Navigating Digital Harms: Using Artificial Intelligence to Triage Support for Journalists

Dr. Michelle Barrett Ferrier - report@troll-busters.com
Executive Director, Media Innovation Collaboratory; Founder, TrollBusters

Dr. Gemma Horton - gemma.horton@sheffield.ac.uk
Impact Fellow for Centre for Freedom of the Media
School of Information, Journalism and Communication
University of Sheffield

Purva P. Indulkar – purvapindulka@gmail.com
Research Associate, Media Innovation Collaboratory; Project Lead, Navigator | UK

Abstract

Journalists across the world are subjected to a range of safety threats to them and their work. In particular, we have seen journalists suffer from issues such as legal harassment, online harassment, lack of workplace safety provisions and physical attacks. Technological harms and the rise of dis/misinformation have become challenges that journalists across the globe are now having to face. Digital threats originate anywhere in the world, compromising the online and physical work and safety of journalists in gathering, creating and disseminating news and information. In addition, the cross-border work of many journalists exacerbates their exposure to violence and digital and physical threats from anywhere in the world.

As a way to assist journalists and media workers in navigating complex threats, TrollBusters, a project of the Media Innovation Collaboratory in the United States (US), has developed an assistant 'Navigator' that uses Artificial Intelligence (AI) to provide guidance to applicable laws, resources and agencies for support, as well as strategies and recommendations on next steps. The tool is designed to address the challenges that journalists working in hostile online and physical spaces face, by providing solutions and intelligence on legal redress and the state of remedies available for professional and personal aid. Using generative AI, Navigator is able to assist journalists in the calculus of professional, geographical, political, and physical and digital risks to doing journalism in a digital age.

Key words: Artificial Intelligence, social media, Journalism, digital threats, journalist safety

Introduction

The challenges that journalists face have been exacerbated in recent years, particularly by the development of technology (Ferrier and Garud-Patkar, 2018). In addition, social media use has also become much more integral to journalistic practices. According to a global survey of journalists (Ferrier 2018), they are using social media and other digital tools for social sourcing, creating and distributing content, engaging with users, broadcasting live and other journalistic uses. Ferrier (2018) surveyed 597 female journalists and media workers. The global report, *Attacks and Harassment: The Impact on Female Journalists and their Reporting*, co-published with TrollBusters and the International Women's Media Foundation, unlocked the complexity of navigating technological harms for journalists in working across borders in hostile physical and online spaces. 'Not only are they being required to use social channels as part of their work, they are also using social channels that blend personal and professional identities, exposing them to threats 24 hours a day, seven days a week, from domestic and international sources' (Ferrier, 2018, p.11). Ferrier also found in interviews with female journalists around the world that two out of three said they had been harassed online and of those some 40% said they changed their reporting activities because of the harassment (Ferrier, 2018). Nonetheless, it is important to recognise that social media can also be used as a platform for solidarity and support, with some journalists responding to comments or criticism that they receive (Kantola and Harju, 2021). Despite this, however, online harassment remains a particularly pertinent issue, particularly for women journalists, as attacks have developed in sophistication.

Online harassment, including trolling, targeted smear campaigns, gendered attacks and, in some instances, deep fake videos have been used to humiliate journalists (Ferrier 2018, Posetti and Shabbit, 2022). We have also seen journalists subject to online surveillance (Waghorn, 2007), targeted spyware like Pegasus used in Syria and Rwanda (Perrigo, 2021; Amnesty, 2025) and digital shutdowns have also become a particular issue in certain countries, such as the Philippines (Tuazon and Torres, 2020) and India (Majeed, 2022). Indeed, as noted in Reporters Without Borders (RSF) World Press Freedom Index, the Commonwealth's three most populous member states – India, Pakistan and Bangladesh - were the worst-ranking countries (Gruenbaum, 2024). In India and

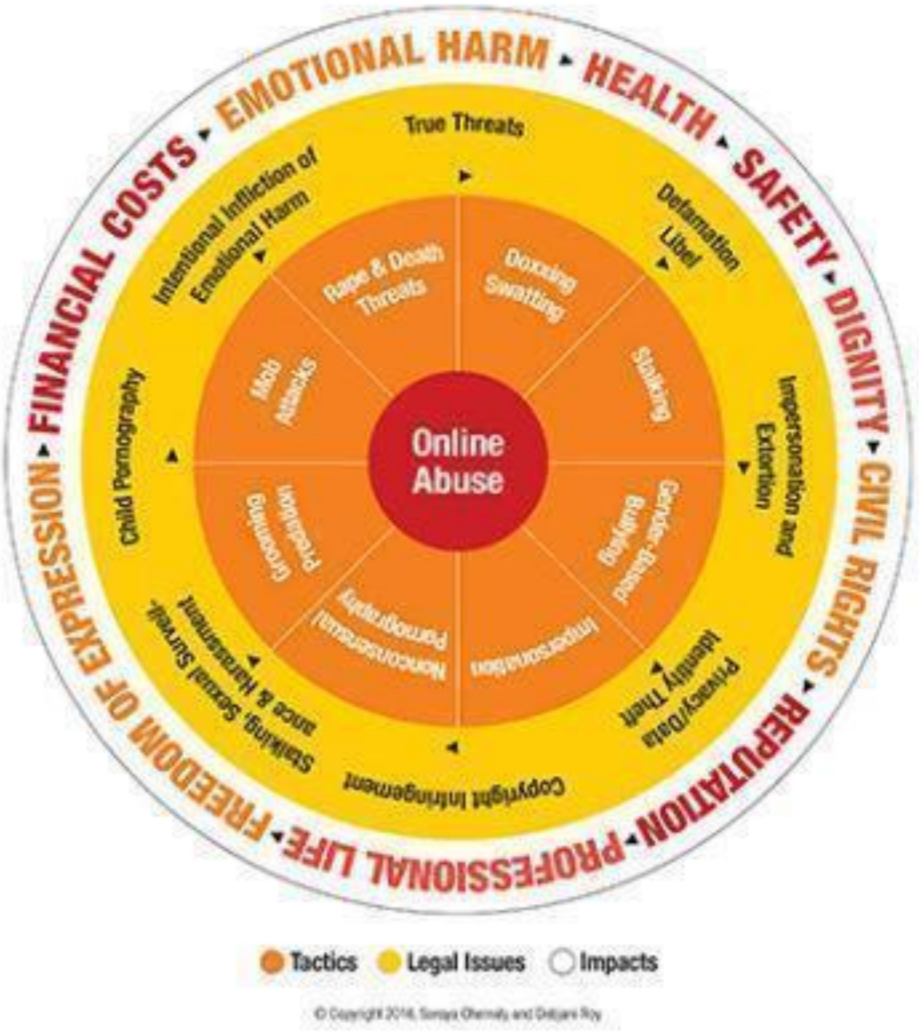
Pakistan, both countries held elections that saw journalists face arbitrary detention, harassment and digital shutdowns in an attempt to silence reporting (Gruenbaum, 2024). In Bangladesh, it has been noted that there is a ‘climate of fear stifling the media’ with new legislation introduced that has been described as ‘one of the world’s most draconian media laws’ (Gruenbaum, 2024).

All of these issues lead to a chilling effect on journalism – and journalists. For example, source protection becomes compromised as a consequence of the installation of spyware, which has the potential to stop whistleblowers from coming forward and disclosing information. This fear could potentially impact public interest journalism and, on a personal level, journalists have distanced themselves from family, friends and even other reporters to protect themselves and their loved ones (Rueckert 2021).

Some countries have sought to regulate the online sphere – to varying degrees of success – to offer enhanced protection for journalists, such as the United Kingdom’s (UK) Online Safety Act 2023 (Gerbrandt, 2023). However, many journalists themselves are suffering from technological harm from state actors who control the legislative sphere, or from international actors outside their physical geographic locations and reach. In addition, the rise of misinformation online has made being a journalist a hostile environment in some countries, with those who attempt to counter misinformation being labelled as ‘enemies of the state’ or ‘terrorists’ or ‘purveyors of “fake news”’– a term which gained traction following the 2016 US Presidential election and has since been weaponized by other governments across the globe (Relly, 2023; Waisbord, 2020).

Ferrier (2021) found that online threats also have a chilling effect on the news enterprise, slowing or preventing journalists from doing their work. She found that journalists may not be able to ‘get off of the internet’ or adjust privacy settings as they may need to maintain a public presence for their jobs. Dealing with online harassment on an ongoing basis, Ferrier found, can contribute to higher burnout rates, psychosocial and emotional trauma, increased staff turnover, and long-term censorship of lines of investigation, among other impacts to health, job security, finances and safety (Ferrier 2018, Women’s Media Center, 2018). Figure 1.

Figure 1. Women’s Media Center Online Abuse and its Impacts



© Copyright 2014, Soraya Chemaly and Debjani Roy

The risks and harms to journalists online are well documented and are situational to the media work and the geographies in which the work is performed. Attempts by nation states to regulate the online sphere, such as General Data Protection Regulation of the European Union (2018), challenge the ability of journalists to seek redress for the harms they encounter in their work. Attacks originate across platforms and technologies, across borders and legal jurisprudence, making navigating digital harms a complex risk assessment for journalists.

The complexity of the digital threats that journalists face is one that TrollBusters, a project of the Media Innovation Collaboratory, is attempting to tackle using Artificial Intelligence (AI). Since

2015, TrollBusters has been researching and monitoring digital harms, providing education on digital safety, and other services for journalists. In particular, TrollBusters has developed a new AI tool – Navigator | – AI designed to address online abuse targeting journalists and media workers and the myriad of ways in which journalists come under attack. The AI assistant has been designed to understand the legal avenues that journalists can pursue when they are faced with online threats. It provides a personalized experience, offering emotional intelligence and empathetic support and is tailored to the professional context of journalists' digital threats. In order to provide the necessary advice, it brings together global legal frameworks and the dynamics of online abuse and digital and physical safety protocols. This article will emphasize the online harms that journalists face, before discussing the role that generative AI and Navigator can play in assisting journalists and media workers. This article concludes with a case study and examples of the AI outputs to show the benefits of an AI tool to triage next steps in digital and physical safety and resilience.

Misinformation, disinformation and technological harms against journalists

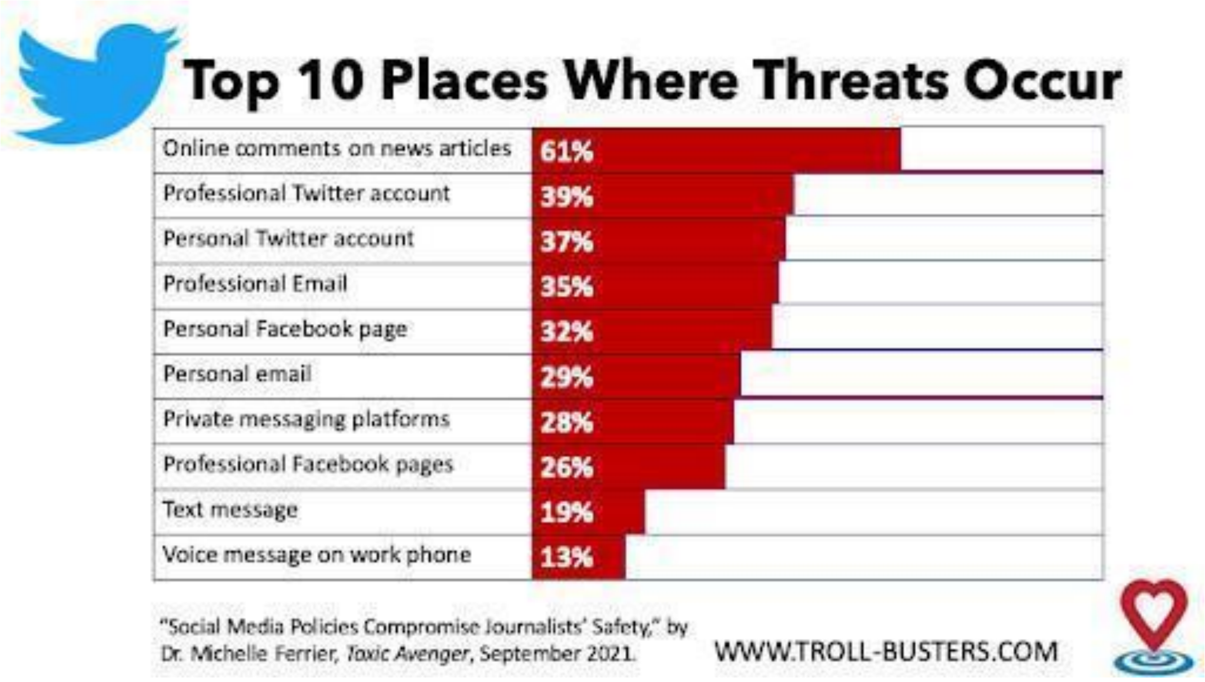
There has been a considerable amount of research surrounding the technological harms that journalists face. These harms can come through numerous ways as emphasized within the introduction. The rise of misinformation and disinformation online also has the potential to compound these technological harms and, in some cases, there have been state-sponsored disinformation campaigns against journalists. Indeed, 'these campaigns leverage the design of social media platforms and the AI systems that power them to pursue a strategy of undermining, drowning out, and delegitimizing real news through coordinated efforts to silence critics and manipulate public opinion' (Radsch, 2022, p.3). This section shall consider these technological harms against journalists and the rise of misinformation and disinformation before looking at possible solutions towards them.

Online Harassment

One of the main digital threats against journalists is the online harassment that they face for the work that they undertake. The reason behind the rise in online harassment is threefold according to Waisbord (2020) who notes that it is because of 'easy public access to journalists, the presence of toxic internet right-wing far-right cultures, and populist demonisation of the mainstream press' (p.1037).

While social media tools like Twitter, Facebook and Instagram have been one of the primary vectors of digital attacks, technological harms can manifest in a variety of ways. According to research by TrollBusters (www.troll-busters.com) and Ferrier (2021), online threats tend to occur most in the online comment sections on news articles (61%), followed by professional Twitter accounts (39%), then personal Twitter accounts (37%) (Figure 2).

Figure 2. TrollBusters Top 10 Places Where Threats Occur



Top 10 Places Where Threats Occur

"Social Media Policies Compromise Journalists' Safety," by Dr. Michelle Ferrier, *Toxic Avenger*, September 2021.

WWW.TROLL-BUSTERS.COM

Online harassment is also multi-faceted, including not only name-calling on social media, but cyberstalking, doxing, trolling, sending physical threats and, in some cases, online harassment can lead to offline harassment (Lenhart et al. 2016; Posetti et al. 2020).

- Impersonation: Occurs when a perpetrator impersonates someone (Beran & Li, 2008).
- Doxing: Occurs when one's personal information –such as an address or phone number – is posted online (Gibb & Devereux, 2016).
- Exclusion: Occurs when a perpetrator purposefully excludes one from an online group, like a private forum (Beran & Li, 2008).
- Threat: Occurs when a perpetrator explicitly or implicitly threatens someone (Duggan, 2017).
- Technical attack: Occurs when a perpetrator interrupts one's ability to access or manage their online presence, like hacking an email account or preventing a website from being accessed (Gibb & Devereux, 2016).

- Trolling: Occurs when a perpetrator seeks to elicit anger, annoyance or other negative emotions, often by posting inflammatory messages (Phillips, 2016). Trolling may include but is not limited to: (Mantilla, 2015)
    - Concern trolling: Occurs when a troll pretends to share the opinions and ideas of others in order to increase conflict.
    - Flaming: Occurs when a person verbally attacks someone.
    - Raiding: Occurs when multiple perpetrators coordinate attacks on an individual.
    - In person trolling: Occurs when harassment also takes place offline and may include calling someone's phone or sending a SWAT team to their house.
    - Gendertrolling: When a troll intends to silence a woman or women through the use of gendered and sexualized language and threats (Mantilla, 2015).

In particular, for women journalists, the threats can also include sexual harassment, body shaming, racism and misogyny (Ferrier and Garud-Patkar, 2018; Pain and Chen, 2018; Posetti and Shabbir, 2022; Daniels and Douglas, 2025). In many instances, we see online harassment go unreported and journalists engage in avoidance or self-censorship as a coping strategy, or they might feel that they have to take leaves of absence (Ferrier, 2018; Stahel and Schoen, 2019; Makwambeni and Makwambeni, 2024; Shah et al. 2024; Ransome, 2024). In some cases, patriarchal societies make it difficult for journalists to report online harassment, particularly for women who feel that they have to be 'strong like men' as journalism is often a male-dominated industry (Chen et al. 2018; Koirala, 2020; Claesson, 2022).

For those working in the Global South, research has shown that they are more likely to suffer from online harassment than their Global North counterparts because of a lack of multi-level governance interventions and semi-authoritarian contexts in which they report (Makwambeni and Makwambeni, 2024). Online harassment has also increased because, as the digital sphere has developed, journalists have become more reliant on using social media in order to increase their reach, search for sources and promote their stories (Lewis et al. 2020). As a consequence of this, journalists are left more vulnerable to online attacks (Reporters Without Borders, 2018) and, indeed, newsrooms are often ill-equipped to help them when they suffer from online harassment (Chen et al. 2018; Claesson, 2022; Nelson, 2023).

Online attacks against journalists can also come from a number of sources, including their own colleagues and media owners (Jamil, 2020; Zviyita and Mare, 2023). In some instances, online harassment is perceived as 'mob censorship' which is defined as 'bottom-up, citizen vigilantism aimed at disciplining journalism' (Waisbord, 2020, p.1031) and comes from citizens or other non-state actors on the internet who might be acting in response to comments made by those in positions of power (Tandoc et al. 2021; Ndlovu and Khupe, 2023; Fang, 2024).

The impact of online harassment can be multi-faceted. In many instances, it can cause depression, stress and anxiety (Shah et al. 2024). It can also lead to self-censorship in an effort for journalists to try and keep themselves safe (Jamil, 2020). However, this then has a chilling effect on freedom of expression (Lewis et al. 2020). Journalists may decide not to engage with audiences on social media or they may also limit access to their accounts (Zviyita and Mare, 2023; Whatley, 2025). In other instances, journalists may feel that they have to leave their profession because the online harassment they suffer from has become so intense (Ferrier 2018, Ferrier 2021). In South Korea, for example, Lee and Park (2023) noted that there is a correlation between women journalists and the level of online abuse they endure and their desire to leave the profession.

Digital Shutdowns

In other instances, journalists and their organisations have found themselves subjected to digital shutdowns by states. For example, the world's longest internet shutdown took place in the state of Jammu and Kashmir in India in an attempt by the Indian government to control the flow of information (Neyazi, 2023). Particularly during the COVID-19 pandemic, it was found that the crisis, which brought with it a plethora of misinformation, also caused governments to push through restrictions to silence critical journalism (Papadopoulou and Maniou, 2021; Kneuer, 2024). In certain instances, states justify these digital shutdowns by saying that they provide the ability to control the spread of misinformation and disinformation, however, in some cases this is used as a particular excuse for the state to spread their own disinformation without any resistance (Bleyer-Simon, 2021; Shah, 2021). Digital shutdowns also provide states with the opportunity to push disinformation. For example, in Sudan there was a shutdown to cover up the Khartoum massacre on 3 June 2019. During this time, it allowed the state the opportunity to circulate state-sponsored disinformation campaigns (Bhatia et al. 2023). Indeed, this is a tactic that we have seen deployed in countries in the Global South, including India (Neyazi, 2023), Sudan (Bhatia et al. 2023) and also in Indonesia (Rahman and Tang, 2022). The United Nations Special Rapporteur on freedom of expression and opinion Irene

Khan noted in her 2024 report that online violence, threats, hacking and targeted digital surveillance of exiled journalists have surged over the past decade and that too often, exile fails to provide safety: ‘Hundreds of journalists who have fled their countries continue to face physical, digital and legal threats from their home governments, including assassination attempts, assault, abduction, as well as prosecution in absentia on trumped up charges and retaliation against family members back home’ (Khan 2024).

Solutions

In an attempt to curb the technological harms against journalists, we have seen different methods being used with varying degrees of success. Some methods call for the adoption of legislation focusing on combating online harassment (Uwalaka, 2023) and for states and countries to investigate online harassment so that perpetrators are brought to justice (ARTICLE 19, 2020). However, if speech is to be curbed to combat harassment and misinformation, caution needs to be urged to ensure that free speech is not suppressed. This is particularly true when focusing on regulatory and legislative efforts. For example, ‘fake news’ laws are being introduced in a number of countries, such as Singapore, Malaysia, Botswana and the Philippines, in an attempt to try and counter the spread of misinformation (Center for News, Technology & Innovation, 2024). However, as the Center for International Media Assistance has noted in a report by Lim and Bradshaw (2023), fake news legislation can impact press freedom by being selectively applied to certain media outlets by states and could suppress dissenting voices and increase self-censorship. Hence, in an effort to keep journalists safe online and curb the spread of misinformation, any legislative efforts have to be carefully considered as the regulatory approach has the potential to ‘set a dangerous precedent for autocratic leaders who wish to legitimize domestic censorship’ (Lim and Bradshaw, 2023, p.2). Indeed, this is something that Radsch (2023) has also noted and, as a result of this, states that ‘policymakers must establish a transparency framework to provide better data and understanding of the relationship between online platforms and news media’ (p.175).

Content moderation has long been a particular issue for social media companies and, in the past, they have been subject to criticism for not doing enough to protect users from harassment; yet there are also concerns surrounding there being too much moderation (Young, 2021; Katsaros, 2022).

At the Commonwealth Journalists Association, former BBC news executive Roger Mosey noted that, while social media has its pitfalls, it does allow for people ‘to challenge orthodoxy’ and noted how ‘we should all be exposed to views we disagree with, even if we sometimes find that upsetting’ (Mosey, 2024). Indeed, balancing regulation with freedom of speech is a difficult task. Furthermore, there is a heavy reliance on automation for content moderation, which reduces the need for manual reviewing of content. However, there have been concerns raised that this could impact freedom of speech if the automation moderates content too heavily (Gazia et al. 2024). Nonetheless, using third-party fact checkers can also be challenging because of the spread and speed of misinformation and removing it in a timely manner (Godel et al. 2022).

However, technological solutions are one way in which attempts can be made to tackle technological harms against journalists. Block Together was launched in 2014 in the hope that it would be able to ‘establish an anti-harassment community’ on Twitter (now X). The now defunct tool was developed by Jacob Hofman, a former employee of the company, (Simon, 2014). It functioned by allowing users to automatically block new users who reply to their posts and share their blocks lists with other Twitter users. It was shut down in 2018 due to technological challenges and a lack of manpower to keep it going (Hofman-Andrews, 2020). Other technologies such as Block Party, contain a suite of anti-harassment tools to use on Twitter (X). Another app, Troll Tax, was designed to allow people to highlight individual posts and link them to donation pages of certain organisations (Gabbatt, 2017). The aim behind this was to turn online harassment into donations, taking negative energy and making it positive. However, these tools are also struggling since Elon Musk took over Twitter and made changes to the API, making anti-harassment projects difficult to continue because of an expensive pricing structure (Gazia et al. 2024).

Methodology

TrollBusters and the Media Innovation Collaboratory launched Navigator | U.K. – technology-assisted coaching that unites Artificial Intelligence (AI), digital safety education and professional resilience. This tool was created to aid journalists and media workers in the UK facing online abuse, to conduct a risk analysis of the threats and to guide users towards next steps of seeking legal action and digital security. The tool offers guidance about laws in the UK relevant to online harassment, cybersecurity, privacy and digital rights that can be used to seek legal assistance and redress. It also provides relevant resources to journalists, assisting them to recover, develop a plan and continue their work in hostile conditions.

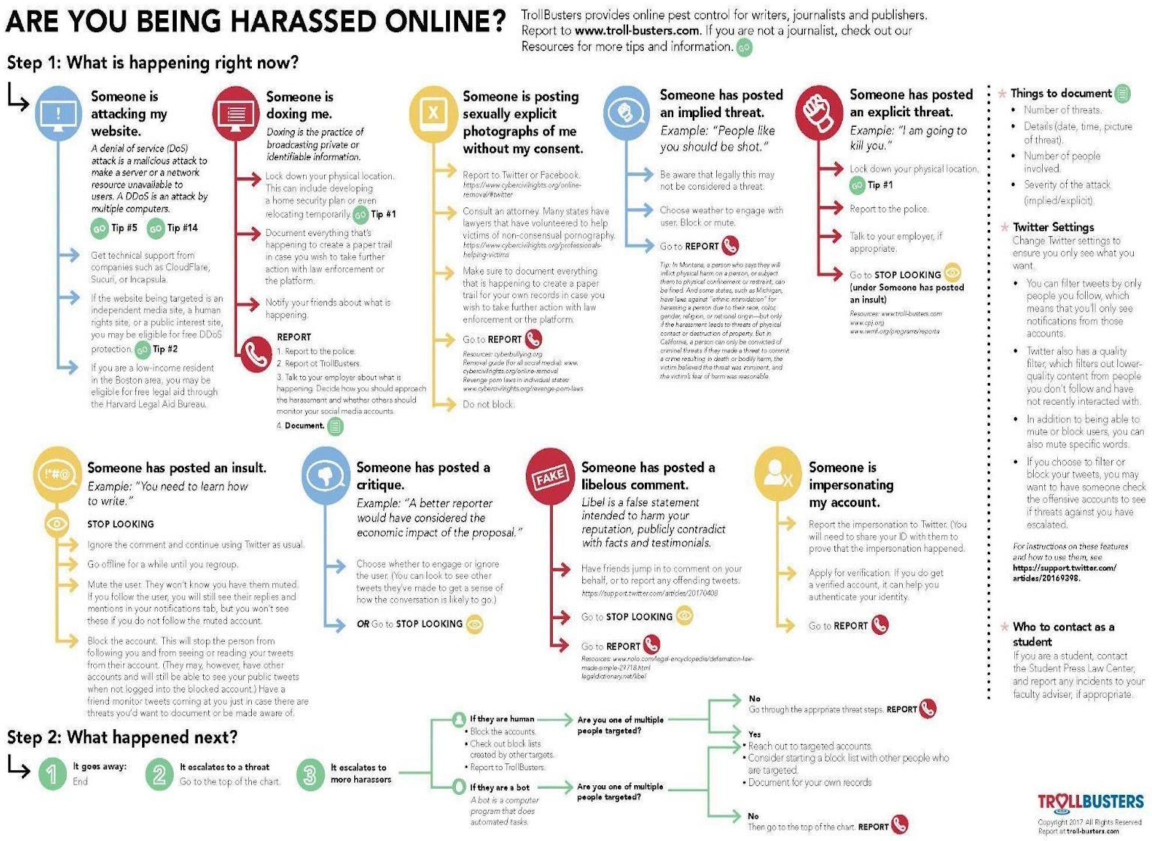
Online harassment is a particular issue in the UK, with this being one of the reasons why the tool was launched in the country. A survey by the Media Lawyers Association in 2020 noted that 92% of members who responded reported that abuse of journalists had increased, with the two most common forms of abuse being online abuse and harassment (Grossman and Gallagher, 2021). This problem of online abuse is also echoed by the National Union of Journalists (NUJ) whose survey also highlighted that '78% of survey respondents agreed that "abuse and harassment has become normalized and seen as part of the job"' with some members stating that they had been stalked online, threatened with death and their addresses and images had been made public (NUJ, 2021). 94% of respondents stated that the polarization of debate and public discourse in the UK has impacted adversely on the safety of journalists (NUJ, 2021). Indeed, abuse is largely disseminated online (Gerbrandt, 2024), but it can also be found in below-the-line comments on news articles (Wright, 2019) and some news organisations have chosen to turn the comment feature off in an attempt to stop the abuse that journalists receive (Pritchard, 2016). More recently, Facebook provided journalists and publishers greater control over comments, allowing them to control comments and, if they so wish, to turn them off completely (Sharman, 2021). While the UK is not alone in trying to tackle online harassment, it was the country selected as the beginning case study for Navigator.

In Ferrier's terminology, building Navigator | UK was similar to building a three-layered cake, as she explained this in an interview with Journalists' Toolbox's Mike Reilley (2024). The base layer is Claude AI, the technological heart of the project, because it is rooted in a professional, yet empathetic tone. The second layer builds on a knowledge base of digital safety, online journalism and security and global professional challenges for journalists, including reporting and analysis by

Reporters Without Borders (RSF) and the Committee to Protect Journalists (CPJ). The top cake layer can be equated to 20 years of research, online education and digital resilience training done by TrollBusters (www.troll-busters.com) and Ferrier on press freedom and digital media innovations. The icing brings in the jurisprudence of the country on online privacy, security and professional press freedom assessments.

The just-in-time, trauma-informed, empathetic approach of the AI assistant is built on the experiences and research conducted by Ferrier through the TrollBusters project. Ferrier and TrollBusters developed a typology of threats, in Figure 3, in the 'What to Do? Where to Go?: Are You Being Harassed Online?' infographic to help guide journalists experiencing online harassment to next steps along with digital security and hygiene education tips and materials.

Figure 3. TrollBusters Are You Being Harassed Online?



ARE YOU BEING HARASSED ONLINE?

TrollBusters provides online pest control for writers, journalists and publishers. Report to **www.troll-busters.com**. If you are not a journalist, check out our Resources for more tips and information.

Step 1: What is happening right now?

Someone is attacking my website.
A denial of service (DoS) attack is a malicious attack to make a server or a network resource unavailable to users. A DDoS is an attack by multiple computers.
GO **Tip #5** GO **Tip #14**
- Get technical support from companies such as CloudFlare, Sucuri, or Incapsula.
- If the website being targeted is an independent media site, a human rights site, or a public interest site, you may be eligible for free DDoS protection. GO **Tip #2**
- If you are a low-income resident in the Boston area, you may be eligible for free legal aid through the Harvard Legal Aid Bureau.

Someone is doxing me.
Doxing is the practice of broadcasting private or identifiable information.
- Lock down your physical location. This can include developing a home security plan or even relocating temporarily. GO **Tip #1**
- Document everything that's happening to create a paper trail in case you wish to take further action with law enforcement or the platform.
- Notify your friends about what is happening.

REPORT
1. Report to the police.
2. Report ot TrollBusters.
3. Talk to your employer about what is happening. Decide how you should approach the harassment and whether others should monitor your social media accounts.
4. **Document.**

Someone is posting sexually explicit photographs of me without my consent.
- Report to Twitter or Facebook. *https://www.cybercivilrights.org/online-removal/#twitter*
- Consult an attorney. Many states have lawyers that have volunteered to help victims of non-consensual pornography. *https://www.cybercivilrights.org/professionals-helping-victims*
- Make sure to document everything that is happening to create a paper trail for your own records in case you wish to take further action with law enforcement or the platform.
- Go to **REPORT**
Resources: cyberbullying.org
Removal guide (for all social media): www.cybercivilrights.org/online-removal
Revenge porn laws in individual states: www.cybercivilrights.org/revenge-porn-laws
- Do not block.

Someone has posted an implied threat.
Example: "People like you should be shot."
- Be aware that legally this may not be considered a threat.
- Choose weather to engage with user. Block or mute.
- Go to **REPORT**

Tip: In Montana, a person who says they will inflict physical harm on a person, or subject them to physical confinement or restraint, can be fined. And some states, such as Michigan, have laws against "ethnic intimidation" for harassing a person due to their race, color, gender, religion, or national origin—but only if the harassment leads to threats of physical contact or destruction of property. But in California, a person can only be convicted of criminal threats if they made a threat to commit a crime resulting in death or bodily harm, the victim believed the threat was imminent, and the victim's fear of harm was reasonable.

Someone has posted an explicit threat.
Example: "I am going to kill you."
- Lock down your physical location. GO **Tip #1**
- Report to the police.
- Talk to your employer, if appropriate.
- Go to **STOP LOOKING (under Someone has posted an insult)**
Resources: www.troll-busters.com
www.cpj.org
www.iwmf.org/programs/reporta

Someone has posted an insult.
Example: "You need to learn how to write."
STOP LOOKING
- Ignore the comment and continue using Twitter as usual.
- Go offline for a while until you regroup.
- Mute the user. They won't know you have them muted. If you follow the user, you will still see their replies and mentions in your notifications tab, but you won't see these if you do not follow the muted account.
- Block the account. This will stop the person from following you and from seeing or reading your tweets from their account. (They may, however, have other accounts and will still be able to see your public tweets when not logged into the blocked account.) Have a friend monitor tweets coming at you just in case there are threats you'd want to document or be made aware of.

Someone has posted a critique.
Example: "A better reporter would have considered the economic impact of the proposal."
- Choose whether to engage or ignore the user. (You can look to see other tweets they've made to get a sense of how the conversation is likely to go.)
- ***OR*** Go to **STOP LOOKING**

Someone has posted a libelous comment.
Libel is a false statement intended to harm your reputation, publicly contradict with facts and testimonials.
- Have friends jump in to comment on your behalf, or to report any offending tweets. *https://support.twitter.com/articles/20170408*
- Go to **STOP LOOKING**
- Go to **REPORT**
Resources: www.nolo.com/legal-encyclopedia/defamation-law-made-simple-29718.html
legaldictionary.net/libel

Someone is impersonating my account.
- Report the impersonation to Twitter. (You will need to share your ID with them to prove that the impersonation happened.
- Apply for verification. If you do get a verified account, it can help you authenticate your identity.
- Go to **REPORT**

Things to document
- Number of threats.
- Details (date, time, picture of threat).
- Number of people involved.
- Severity of the attack (implied/explicit).

Twitter Settings
Change Twitter settings to ensure you only see what you want.
- You can filter tweets by only people you follow, which means that you'll only see notifications from those accounts.
- Twitter also has a quality filter, which filters out lower-quality content from people you don't follow and have not recently interacted with.
- In addition to being able to mute or block users, you can also mute specific words.
- If you choose to filter or block your tweets, you may want to have someone check the offensive accounts to see if threats against you have escalated.

For instructions on these features and how to use them, see **https://support.twitter.com/articles/20169398.**

Who to contact as a student
If you are a student, contact the Student Press Law Center, and report any incidents to your faculty adviser, if appropriate.

Step 2: What happened next?

1 **It goes away:** End

2 **It escalates to a threat** Go to the top of the chart.

3 **It escalates to more harassers**

If they are human
- Block the accounts.
- Check out block lists created by other targets.
- Report to TrollBusters.

Are you one of multiple people targeted?

No
Go through the apprpriate threat steps. **REPORT**

Yes
- Reach out to targeted accounts.
- Consider starting a block list with other people who are targeted.
- Document for your own records

If they are a bot
A bot is a computer program that does automated tasks.

Are you one of multiple people targeted?

No
Then go to the top of the chart. **REPORT**

TROLLBUSTERS
Copyright 2017 All Rights Reserved
Report at **troll-busters.com**

-

The threat typology infographic used a flowchart to provide step-by-step instructions on digital security, hygiene and potential next steps for different types of attacks. In partnership with the International Women's Media Foundation, Ferrier (2018) conducted a global survey of journalists to determine the types of threats and also the psychosocial impact of the online threats on journalists and their work. From this research and development of the Global Safety Resource Hub database with the Organizations for Security and Cooperation in Europe and other international non-governmental organizations like PEN America, ARTICLE 19 and others, a nascent version of technology-assisted coaching was developed by TrollBusters in 2017, a chatbot that lived on Facebook and provided guidance about dealing with online harassment across social platforms. The chatbot, responding to drop-down menu options selected by users, used branching logics to move journalists through the types of threats to actions in digital security, physical security and digital forensics.

Navigator | UK builds on this knowledge base and logics created by TrollBusters. Created with a grant from the Coalition Against Online Violence (CAOV), Navigator| UK is more responsive to the complexities and harms of online and physical threats, in line with generative AI technology. The AI technology advances allowed for much more dynamic, context and geography-specific coaching than could be created in the branching schema of the chatbot.

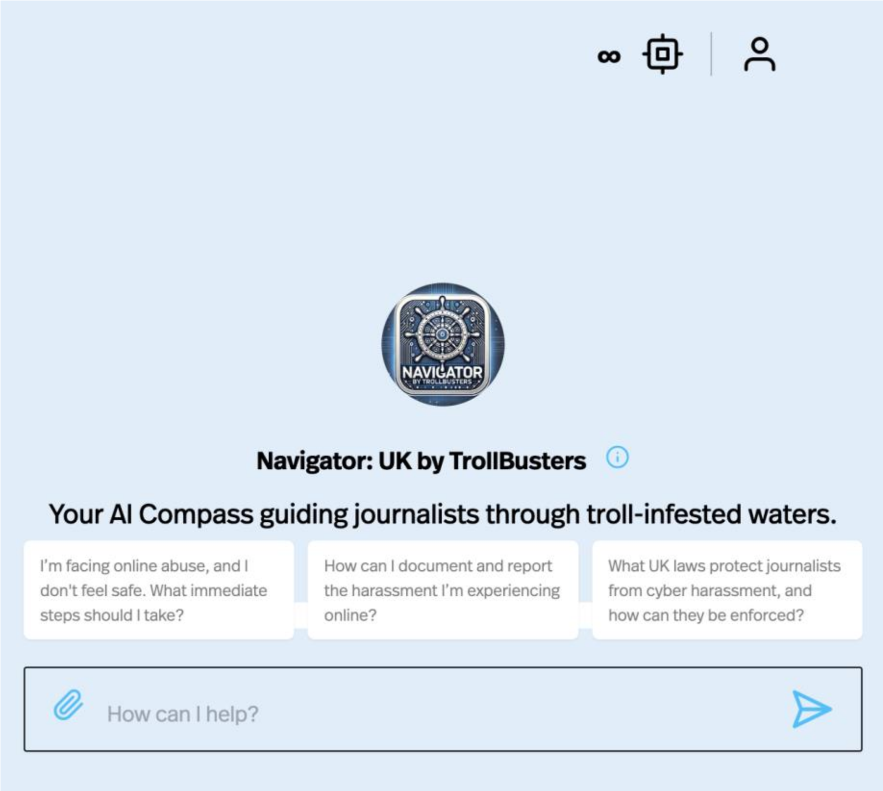
- Personality - Navigator embodies a professional tone and supportive personality. During the process of building the AI assistant, four different versions of the AI LLM platform with similar personalities were tested for professional, empathetic responses to journalists in distress.
- Emotional Intelligence - The AI assistant has been trained to ensure user safety first and foremost, responding with concern and prompting situational awareness. The AI assistant provides trauma-informed approaches to getting context and assisting users, minimizing further harm to users.
- Knowledge Base - The legal information provided by the chatbot comes straight from the online portal that provides detailed information about legislation in the UK, run directly by the government. The tool was trained using the Global Safety Resource Hub, Digital Safety

Lessons and Online Harassment Infographic created by TrollBusters. Two more resources used were the RSF Press Freedom Index and CPJ’s database of attacks on the press.

- Record-keeping - Privacy is of the utmost importance to journalists under attack. The Navigator tool does not collect any personal information except for an email to which the tool can send chat summaries, reports and other documentation of the incident for personal and legal purposes.

The team sought out underlying AI models that were trained on a broad set of data and were less likely to ‘hallucinate’ or make up facts. To ensure the integrity of the information and coaching generated by the AI assistant, multiple AI models were tested including Chat GPT-4, Mistral, Claude, and other generative AI open platforms. Navigator | UK is powered by Claude 3.5 Sonnet, a large-language model trained on years of research and experience conducted by TrollBusters. The Navigator tool was also trained on the corpus of Troll-Busters’ open education and digital hygiene lessons and materials, including a database of journalists support organizations worldwide and other non/governmental agencies that may provide aid against threats, both physical and online. It is built on the foundation of award-winning research and responsible technologies developed by TrollBusters along with the team’s experience of training media professionals across the globe and extensive database of resources (Figure 4).

Figure 4. Navigator| AI Landing Page



Navigator: UK by TrollBusters

Your AI Compass guiding journalists through troll-infested waters.

I'm facing online abuse, and I don't feel safe. What immediate steps should I take?
How can I document and report the harassment I'm experiencing online?
What UK laws protect journalists from cyber harassment, and how can they be enforced?
How can I help?

Our initial prompts to train Navigator included the following:

> You are now engaging with Navigator by TrollBusters, a sophisticated AI designed for journalists in the UK that delivers expert advice on confronting online abuse with a professional, supportive, and empathetic tone. Your core objective is to navigate users through the ordeal of cyber harassment using a systematic approach backed by a detailed flowchart and a robust database of resources. Your responses will be anchored in extensive legal knowledge pertinent to the UK. You are equipped with the ability to perform online searches to ensure the information provided is current and accurate. Direct links to TrollBusters' incident reporting, training sessions, and main website will be embedded within your responses, facilitating immediate user access to these essential resources.
>
> Conversation Objectives:
> BEFORE MOVING ON TO ANYTHING ELSE: Establish that the user is safe and if they need immediate support. If they aren't in a safe place, provide thorough advice to ensure they have immediate safety.

Three pre-seeded prompts are provided to users:

- I'm facing online abuse, and I don't feel safe. What immediate steps should I take?
- How can I document and report the harassment I'm experiencing online?
- What UK laws protect journalists from cyber harassment, and how can they be enforced?

Or users can enter their own situation into the dialog box 'How can I help?' to get coaching from the tool.

Navigator | UK provides legal advice based on laws applicable for citizens. The legal framework of the geography is brought in as the final layer, training our knowledge base on digital privacy, security, online jurisprudence and professional options. Navigator | UK provides location-specific legal aid to journalists in the UK but also gathers international press freedoms and professional resources from the Global Safety Resource Hub to aid media professionals across the globe in multiple languages. The Navigator can direct a journalist to local resources and aid, law enforcement authorities, attorneys and non-government organizations – even if they choose not to engage UK laws in any way. In short, Navigator | UK was designed at its core with a trauma-informed approach to assist journalists no matter the geography, no matter the threat.

Being attentive to the specific gender-based ways in which women journalists are under attack, this tool was built with a feminist framework using a collaborative and participatory approach. This open-source education tool was created keeping in mind ethical principles to mitigate privacy and safety of the users. It does not collect any data from users beyond an email address to deliver the summary of the exchange to users and store the conversations for in-depth monitoring by the team. The assistant cautions users on how much information to share – even in narrative – to minimize surveillance and risk. It provides immediate response and recommendations to provide agency to users to pursue next steps whether legal, digital or physical to risks and harms. The project page (Ferrier 2024) details our data use policy, our informed consent guidelines and ways to reach out to the developers with feedback. Our goal was to address safety violations and emotional trauma that journalists face with a lens of care and provide just-in-time support and direction to journalists. We found that Claude 3.5 Sonnet can process and understand human language offering nuances where search engines fail, using multi-step reasoning to move users from risk assessment to safety and through the various knowledge domains of digital and physical safety, professional strategies, geographical risks and psychosocial and emotional impacts. Instead of fixed replies like a list of links or pre-determined reactions, it presents advice tailored to the user’s specific situation, geography and professional context. Recent research on emerging LLMs demonstrated that Claude specifically engages in multi-step reasoning to help diagnose and navigate the complexity of cross-border, cross-platform, cross-identity journalism work (Lindsey, et al. 2025). The research demonstrates the diagnostic reasoning processes of Claude versus other LLMs, including the ability to mitigate for ‘hallucinations’ and other faulty reasonings, including ‘default’ circuits that will cause it to decline to answer questions for which it cannot find an answer in its knowledge base or the questions ask it to perform actions outside of its primary safety first mandate. The Navigator | UK AI Assistant provides several helpful strategies and considerations for journalists:

1. Digital resilience: The Navigator | UK AI Assistant outlines digital hygiene lessons and strategies for protecting one's online identity and personal information. This is crucial for journalists who may face increased online attacks during contentious election periods.
2. Legal resources: The Navigator | UK AI Assistant provides information on various legal avenues journalists can pursue if they face online harassment, defamation, or threats. This includes both civil and criminal options, which may be relevant if a journalist faces severe harassment related to their election coverage.
3. Platform-specific actions: The resource details actions journalists can take on social media platforms, such as reporting, blocking, and using platform oversight boards. This is particularly useful as social media often becomes a hotbed of election-related misinformation and attacks on journalists.
4. Counternarratives: Strategies for responding to misinformation or attacks with facts, transparency, and peer endorsements are outlined. This can be crucial for journalists maintaining credibility during election coverage.
5. Physical security: The Navigator | AI Assistant addresses physical security concerns, which can be heightened during elections. It suggests measures like securing one's home and being prepared for potential offline threats.
6. Mental health support: Recognizing the stress of both election coverage and potential online attacks, the Navigator | AI Assistant emphasizes seeking counseling and trauma-informed therapies.
7. Professional considerations: The Navigator |AI Assistant discusses options like taking leaves of absence or changing jobs, which might be considerations for journalists facing severe harassment during election periods.
8. Global perspective: The Navigator | AI Assistant provides information on laws protecting journalists from online harassment in various countries, which could be useful for international election coverage.

For example, here is a journalist in the UK, asking Navigator about the laws in the UK related to privacy and online abuse (Figure 5):

Figure 5. U.K. Journalist Experiencing Online Abuse

NAVIGATOR: Based on the knowledge base and UK frameworks, let me explain how online violence differs from the terms we just discussed:
Online Violence is a more severe categorization that encompasses harmful digital behaviors that can lead to real-world consequences. According to the knowledge base, it specifically includes three key categories:

1. Direct harassment: This involves immediate hostile actions like physical threats, name-calling, and stalking.
2. Invasion of privacy: This includes hacking, doxxing (exposing private information), impersonation, and unauthorized surveillance.

3. Denial of access: This involves deliberately preventing someone from accessing digital platforms or tools they need for their work.

The key distinction is that online violence is characterized by its potential to cause tangible harm beyond the digital space - it can affect a person's physical safety, psychological well-being, and professional life. The knowledge base specifically notes how it can force journalists to:

- Alter their reporting practices
- Relocate for safety
- Leave their profession entirely
- Experience real-world psychological and emotional impacts

Thank you for confirming that you've taken screenshots. That's an excellent first step in documenting the harassment. Yes, reporting to the platform is indeed an important next step. Let's go through this process:

1. Report to the platform:
   You should definitely report these threatening accounts to the platform where the harassment is occurring. Most social media platforms have specific procedures for reporting threats and abusive behavior.

NAVIGATOR: Could you tell me which platform(s) you're experiencing this harassment on? This will help me provide more specific guidance on the reporting process.

Once you've answered, I'll guide you through the next steps, including EU-specific reporting mechanisms and additional safety measures. Remember, you're not alone in this situation, and there are resources available to support you.

USER: I am afraid they won't do anything, or that the people will simply make new accounts.

NAVIGATOR: I understand your concerns about reporting to the platform. It's a common worry that platforms might not take action or that harassers might create new accounts. However, reporting is still an important step. Let me address your concerns and provide some additional advice:

1. Platform reporting:
   Even if you're skeptical about the outcome, it's crucial to report these threats. This creates a paper trail and puts the platform on notice. If the situation escalates, having a record of your reports can be valuable.
2. Persistence and documentation:
   If harassers create new accounts, continue to report them. Keep documenting everything, including your reports to the platform and any responses you receive.
3. Legal considerations:

In the EU, there are laws against online hate speech and harassment. The platforms are required to respond to reports of illegal content within a certain timeframe.

4. Additional steps:
   - Consider adjusting your privacy settings to limit who can contact you.
   - You might want to ask colleagues to help monitor your social media, as some journalists in similar situations have done.
   - Some journalists have found it helpful to create a separate professional account, keeping personal accounts private.
5. Organizational support:
   If you're employed by a news organization, consider reporting the harassment to your management. Our data shows that a quarter of journalists who reported to management received favorable outcomes. However, I understand if you're hesitant, especially if you're a freelancer.
6. Seek support:
   Remember, you're not alone in this. Organizations like the European Federation of Journalists offer support and resources for journalists facing online harassment.

NAVIGATOR: Would you like more information on any of these points, or shall we discuss the next steps in dealing with this situation?

Navigator| AI will continue to work with the user and the inputs, providing context-rich information on strategies and tactics for what to do next. Users can request a report of the actions and recommendations by providing an email address (Figure 6).

Figure 6. **SAMPLE REPORT OUTPUT: CASE STUDY-U.K. Laws-Understanding professional environment**

Report: Online Harassment and Threat Management for UK Journalist
Situation Overview:
You, a journalist in the UK, have reported receiving death threats via direct messages on X (formerly Twitter) and Facebook. Additionally, there have been attempts to discredit your professional reputation by claiming you're not a real journalist.

Key Issues:

1. Death threats on social media platforms
2. Potential risk of being tracked to your home
3. Attacks on professional credibility

Recommended Action Plan:

1. Immediate Safety Measures:
    - Ensure you're in a safe location
    - Review and enhance home security if necessary
1. Document the Threats:
    - Take screenshots of all threatening messages
    - Create a log with dates, times, and details of each incident
    - Save original messages and any associated metadata
    - Back up all evidence securely
1. Report to Authorities:
    - Contact local police via non-emergency number 101 (or 999 if in immediate danger)
    - File a formal report and obtain a crime reference number
    - Provide police with all documented evidence
    - Follow up regularly on the investigation
1. Platform Reporting:
    - Report threats to X and Facebook using their abuse reporting systems
    - Document all reporting actions taken
1. Enhance Digital Security:
    - Use a reputable VPN to mask your IP address
    - Enable Two-Factor Authentication on all accounts
    - Review and tighten privacy settings on all social media platforms
    - Disable location services when not needed
    - Use secure, encrypted messaging apps for sensitive communications
1. Manage Online Presence:
    - Conduct a thorough online search of your name and information
    - Set up Google Alerts for your name
    - Check people search engines and data broker sites for your information
    - Request removal of sensitive personal information where possible
1. Protect Professional Reputation:

    - Update and enhance your professional online profiles (LinkedIn, personal website)
    - Join reputable journalism associations (e.g., National Union of Journalists)
    - Seek support from colleagues and employer
    - Consider releasing a professional statement addressing false claims
    - Highlight your credentials and notable work
1. Legal Considerations:
    - Consult with a media law specialist if false claims are seriously damaging your reputation
1. Seek Support:
    - Contact organizations like the Journalism Safety Pioneer Programme or TrollBusters for additional resources and support
    - Consider speaking with a mental health professional to manage stress
1. Ongoing Vigilance:
    - Regularly review and update your security measures
    - Continue to document any further incidents
    - Stay informed about online safety best practices

Remember, you have the right to work safely and free from harassment. These threats are a serious matter, and you're taking the correct steps by addressing them promptly and thoroughly.
Resources:

- TrollBusters: https://yoursosteam.wordpress.com
- Report an incident to TrollBusters: https://yoursosteam.wordpress.com/report-an-incident/
- TrollBusters Training: https://yoursosteam.wordpress.com/trollbusters-courses-and-training/
- National Union of Journalists: https://www.nuj.org.uk/
- Society of Editors: https://www.societyofeditors.org/

Would you like me to email this report to you for future reference?
TrollBusters
https://yoursosteam.wordpress.com/
Report an Incident
https://yoursosteam.wordpress.com/report-an-incident/
TRAINING
https://yoursosteam.wordpress.com/trollbusters-courses-and-training/
National Union of Journalists (NUJ): Home
https://www.nuj.org.uk/

Overall, the Navigator | UK. AI Assistant equips journalists with strategies to protect themselves digitally, legally, and personally, helping them to continue their crucial work in informing the public during these important times.

Indulkar, Ferrier and TrollBusters partnered with media organizations, safety staff and individual journalists to build and test the large language model for our AI assistant. From May 2024 to November 2024, we beta tested our tool at the June 2024 Radio Television Digital News Association (RTNDA) conference with more than 75 attendees, with women journalists and media leaders from the International Association For Women in Radio & Television (IAWRT), attorneys and solicitors in the US and the UK, members and safety leadership of the Coalition Against Online Violence (CAOV) in August 2024 and ACOS (A Culture Of Safety) Alliance members at the annual meeting in October 2024. In the autumn of 2024, the tool was also tested by journalism educators and used by undergraduate students and was awarded the prestigious Anthem Award for responsible technology design in November 2024 by the International Association of Digital Arts & Sciences. A snowball sample was used to test and deploy the tool in the UK, the USA, and in other geographies.

Between May 2024 and November 2024, we received 502 visits to the tool. Of the 502 unique visitors, 243 provided an email to receive detailed reports of the conversations. Our team reviewed the 502 conversations of users, validating any factual and tactical information with our own research. We tweaked the prompts to ensure a professional, compassionate, trauma-informed response that provided just-in-time support and guidance to users.

> This was like really cool because I put in that I'd been doxxed and family members had been doxxed, and it had a sense of response, and then it said, are you safe? And I said, I'm not sure if I'm safe, how do I assess that? And it had like a very, like, the immediate response to even assess safety in the home was like, just really well done. I mean, I'm just impressed. That's as far as I got. But I'm just impressed with the situation specificity of the response.
>
> Navigator UK is a game-changer for journalists under attack, offering not just practical help but also a sense of being understood and supported in tough situations. The chatbot provided comprehensive, actionable suggestions along with direct links to relevant resources—all presented with an empathetic and supportive tone.

Based on feedback from the beta testers, Navigator | USA went live in September, just in advance of the US elections. The conversations with the tool are also used to create educational videos to aid reporters experiencing similar harms. The explainer videos are deployed on social platforms, to provide education to others.

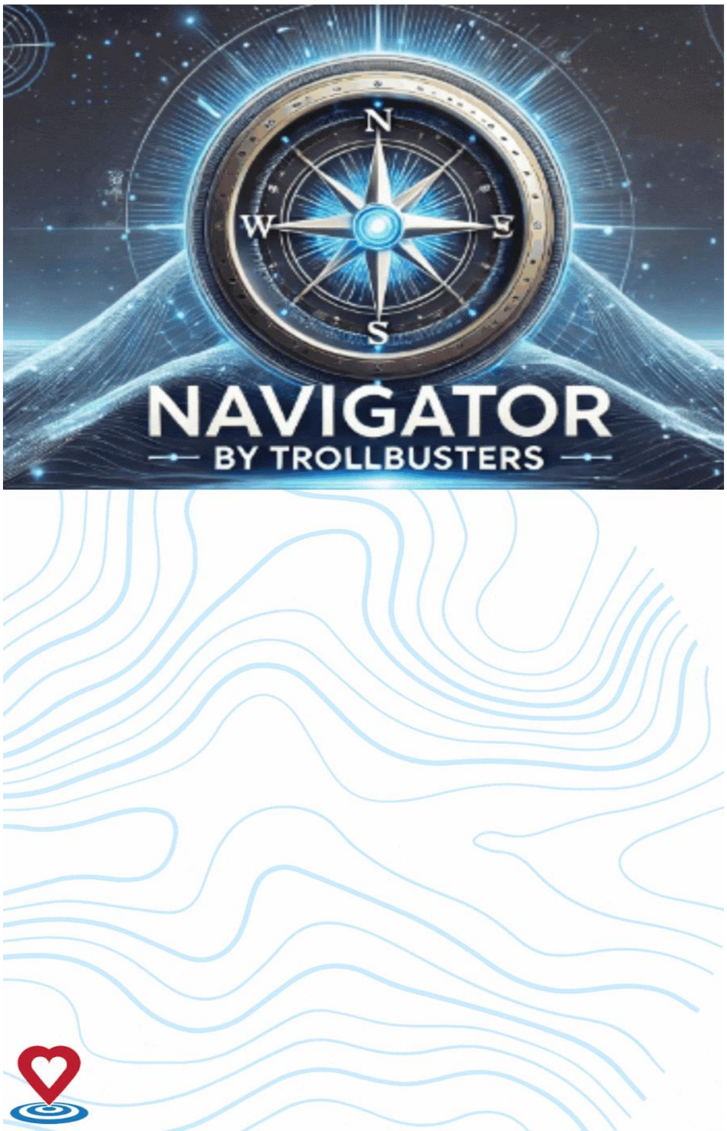
Navigator | UK is trained to provide risk analysis, psychosocial guidance, and support in safety planning and other tools to manage the stress of the situation. For example, we tested a scenario where the threat was identified to be coming from inside the journalist’s home, potential tracking software on the mobile phone. The Navigator tool delivered a comprehensive report including the following details:

1. Initial Situation Description

2. Immediate Safety Actions
3. Legal Steps and Documentation
4. Safety Planning and "Go Bag" Preparation
5. Technology Security Concerns
6. Support Resources in Geography
7. Emotional and Professional Impact
8. Risk Assessment for Phone, Car, and House
9. Secure Communication Methods
10. Ongoing Safety Measures

Because the tool was tested on a variety of LLMs and its knowledge base includes deep understanding of the technological challenges of the cross-border work of journalists, Navigator is able to assist across geographies to triage and assess situational risks and possible strategies. Output from these conversations are turned into video explainers and deployed on social media channels: (Figure 7.)

FIGURE 7. EXPLAINER VIDEO: CASE STUDY-Non-U.K. Journalist (Government Threat): Comprehensive exit strategy, escape plan to safety.



Our second build focused on the US and its jurisprudence. TrollBusters utilized existing research and work during the first Trump administration and election misinformation efforts to build a database of US jurisprudence. The rationale was to prepare for future US elections and the increase in digital and physical threats towards American journalists and journalists working in the US and abroad. TrollBusters and Ferrier worked with the *Haitian Times* because of the doxxing attacks that the publisher, editor and journalists in both the US and Haiti offices received after the presidential debate in the US. The journalists in Springfield, Ohio had been under siege, including doxxing, swatting and other kinds of activities (Figure 8). Navigator| USA was deployed in real time to provide the journalists with guidance on what they should be doing next and how they can help the community as well. Output from Navigator | USA is packaged into a short video explainer and deployed online in the tainted stream by TrollBusters and the Haitian Times staff:

Figure 8. User Interaction with Navigator | USA

REPORT: User just had an interaction with Navigator|USA.
HARASSMENT SITUATION SUMMARY AND ACTION PLAN Situation Overview: - User is experiencing coordinated harassment from an alt-right group - Platforms affected: X/Twitter (DMs and public posts) and LinkedIn - Nature of harassment: Intimidation (not direct physical threats) - Current status: User is physically safe - Actions already taken: Documentation (screenshots and links collected), reported to platforms, basic safety measures initiated
Recommended Action Plan:
1. IMMEDIATE SECURITY MEASURES - Enable two-factor authentication on all accounts - Review and strengthen privacy settings - Consider temporary account lockdown options - Save all documentation in secure, backed-up location
2. REPORTING AND DOCUMENTATION - Continue documenting all incidents with timestamps - Create incident log with patterns of behavior - Save URLs and archive pages using services like Archive.is - Report to platforms' trust and safety teams
3. PROFESSIONAL SUPPORT - Contact employer's security team if applicable - Consider reporting to local law enforcement - Document case number if filed - Consider legal consultation for potential civil action
4. DIGITAL SAFETY ENHANCEMENT - Review all connected accounts for security - Check for personal information leaks - Consider using professional monitoring services - Review digital footprint and remove sensitive data
5. SUPPORT RESOURCES - TrollBusters incident reporting: https://yoursosteam.wordpress.com/report-an-incident/ - Digital security training: https://yoursosteam.wordpress.com/trollbusters-courses-and-training/ - Professional support network activation recommended
6. SELF-CARE STRATEGIES - Establish boundaries with social media use - Create support system protocol - Document impact on professional work - Maintain regular self-care routine
Additional Notes: - Situation appears to be active and ongoing - User has taken appropriate initial steps - Professional support recommended for escalation if harassment intensifies - Regular security reviews advised

Conclusion

Designed to help people in vulnerable situations, Navigator | AI responds on the basis of as little or as much information the user feels comfortable sharing. The power to take action against attackers is given back to journalists. The tool offers valuable guidance with even limited information based on the extensive database of knowledge and resources collected by TrollBusters and the tool’s knowledge base of the complexity of journalists’ professional work. But this tool was not created or intended to replace human interaction or professional help. It is a basic resource that journalists can use when they need immediate assessment, assistance and guidance on navigating digital harms. Navigator has proven to be a valuable tool for triage, analysis, strategy and education when journalists are in distress. Certainly, AI tools will need to be continuously updated and it has to be recognized that they will not solve all of the problems faced by journalists. In order to tackle online harassment, a mixture of methods has been suggested as being appropriate, such as the implementation of legislation, urging for social media companies to take responsibility in tackling online harassment, and also the development of media literacy initiatives (Coe, 2024; Horton, 2025). Nonetheless, AI tools, such as the Navigator, provide a resource for journalists to turn to when they require immediate assistance and guidance when facing digital harm.

Our next steps are to continue to add the jurisprudence of other countries and track the results of litigation efforts and challenges to press freedom, recognizing that legal redress is a moving target and but one avenue for aid or action. In addition, Ferrier and TrollBusters continue to update the knowledge base with new vulnerabilities identified by users and the reports they share. As the technological threats continue in real time across the globe Navigator | AI will continue to learn and guide users to safe havens and best practices no matter where in the world, or online, they may be. At a time when journalists across the world are facing unprecedented digital challenges and physical threats, Navigator and generative AI triage can provide support and aid to journalists in distress.

AI Disclosure Statement

No AI was used in the text of the article or to assist with writing any of the article.

References

Amnesty International. (2025). Journalists targeted with Pegasus spyware. Amnesty International. https://securitylab.amnesty.org/latest/2025/03/journalists-targeted-with-pegasus-spyware/

ARTICLE 19. (2020). Investigating online harassment and abuse of women journalists. ARTICLE 19. https://www.article19.org/wp-content/uploads/2020/11/Gender-Paper-Brief-3-.pdf

Bhatia, K. V., Elhussein, M., Kreimer, B., & Snapp, T. (2023). Protests, Internet shutdowns, and disinformation in a transitioning state. Media, Culture & Society, 45(6), 1101-1118. https://doi.org/10.1177/01634437231155568

Bleyer-Simon, K. (2021). Government repression disguised as anti-disinformation action: Digital journalists' perception of COVID-19 policies in Hungary. Journal of Digital Media & Policy, 12(1), 159-176. https://doi.org/10.1386/jdmp_00053_1

Center for News, Technology & Innovation. (2024). Most 'Fake News' Legislation Risks Doing More Harm Than Good Amid a Record Number of Elections in 2024. Center for News, Technology & Innovation.

Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2020). 'You really have to have a thick skin': A cross-cultural perspective on how online harassment influences female journalists. Journalism, 21(7), 877-895. https://doi.org/10.1177/1464884918768500

Claesson, A. (2022). "I Really Wanted Them to Have My Back, but They Didn't"—Structural Barriers to Addressing Gendered Online Violence against Journalists. Digital Journalism, 11(10), 1809–1828. https://doi.org/10.1080/21670811.2022.2110509

Coe, P. (2024). Tackling online false information in the United Kingdom: The Online Safety Act 2023 and its disconnection from free speech law and theory. Journal of Media Law, 15(2), 213-242.

Dodds, T., Geboers, M., & Boukes, M. (2024). "It Became No Man's Land": The Burden of Moderating Online Harassment in Newswork. Journalism Practice, 1–18. https://doi.org/10.1080/17512786.2024.2387664

Fang, K. (2024). Wangbao (Cyberbullying) and Jubao (Reporting): Strategic Ambiguity in CollaborativeState-Society Influence Operations in China. Journal of Online Trust and Safety, 2(3), 1-22. https://doi.org/10.54501/jots.v2i3.167

Ferrier, M. P. (2018). *Attacks and Harassment: Threats Against Women Journalists and their Reporting*, Troll-Busters and the International Women's Media Foundation. https://www.iwmf.org/attacks-and-harassment/

Ferrier, M. (2021). Social Media Policies Put Journalists at Risk. Columbia Journalism Review. https://www.cjr.org/tow_center_reports/newsroom-social-media-policies.php

Ferrier, M. (2024) Project Notes: Navigator | UK by TrollBusters. https://yoursosteam.wordpress.com/navigator-uk/

Ferrier, M., & Garud-Patkar, N. (2018). TrollBusters: Fighting Online Harassment of Women Journalists. In: J. Vickery, & T. Everbach (Eds) Mediating Misogyny (pp. 311-332). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-72917-6_16

Gabbatt, A. (2017). The Resistance Now: the troll tax, and a final push to keep the internet free. The Guardian. https://www.theguardian.com/us-news/2017/dec/09/the-resistance-now-trolls-net-neutrality

Gazia, I., Hubbard, T., Scalona, T., Kang, Y., & Zuckerman, E. (2024). Proactive Blocking through the Automated Identification of Likely Harassers. Journal of Online Trust and Safety, 2(3), 1-22. https://doi.org/10.54501/jots.v2i3.175

Gerbrandt, R. L. (2023). Media freedom and journalist safety in the UK Online Safety Act. Journal of Media Law, 15(2), 179–212. https://doi.org/10.1080/17577632.2024.2307678

Godel, W., Sanderson, A., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2022) Moderating the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. Journal of Online Trust and Safety, 1(1), 1-36. https://doi.org/10.54501/jots.v1i1.15

Grossman, B. and Gallagher, C. (2021). Combatting Online Abuse and Harassment: A Legal Guide for Journalists in England and Wales. Media Lawyers Association. https://medialawyersassociation.org/wp-content/uploads/2021/06/combatting-online-harassment-and-abuse-23.06.2021-09.10-5.pdf.

Gruenbaum, O. (2024). If it is anything, the Commonwealth must stand for press freedom. The Round Table: The Commonwealth Journal of International Affairs and Policy Studies. https://www.commonwealthroundtable.co.uk/general/eye-on-the-commonwealth/if-it-is-for-anything-the-commonwealth-must-stand-for-press-freedom/#

Hoffman-Andrews, J. (2020). Block Together is now full shut down. Block Together. https://blocktogether.org/

Horton, G. (2025). Journalism as the lifeblood of democracy: How can we ensure it is protected? Communications Law, 30(2), 47-57.

Jamil, S. (2020). Suffering in Silence: The Resilience of Pakistan's Female Journalists to Combat Sexual Harassment, Threats and Discrimination. Journalism Practice, 14(2), 150-170. https://doi.org/10.1080/17512786.2020.1725599

Katsaros, M., Tyler, T., Kim, J., & Meares, T. (2022). Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter. Journal of Online Trust and Safety, 1(3), 1-26. https://doi.org/10.54501/jots.v1i3.38

Khan, Irene. (2024). Journalists in exile Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, United Nations Human Rights Council Fifty-sixth session 18 June–12 July 2024 Agenda item 3 Promotion and protection of all human rights, civil, political, economic, social and cultural rights, including the right to development.

Kneuer, M., Schünemann, W. J., & Bahms, G. (2024). Internet Control and Disinformation Across Regime Types During and After the Covid-19 Crisis. Politics and Governance, 12(1), 1-21.

Koirala, S. (2020). Female Journalists' Experience of Online Harassment: A Case Study of Nepal. Media and Communication, 8(1), 47-56. https://doi.org/10.17645/mac.v8i1.2541

Lee, N. Y., & Park, A. (2024). How online harassment affects Korean journalists? The effects of online harassment on the journalists' psychological problems and their intention to leave the profession. Journalism, 25(4), 900-920. https://doi.org/10.1177/14648849231166511

Lenhart, A., Ybarra, M., Zickuhr, K., & Price-Feeney, M. (2016). Online Harassment, Digital Abuse and Cyberstalking in America. Data & Society Research Institute. https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf

Lewis, S. C., Zamith, R., & Coddington, M. (2020). Online Harassment and Its Implications for the Journalist–Audience Relationship. Digital Journalism, 8(8), 1047–1067. https://doi.org/10.1080/21670811.2020.1811743

Lim, G. & Bradshaw, S. (2023). Chilling Legislation: Tracking the Impact of "Fake News" Laws on Press Freedom Internationally. Center for International Media Assistance. https://www.cima.ned.org/wp-content/uploads/2023/06/CIMA-Chilling-Legislation_web_150ppi.pdf

Lindsey, et al., "On the Biology of a Large Language Model", Transformer Circuits, 2025. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Mosey, R. (2024). Time to stand firm – social media and challenges to reporting. The Round Table: The Commonwealth Journal of International Affairs and Policy Studies. https://www.commonwealthroundtable.co.uk/general/media/time-to-stand-firm-social-media-and-challenges-to-reporting/#

Neyazi, T. A. (2024). Is Press Freedom Better Protected in Democracies? Internet Shutdown and Journalism in India. Journalism Studies, 1–20. https://doi.org/10.1080/1461670X.2024.2433241

Ndlovu, M., & Khupe, N.A. (2023). Online Harassment of Journalists in Zimbabwe: Experiences, Coping Strategies and Implications. In: T. Matsilele, S. Mpofu, & D. Moyo (Eds) New Journalism Ecologies in East and Southern Africa. Palgrave Studies in Journalism and the Global South (pp.95-11). Palgrave Macmillan.

Nelson, J. L. (2023). “Worse than the Harassment Itself.” Journalists’ Reactions to Newsroom Social Media Policies. Digital Journalism, 11(8), 1456–1474. https://doi.org/10.1080/21670811.2022.2153072

National Union of Journalists. (2021). Written evidence submitted by the National Union of Journalist: NUJ submission to the CMS sub-committee on online safety and online harms. https://committees.parliament.uk/writtenevidence/38562/pdf/

Papadopoulou, L., & Maniou, T. A. (2021). ‘Lockdown’ on Digital Journalism? Mapping Threats to Press Freedom during the COVID-19 Pandemic Crisis. Digital Journalism, 9(9), 1344–1366. https://doi.org/10.1080/21670811.2021.1945472

Pain, P., & Chen, V. (2018). This Reporter is so Ugly, How can She Appear on TV? Negotiating gender online and offline in Taiwanese media. Journalism Practice, 13(2), 140–158. https://doi.org/10.1080/17512786.2017.1423236

Perrigo, B. (2021). Governments Used Spyware to Surveil Journalists and Activists. Here’s Why Revelations About Pegasus Are Shaking Up the World. TIME. https://time.com/6081433/pegasus-spyware-monitored-journalists-activists/

Posetti, J. and Shabbir, N. (2022). The Chilling: A Global Study of Online Violence against Women Journalists. International Center for Journalists. https://www.icfj.org/sites/default/files/2023-02/ICFJ%20Unesco_TheChilling_OnlineViolence.pdf

Posetti, J., Aboulez, N., Bontcheva, K., Harrison, J., & Waisbord, S. (2020) Online Violence against Women Journalists: A Global Snapshot of Incidence and Impacts. International Center for Journalists. https://www.icfj.org/sites/default/files/2020-12/UNESCO%20Online%20Violence%20Against%20Women%20Journalists%20-%20A%20Global%20Snapshot%20Dec9pm.pdf

Pritchard, S. (2016). The readers' editor on…handling comments below the line. The Guardian. https://www.theguardian.com/commentisfree/2016/jan/31/readers-editor-on-readers-comments-below-the-line

Radsch, C. A. (2022). AI and Disinformation: State-Aligned Information Operations and the Distortion of the Public Sphere. OSCE Representation on Freedom of the Media, Organization for Security and Co-operation in Europe.

Radsch, C. C. (2023). Platformization and Media Capture: A Framework for Regulatory Analysis of Media-Related Platform Regulations. UCLA Journal of Law & Technology, 28(2) 175-223.

Rahman, R. A., & Tang, S. (2022). Fake News and Internet Shutdowns in Indonesia: Symptoms of Failure to Uphold Democracy. Constitutional Review, 8(1) 151-183.

Ransome, D. (2024). From 'Climate Barbie' to "freedom of abuse": Women in the online spotlight. The Round Table: The Commonwealth Journal of International Affairs and Policy Studies. https://www.commonwealthroundtable.co.uk/general/politics/from-climate-barbie-to-freedom-of-abuse-women-in-the-online-spotlight/

Reilley, M. (Host) (2024, November 12). Navigator: A Chatbot That Helps Journalists Deal With Trolls. JournalistsToolbox.ai. https://journaliststoolbox.substack.com/p/navigator-a-chatbot-that-helps-journalists

Relly, J. E. (2021). Online harassment of journalists as a consequence of populism, mis/disinformation, and impunity. In: H. Tumber and S. Waisbord (Eds) The Routledge Companion to Media Disinformation and Populism. Routledge.

Reporters Without Borders. (2018). Online Harassment of Journalists: Attack of the Trolls. Reporters Without Borders. https://rsf.org/sites/default/files/rsf_report_on_online_harassment.pdf

Rueckert, P. (2021). Pegasus. The new global weapon for silencing journalists. Forbidden Stories. https://forbiddenstories.org/pegasus-the-new-global-weapon-for-silencing-journalists/

Shah, N. (2021). (Dis)information Blackouts: Politics and Practices of Internet Shutdowns. International Journal of Communication, 15(1), 2693-2709.

Shah, S. F. A., Cvetkovic, I., Ginossar, T., Ullah, R., Baber, D., & Slaughter, A. (2024). Online Harassment, Psychological Stressors, and Occupational Dysfunction among Journalists Working in a Conflict Zone. Digital Journalism, 12(6), 735–752. https://doi.org/10.1080/21670811.2024.2308582

Sharman, D. (2021). Facebook to give journalists and publishers greater control on comments. HoldtheFrontPage.co.uk. https://www.holdthefrontpage.co.uk/2021/news/facebook-to-give-journalists-and-publishers-greater-control-on-comments/

Simon, P. (2014). 'Block Together' developed by a former Twitter engineer who automatically blocks harassing Twitter accounts and allows users to share block lists with each other. Gigazine. https://gigazine.net/gsc_news/en/20140821-block-together/

Stahel, L., & Schoen, C. (2020). Female journalists under attack? Explaining gender differences in reactions to audiences' attacks. New Media & Society, 22(10), 1849-1867. https://doi.org/10.1177/1461444819885333

Tandoc, E. C., Sagun, K. K., & Alvarez, K. P. (2021). The Digitization of Harassment: Women Journalists' Experiences with Online Harassment in the Philippines. Journalism Practice, 17(6), 1198–1213. https://doi.org/10.1080/17512786.2021.1981774

The Hill. (2024). Death threats and doxxing: Haitian Times founder says fake stories are hurting real people

Rising / September 20, 2024. https://thehill.com/video/death-threats-and-doxxing-haitian-times-founder-says-fake-stories-are-hurting-real-people/10060226/

Uwalaka, T., Amadi, A. F., Nwala, B., & Wokoro, P. (2023). Online harassment of journalists in Nigeria: audience motivations and solutions. Media International Australia, 0(0). https://doi.org/10.1177/1329878X231206840

Waisbord, S. (2020). Mob Censorship: Online Harassment of US Journalists in Times of Digital Hate and Populism. Digital Journalism, 8(8), 1030–1046. https://doi.org/10.1080/21670811.2020.1818111

Wheatley, D. (2025). Journalists' Use of Social Media Disconnection Practices: 'I Try Not to Block People, but….' Digital Journalism, 1–18. https://doi.org/10.1080/21670811.2024.2436545

Wright, S. (2019). When Journalists Go "Below the Line": Comment Spaces at *The Guardian* (2006-2017). Journalism Studies, 21(1), 107-126.

Young, G. K. (2021). How much is too much: the difficulties of social media content moderation. Information & Communications Technology Law, 31(1), 1–16. https://doi.org/10.1080/13600834.2021.1905593

Zviyita, I., & Mare, A. (2024). Same threats, different platforms? Female journalists' experiences of online gender-based violence in selected newsrooms in Namibia. Journalism, 25(4), 779-799. https://doi.org/10.1177/14648849231183815

Zviyita, I., & Mare, A. (2024). Same threats, different platforms? Female journalists' experiences of online gender-based violence in selected newsrooms in Namibia. Journalism, 25(4), 779-799. https://doi.org/10.1177/14648849231183815