Maria Ressa

At the core of an online violence storm

First, I'm attacked for being a journalist, second I'm attacked for being a woman.

Maria Ressa

Death threats. Rape threats. Doxxing. Racist, sexist, and misogynistic abuse. In text, image and memes. These are just some of the features of the online violence that Maria Ressa, the Filipino-American journalist who founded the Manila-based news site Rappler, has faced daily since the Philippines' 2016 election. Ressa is an emblematic case study in the global scourge of online violence against women journalists. She lives at the core of a very 21st century storm. It is a fusillade of disinformation and attacks - one in which credible journalists are subjected to online violence with impunity; where facts wither and democracies are put under severe strain. The former CNN war correspondent and laureate of the 2021 UNESCO World Press Freedom Prize says none of her experiences in the field prepared her for the massive and destructive campaign of gendered online abuse, threats, and harassment directed at her over the past five years. At one point, in response to an investigative series on State-linked disinformation, she recorded receiving more than 90 hate messages an hour on Facebook.

This collaborative case study[14] represents a forensic analysis of hundreds of thousands of Facebook and Twitter posts between 2016 and 2021 directed at the internationally celebrated digital media pioneer. It highlights the intensity and ferocity of attacks designed not only to vilify a journalism icon, but to discredit journalism itself, and destroy 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 decades in jail.

---

[14] A detailed version of this Participatory Action Research case study (including methodology and dataset descriptions) is available here: https://www.icfj.org/our-work/maria-ressa-big-data-analysis

Ressa is not only attacked for being a journalist. She is attacked for being a woman. For the colour and texture of her skin. For her dual citizenship. And for her sexuality. The attacks against Ressa are enabled by a climate in which President Rodrigo Duterte has publicly condemned Ressa - while also once musing that journalists are not exempt from assassination.

The worst attacks appear to have been orchestrated.

Despite belated and largely ineffective attempts to address hate speech within the social media ecosystem, the technology companies concerned serve as the vectors for these attacks, which in turn create a facilitative environment for State-led legal actions against Ressa. This 'lawfare,' as she refers to it, led to her conviction in mid-2020 on a 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, Ressa is currently fighting nine separate cases. If she is convicted on all charges, she could spend the rest of her life in jail. **Caoilfhionn Gallagher QC**, the co-lead of Ressa's international legal team, states that for having published journalism, her client now faces a barrage of baseless lawsuits that seek to criminalise her work and expose her to a century in prison:

> " *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.* "

This case study combines big data analysis - using Natural Language Processing (NLP) techniques (merging linguistics, computer science and Artificial Intelligence) and network analysis - with deepdive interviews and extensive desk research. Our analysis of the online violence Ressa experiences helps us understand how and why she is targeted, and how the threats spread. It also gives us insights into the role of State actors in manufacturing consensus by fuelling the behaviour 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 said. "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."

None of her experiences in the field prepared her for the massive and destructive campaign of gendered online abuse, threats, and harassment directed at her over the past five years.

12 KEY FINDINGS

**from our examination of online violence
against Ressa over a five-year period (2016 to 2021)**

Attacks designed to discredit Ressa's journalism:

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

// Credibility or reputation-based attacks frequently deployed disinformation tactics and abuse conflating Ressa and her journalism with "fake news".

Attacks designed to undermine Ressa's personal dignity:

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

// The most prevalent type (34%) of personal abuse against Ressa is sexist, misogynistic and explicit. Racist and homophobic abuse accounted for 5% of personal abuse.

// 14% of all of the abuse studied is sexist, misogynistic and explicit.

Networked and technically advanced abuse:

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

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

// There is evidence that the worst of the attacks on Ressa involve coordination or orchestration - often associated with disinformation campaigns led by State actors.

Sources, triggers and impacts of online violence against Ressa:

// Much of the abuse and threats are fuelled by statements and messages from political leaders and partisan bloggers/social media influencers who demonise Ressa and Rappler as unpatriotic criminals. It is also amplified by pro-government news media.

// 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**.

The role of the platforms:

// Facebook is the main vector for the online violence Ressa faces. It is also the most used social media site in the Philippines.

// Both Facebook and Twitter have promised to address the attacks on Ressa but during the period under review, Facebook failed "woefully", she said, to effectively stem the tide of hate against her. However, she now feels "significantly safer" on Twitter.

A climate of impunity

The Philippines is routinely ranked as the most dangerous country in Southeast Asia for journalists, and UNESCO's Observatory of Killed Journalists identifies 109 fatalities since 2006, with 83% of cases currently unresolved. Targeted online violence attacks against journalists like Maria Ressa thus need to be examined in the light of the impunity situation. 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 30 journalists and media workers[15] in an "orgy of political violence." Immediately before President Rodrigo Duterte took office in 2016, he issued threats against journalists and a year later, he criticised Ressa in a State of the Nation Address. Since 2017, political influencers have been 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."

In a country like the Philippines, where extrajudicial killings continue to be condemned by UN actors, 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. There was this female activist who was first blasted online and shot while on her way home. It's real. Especially against women."

i. "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 56,000 posts and comments published on Facebook between 2016 and 2021.

The tweets were gathered and analysed using Natural Language Processing (NLP) techniques. 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.

UNESCO OBSERVATORY OF KILLED JOURNALISTS

83%

of cases currently unresolved in the Philippines

109

journalists killed in the Philippines since 2006

---

[15] UNESCO officially records 30 journalists murdered during this massacre: https://en.unesco.org/themes/safety-journalists/observatory/country/223790

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 over 47,000 other public Facebook posts mentioning Ressa extracted from a massive database that Rappler maintains, which they call 'Sharktank.' According to Ressa, 'Shartank' 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. The same NLP techniques were applied to both sets of 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.

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 and save for big data analysis, it is also more difficult to report because automated abuse reporting systems have very limited multimedia capability and they do not accommodate nuance. Similarly, public abuse is often subtle and potentially designed to 'slip under the radar.'

They were also calling for her to be sexually assaulted, killed and even "raped repeatedly to death".
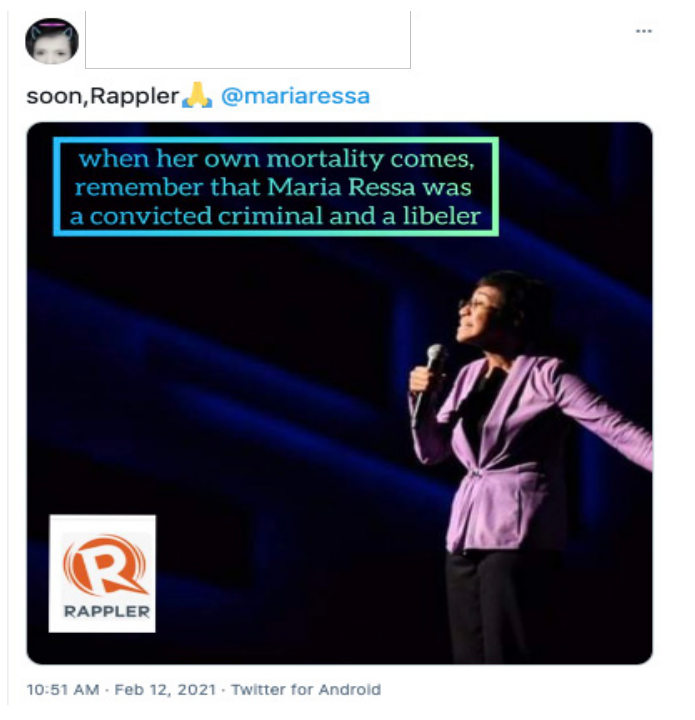
A death threat sent to Maria Ressa via Twitter on February 12, 2021. The threat was finally removed by Twitter in mid March 2021.

A classic example of this is a death threat sent to Ressa on Twitter on 12 February 2021. This is subtle partly because it carries 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 finally deleted in mid March 2021 after being reported to the platform by the lead author and others.

> *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.*

Glenda Gloria
Rappler's Executive Editor

ii. METHODS, THEMES AND TROPES DEPLOYED BY THE ATTACKERS

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. An examination of the combined datasets surfaced the following findings:

● **Dominant themes and tropes**

// **Abuse damaging to Ressa's professional reputation or credibility includes disinformation designed to discredit her as a journalist and erode trust in her journalism. It features calls for her to be charged, tried, raped, imprisoned etc. for her work. This abuse frequently involves false claims that she is a purveyor of "fake news" and includes the pernicious hashtag #presstitute.**

// **Misogynistic, sexist and explicit abuse includes abuse targeting Ressa's physical appearance (emphasising her skin condition) and manipulated photographs depicting her head associated with male genitalia.**

// **Racist abuse and memes constitute 3% of the personal abuse Ressa receives.**

// **Homophobic slurs designed to question her sexuality and increase her vulnerability were determined to represent 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.**

● Triggers for attacks

The attacks spike in association with:

// Rappler's investigative journalism focused on the 'drug war', and the erosion of democracy in the Philippines.

// Ressa's reporting and commentary on disinformation and the government.

// Ressa's high-profile media appearances.

// Ressa's international awards and civil society statements of support.

// Ressa's court appearances.

● Typical methods of attack

// Key significant attacks appear to be orchestrated (with the detectable 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 inconsistent and "woefully inadequate," in her words.

// Hashtags designed to encourage swarms of attackers and fuel 'patriotic trolling' are frequently used, and sometimes include abuse 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 is used to motivate Ressa's online attackers to attack her offline as well.

Spikes and triggers

The largest surge of attacks we identified against Ressa on her Facebook page occured in October 2016, when Rappler published a three-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 the '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), her high-profile industry awards, and her arrests, detentions, and trials are also associated with increased attack spikes in the Facebook data.

An examination of the Twitter data detailing frequency of abuse 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, and the error was used to attack her credibility. The next biggest spike came in June 2020, when a Manila court delivered a guilty verdict against Ressa in a criminal cyberlibel charge prosecuted by the government. Her chastisement of the government due to the Philippines being recorded as having the highest proportion of COVID-19 cases in Asia in August 2020 represented the third-biggest spike.


The 100 most prevalent abuse terms in the analysed twitter data. Term size reflects frequency of occurrence.

iii. SCROTUMS, MONKEYS, SWARMS AND LIES

Attacks against Ressa and Rappler dominated the comments on her professional Facebook page, which she established in 2015. According to Ressa, Facebook had 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 to 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 revealed that approximately four of these 14 abusive comments would constitute personal attacks, focusing on her appearance, nationality, gender and sexuality.

54%
attack clusters

4%
supportive comments
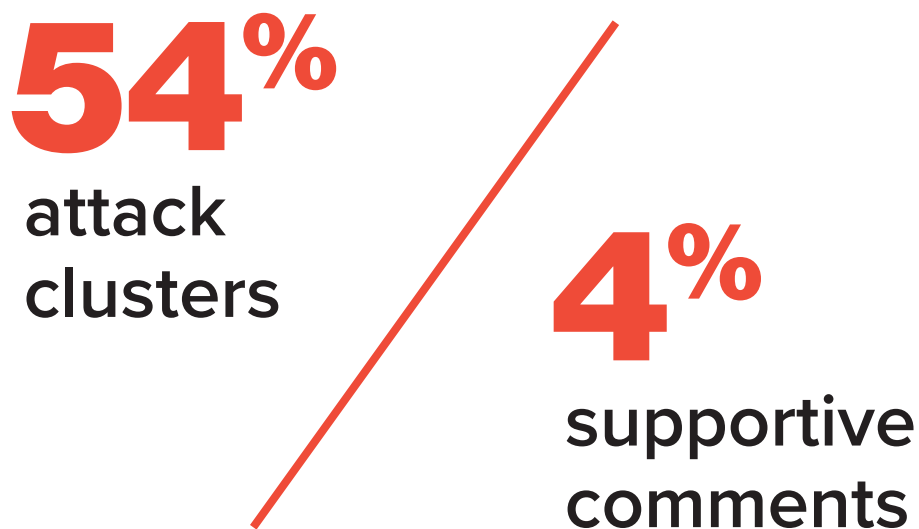
Attacks designed to discredit Ressa's journalism and erode public trust in facts

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 a smaller 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 f**king [eds**] Bitch (sic) "

" Stop spreading Lies you Piece of SHITS (sic) I wish you Rotten (sic) in Jail ... "

" Maria Ressa get the f**k [eds' asterisks] out of our country Philippines! dont (sic) mislead the people with your fake news ... "
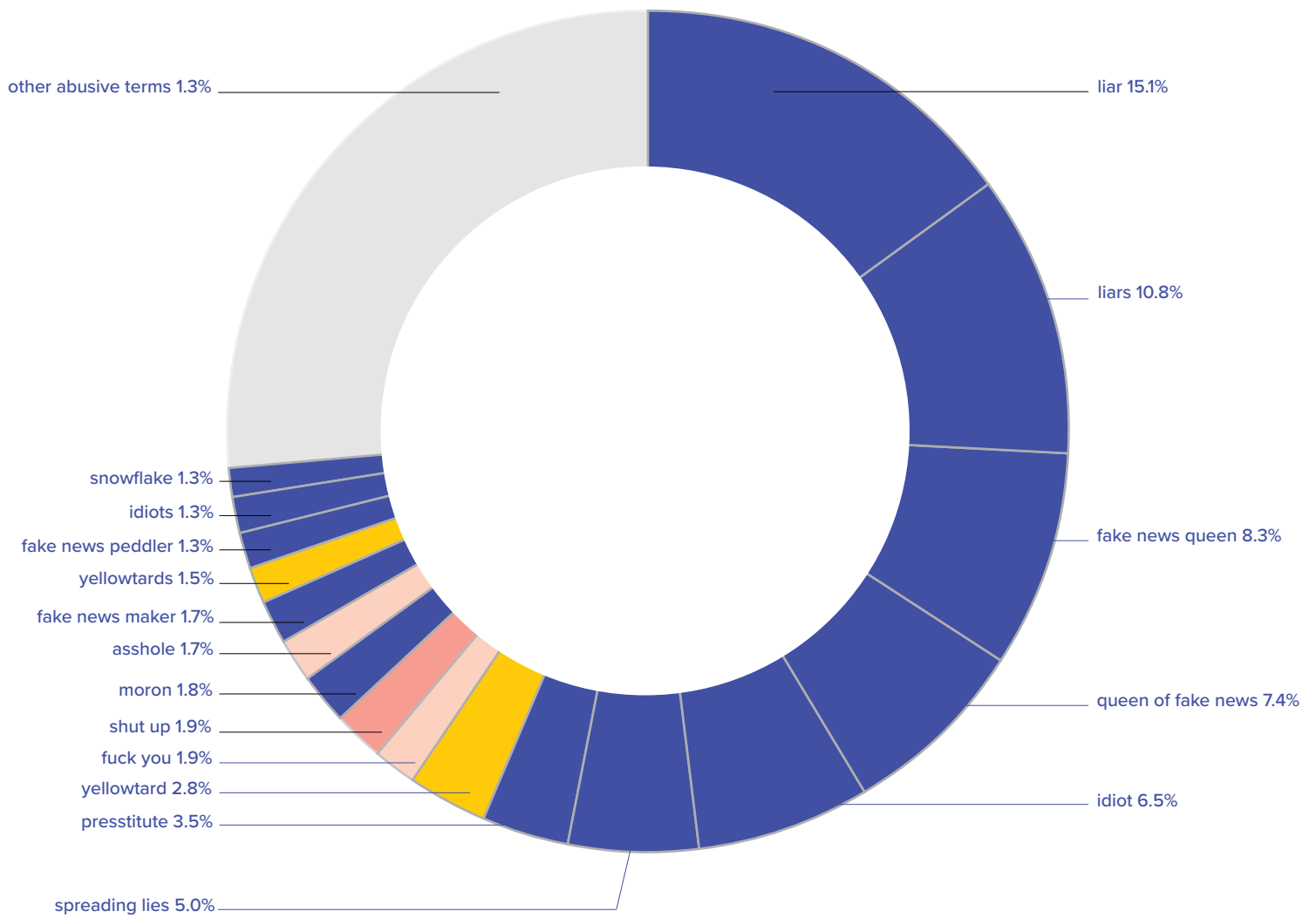
The most frequently used abusive terms in this data subset from Ressa's Facebook page were words designed to ridicule, silence and discredit her while simultaneously undermining public trust in her critical journalism. The top ranked words were "idiot," "shut up," "presstitute," and variations on "liar." Around 20% of the attacks on her credibility were related to disinformation - either equating her with it, or falsely accusing her of peddling it. Ressa has assessed that: "Lies spread faster than facts. And lies laced with anger and hate spread faster and further than facts."

Disinformation narratives deployed against Ressa were also prevalent in the larger multilingual dataset extracted from her professional Facebook page. These include repeats of 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 a prominent Philippine blogger, known as Thinking Pinoy, who tried to get the hashtag #ArrestMariaRessa to trend in May 2017. Two years later Ressa was in fact arrested, and within three years she would be convicted on a criminal cyberlibel charge prosecuted by the State.
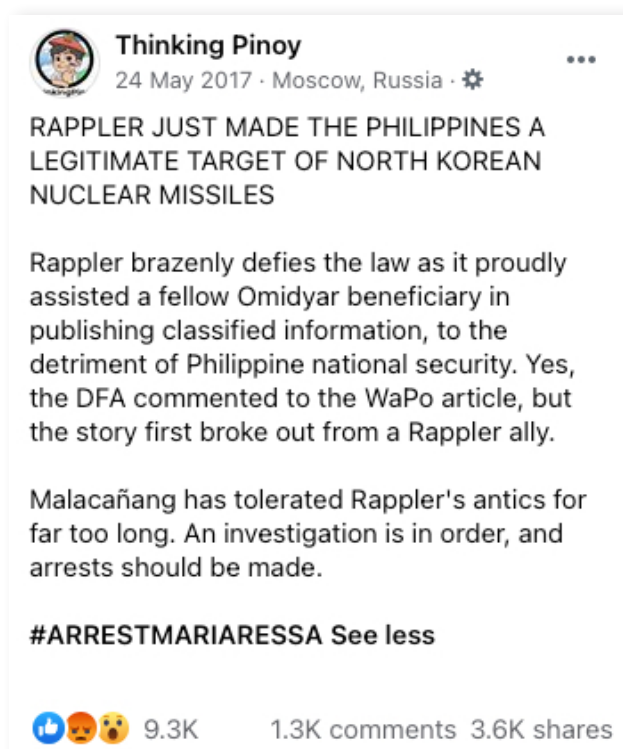
A screengrab of a Facebook post published by the. pro-government blogger 'Thinking Pinoy'[16].

---

[16] 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-government influencers targeting Maria Ressa and Rappler.

other abusive terms 1.3%

liar 15.1%

liars 10.8%

fake news queen 8.3%

queen of fake news 7.4%

idiot 6.5%

snowflake 1.3%
idiots 1.3%
fake news peddler 1.3%
yellowtards 1.5%
fake news maker 1.7%
asshole 1.7%
moron 1.8%
shut up 1.9%
fuck you 1.9%
yellowtard 2.8%
presstitute 3.5%

spreading lies 5.0%

This 'donut chart' visualises the frequency breakdown of abusive terms that appear more than once in Facebook posts mentioning Maria Ressa contained within the large 'Sharktank' dataset ("other abusive terms" includes those which appear between one and 15 times in the data). Personal attacks consist of sexist, misogynistic and explicit sexual terms (light pink); and other kinds of personal insult (dark pink). Political attacks (yellow) use terminology associated with (real or imagined) political affiliation. Credibility attacks are coloured in blue.

Online violence against Ressa undoubtedly impacts on decisions in the progress of the cases and charges she faces according to her international lawyer, **Caoilfhionn Gallagher**, who said that when the administration is making decisions in that environment, the fact that Maria is a hate figure online is enabling those decisions to be taken. Referring to the decision to charge Ressa with a third count of criminal cyberlibel in January 2021, Gallagher said:

> " *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 State "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."

According to Rappler's Managing Editor Glenda Gloria, Ressa transformed into a "warrior" in response: "…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 said.

Foreign actors

Foreign States also feature extensively in the tweets identified as containing highly explicit abuse, as well as a conspiracy theory that Ressa was in league with some of them who were allegedly paying her to spread lies. This was also a theme evident in the Facebook data. On that platform, Ressa was targeted in influence operations originating in China designed to foment popular support for political campaigns in the Philippines. That disinformation network was removed by Facebook based on evidence of 'coordinated inauthentic behavior'.

Networked gaslighting

Another feature of 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 political interests. These practices are features of what we called networked gaslighting. In Ressa's case, the target of the attack is falsely accused of practicing the behaviour of the attackers.

One early example involves Ressa's 2016 investigation Propaganda War: Weaponizing the Internet examining 'astroturfing'[17] and 'sock puppet networks' of fake accounts linked to President 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' swarmed Ressa's social media zone, often prompted by partisan bloggers who encourage their followers to prove that they are not paid trolls.

"Even if some of the people who send Maria horrendous, abusive images have no known direct link to the State," Caoilfhionn Gallagher QC said, "they are likely to be doing this, or feel emboldened in doing this, because the State is sending the message that this person is fair game." These campaigns have 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-government social media activists came to the Rappler newsroom in person. They bypassed security 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.

---

[17] The act of manufacturing consent through influence operations designed to create the false impression of a groundswell of support within online communities.
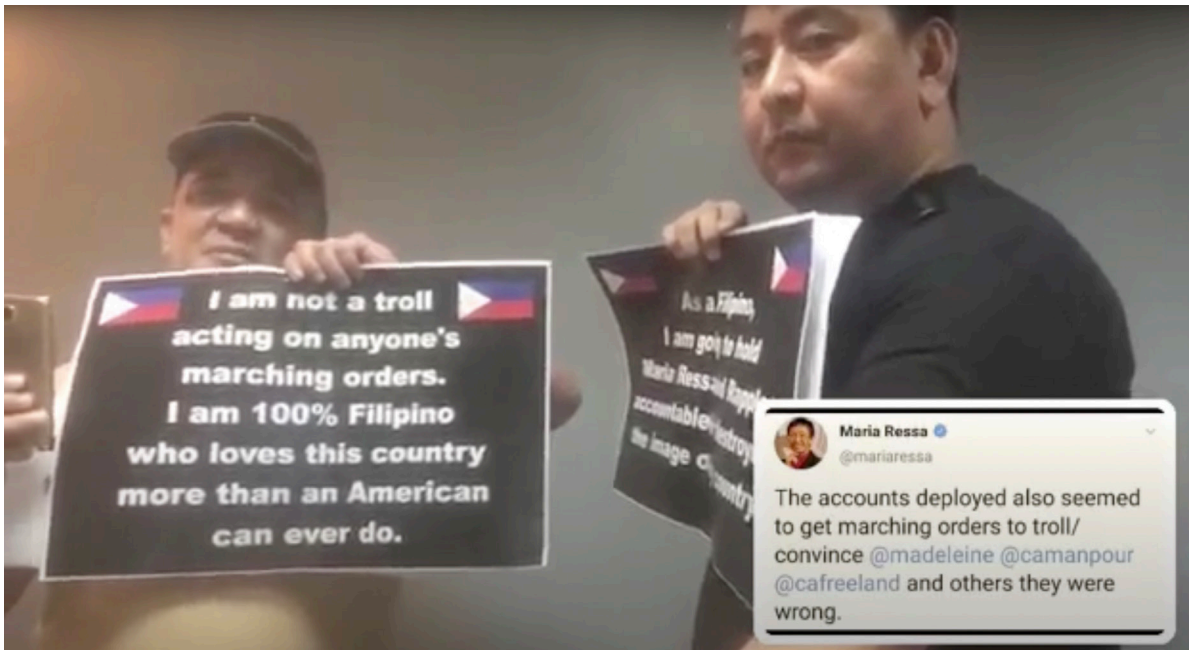
The online abuse targeting Ressa and Rappler spills offline in February 2019. These men came to the Rappler newsroom in Manila and held their 'posts' up to the glass windows of the office when they were barred entry by security (image: Rappler).

Personal attacks designed to shame, humiliate and silence

A total of 40% of all the abusive posts studied could be categorised as 'personal attacks' and these are frequently sexist, misogynistic, racist and vulgar in combination, focusing on Ressa's physical appearance. The biggest cluster 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. Terms like "scrotum face," "idiot," "imbecile," "bobo" (dumb or stupid), "moron," and "psycho" were prevalent.

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 "f**k [eds' asterisks] you" and "putang ina" (son of a bitch) were also prevalent within these clusters. 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, or be killed.
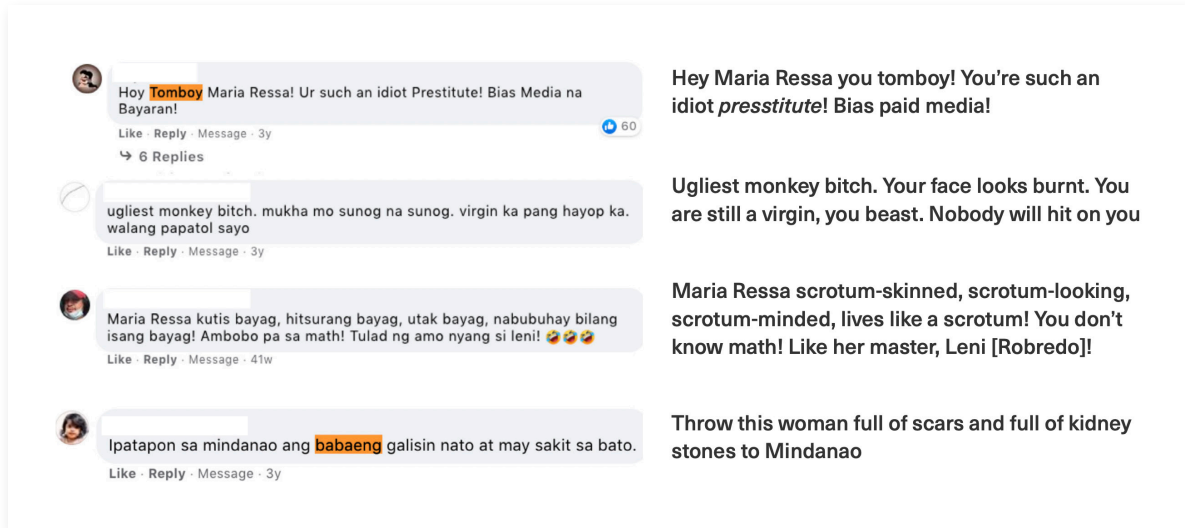
Samples of Facebook comments with personal attacks against Maria Ressa. In the comments above, Ressa is called a "lesbian" and "monkey," and her skin was compared to a scrotum. She is also often condescendingly referred as "this woman".

"I've always known online violence leads to real world violence," Ressa said. 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." Gallagher represents Caruana Galizia's bereaved family, who issued a statement highlighting the parallels between the cases when Ressa was convicted in 2020. "I'm aware of where this can go," Ressa said. "But at the same time, that's also why I'm very vocal. I think that the only defence 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."

iv. DEEP DIVE: NETWORK ANALYSIS OF TWITTER REACTIONS TO RESSA'S JUNE 2020 CONVICTION

According to Rappler's Executive Editor Glenda Gloria, Maria Ressa's June 2020 conviction on a criminal cyberlibel charge "...really provided the trolls [with] 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!'"
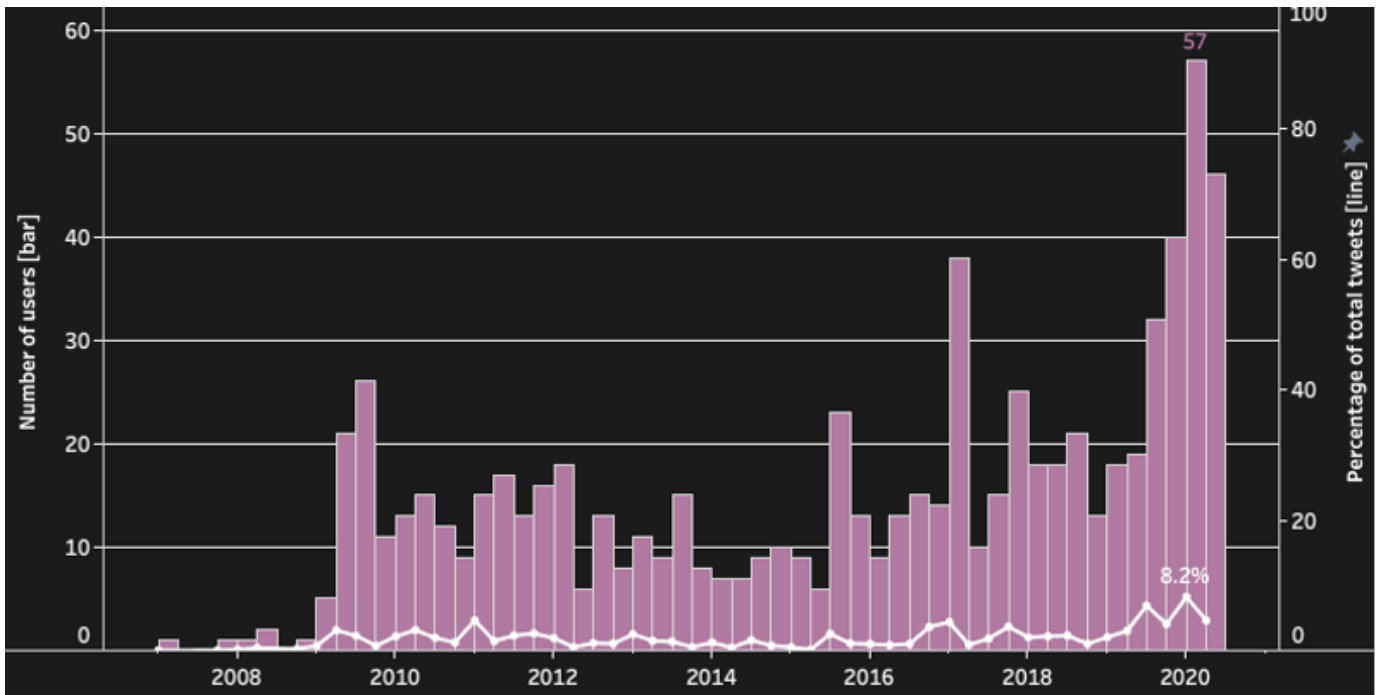
Distribution of pro-government account creation dates for Twitter accounts targeting Ressa (source: Graphika).

In parallel with our research, 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. While '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 organisations - dominated, pro-government accounts provided evidence of an orchestrated response. These accounts largely celebrated her conviction while attacking Ressa based on her dual Filipino-American citizenship.

Here we find strong pointers to organised "trolling", with over 40 accounts from the pro-government 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.

Top false narratives deployed by the pro-government 'troll army' in the immediate aftermath of Ressa's 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 / She is a foreigner (she is a dual national), 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).

Pro-government accounts showed the strongest tendency to mention other users in regard to this narrative. Thus, 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-government accounts were also fairly recently created. 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-government messaging and target government critics.

v. THE ENABLING ROLE OF THE PLATFORMS IN MARIA RESSA'S ABUSE

As our big data analysis demonstrates, while a coordinated and vitriolic mob of digital aggressors 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 President Duterte rode the platform to victory in 2016. Ressa and the Rappler team have consistently flagged the online attacks with Facebook, which has largely done little to deal with the problem, they said. Ressa's Facebook page was overwhelmed by attacks for years, and posts that incite violence, despite violating Facebook's community standards, remained visible on the page at the time of writing. Rappler responded to the attacks with stricter comment moderation policies, but the abuse on Ressa's page only dropped when she wound down posting in the second half of 2018.

In September 2018, Ressa told Facebook executives: "If you don't change what you're doing, I could go to jail". After she was convicted in June 2020, she apportioned some blame to the company for her conviction, and she continues to argue that Facebook has enabled the destruction of democracy in the country. Ressa is very 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 said.

"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 said. She believes the platforms need radical renovation - 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 said. "It's a perfect comparison to the COVID-19 virus, because it is very contagious. And once you're infected, you become impervious to facts."

Twitter is also a significant distributor of abuse against Ressa, although she says she feels "much safer" on that platform - especially since the company appears, in her view, to have begun working harder over the past few years to protect women journalists and human rights defenders among its users. "Their reporting tool aggregates similar tweets, takes less time, and is far more effective in takedowns," Ressa said.

She has long spoken about being the 'canary in the coalmine,' warning that the Philippines' information ecosystem represents the West's dystopian future. She is extremely pessimistic about the prospect of the platforms responding quickly and effectively enough without changes to accountability and liability:

> **" 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 impacts of their inaction... "**

Ressa continues to work at the intersection of investigative journalism, research and policy to respond proactively to the information ecosystem crisis. And Rappler 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.

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. Well, this is the same thing. They have enabled these attacks. They've certainly changed my life in many ways.

Maria Ressa