English or hybrid English-Tagalog). The largest peak was in early May 2020, and it was triggered by an interview with ABC Australia in which Ressa misspoke - the error was used to attack her credibility. The third biggest spike came when a Manila court delivered a guilty verdict against Ressa in a criminal cyberlibel charge prosecuted by the government (June 2020). Her chastisement of Duterte and his government due to the Philippines being recorded as having the highest proportion of COVID cases in Asia (August 2020) represented the second-biggest spike. These timeline spikes in targeted attacks on Ressa during 2020-2021 were also reflected in the Facebook data.

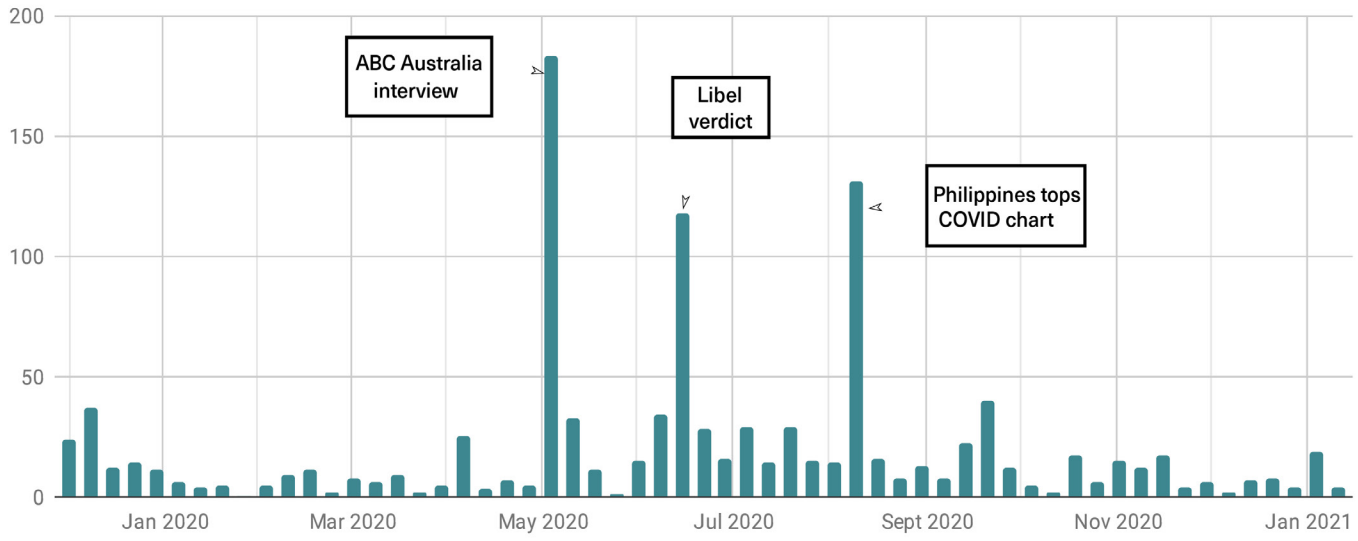

FIGURE 4

Spikes in online abuse towards Ressa in the Twitter data.

The spike in attacks on Twitter associated with Ressa's appearance on ABC TV in 2020 also reflect peaks associated with other international media coverage. Two of the biggest spikes in online violence against Ressa seen on Twitter in 2019 by Rappler's social media team were associated with a New York Times Magazine feature on Ressa, and a CBS 60 Minutes story about her harassment by the State.

Scrotums, monkeys, swarms and lies

Attacks against Ressa and Rappler dominated the comments on her professional Facebook page, which she established in 2015 at the behest of the company, she says. Ironically, Facebook recommended she start the page to help better manage her comments. But the harassment she experienced on the page soon became overwhelming. She has not posted to the page since early 2019, and it now lies effectively dormant.

Of the 9,433 comments from Ressa's professional Facebook page - spanning the period 2015-2018 - 54% fall under 'attack clusters', while supportive comments represented only 4% of the data. **This means that for every one comment supportive of Ressa, there were about 14 comments attacking her.** And a more granular analysis reveals that approximately four of these 14 abusive comments would constitute personal attacks, focusing on her appearance, nationality, gender, and sexuality.

The themes of the abuse which constituted the three 'attack clusters' identified within the dataset are:

1. **Attacks on Maria Ressa's professional credibility (25% of all comments).** This abuse is designed to undermine Ressa's professional reputation and trust in independent journalism. It uses disinformation tactics to accuse her of peddling 'fake news', and it includes false claims of corruption.

2. **Personal attacks against Maria Ressa (14% of all comments).** These include: sexist, misogynistic and explicit abuse; threats of sexual and physical violence; racist abuse; homophobic abuse.

3. **'Dilawan' and bias - politically themed attacks (15% of all comments).** The Tagalog word 'dilawan' is borrowed from the word 'dilaw', which means yellow, and is used by Duterte supporters as a derogatory term to describe members of the opposition, while 'dilawan' is used to accuse journalists of being 'pro-opposition', challenging the core professional tenet of impartiality.

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%

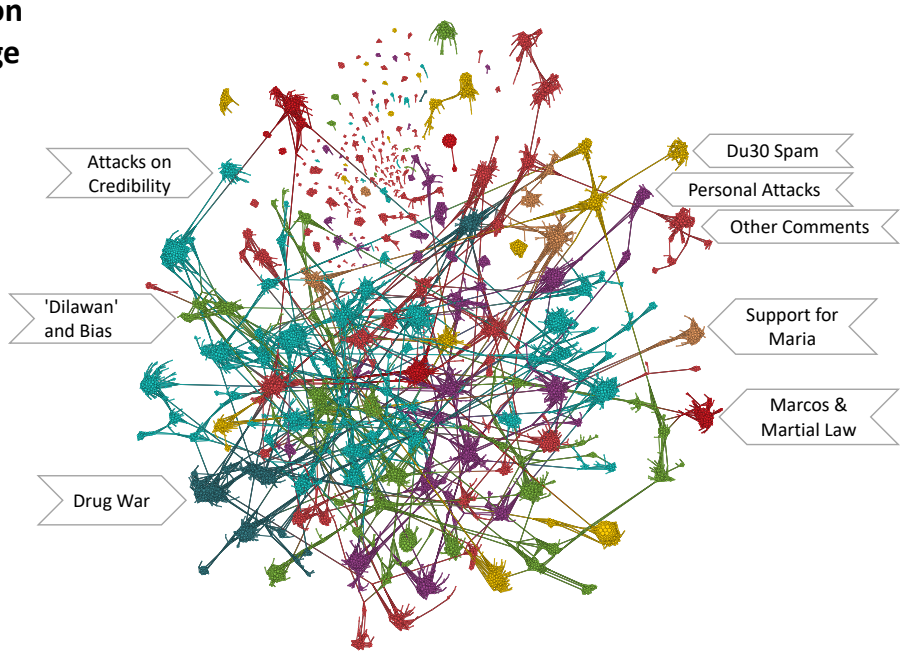

FIGURE 5

This illustration maps the clusters of topics identified through NLP analysis on 9,433 comments made on Maria Ressa's Facebook page between January 2016-February 2021. The connections between clusters indicate topical similarities.

We then submitted a subset of English-language comments on Ressa's Facebook page determined to be *highly explicit abuse* to deeper NLP analysis for this case study. The 263 comments we identified in this category were classified into three main types:

1. **Attacks on Maria Ressa's credibility (45%).** We found that just over 45% of theabuse in this data set could be categorized as attacks on Ressa's journalistic credibility, designed to undermine trust in her reporting.

2. **Personal attacks against Maria Ressa (50%).** Within this category, 45% of the personal abuse contained sexist, misogynistic, or explicit content; 2% was classified as abuse; and 53% was sorted as other (more general) abuse (which typically contains damaging, though slightly less offensive insults, such as "moron" and demands that Ressa "shut-up").

3. **Politically themed abuse against Maria Ressa (5%).** This abuse included references to her being a 'yellowtard' or a traitor (in line with the category in the bigger dataset labeled 'Dilawan and Bias').

FIGURE 6

Attacks on credibility (colored in blue) aim to undermine Ressa's reputation as a journalist. Personal attacks are colored in shades of red and consist of sexist, misogynistic and explicit sexual terms (beige); racist terms (darker red); and other kinds of personal insult (pink). Political attacks (yellow) use terminology associated with (real or imagined) political affiliation.

Next, we examined 1,387 public Facebook posts published across the platform (as distinct from those comments on Ressa's professional Facebook page), extracted from the vast Rappler 'Sharktank', which we determined to be *highly explicit abuse* (largely in English or hybrid English-Tagalog) directed against Ressa by name. These posts and comments were published between January 1st 2016 and February 20th 2021.

Again, with this Facebook 'Sharktank' dataset, we can divide these attacks into three main classifications:

1. **Attacks on Maria Ressa's credibility (70%).** The vast bulk (n. 943)[2] of these highly abusive posts were categorized as forms of attack on Ressa's professional credibility.

2. **Personal attacks against Maria Ressa (26%).** These (n: 351) were identified as forms of personal attack (including sexist, misogynistic, explicit, racist, and homophobic abuse; and threats of sexual and physical violence).

3. **Politically themed abuse against Maria Ressa (4%).** These (n: 47) were determined to be attacks designed to imply political bias or opposition affiliation.

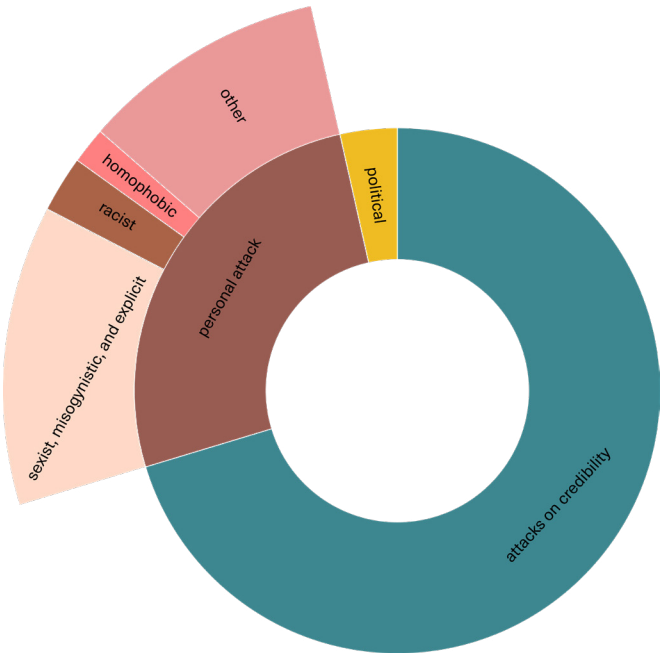

FIGURE 7

Attacks on credibility (colored in blue) aim to undermine Ressa's reputation as a journalist. Personal attacks are colored in shades of red and brown and consist of sexist, misogynistic and explicit sexual terms (beige); racist terms (brown); homophobic terms (darker pink); and other kinds of personal insult (bright pink). Political attacks (yellow) use terminology associated with (real or imagined) political affiliation.

[2] n.943 represents the number of posts deemed highly abusive.

In analyzing the Twitter data, we used the automated NLP tools to classify the attacks detected into two main types, which are closely associated with the categories of abuse identified in the Facebook data:

1. **Attacks on Maria Ressa's reputation/credibility (56%).** This category includes terms specifically designed to undermine her professional reputation and the credibility of her journalism, such as "liar" and "fake news queen."

2. **Personal attacks against Maria Ressa (44%).** We broke these down into **five subcategories:**

   - Sexually explicit attacks (includes terms referring to sexual acts and body parts, e.g., "pussy" and "go fuck yourself")
   - Sexist/misogynistic attacks (refers to terms such as "witch", "whore", and "bitch")
   - Homophobic attacks
   - Racist attacks
   - Other abuse (includes terms such as "moron" and other broadly demeaning comments)

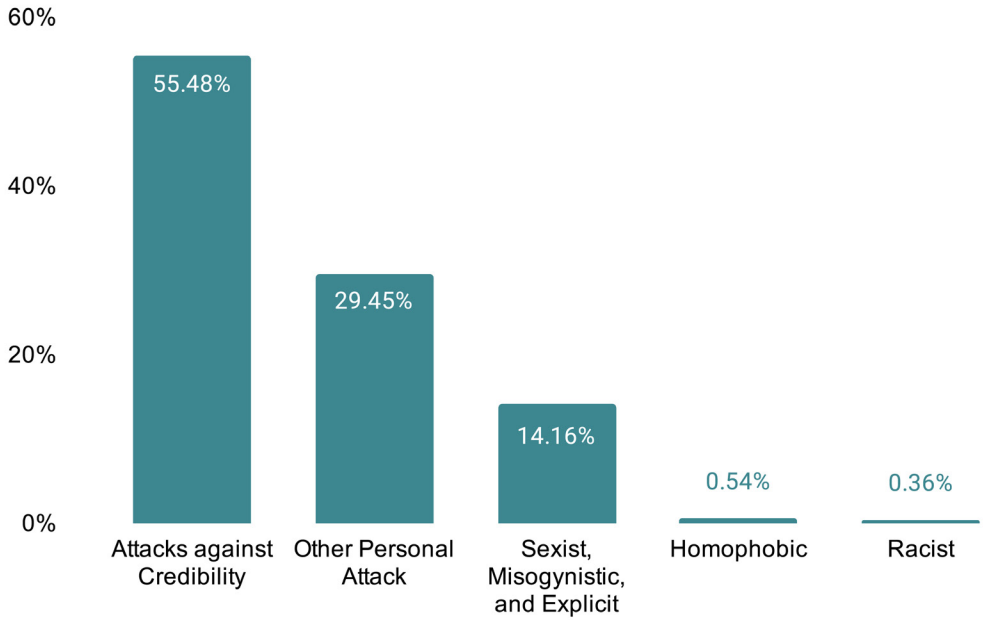

FIGURE 8

Percentages of categories of abuse in the Twitter sample.

Attacks designed to discredit Ressa's journalism and erode public trust in facts

As indicated above, the dominant theme of the online violence waged against Maria Ressa involves damaging her professional credibility, by extension her reportage, and by association Rappler's. This is clearly evident within the subset of *highly explicit* abusive comments (mostly in English or English-Tagalog) gathered from her professional Facebook page, with nearly half of that abuse falling into the category of attacks on her professional reputation.

Among these comments were disinformation-laced attacks, including accusations that she was a 'fake news' peddler, like these:

> *You are the Queen of Fake News fucking Bitch*
>
> *Stop spreading Lies you Piece of SHITS I wish you Rotten (sic) in Jail ..*
>
> *Maria Ressa get the fuck out of our country Philippines! dont (sic) mislead the people with your fake news...*

The most frequently used abusive terms in this data subset were words designed to ridicule, silence and discredit Ressa while simultaneously undermining public trust in her critical journalism. The top ranked words were 'idiot,' 'shut up,' 'presstitute' (a portmanteau of 'press' and 'prostitute' often used by Duterte supporters and popularized by pro-Duterte blogger turned- government official Mocha Uson, which also features in our analysis below of sexist and misogynistic attacks against Ressa), and variations on 'liar.' Around 20% of the attacks on her credibility were related to disinformation - either equating her with it, or falsely accusing her of peddling it. As Ressa says, "Lies spread faster than facts. And lies laced with anger and hate spread faster and further than facts."

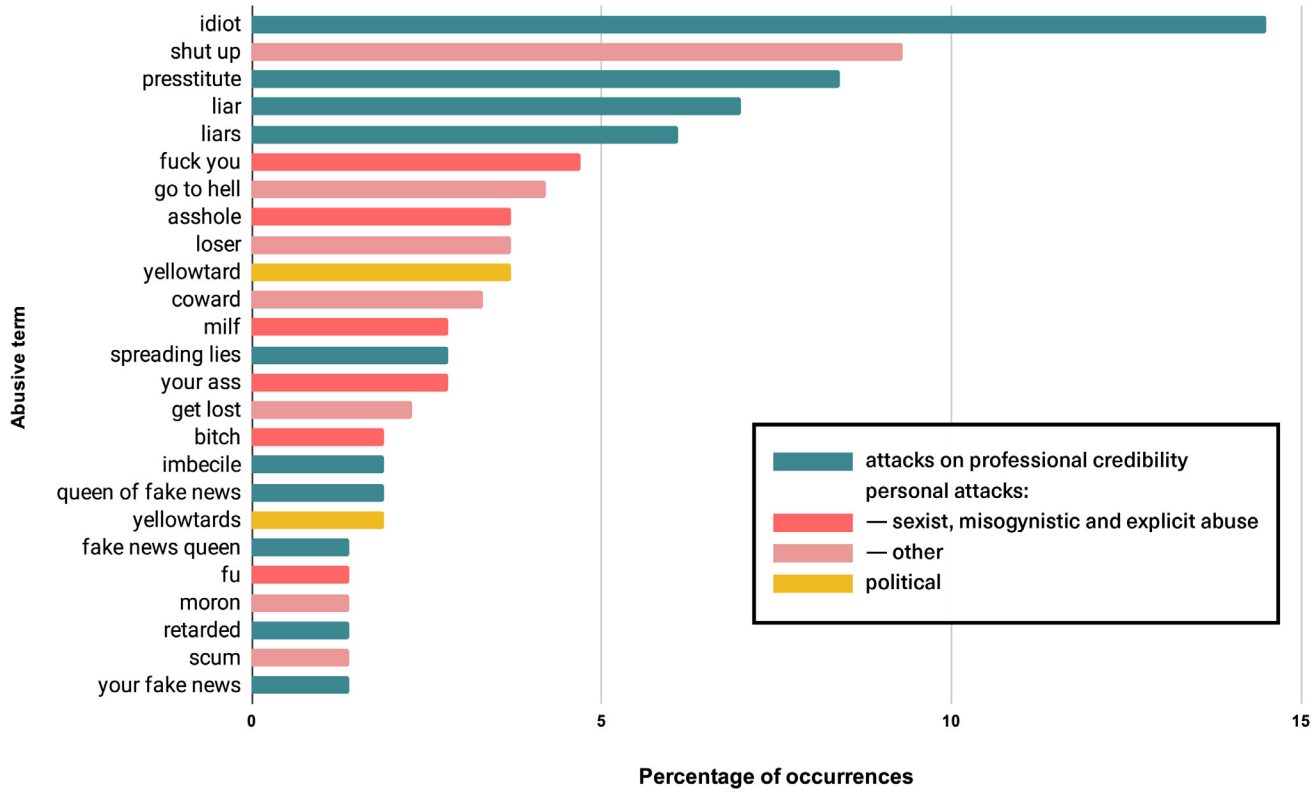

FIGURE 9

Most frequent abusive terms found in the study of comments on Maria's professional Facebook page.

Disinformation narratives deployed against Ressa were also prevalent in the larger multilingual dataset (n. 5,093) extracted from her professional Facebook page. These include repeated claims that she is a "liar"; the "Queen of Fake News"; "Bayaran" (a Tagalog term for a corrupt journalist who takes payment for favourable coverage); a "presstitute"; and a "national security threat" or terrorism supporter, echoing narratives from prominent pro-Duterte blogger RJ Nieto, known as Thinking Pinoy. He tried to get the hashtag #ArrestMariaRessa to trend in May 2017, when he and Mocha Uson accompanied Duterte on a state visit to Russia. Two years later Ressa was in fact arrested, and within three years she would be convicted of a trumped-up criminal cyberlibel charge prosecuted by the State.

FIGURE 10

A screengrab of a Facebook post published by the pro-Duterte blogger 'Thinking Pinoy' who accompanied the Philippine President on a state visit to Moscow in May, 2017.[3] This Facebook post remained live until just days before this report was published.

The role of disinformation tactics in the attacks on Ressa and Rappler reflects the findings of an ICFJ-UNESCO survey (conducted late 2020) of nearly 1000 international journalists which concluded that 41% of women respondents had experienced online violence that they believed was connected to orchestrated disinformation campaigns.

Online violence against Ressa, seeded by pro-Duterte digital influencers and fanned by the State, undoubtedly impacts on decisions in the progress of the cases and charges she faces according to her international lawyer, Caoilfhionn Gallagher:

> **I think that when the Duterte administration is making decisions in that environment, the fact that Maria is a hate figure online is enabling those decisions to be taken."**
>
> Caoilfhionn Gallagher QC

[3] In line with academic research ethics protocols, we have obscured the identities of most social media users. In this instance, however, we have not redacted the user's identity as a matter of public interest. 'Thinking Pinoy' is identified in the data and in media coverage linked throughout this report as one of the key pro-Duterte influencers targeting Maria Ressa and Rappler, and he is closely associated with the administration.

Referring to the decision to charge Ressa with a third count of criminal 'cyberlibel' in January 2021, Gallagher says: "It's impossible not to be aware of the fact that Maria was the subject of a very large amount of viral abuse throughout that time, when the decision was being made. There was a coincidence in time between when the prosecutor was making that decision about the third charge, and a spike in abuse linked to the arrest warrant for Maria in respect of the second cyberlibel charge."

In Ressa's view, the Duterte propaganda machine has accomplished its goal: "They pounded opponents and journalists to silence to create a bandwagon effect for seeded meta-narratives of bias, incompetence, criminality, and corruption to be leveled against them." Rappler's Head of Digital Strategy Gemma Mendoza has spent nearly as much time as Ressa swimming in the 'Sharktank' that the news outlet maintains, and she can easily pinpoint what triggered the tsunami of online violence against her CEO. "When Maria wrote that story about the propaganda machine online, that was when all the abuse was directed at her - it was like 'here's the lightning rod.' And it never ceased," Mendoza says.

This abuse doesn't just discredit Ressa and enable her legal harassment. It also serves to further erode public trust in independent journalism, and facts in general, sowing confusion and undermining the democratic pillar of press freedom in the Philippines.

FIGURE 11

Thematic clusters of abuse thrown at Ressa on her professional Facebook page.

Again, when we break down the categories of abuse identified in the public Facebook posts collected by Rappler from across the platform (i.e., the 'Sharktank' dataset), we see the predominance of reputationally damaging terms such as "liar" and "fake news queen", underlining the use of disinformation tactics to tarnish the credibility of Ressa's critical journalism (to the advantage of the Duterte administration) and public trust in facts. The irony involved with disinformation agents accusing a target of disinformation is noteworthy.

When she realized how State-led legal harassment, viral disinformation, online violence, and 'platform capture' converged in her case, Ressa said the impact was explosive.

This is like an atom bomb has gone off in our information ecosystem. This is how bad it's going to get. And it did get that bad."

Maria Ressa

And the attacks really bothered her, according to Rappler Managing Editor Glenda Gloria. "But she also wanted to convert that into knowledge, and a clear understanding of the tactics and the strategy of the enemy. So it's really a war in that sense," Gloria says. "And she was a warrior trying to think both of strategy and tactics, and the soldier being hit and being machine-gunned." But the global support Ressa mustered through investigative journalism and press freedom advocacy had an international impact and that was a source of tremendous psychological benefit for Ressa, because "It showed her that there's hope," Gloria says.

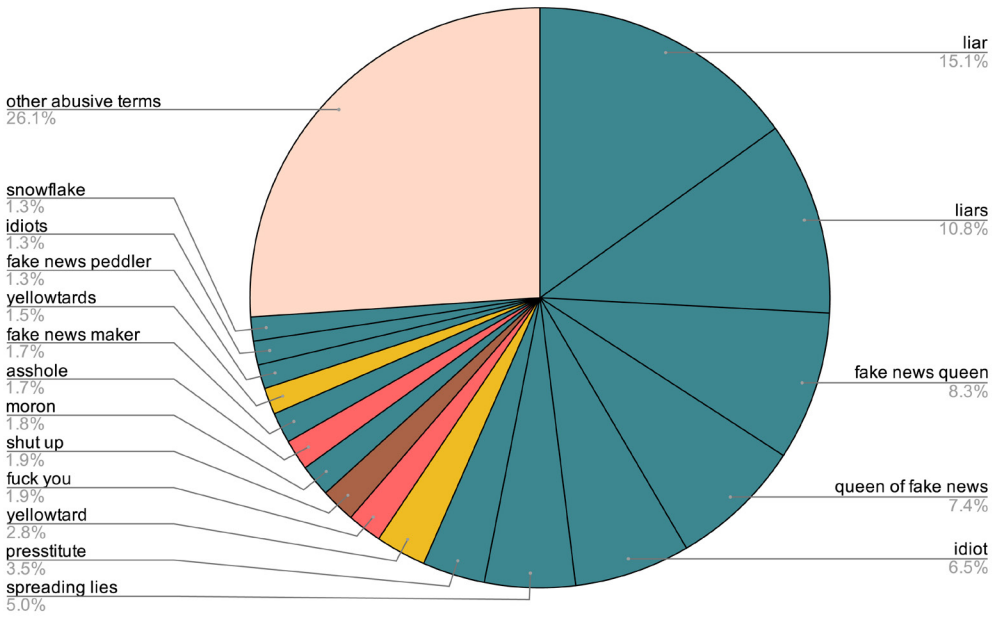

FIGURE 12

Frequency breakdown of abusive terms that appear more than once in Facebook posts mentioning Maria Ressa ("other abusive terms" includes those which appear between 1 and 15 times in the data). Personal attacks consist of sexist, misogynistic and explicit sexual terms (red); and other kinds of personal insult (beige). Attacks against professional credibility are colored in blue. Political attacks (yellow) use terminology associated with (real or imagined) political affiliation.

The same patterns of assaulting Ressa's credibility using disinformation tactics - including falsely accusing her of being a liar and a disinformation purveyor - were evident in the detailed analysis of the Twitter data. Here, we see the overwhelmingly dominant use of reputational slurs such as "liar" and variants such as "spreading lies" and "lying bitch," as well as terms relating to the spreading of disinformation such as "fake news queen" and "fake news peddler." (See figures 14 & 15 below).

FIGURE 13

This tweet demonstrates the ways in which Ressa is demonized as a 'fake news' peddler in disinformation campaigns. It also leverages a minor error she made in an international TV interview.

The word cloud below (figure 14.) shows the 100 most frequently occurring abusive terms. Terms such as "liar" are often entirely capitalized in the messages, indicating strength of emotion - these slurs are being 'shouted.'

Looking at breakdown by frequency, we see from the pie chart below (figure 15.) that the term "liar" alone accounts for more than 15% of all abusive terms found among the tweets analysed, with "liars" accounting for another 3%, not to mention variants on this theme. The terms "queen of fake news" and "fake news queen" together make up 17% of abusive terms (again, with other similar terms relating to 'fake news' accounting for a similarly high proportion. Also evident are sexist terms (e.g., "bitch," "slut"); sexual terms ("go fuck yourself," "pussy," "scrotum face," "asshole"); homophobic terms (e.g., "lesbo"); and racist terms (e.g., "gringo").

FIGURE 14

The 100 most prevalent abuse terms in the analyzed tweets. Term size reflects frequency of occurrence.

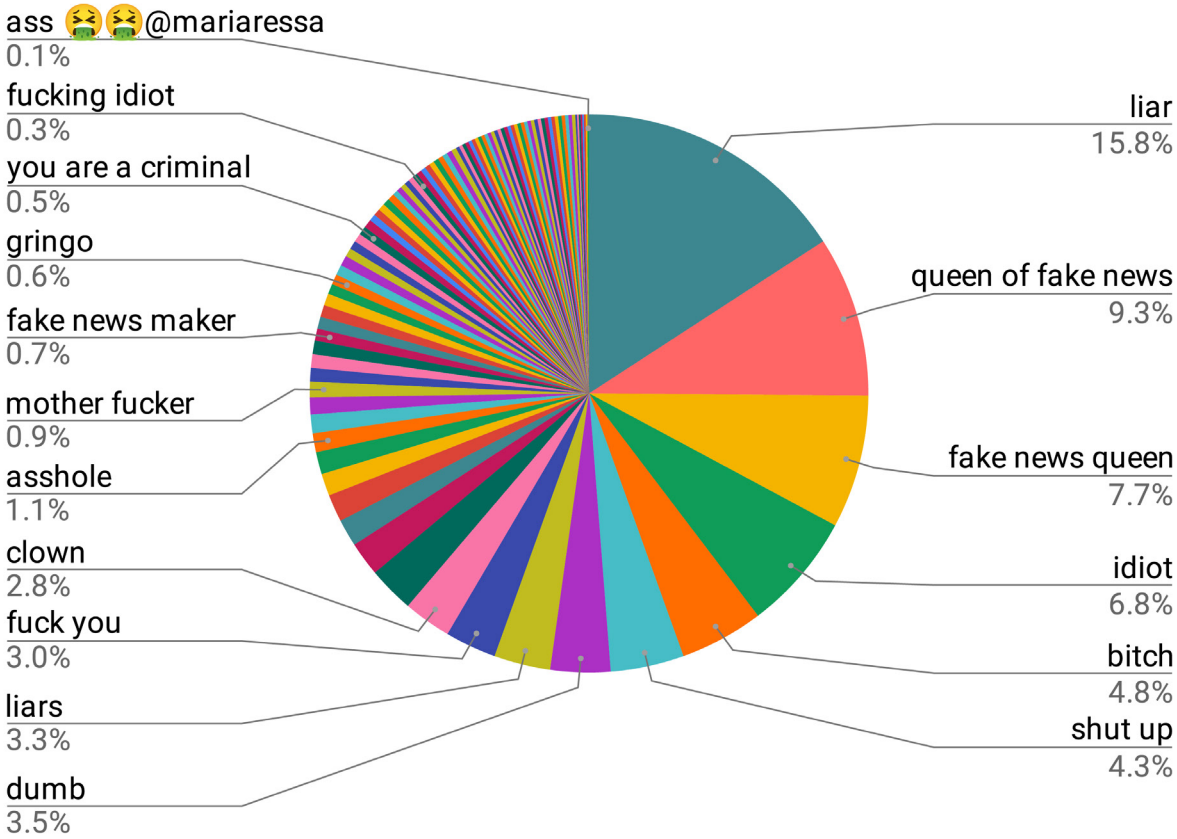

FIGURE 15

Frequency breakdown of the 100 most abusive words and phrases in the Twitter dataset.

#Hashtagged abuse

As with the analysis of abusive terms, we see many hashtags related to undermining Ressa's reputation and professional credibility on Twitter. These include #liar, #fakenewsmaker, #yestoshutdownrappler, #rapplerfakenews, #presstitutes, as well as some sexual terms and other generally abusive ones like #dicksucker and #moron.

FIGURE 16

Top hashtags in the Twitter dataset.

We also see a number of hashtags which support Ressa, Rappler and press freedom, such as #HoldTheLine, #DefendPressFreedom, #CourageON, #istandwithmariaressa, #athousandcuts, #dontshutup. However, these mostly indicate the weaponization of such terms, since these hashtags typically occur as part of a quoted tweet, as in the example below.

FIGURE 17

Strong support for Ressa from the international community is shouted down by pro-Duterte bloggers and social media influencers who counter with disinformation narratives designed to discredit both Ressa and her journalism.

Hey @HNeumannMEP You sponsored that EU resolution against Ph and President Duterte How stupid of you to believe all lies spoon fed by @mariaressa She convicted criminal Fighting other criminal charges tax evasion cyber libel anti dummy law NO SUPPRESSION OF PRESS FREEDOM there!!!

In the Facebook dataset extracted from the 'Sharktank,' we also see hashtags, including a number of derogatory ones such as #yellowtard, #crappler, #liar, #presstitute, and variations on the #fakenews theme.

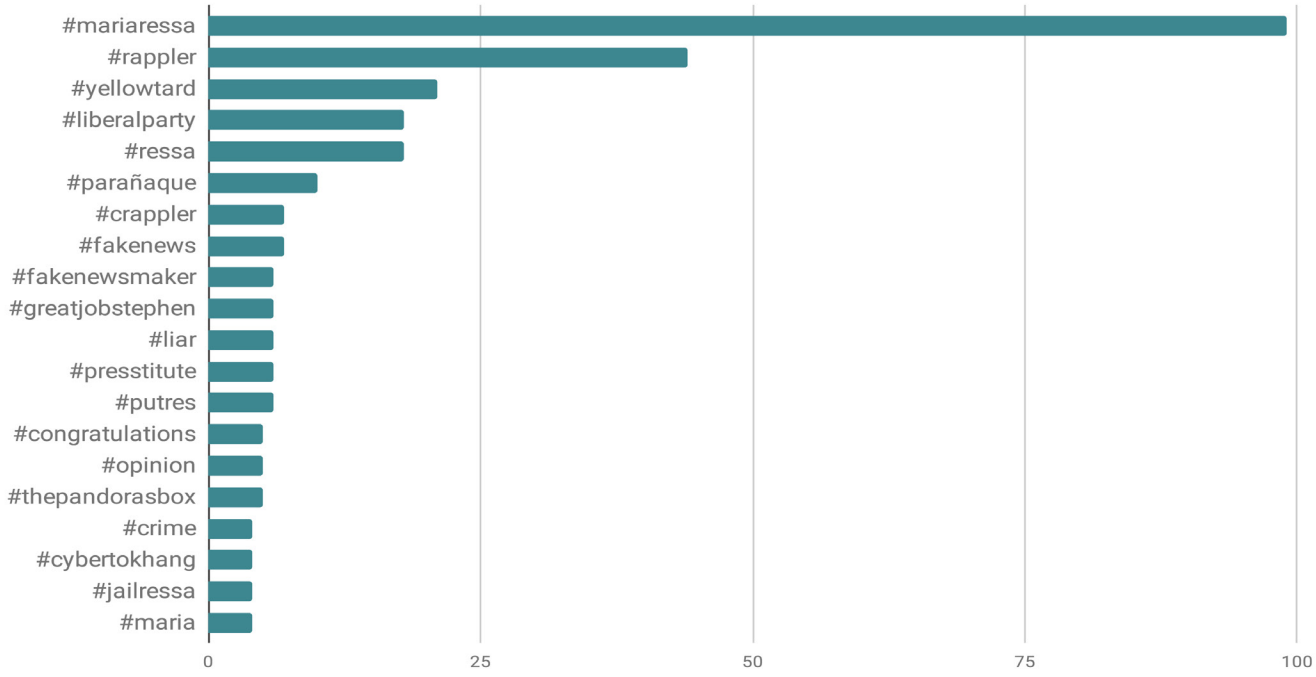

FIGURE 18

Top 20 hashtags in the `Sharktank' Facebook dataset.

Foreign actors and networked gaslighting

China also features extensively in the tweets identified as containing *highly explicit abuse*, due to links with Duterte as well as a conspiracy theory that Maria was in league with the Chinese who were paying her to spread lies. This was also a theme evident in the Facebook data. On that platform (as discussed above), Ressa was targeted in influence operations originating in China designed to foment popular support for the political campaigns of Duterte and Marcos family members. That disinformation network was removed by Facebook based on evidence of 'coordinated inauthentic behaviour'. Another feature of these credibility-based attacks is the orchestration of disinformation laden 'pile-ons' aimed at Ressa, which are designed to discount her investigations into disinformation campaigns associated with the Duterte administration. These practices could be labelled '*networked gaslighting*': the target of the attack is falsely accused of practicing the behavior of the attackers.

One early example involves Ressa's 2016 investigation Propaganda War: Weaponizing the Internet examining 'astroturfing' (the act of manufacturing consensus through influence operations designed to create the false impression of a groundswell of support within online communities) and 'sock puppet networks' of fake accounts linked to Duterte's election campaign, practices also later associated with the extrajudicial killings connected to the so-called 'drug war.' In response to Ressa's and Rappler's reporting on what they describe as "government-sponsored information operations," 'patriotic trolls' were encouraged to flood Ressa's social media zone to prove that they were "not fake accounts or bots." These swarms of comments were often prompted by pro-Duterte bloggers who encourage their followers to prove that they are "not trolls." (See the "We're not fake accounts" cluster map in the illustration above. Figure 11.).

Such campaigns have also spilled offline, increasing the physical threats Ressa is facing. In one instance involving the doxxing of Ressa with her email and office address published online, pro-Duterte social media activists came to the Rappler newsroom in person. They bypassed security designed to keep them away from the floor of the high-rise building housing the news outlet, and protested outside the glass walls of the newsroom, while holding up signs replicating some of the offensive hashtags and narratives swirling on social media.

Nevertheless, Ressa's biggest concern remains the broader impacts for society and democracy: "If you don't have facts and you don't have a shared reality, then it will be impossible to have democracy. It will be impossible to deal with the 21st century problems that face us - the virus, climate change, governance...", she trails off.

FIGURE 19

The online abuse targeting Ressa and Rappler spills offline in February 2019. These men came to the Rappler newsroom in Manila and held their 'posts' up to the glass windows of the office when they were barred entry by security.

Personal attacks designed to shame, humiliate and silence

Fourteen percent (14%) of all the posts mentioning Maria Ressa in the larger multilingual dataset retrieved from her professional Facebook page could be categorized as 'personal attacks.' These include: sexist and misogynistic abuse; the use of explicit language and images; racist abuse; homophobic abuse; threats of sexual and physical violence (including death threats); and other forms of harassment such as demeaning comments about her physical appearance associated with her skin condition (she has eczema), intellectual capacity, or mental health - using terms like "scrotum face," "imbecile," "moron," and "psycho."

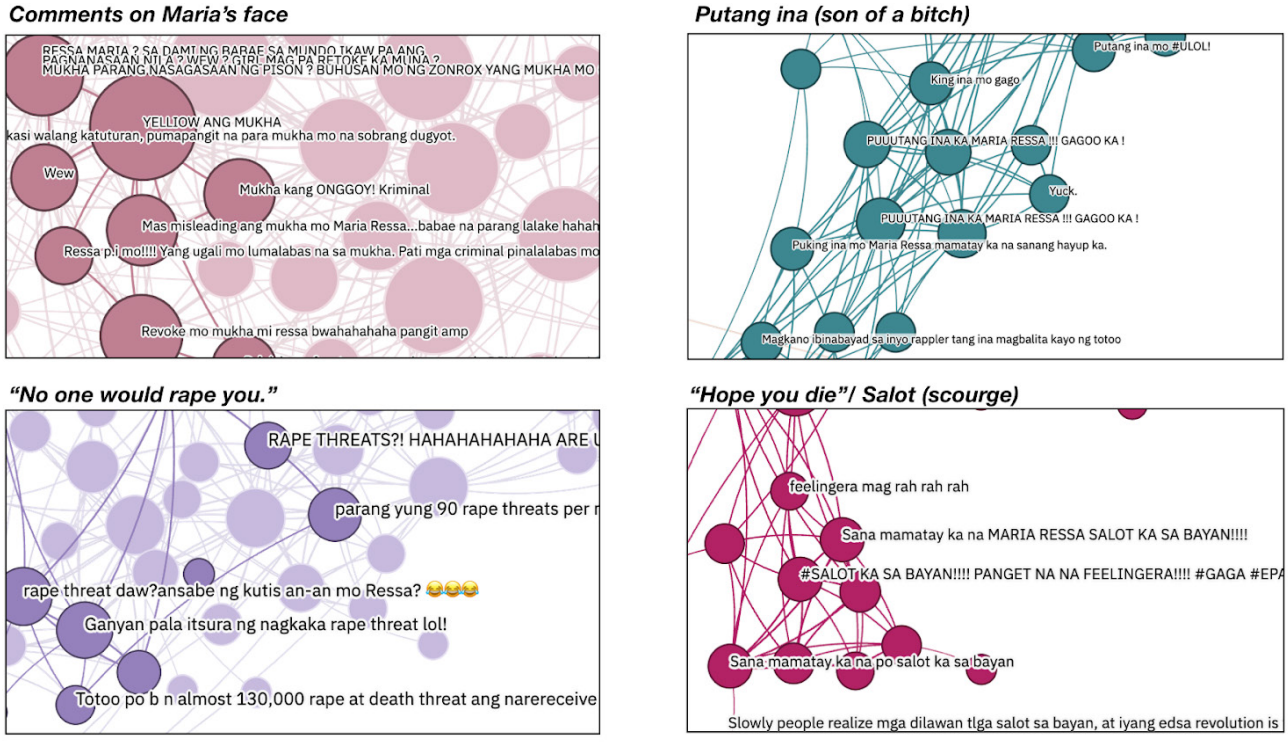

FIGURE 20

Types of personal attack abuse identified in Facebook comments to Maria Ressa.

There were substantial clusters of comments on "rape" in this dataset, as Duterte supporters took cues from pro-Duterte bloggers who falsely claimed that Ressa said she was receiving "90 rape threats every minute." In fact, she actually said she was receiving more than ninety hate messages an hour. Additionally, comments in these clusters typically insulted her physical appearance.

More disturbingly - especially in a country which remains one of the world's deadliest for journalists - outright threats to Ressa were detected in this dataset, too, with commenters saying she should be sexually assaulted, die, be killed and even "publicly raped to death."

"I've always known online violence leads to real world violence," Ressa says. This is a view shared by Caoilfhionn Gallagher QC, who points to chilling similarities between Maria Ressa's case and that of the murdered Maltese journalist Daphne Caruana Galizia, who was brutally attacked online - with misogynistic references to 'witch burning' - before she was killed with a car bomb in 2017. "There are shocking similarities between Maria and Daphne's cases, including a long period of time in which they both experienced a combination of attacks, from multiple different sources, online and offline – State facilitated and State fuelled," she says. Gallagher represents Caruana Galizia's bereaved family, who issued a statement highlighting the parallels between the cases when Ressa was convicted of criminal 'cyberlibel' in 2020.

Ressa is cognizant of the risks. "I'm aware of where this can go," she says. "But at the same time, that's also why I'm very vocal. I think that the only defense is to shine the light. I continue to do my job even better. My job in Rappler is really to hold up the sky so our team can work."

Sexism fuels the flames

Ressa was also frequently called "bobo" (dumb or stupid) in these comments, while "idiot" was a top keyword. Additionally, she was condescendingly referred to as "ang babae", or "this woman," demonstrating the sexist undertones of much of the abuse that targets her. Comments about Ressa's sexuality, including homophobic slurs (e.g., 'Tomboy' is slang for 'lesbian' in the Philippines) and outright profanities like "fuck you" and "putang ina" (son of a bitch) were also prevalent within these clusters.

Hey Maria Ressa you tomboy! You're such an idiot *presstitute*! Bias paid media!

Ugliest monkey bitch. Your face looks burnt. You are still a virgin, you beast. Nobody will hit on you

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

Throw this woman full of scars and full of kidney stones to Mindanao

FIGURE 21

Samples of Facebook comments with personal attacks against Maria Ressa. In the comments above, Ressa is called a "lesbian" and "monkey," and her skin was compared to a scrotum. She is also often condescendingly referred as "this woman."

Attacks on Ressa are frequently sexist, racist and vulgar in combination, focusing on her physical appearance. The biggest cluster under this theme of 'personal attacks' were comments related to her face ("mukha"). She was often compared to animals like monkeys and dogs (classic racist and sexist tropes) and in several instances, her eczema was compared to a scrotum - a form of abuse which has more recently grown into a viral meme that jumps platforms.

FIGURE 22

An example of a threatening (and deeply racist) meme which flies under the radar of automated detection and analysis tools. The tweet translates as: "I will give a reward to whoever captures this."

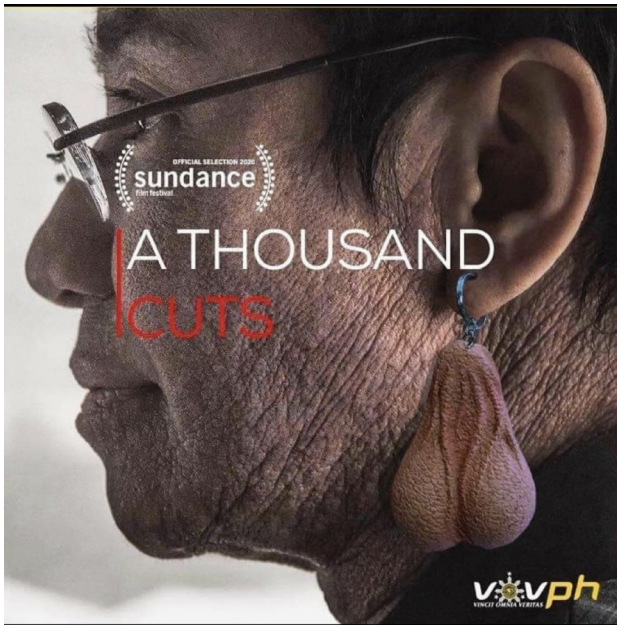

FIGURE 23

An offensive misogynistic meme manipulating Ressa's photo to mock her skin condition. VovPH is a Facebook page that spread "scrotum face".

Who are the abusive tweeters?

Of the 1,218 tweets we identified as *highly explicit abuse*, there are 878 unique authors. Of the serial abuse senders, the most prolific sent 13 tweets, while most sent fewer than five tweets. However, further qualitative analysis of the activities of these accounts revealed additional and more subtle trolling conducted by the same accounts, supporting the argument that troll accounts are targeting Ressa systematically.

Out of the 878 authors of abusive tweets in the dataset, as of 15 February 2021, 109 of these had deleted their accounts, and 51 had their accounts suspended. In total, this means that just over 18% of the accounts were no longer active. For comparison, the entire dataset of over 414,000 tweets has been authored by 112,584 distinct Twitter accounts. Among them, 3,000 (2.67%) were suspended accounts as of 15 February 2021; 14,931 (13.26%) were deleted (by the user themselves or Twitter); and the remaining 94,653 (84.07%) were still live.

Comparing the account creation dates of the live accounts, authors of abusive tweets have more recently established accounts than the other tweet authors in our dataset (average length of time from account creation until the end of the data collection is 1,795 days, compared with 2,434 days). On average they also have fewer followers, follow fewer users, and post fewer tweets (an average of 3.76 tweets/day compared with 12.8).

Deep dive: Network analysis of Twitter reactions to Ressa's June 2020 conviction

According to Rappler's Executive Editor Glenda Gloria, Maria Ressa's 2020 conviction on a criminal 'cyberlibel' charge "...really provided the trolls a powerful hashtag, because this was like a court already saying what the troll army believed and shared to be true. And so that gave them an editorial agenda - 'it's not just us saying that she's a criminal, it's the court!'"

The data analytics company Graphika conducted a detailed analysis of 196,000 tweets from 80,886 distinct users featuring citations of @mariaressa and the term "Maria Ressa" posted between June 9th and June 17th 2020. Activity began accelerating in the days leading up to the court's verdict, peaking on June 15th 2020, the day of the decision.

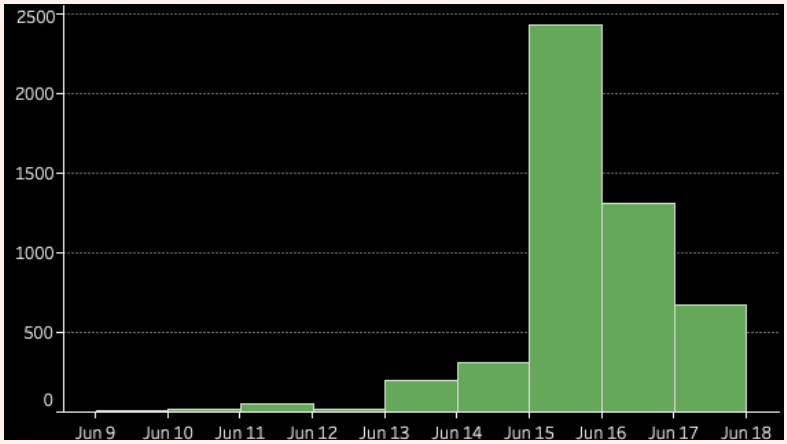

FIGURE 24

Activity (total tweets) mentioning @mariaressa or "Ressa" from pro-Duterte accounts previously identified in Graphika Maps.

Indicators of orchestrated online attacks

Mentions expressing shock at the verdict and support for Maria Ressa flooded the zone on the day of her conviction - including solidarity messages from international journalists and civil society organizations - represented the dominant group of tweets. But pro-Duterte accounts provided evidence of an orchestrated response. These accounts largely celebrated Ressa's conviction while attacking her based on her dual Filipino-American citizenship. Top hashtags shared by the group included #ISaluteJudgeMontesa - praising the judge who convicted Ressa.

Here we find some evidence of organized 'trolling' of Ressa, with over 40 accounts from the pro-Duterte segment constantly mentioning @mariaressa or citing the term "Ressa" over 30 times each within the period. Besides directly targeting her (@mariaressa), these accounts were also predominantly retweeting anti-Ressa messaging pushed by a select few accounts. As indicated by the selection of Twitter messages below, top false narratives deployed by the pro-Duterte 'troll army' in the immediate aftermath of her June 2020 conviction were:

1. She is now proven to be the criminal we said she was.

2. Disinformation about the role of the State in her prosecution (i.e. they falsely argue that the case was prosecuted by a private citizen but it was a criminal prosecution waged by the State).

3. False claims that she is a foreigner (she is a dual national but was born in Manila), subject to foreign masters (this theme helps prosecute the false argument that Rappler is foreign-owned, which is attached to a string of cases designed to shut down the news publisher).

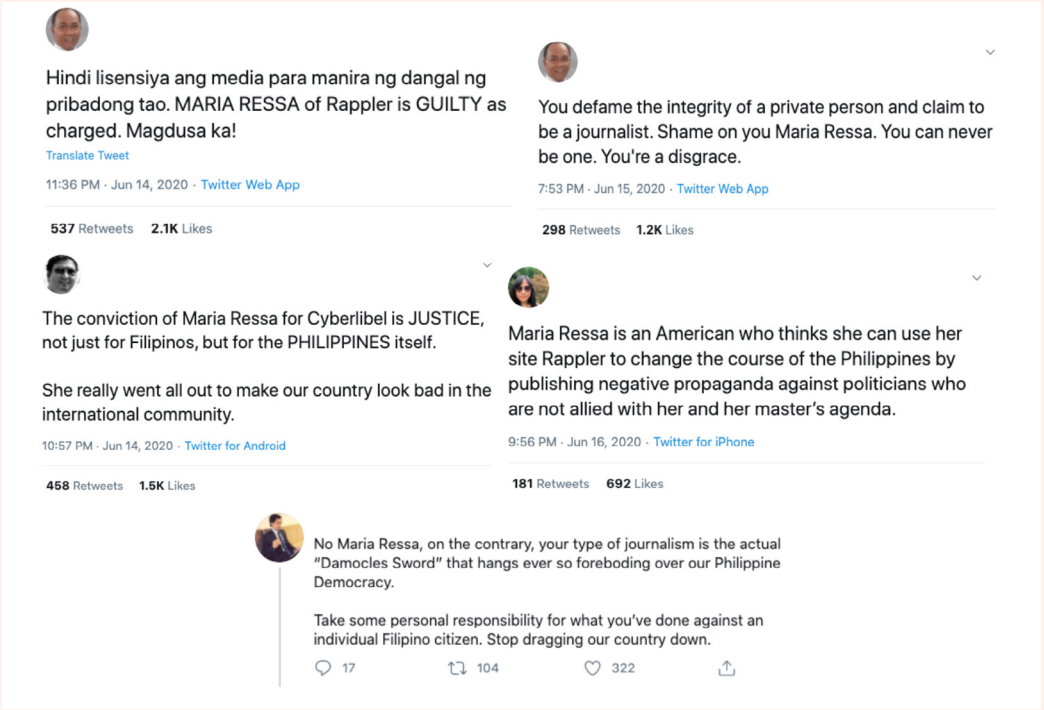

FIGURE 25

Top Retweets shared by pro-Duterte influencers who seed the meta-narratives described above (source: Graphika).

Indeed, pro-Duterte accounts showed the strongest tendency to mention other users within the collection range. Almost 60% of these accounts referenced at least four other users within the dataset in this time period. This level of interactivity is uncommon and indicative of a possibly aligned/coordinated harassment campaign in which users aim to amplify attacks. A large number of the pro-Duterte accounts were also fairly recently created; for instance, 103 of these accounts (out of almost 600) were created in 2020. Accounts created after 2019 also tended to be more active (see activity line at the bottom of the chart below).

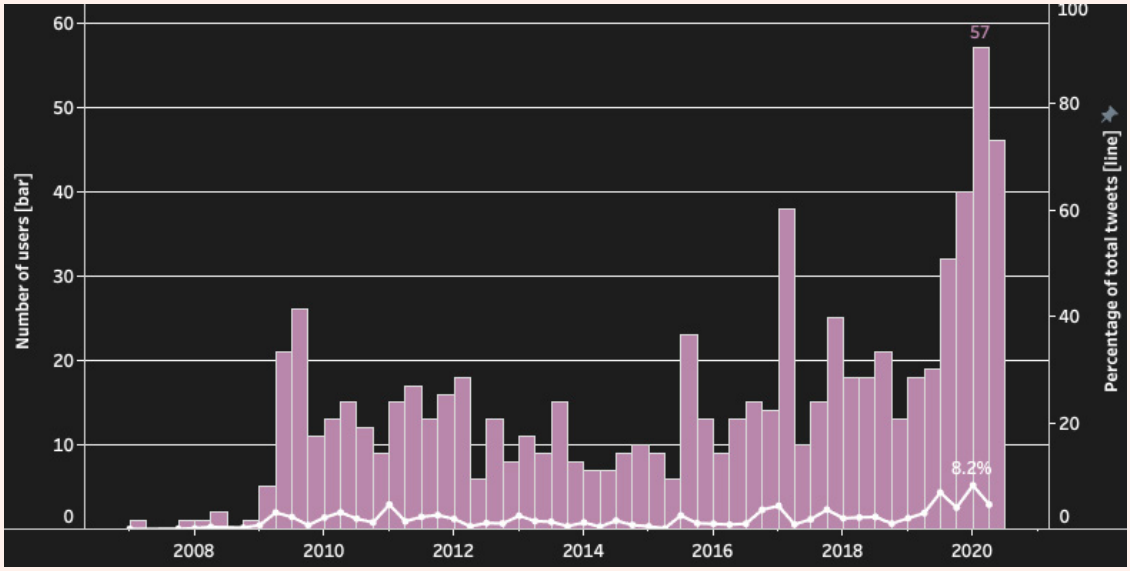

FIGURE 26

Distribution of pro-Duterte account creation dates for accounts targeting Ressa (source: Graphika).

These accounts posted a little over 2,400 times and appear to have been quite active in their targeting of Maria. Including retweets, 186 accounts from this set, for instance, mentioned @mariaressa or "Maria Ressa" four or more times, and 59 accounts did so 10 or more times. Over 25% of these accounts (180 total) were created within 2020, and 18% (126) were created from April on. The accounts created during or after April 2020 were also the most active within this set - producing about 27% of the total activity (662 tweets). Finally, a significant number of these accounts have low follower counts with 5% having zero followers and a little over 25% having 10 or less followers.

This combination of high activity, low follower count, and recent creation date are together possible indicators of accounts created in bulk to amplify pro-Duterte messaging and target government critics.

The role of the platforms

As our big data analysis demonstrates, while President Duterte and his digital mob instigate and fuel the attacks on Maria Ressa, Facebook - which is equated with the internet in the Philippines - is the major vector for the disinformation-laced online violence she experiences.

Rappler was 'born' on Facebook and Duterte rode the platform to victory in 2016. Ressa, Duterte and Facebook are intertwined combatants in the Philippine 'disinformation war'.

Ressa and the Rappler team have consistently flagged the online attacks with Facebook, which has largely done little, they say. Ressa's page was overwhelmed by attacks for years, and posts that incite violence, despite violating Facebook's community standards, incited more attacks. Today, most of the abuse remains visible on Ressa's Facebook page.

FIGURE 27

Samples of comments on Ressa's Facebook page that incite harm.

Rappler responded to the attacks with stricter comment moderation policies, but the attacks on Ressa's page only dropped when she wound down posting in the second half of 2018. Her last post was in mid 2019.

In September 2018, when she was still both Executive Editor and CEO of Rappler, Ressa was one of the speakers at a Facebook-convened meeting called to address the problem of hate speech on the platform. She told the Facebook executives present: "If you don't change what you're doing, I could go to jail." After she was convicted in June 2020, she apportioned blame to the company for her conviction, and she continues to argue that

Facebook has enabled the destruction of democracy in the country.

"The design of social media turned 'wisdom of the crowds' into the mob. It's the chaos of a mob. And beyond that, it's actually pumping hate into the system," Ressa says. She believes the platforms need radical renovation and regulation - of business models and design - to stop the toxicity that overruns them. "I don't think anything is possible until we clean up the information ecosystem, until you stop the virus of lies," she says. "It's a perfect comparison to the COVID-19 virus, because the virus of lies is very contagious. And once you're infected, you become impervious to facts. You have to be healed. You have to be rehabilitated."

Ressa is particularly critical of Facebook's moderation policies and practices, saying that their automated reporting systems just do not work when it comes to dealing with online violence against women journalists. "I have very rarely had anything taken down when trying to report attacks using Facebook's standard online reporting system," she says.

"The only times Facebook has done something about the attacks against me is when I have gone directly to people I know inside the company, many of whom have nothing to do with content moderation! Their systems need to be clearer, faster and more responsive to the impact of their inaction."

Twitter is also a significant distributor of abuse against Ressa, although she says she feels "much safer" on that platform - especially since the company began working harder over the past few years to protect women journalists and human rights defenders among its users. "I feel like Twitter is prioritizing human rights activists and journalists under attack - and I've spoken to others who feel the same way," Ressa says. "Their reporting tool aggregates similar tweets, takes less time, and is far more effective in takedowns."

Ressa has long spoken about being the 'canary in the coal mine,' warning that the Philippines' corrupted information ecosystem represents the West's dystopian future. That's a prediction that seems much more real after the January 2021 storming of the U.S. Capitol. She is also extremely pessimistic about the prospect of the platforms responding quickly and effectively enough to address the crisis illustrated through her lived experience.

She sees regulation, accountability and liability as key elements of any serious response: "The only way it will stop is when the platforms are held to account... because they allow it. It's kind of like if you slip on the icy sidewalk of a house in the US, you can sue the owner of the house," she says. "Well, this is the same thing. They have enabled these attacks. They've certainly changed my life in many ways."

Rappler's Executive Editor Glenda Gloria goes further, illustrating the role of 'platform capture' in creating the enabling environment for the online violence experienced by Ressa and her staff: "They're like the government, the UN rolled into one - but you can't withdraw from it," she says.

Rappler now works at the intersection of investigative journalism, advocacy,

research and policy to respond proactively to the information ecosystem crisis. It continues to be a Facebook fact-checking partner in the Philippines, and a collaborator on investigations into disinformation networks which, in some cases, have resulted in the company removing clusters of inauthentic accounts from the platform. But, as Glenda Gloria explains, Rappler management remains highly critical of Facebook:

"I have to be critical because my government runs my life, Facebook runs my life and Rappler's in so many ways. And so the critical lesson is not just to engage the platforms, not just to make them accountable, but to have an institutional approach to it, and that is not just Rappler or the Philippines, it has to be global."

Glenda Gloria

There is increasing debate internationally about platform regulation and legal remedies in response to the intersecting threats of viral disinformation and hate speech, however there is also a need to balance such responses against the threat they represent if operationalized to undermine press freedom.

FIGURE 28

Ressa tweets from a 2018 conference where she sat on a panel with Facebook's Public Policy Director, Katie Harbath.

However, the role of Duterte and his 'troll army' in inciting and fueling attacks on Ressa and Rappler should not be underestimated. In the days before this study was finalized, Rappler was 'red-tagged' by the spokesperson for the National Task Force to End Local Communist Armed Conflict (NTF-ELCAC) - an anti-communist agency launched by Duterte. 'Red-tagging' is a peculiarly Filipino approach to attacking critics, including journalists, by labelling them as communists, or 'reds'. After Rappler published two fact-checks which debunked other 'red-tagging' efforts by the Task Force, NTF-ELCAC spokesperson Lorraine Badoy labeled Rappler "a friend and ally" and "mouthpiece" of communist rebels on Facebook. She also suggested to her followers that action would follow.

"The next coming days will lend even greater clarity to this claim of mine," she wrote. Considering designation as a terrorist sympathizer could result in summary detention under the Philippines' new anti-terror law, this disinformation-laced attack on Rappler represents a significant threat to Ressa. "When you're labeled a terrorist, you can be arrested under this law without a warrant and held for up to twenty-four days. So these are not idle threats," she says.

The National Union of Journalists in the Philippines issued a statement condemning the attacks on Rappler and, on this occasion, Facebook has removed the offending content and blocked the NTF-ELCAC spokesperson's account.

The day we sent this publication to press, nine activists who had been labelled as communists were killed in a crackdown that human rights organizations called 'Bloody Sunday'. This came two days after Duterte ordered the military and police to kill insurgents without regard for human rights if they were carrying weapons.

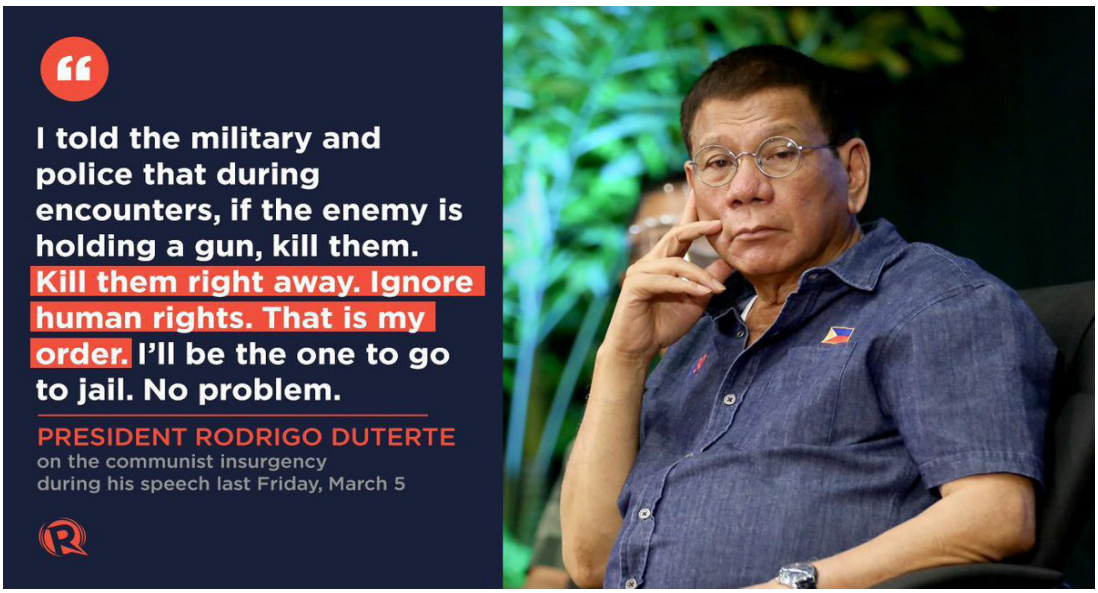

FIGURE 29

Image attached to a tweet sent by Rappler on Bloody Sunday - March 7th, 2021

Conclusion and recommendations

This case study has presented a multidimensional picture of what it is like to be a woman journalist under fire on social media platforms. It also represents the first major case study of its kind focused on the torrent of online violence facing women journalists who work on the new front line of journalism safety - at the epicenter of Digital Age risks. They are attacked, threatened, abused, discredited and harassed on a daily basis, often in targeted assaults, or as part of disinformation campaigns designed to chill their critical reporting and shut them up. Frequently, these attacks are instigated or licensed by political actors and fuelled by coordinated 'patriotic trolls' or misogynistic mobs who are facilitated by social media platforms. They are also increasingly spilling offline. This is the perfect enabling environment for the sort of State-led persecution and prosecution that Maria Ressa faces in the Philippines. But this toxic phenomenon knows no borders.

This report was produced under the umbrella of a larger multi-country study into online violence against women journalists, commissioned by the United Nations Educational, Scientific and Cultural Organization (UNESCO), to be published in 2021. The aim of the broader project is to trigger action to protect women journalists online. However, this report is the exclusive responsibility of ICFJ.

"Westerners tend to think strength is like an oak tree. It's big, it's strong. It looks imposing. I like to think of strength as bamboo. Because when there's a really big storm, the oak tree can't bend with the wind. And so it's uprooted and dies. But bamboo sways and bends with the typhoon, weathering the storm. And in the aftermath, there it remains, standing strong!"

Maria Ressa, Rappler CEO and co-founder

Below, we offer **10 key recommendations** for action based on this case study:

1. Political actors should be held accountable for inciting online violence against women journalists - by their political parties, parliaments, the government agencies responsible for administering and implementing human rights frameworks, and the international organizations they have obligations within.

2. States should review and adapt (where necessary) legal and regulatory frameworks designed to uphold freedom of expression and equality, to ensure they can be applied effectively online as well as offline.

3. States should support truly independent and transparent social media councils and/or national ombuds facilities to give victims of online abuse recourse to independent arbitration.

4. News organizations should provide integrated digital and physical security, as well as psychological and peer support, for women journalists targeted online. They should also provide gender-sensitive training, guidelines, policies and resources designed to mitigate the impacts of online violence.

5. News organizations should hold the platforms to account through critical independent journalism in the same ways that they hold governments to account, and advocate for the protection of their staff, regardless of any commercial ties with social media companies.

6. Social media companies should create specialist teams to respond swiftly to attacks on women journalists, recognizing the particular journalism safety and freedom of expression threats facilitated by their platforms (as noted by the UN). Such teams should provide accessible, human contact points in every language in which the company functions.

7. Governments and international organizations should work with platforms to ensure better protection of journalists and human rights defenders from online abuse and to co-develop a culture of transparency, trust and accountability. Regulatory responses may be necessary to ensure this.

8. Social media companies should provide researchers with privacy-preserving, equitable access to key data from the platforms (without strings attached), to enable independent analysis of the incidence, spread and impact of online abuse on female journalists, and to contribute to technological and policy solutions.

9. Law enforcement and the judiciary should ensure gender-sensitive and digitally-literate responses to online violence against women journalists. This could be aided by access to appropriate expertise and training for law enforcement agents and members of the judiciary.

10. Women journalists under attack should be empowered to fight back against online violence, but they should not be made to mitigate the abuse they face online independently, nor should they be made to feel responsible for being targeted.

Further resources:

- A global snapshot of the incidence and impacts of online violence against women journalists based on a survey conducted by ICFJ and UNESCO.
- A PBS Frontline conversation about the documentary film A Thousand Cuts, and the role of online violence in María Ressa's ordeal. The discussion features Ressa, Frontline's Executive Producer Raney Aronson, filmmaker Ramona Diaz, lead author of this study Julie Posetti, and UNESCO Freedom of Expression Director Guy Berger. The panel was moderated by Lana Wilson.
- A profile on Maria Ressa's fight back against online harassment and abuse published by UNESCO.
- A comprehensive set of recommendations for action on the problem of online abuse, from Article 19.
- A newsroom protocol for responding to online abuse against journalists from the International Press Institute (IPI).
- Pen America's Online Harassment Field Manual.
- IWMF-ICFJ Online Violence Response Hub (launching 2021).

About the authors:

Dr. Julie Posetti is Global Director of Research at the International Center for Journalists (ICFJ). She is an award-winning journalist who is also academically affiliated with the University of Sheffield's Centre for Freedom of the Media (CFOM) and the University of Oxford's Reuters Institute for the Study of Journalism. She leads the Online Violence Against Women Journalists Project for ICFJ, under commission from UNESCO.

Dr. Diana Maynard is a Senior Research Fellow at the University of Sheffield and a member of the Centre for Freedom of the Media (CFOM), specialising in Natural Language Processing, social media analysis, and media freedom issues.

Professor Kalina Bontcheva is a Research Professor of Text Analytics at the University of Sheffield and member of the Centre for Freedom of the Media (CFOM), specialising in analysing online abuse and disinformation.

Rappler social media analytics staff **Kevin Hapal** and **Dylan Salcedo** contributed to data collection, analysis and visualisation.

Cover photo credit: Franz Lopez/Rappler

ICFJ International Center for Journalists

The University Of Sheffield.