

Guardian: Tool for Protection against Social Media Creeps

Shubhi Katiyar¹, Vartika Choudhary¹, Bhargavee Singh¹, Aditi Srivastava¹, Daood Saleem²

^{1,2}*School of Computing Science and Engineering, VIT Bhopal University, Sehore, India*

Abstract - Online harassment has become a critical issue on social media. Women repeatedly face the risk of various abusive messages. The existing system generally at platforms like Instagram depends on user reporting or basic keyword filtering which make them incapable of detecting repetitive forms of harassment. To bridge this gap, Guardian is developed as an AI powered, privacy-preserving Chrome extension which preemptively detects harmful messages during live Instagram chat. The system uses a fine-tuned DistilBERT transformer model for multiclass text classification, categorizing messages as safe, creepy, or different levels of harassment. Total privacy is guaranteed as all the processing occurs locally on the user's device, making sure that no message content is stored externally. A strike-based tracking system keeps track of repeat offenders. It is activated after three times of continuous harassing behaviour. After that an alert pop up comes automatically bifurcating the type of harassment being done and also highlighting the text at which it's been done. Also, it auto-hides the chats after multiple offenses. It is hooked into using content scripts and a local FastAPI server. Evaluation shows strong model performance, achieving 78% accuracy and for most critical categories of stalker behaviour and severe harassment, both precision and recall are very high. Further, real-time tests confirm this system indeed hides harmful messages while keeping normal conversations untouched.

Keywords - Online harassment detection, privacy-preserving classification, strike-based offender tracking, real-time message moderation, transformer-based text analysis.

1. INTRODUCTION

The fast expansion of digital communication tools and social media platforms has changed how people in society today build relationships and exchange information and show their identity. People can use Instagram and various online communities to create extensive opportunities for communication and networking and learning and self-expression. The platforms enable users to create connections with others who live in different locations while they engage in social conversations and display their artistic work. The widespread ability to connect with others has led to an increase in online harassment [1] which now stands as a major societal problem. The digital platforms which people use nowadays have

resulted in increased exposure to harmful interactions among users from various age groups and gender identities and social backgrounds. Online users from different demographic groups experience cyber harassment as a widespread problem which damages their mental health and social relationships and their ability to use online services.

Cyber harassment includes a wide range of dangerous activities which involve sending abusive messages and trolling and cyberstalking and using fake identities and spreading disinformation and making threats and sending unwanted messages. The target of these actions experiences emotional distress because they create a personal attack which generates fear and insecurity. The studies on gender-based online abuse demonstrate that women and marginalized groups experience [2,3] greater online abuse than other demographics. The psychological effects of harassment increase because of repetitive abuse which combines with escalating patterns of abuse to create more severe long-term effects on victims. Cyber harassment occurs as a sequence of abusive events which repeat themselves because they represent social and cultural and technological forces that operate in society.

1.1 Background

1.1.1 Nature and Forms of Cyber Harassment

Cyber harassment exists on various digital platforms which include social networking sites and messaging apps and online gaming environments and forums and professional networks. Users commonly encounter identity-based abuse and cybercrime activities [4,5] which include offensive messages and threats and unauthorized sharing of personal information and online impersonation. Harassment can occur through two types of behavior which include minor exclusion and disrespectful comments and major threats and forced disclosure of private details.

Recent studies highlight the increasing role of technology-facilitated violence [6] which demonstrates how digital tools and bots and automated systems are used to track and attack specific people. The internet permits users to view objectifying content and harmful media content which teaches them to adopt abusive behavior while it builds social prejudice against victims and reduces their

empathy. Research on cyberbullying and online aggression [7,8] shows that these two behaviors link with each other while they create community standards which result in ongoing harassment patterns that impact multiple users at once.

1.1.2 Progression and Severity of Online Abuse

Online harassment usually starts with minor harmful acts which eventually develop into severe forms of online abuse. Minor negative interactions, such as unsolicited comments or subtle insults, can escalate into persistent abuse, which develops through the process of continuous targeting [9,10] and increased harassment activity by offenders. The severity and impact of cyber harassment are influenced by multiple factors, including the type of abuse, duration, frequency, and the victim's online presence. Studies on abuse intensity and impact variation [11-13] show that repetitive harassment exposure causes people to experience higher psychological stress which results in increased anxiety and social withdrawal. The identification of dangerous situations depends on understanding these patterns, which also helps to create preventive measures and educational programs and policy frameworks.

1.1.3 Contextual and Behavioural Perspectives

The process of identifying and managing cyber harassment offenses requires complete knowledge about actual online communication [14] settings which includes all elements of prior interactions together with their associated emotional expressions and their intended meanings and the distinctive communication methods that each online platform permits. The ability to identify contextual elements enables people to recognize between normal conflicts and dangerous conduct which assists authorities and moderators in making correct decisions.

The online environment depends on user engagement because it establishes the framework for digital spaces. Research on bystander behavior and intervention patterns [15,16] shows that online communities influence the spread and mitigation of harassment. Bystanders who recognize abusive behavior can intervene by reporting incidents or supporting victims, whereas inaction can reinforce harassment. The studies on user interaction and engagement behavior [11,12] demonstrate that active participation together with community moderation efforts results in decreased harmful interactions which happen in online spaces.

1.1.4 Socio-Cultural Influences on Harassment

Cyber harassment shows a connection to the cultural and societal standards which exist in society. In areas which experience increasing cybercrime attacks against women users face greater risk of

online harassment especially women and vulnerable populations. The combination of social status systems and gender roles and male-dominated social systems leads to increased harassment incidents which create safe spaces for attackers because their victims remain silent.

Research on online silencing and gender inequality [18,19] demonstrates how societal expectations restrict women's ability to express themselves freely. Broad studies which use social and theoretical approaches to research [20-22] show that cultural norms and power imbalances together with institutional structures determine how cyber harassment becomes accepted and continues to exist. The understanding of these socio-cultural factors provides essential knowledge which enables the creation of effective educational programs and awareness campaigns and intervention strategies that focus on solving the main problems behind harassment instead of treating its visible effects.

1.1.5 Psychological and Social Impact

Cyber harassment creates effects that extend beyond digital platforms because it disrupts victims' psychological and social health. Researchers have discovered that people who experience stress together with anxiety and psychological distress [20,21] will face difficulties in maintaining focus and they will experience sleep problems and decrease their time spent engaging in both virtual and real-life activities.

The long-term social effects of harassment emerge when some victims choose to limit their online activities or they decide to stay away from particular social media platforms. Studies on user safety concerns and behavioral changes [17,18] indicate that repeated exposure can lead to social withdrawal, reduced self-confidence, and even changes in professional or educational engagement. Research demonstrates that cyber harassment causes emotional and social and mental health effects [21,22] because its impact creates multiple effects that continue to exist.

1.1.6 Repetitive Patterns and External Factors

Cyber harassment occurs through established repeating patterns which take place in structured time intervals. Offenders usually select their victims through systematic approaches which result in ongoing abusive behavior. Studies on trolling and repeated harassment behavior [8,9] show that such actions are intentional, coordinated, and recurring rather than isolated incidents.

The occurrence and severity of online harassment increase because of both environmental elements and external factors which include times when people use social media more and important global events and emergency situations. The research

shows that outside pressures and online patterns of behavior which users encounter lead to higher technology-enabled harassment rates [9,10] in these specific environments.

1.1.7 Legal and Protective Perspectives

Cyber harassment needs both strong legal systems and effective protective systems to achieve successful resolution. Research highlights the importance of laws, policies, and institutional safeguards [23-25] that aim to protect users, provide recourse to victims, and deter potential offenders.

The existing legal system faces operational difficulties because people lack understanding of the law and because enforcement agencies operate weakly and because people face obstacles when trying to report violations. The research studies on legal frameworks and justice systems [25-27] demonstrate that successful implementation requires educational programs and platform-based solutions to enable laws to deliver actual protection and justice for victims.

1.2 Significance of the Study

Cyber harassment has become more common in society which requires complete research for its root causes and its development process and all its consequences. The research investigates victim response patterns where online abuse occurs to show how harassment develops and progresses and impacts victims through psychological and social effects.

The study uses victim experience findings and justice system research from [27,28] together with cyberbullying environment and institutional response research from [29,30] to create a complete understanding of the subject. The research establishes safer digital environments which guide policy development and promote responsible online behavior to diminish the negative effects of cyber harassment on individuals and communities.

2. PROBLEM STATEMENT

The ability to communicate and connect through social media is available to users; however, the platform exposes many users (especially women) to online abuse. Today, we are able to implement reactive solutions that can help provide an efficient and user-focused solution for maintaining safety during communication in particular, when communicating with unknown individuals.

Key Issues Identified:

- Many women have received inappropriate, creepy, and/or abusive messages via the internet. Harassment can be directed toward women through different methods; including minor inappropriate commentary as well as severely threatening/stalking behaviors.
- Experiences of online abuse can negatively affect women's emotional wellbeing and overall self-esteem.
- Many women do not use social media or use social

media less than they would if they had not had negative experiences online.

- The effectiveness of most current solutions; such as reporting and blocking; is that they only address the issue of abuse after the fact.
- There are many forms of indirect/subtle harassment that are hard for women to see and consequently demonstrate that they occurred.
- Women have nowhere to go for assistance after receiving harmful messages.
- There is currently no system for providing timely assistance to women who have received harmful messages.
- A solution is required that promotes and enhances the online safety of women in an effective and user-friendly manner.

2.1 Objectives of the Project

The main aim of this project is making a system that offers a more secure, hassle-free and harassment-free experience for women on social media platforms. The system uses artificial intelligence for detecting, classifying, and hiding abusive or creepy messages in real time, giving users control over their digital interactions again. Specific Objectives of this project are:

- The focus is to build an AI-based Chrome extension which can monitor and analyse, in real time, Instagram DMs for messages that are potentially abusive, creepy, or harassing. In this case, this task would involve the multiclass classification of incoming messages into categories such as safe, creepy, mild harassment, medium harassment, and severe harassment, based on linguistics and emotional tone.
- It provides total user privacy since all data is processed directly on the user's device, without sending any information to any server or cloud platforms.
- To provide full control for the user over enabling and disabling the extension, thus being able to decide when the tool is working and how it treats the recognized messages.
- Implement a strike system that monitors repeat offenders, so that whenever a user has been identified for multiple strikes, their messages would be automatically deleted or flagged, or both - creating assurance of long-term safety for USERS.
- Let's implement action to reduce stress on the emotional and psychological levels for users being harassed online, and let them communicate freely and without fear.
- Encourage the further responsible and ethical use of AI in applications related to social safety, while making a case for continued research and innovation in digital safety tools.

With these goals, Guardian hopes to protect users

not only from harassment but to change the way technology is used in order to make digital ecosystems that thrive on empathy. This is a project that imagines a world where social media continues to be a platform for staying connected and creative-not a source of fear or anxiety.

2.2 Scope and Limitations

Scope:

- The system is designed to detect and filter abusive or harassing text messages on the Instagram web interface.
- It uses machine learning-based natural language processing to classify messages on a severity scale ranging from safe to severe harassment.
- All processes run locally at the user's device, thus providing data privacy.
- Users have full control over when they want to enable or disable this tool.
- This system tracks the behaviour of senders to protect users against repeated patterns of harassment.

Limitations:

- The system currently accepts only text-based harassment; images, emojis, and videos which may contain inappropriate content are not analysed.
- A significant challenge that is presented is regarding the diversity of language- the model may have a low degree of accuracy with slang, geographically specific phrases, and text that is purely non-English.
- Currently, the project is focused solely on the web-based version of Instagram and does not include mobile applications or other social networking sites.
- Real-time processing relies completely on the user's system capabilities, which I have found can impact its speed and responsiveness.
- The quality and balance of the training dataset determine the accuracy of the detection system; as online language evolves, it is a process that needs constant improvement.

2.3 Applications

2.3.1 Real-world relevance

The Guardian Project has a particularly noteworthy significance in today's digital environment in which online communication has become commonplace. The sharp increase in users of social media has created an environment in which digital harassment, stalking and verbal abuse are prevalent, more so among women. Current reporting and moderation systems formal or informal are often either delayed or incapable of identifying subtle but no less harmful, manipulative, or repeated harassment. Guardian fills the gap by providing a smart, proactive and privacy-centric approach to

equipping users with tools that allow them to assume control of their online experience.

To put this in context, the extension provides immediate benefit for social media users that deal with unwanted or uncomfortable messages frequently- a difficult experience to navigate, let alone secure mental peace and safety from. It reduces emotional exhaustion generated from continual exposure to toxic content and enables users to engage more fully in processes that require confident interaction in online spaces.

Lastly, Guardian tends to the individual user's safety while contributing to larger social designs for digital well-being, ethical online experience, and gender presence in technology inclusive products. Furthermore, it serves to make online environments safer, contributing to more women and marginalized individuals fearless, open, and full participants of their social, educational, and professional networks.

Ultimately, Guardian is not simply a technology company offering an extension, it is a social design initiative formed out of a need for so many users to foster a respectful and harassment-free online culture.

2.3.2 Potential Domains

While Guardian was conceptualized for the needs of Instagram direct messages, its architecture and design principles are extensible to multiple platforms and domains. It might be applied in the following areas:

- **Other Social Media Platforms:**

The very same detection and filtering system can be applied to the integration of Facebook with Twitter (X), WhatsApp, Snapchat, or Telegram to safeguard users against abusive or harassing messages across platforms.

- **Email and Messaging Applications:**

This capability could be integrated into email programs or work messaging systems, such as Gmail, Slack, or Microsoft Teams, and could also automatically filter messages to automatically delete spam messages in a work environment.

- **Online Gaming Communities:**

Many (but especially women) experience toxic chats while playing multiplayer games. An AI similar to a guardian may be developed to oversee a game's chat systems and could automatically hide or censor inappropriate language or harassment, given prior recommendations.

- **Learning environments:**

Moreover, educational environments may be utilized by this technology to help moderate discussion boards, chat rooms, virtual classrooms, and other learning environments and establish the expectations of students being respectful and polite to one another.

- **Corporate/Customer Support Settings:**

Employees in the roles of customer support often

endure rude or inappropriate language from customers. The results from Guardian can be used to filter out inappropriate communication in customer support chats to make for a better workplace.

Public Forums and Comment Sections:

It can also be extended to news websites, blogs, and community forums to automatically hide the harassing comments which will help in reducing online toxicity. In all, the Guardian system shows wide-ranging versatility and can easily be extended from social media into any text-based communication platform requiring real-time, privacy-safe, intelligent text moderation. It constitutes a step forward in the integration of artificial intelligence with ethical computing to advance digital safety and user empowerment.

3. LITERATURE REVIEW

3.1 Cyber bullying and Online Harassment Targeting Women

One big problem today? Harassment online has grown fast alongside social networks. Studies show tech makes it easier to target women, with research on technology-facilitated violence against women [1,2] showing how harm spreads further than before. What stands out is how messages never stop - popping up here and there without warning. Being hidden behind a screen lets people say things they might not in person. This constant pressure wears down mental well-being more deeply over time.

Women facing cybercrime often encounter old biases dressed in new forms, as highlighted in [3,4] studies on patterns of cybercrime against women. Power imbalances shaped by gender play a big role behind these attacks. Online abuse tends to mirror how disrespect shows up in everyday life, just shifted into digital spaces. Research spanning many countries points out that bullying travels easily between different kinds of apps, with global studies on cyberbullying trends [5] showing its spread across platforms. It pops up where people chat privately, post publicly, even when they work together online. The pattern stays consistent - no single site contains it.

Looking at how people react when exposed to degrading media shows a pattern - over time, it makes disrespect seem ordinary online, especially in [6] research on media-induced harassment and objectification. Studies tracking what targets feel during digital attacks reveal deeper wounds: unease creeps in, self-assurance fades, fear takes hold, participation drops off, which is strongly supported by [7,8] findings on user experiences of cyber harassment.

When images reduce humans to props, behavior shifts without notice. Repeated scenes of exploitation quietly reshape expectations around interaction. Those caught in abusive exchanges often pull back,

voice dimming. What feels like background noise in entertainment may feed real distress far away from screens. Emotional weight piles up silently, altering who speaks and who stays quiet.

When global crises hit, studies find more online attacks targeting women - linked to longer screen time, mixed with increased reliance on chat apps and virtual meetings, particularly observed in [9] research on pandemic-related gender-based violence. Devices dominate daily life now, so risks sneak in quietly, which makes real-time monitoring essential. The complexity surprises most: digital routines, emotional stress, and social patterns knot together, defying fast solutions.

3.2 Online Abuse Patterns and How They Intensify Over Time

Figuring out how online abuse changes over time helps create better tools to catch it. Studies of nasty online groups show mean actions usually grow slowly, with escalation patterns in harassment communities [10] showing how behavior intensifies over time. Starting off with light mockery, things might shift toward focused bullying, danger signs appear later. If nobody steps in, small jabs turn into organized cruelty.

Noticing how people mean what they say matters just as much as the words they type, which is emphasized in [11] intent detection research. A comment can sting even if it seems small - its purpose shapes its impact. Over time, watching users interact reveals patterns nobody planned at first. When someone spends more hours scrolling through feeds, their chance of facing cruelty - or acting cruel - tends to rise without warning, a trend supported by studies on [12] social media usage and cyberbullying behavior. What begins casually might shift into something heavier by slow degrees.

One key step forward involves sorting cyberbullying by how serious it is, as seen in severity-based classification approaches [13], instead of just flagging it as present or absent. That shift brings clearer insight into digital harassment. Another angle looks at the talk around each message, because single lines out of nowhere can miss the point entirely, which aligns with [14] context-aware detection methods. Understanding what came before often changes everything.

How people act around each other shapes how often harassment happens. When others step in - or stay quiet - it can make a situation worse or help stop it, according to [15] studies on bystander behavior. Looking at women's experiences, ongoing harassment makes them more careful, say less, and join fewer conversations online - findings tied to [16] perceived safety and behavioral adaptation. Spotting clear insults matters, yet systems must also notice subtle actions and context if they are to respond well.

Lately, studies have turned attention to how varied languages and casual ways of talking shape online chats. On social networks, people toss together different tongues, shortcuts, made-up words, symbols, plus loose rules for sentences - this mix trips up most software trying to keep track. Meanings hide between lines, needing more than just word-for-word reading to catch them. Older methods struggle because they miss context, failing to label these messages right. Finding meaning beneath the words matters more now, so today's methods build models aware of surroundings, ignoring language barriers. Because people speak differently, systems must adapt - working well no matter who uses them, where they're from, or how they express themselves.

3.3 Social Laws and Mind Effects

Studies focusing on how safe women feel using internet platforms, particularly research on digital safety concerns [17], show just how hard it is to keep personal information protected. Without clear rules set by authorities, handling these attacks becomes even tougher.

When people face online abuse, they tend to step back - maybe by posting less or leaving apps altogether, as seen in [18] studies on user responses to harassment. Women especially pull away from speaking out in open forums after being targeted, which is highlighted in [19] research on gender-based silencing. This pulling back feeds into deeper imbalances over time. Silence grows where voices should be heard.

A heavy toll on the mind often follows cyberbullying, with research on mental health impacts [20] showing how digital attacks fuel stress, anxious thoughts, sadness, along with lasting inner wounds. Work centered on India's experience with online mistreatment, particularly regional studies on cyber harassment [21], adds depth, uncovering social norms and beliefs shaping who gets targeted, how it's seen.

When it comes to online abuse aimed at specific genders, findings from studies on gender-targeted cybercrime [22,23] show rules are needed. Still, reports about how those rules work - or do not - suggest laws lag behind new ways people talk online, as discussed in [24,25] legal protection and enforcement research. Because of this delay, tools that watch and respond instantly might help close what's missing.

3.4 The Need for Smart Automatic Detection

One reason more people want smart tools is how tricky online abuse has become. Because laws and tech must work together, machines might help back up new rules, with studies on integrating legal and technological solutions [26] supporting this approach.

Feminist takes on cybercrime [27] unpack how online abuse is shaped by systemic forces, showing why tech must respond to gendered risks. From another angle, research into user perceptions of harm and justice [28] reveals that responses should speak plainly - making outcomes clear matters just as much as fixing issues.

Early warning tools matter more when dealing with digital abuse, as findings on harassment, stalking, and blackmail [29] suggest the importance of early intervention. Instead of waiting, watching patterns helps - research into risk factors and institutional responses [30] shows steady oversight cuts harm from online bullying.

Most new research points to one thing clearly: how well a system can change matters just as much as its starting design. When language shifts - slang spreads, shortcuts pop up, hidden meanings emerge - old methods often fall short. Because online abuse rarely stays the same, tools stuck in place lose strength fast. A fresh update here or there can help, yet its value vanishes if the system refuses to adapt with each change. Sharpness comes not from single fixes, rather through steady tweaks that question old assumptions. Today's method may fail tomorrow's unseen shifts unless choices bend without breaking. Performance across forums, apps, or chats depends less on size and more on readiness to evolve.

3.5 Technical Limits in Current Detection Methods

Even so, knowing more about online bullying hasn't fixed all tech flaws yet. Most older tools just hunt keywords without seeing how words fit together now. On top of that, those smart algorithms need hand-crafted traits, making them stiff when faced with new types of data.

Though deep learning has boosted how machines grasp context, it usually demands heavy computing power along with vast amounts of labeled data. Because of their design, several current systems struggle to operate in real time, limiting usefulness when quick responses matter. Running on cloud infrastructure introduces issues too - delays creep in, personal information may be exposed, and keeping data safe grows harder.

Most times, things stay just beyond grasp when there is no way to touch or turn a knob. Off-the-shelf gadgets often do their thing without asking, making changes nearly impossible. What stands clear is a growing demand - simple designs, built around people, working fast without delay.

Beyond tech hurdles, how people interact with tools shapes how well bullying detectors work. When interfaces show clear reasons behind flags and let users adjust settings, confidence grows. Picture software that rates threat levels plainly while nudging gently instead of shouting alarms - people tend to

prefer that. Too many mistaken red flags wear down attention fast, causing folks to ignore warnings altogether. Smooth blends into daily app habits matter just as much as correct spotting, pushing design toward calm reliability rather than perfect scores alone.

3.6 Research Gaps and Motivation

Still, after many studies, some pieces stay missing when spotting online bullying. What stands out? Tools that catch abuse right as it happens just do not exist yet. These delays matter - most setups only respond after harm occurs instead of stopping it early.

Few tools handle how badly someone might be targeted, stuck instead with yes-or-no labels. What shows up next could miss mockery or hidden slurs unless systems learn nuance better.

Fears over private details living on distant servers keep growing - this pushes demand for local handling. People now watch closely where information goes, so tools that guard personal content matter more than ever.

Facing such issues head-on, the new Guardian setup uses a tweaked DistilBERT brain inside a web browser add-on for live warnings based on what's happening. Instead of sending data away, everything gets handled right on your device - keeping things private. By sorting abuse into different intensity buckets, it tells apart mild nudges from serious threats. Unlike older tools, this one moves quicker, grows easier, yet feels simpler to use without weighing users down.

4. METHODOLOGY

4.1 Technologies, Tools and Data Selection

Guardian is based on a hybrid mix of state-of-the-art web technologies and the machine learning ones, which have been trained to work together flawlessly inside the user's browser. The front-end of the system is being developed in the form of a Google Chrome extension, in which the latest Chrome Manifest Version 3 not only plays a major security role but also enhances performance.

This choice allows the extension to directly interact with the web-based interface of Instagram, reading the message content without alienating the user by requesting the permissions on the intrusive side.

The AI model is running in the local server designed with the FastAPI as the working backend system and Python as the programming language. The FastAPI is responsible for the system's ability to promptly process messages as they come in, thereby minimizing latency and guaranteeing responsiveness that is close to real-time.

The AI model is being developed using a dataset that consists of Instagram direct messages which have

been processed with great care and anonymity. The dataset includes a wide variety of communication types. Labeling was done by humans to categorize the messages into safety categories from harmless to various levels of harassment and stalking. Data cleaning was a major component of the process and included the removal of duplicates, normalization of textual variations, and addressing of class imbalance to avoid bias during training.

4.2 Principle behind the Detection Model

At the heart of Guardian's AI skills is a DistilBERT model, which has been tuned to perfection-an optimized and compact version of the original BERT transformer model. DistilBERT is able to handle a lot more of BERT's language understanding capabilities, but with fewer parameters that reduce the inference time.

The model was developed as a sequence classification neural network, and the annotated dataset provided it with a large variety of examples across many risk categories.

Hugging Face's Trainer API played a crucial role in the fine-tuning process by managing epochs, batch sizes, and learning rates in an efficient manner so as to increase the prediction accuracy. Regular validations on the subset of a validation dataset ensured the model's strength and that it could generalize to new messages.

The method allows for very sophisticated detection, revealing even the most discreet linguistic signs like threats, stalking behaviours, unwanted attention and rude language that vary in severity, among others.

4.3 Workflow

Data Collection and Labelling:

The training corpus was created in due course of the systematic extraction and anonymization of raw Instagram direct messages. Human annotators assigned several classification tags indicating the safety level of the message and the type of harassment over the message, which was then followed by preprocessing stages that included tokenization, normalization, and implementation of balancing techniques designed to ensure equal representation across the different categories.

Model Training and Evaluation:

The DistilBERT transformer underwent fine-tuning on this dataset through the application of supervised learning paradigms. Model parameters were iteratively modified and validated by means of cross-validation techniques to achieve the best classification performance while preventing overfitting.

Local AI Service Setup:

After the training, the distilled model was exported and incorporated into a Python FastAPI service that was running locally on the user's computer. With this configuration, it is possible to make fast and private predictions without the need for external API calls or any cloud dependency.

Extension Development and Integration:

The content script that is part of the Chrome extension continually tracks the document object model (DOM) of the Instagram web interface to almost instantly capture the received DM text.

The extracted messages are sent through HTTP requests from the background service worker of the extension to the local FastAPI endpoint for the classification process.

The classification results come back right away, leading to visual cues such as color-coded highlights that show the risk levels of the messages.

A strike system counts the flagged messages for each sender. When four infractions are reached, the sender's messages will be hidden automatically, and the user will receive a notification alert. This feature is unique as it combines automated protection and user awareness perfectly.

Data Persistence and User Interface:

All the metadata consisting of the strike counts and hidden users is stored using Chrome's Local Storage API. This allows the retention of data through the browser sessions without depending on external storage. The extension's popup interface shows the users currently hidden senders and gives them the option to manually unhide contacts, thus offering complete control over moderation.

Deployment, Testing, and Validation:

The whole system is installed by utilizing the Chrome developer mode to load the unpacked extension, which opens the door for quick iterations and easier debugging. The team went through a series of tests that were aimed at checking the system's responsiveness, classification accuracy, and usability while at the same time assuring the system's robustness and friendliness.

Guardian, the cutting-edge and privacy-respecting moderation tool, and its users are together fighting the battle against harassment and creepy messages in a way that is ultimately ineffective and non-translucent, relying solely on the powerful yet simple combination of lightweight transformer models, local inference, and smart UI integration.

5. IMPLEMENTATION

5.1 System Architecture

Guardian is designed in a modular system architecture comprising two major components: Frontend (Chrome Extension) and Backend (Local ML Server), which interact in real time to analyse and filter social media DMs. The text is pre-processed and passed to a fine-tuned DistilBERT model trained for multi-class harassment detection.

The model classifies messages into six categories ranging from safe content to different levels of harassment such as creepy flirt, mild harassment, stalking behavior etc.

Based on the severity, the system either allows the message or instantly alerts the user.

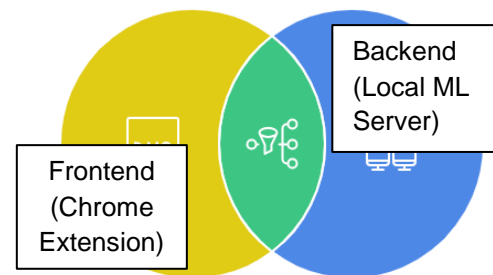


Figure 5.1: Two modules of the project: Chrome Extension, ML server.

5.1.1 Frontend

1. Social Media Web Interface

- This is the front-end platform or, in other words, Social Media's website, where the user views and receives messages.
- Since Chrome extensions have the ability to interface with webpages, the Guardian extension "listens" for new messages that appear in the DOM.
- Here is where the system starts - the content script is injected into this page to read messages.

2. Content Script

- This content script runs inside the context of the Instagram web page.
- It continuously monitors the message area using DOM selectors.
- Each time a new message appears, it:
 - Extracts the text content.
 - Sends it to the background service worker for analysis.

NEW IG DM DETECTED from unknown_ig_user: Hii
NEW IG DM DETECTED from unknown_ig_user: How r u?
NEW IG DM DETECTED from unknown_ig_user:
NEW IG DM DETECTED from unknown_ig_user: ??
NEW IG DM DETECTED from unknown_ig_user: Yoh are so stupid
NEW IG DM DETECTED from unknown_ig_user: I hate you
NEW IG DM DETECTED from unknown_ig_user: Go away loser
NEW IG DM DETECTED from unknown_ig_user: Hello
NEW IG DM DETECTED from unknown_ig_user: Hii
NEW IG DM DETECTED from unknown_ig_user: How r u?
NEW IG DM DETECTED from unknown_ig_user:
NEW IG DM DETECTED from unknown_ig_user: ??
NEW IG DM DETECTED from unknown_ig_user: Yoh are so stupid
NEW IG DM DETECTED from unknown_ig_user: I hate you
NEW IG DM DETECTED from unknown_ig_user: Go away loser
NEW IG DM DETECTED from unknown_ig_user: Hello

Figure 5.2: Detection of incoming texts using Content Script.

3. Background Service Worker

- This is the central controller of the Chrome extension.
- It receives messages from the content script.
- For every message:
 - It sends the text to the ML Inference API (Fast API backend).
 - Gets the predicted label, such as Safe, Creepy, Harassment.
 - Updates strike counts for each sender.

Server response					
Code	Details				
200	<p>Response body</p> <pre>{ "text": "You are beautiful", "prediction": "safe" }</pre> <p>Response headers</p> <pre>access-control-allow-credentials: true access-control-allow-origin: * content-length: 48 content-type: application/json date: Sat,08 Nov 2025 09:44:05 GMT server: uvicorn</pre> <p>Responses</p> <table border="1"> <thead> <tr><th>Code</th><th>Description</th></tr> </thead> <tbody> <tr><td>200</td><td>Successful Response</td></tr> </tbody> </table>	Code	Description	200	Successful Response
Code	Description				
200	Successful Response				

Figure 5.3 (a): Server response as 'Safe'

Server response					
Code	Details				
200	<p>Response body</p> <pre>{ "text": "You are stupid", "prediction": "harassment_severe" }</pre> <p>Response headers</p> <pre>access-control-allow-credentials: true access-control-allow-origin: * content-length: 58 content-type: application/json date: Sat,08 Nov 2025 09:46:16 GMT server: uvicorn</pre> <p>Responses</p> <table border="1"> <thead> <tr><th>Code</th><th>Description</th></tr> </thead> <tbody> <tr><td>200</td><td>Successful Response</td></tr> </tbody> </table>	Code	Description	200	Successful Response
Code	Description				
200	Successful Response				

Figure 5.3 (b): Server response as 'Harassment'

5.1.2 Backend

- AI backend is running locally at <http://127.0.0.1:8000>.
- It hosts the fine-tuned DistilBERT model, trained on the harassment/DM dataset.
- When the background script sends a message:
 - API predicts the category of the message.
 - Returns a label such as:
 - Safe
 - Creepy flirt
 - Mild harassment
 - Medium harassment
 - Severe harassment

1. Message Classification Result

- Once the classification is received:
- The background worker sends the result back to the content script.
- The content script visually highlights or hides that message in the Instagram UI.

Example:

Safe → No change

Creepy flirt → Yellow highlight

Harassment → Red highlight or blurred message

2. User Alerts & Strike Tracking

- Each sender has a strike counter stored in `chrome.storage.local`.
- When the user repeatedly sends toxic messages (for instance, 4 strikes):
 - It automatically hides the sender.
 - The user can unhide them at any time using the extension popup.

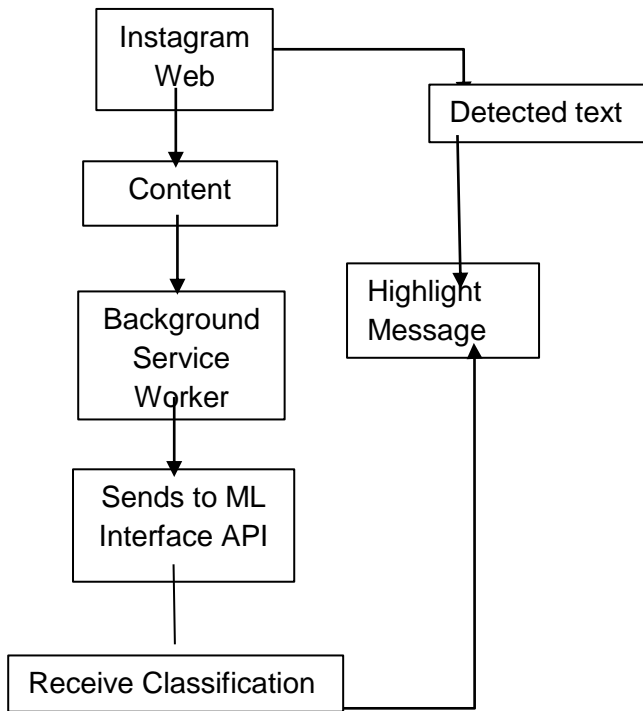


Figure 5.4: System Architecture for Multi class Harmful Content Classification

5.2 Tools, Platforms and Libraries used

Table 5.1: Tools, platforms, and Libraries used in project.

COMPONENTS	TOOLS/TECHNOLOGIES USED
Language	Python (Backend), JavaScript (Frontend)
ML Model	DistilBERT (HuggingFace Transformers)
Backend Framework	FastAPI
Browser Extension Framework	Chrome Manifest V3
Dataset	Custom labeled dataset
Frontend Libraries	HTML, CSS, JavaScript
Storage	Chrome Local Storage
Testing Platform	Chrome (Social Media Web Interface)
IDE	VS Code

6. RESULTS AND ANALYSIS

6.1 Output Results

The Guardian Chrome Extension successfully integrates with the Instagram web interface and classifies Direct Messages (DMs) in real time.

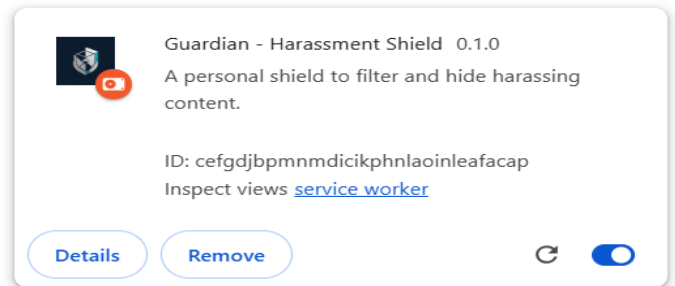


Figure 6.1: Guardian: Chrome Extension.

Observed outcomes:

- Safe messages remain unaltered.
- Creepy or flirty messages are highlighted in yellow.
- Harassment or toxic messages are highlighted in red.
- Repeated offenders are automatically hidden, and the user is alerted through a pop up notification.

```

    Classifying: You re so stupid background.js:46
    ML Response: background.js:13
    {text: 'You re so stupid', prediction: 'harassment_med'}
    Prediction: harassment_med background.js:49
    BG received: background.js:24
    {type: 'NEW_DM_MESSAGE', text: 'Can i drop u home?'}
    Classifying: Can i drop u home? background.js:46
    ML Response: background.js:13
    {text: 'Can i drop u home?', prediction: 'creepy_flirt'}
  
```

Figure 6.2 (a): Classification of messages received.

```

    BG received: {type: 'NEW_DM_MESSAGE', text: 'U r a bitch'} background.js:24
    Classifying: U r a bitch background.js:46
    ML Response: background.js:13
    {text: 'U r a bitch', prediction: 'harassment_severe'}
    Prediction: harassment_severe background.js:49
    BG received: {type: 'NEW_DM_MESSAGE', text: 'F**k u'} background.js:24
    Classifying: F**k u background.js:46
    ML Response: background.js:13
    {text: 'F**k u', prediction: 'harassment_severe'}
  
```

Figure 6.2 (b): Classification of messages.

- When some harassing text/message is noticed, after 3 strikes, the warning message is popped up.

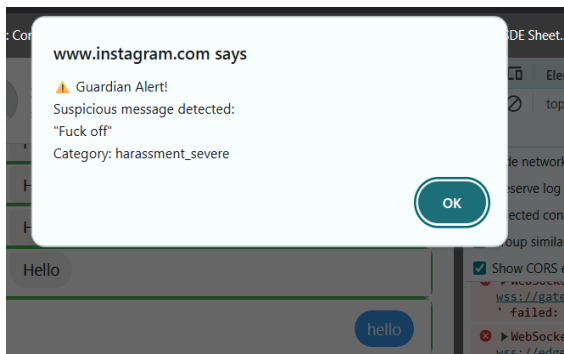


Figure 6.3 Pop-up warning message

- After the popping up of the warning message that particular harassing message is highlighted with different colours according to the severity of the text.

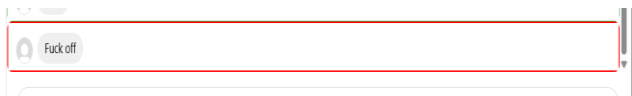


Figure 6.4: Message highlighted according to severity.

6.2 Classification model evaluation

6.3

The overall performance of the model on the test data set, consisting of 1963 samples, stands at 0.78 or 78%. The succeeding analysis provides a snapshot of the performance across all six classes of output, using the standard classification metrics of Precision, Recall, and F1-Score.

It has shown very good performance in classifying the most distinctive and critical categories: **Stalker Behaviour**: The model performs highly reliably in the identification of stalker behaviour, with an extremely high F1-score of 0.96, precision 0.97, and recall 0.95, hence robust, which indicates very effective feature engineering for this class.

The major **"non-harmful"** class performs very well and includes an F1-score of 0.90. Besides that, precision and recall equate at 0.90, which implies that the model captures safe interactions quite well and the misclassification of other types as safe is rare.

Another critical category that does well would include **severe harassment**, with an F1 of 0.76 (Precision: 0.79, Recall: 0.73). Precision is high here; if it identifies something as severe harassment, it will be correct.

	precision	recall	f1-score	support
creepy_flirt	0.79	0.83	0.81	36
harassment_med	0.69	0.69	0.69	35
harassment_mild	0.80	0.77	0.79	31
harassment_severe	0.80	0.77	0.78	26
safe	0.90	0.90	0.90	31
stalker_behavior	0.97	0.97	0.97	37
accuracy			0.83	196
macro avg	0.83	0.82	0.82	196
weighted avg	0.83	0.83	0.83	196

Figure 6.5 Classification report of the model

7. CONCLUSION

7.1 Summary of Findings

Guardian, the project revealed some necessary information about the effect of online harassment on women and why options have not yet met the bar. As the study progressed, one fact became clear: women are repeatedly harassed on social media platforms. These types of attacks are not necessarily loud or obvious. But they're little constant nudges that can make a person feel uneasy or unsafe, or chipped away at over time. The majority of the well-established methods of defense only act after damage has been inflicted. This emphasized the need for an exposure-reducing, pre-damage intervention.

Testing Guardian on Instagram in real time proved that proactive protection is more than a dream it really is achievable. The program went beyond picking out blatantly abusive messages to discovering unsettling trends that often remain hidden in plain sight. That included what team members described as unwanted flirting that becomes more aggressive, manipulative language, threats and other repeated creepy conduct. These exchanges can make women feel trapped, even when no single message is so awful that it should be flagged by established moderation. It demonstrated how contextual AI can solve this problem, rather than just blocking keywords.

Another realization we had was that users care a lot about privacy. For even the most privacy conscious among us, there are lingering doubts about not only where your personal messages go during a sensitive exchange, but also who gets to look at them, who

stores them and who owns that information. Guardian was built with these issues in focus. Because all processing is done on the user's device, no data is transmitted. Nothing is saved, shared or tracked by outside sources. This made it possible for users to feel safe, but not surveilled. That's not something existing social media tools have. While testing, we realized that control mattered to feeling safe online. Elements such as the strike system, gentle warnings, and the capacity to bypass a moderation gave users the impression that they were in control.

Rather than restricting how they communicate, Guardian gave users the ability to define their own limits. It was not about silencing other people; it was about arming users with the means to choose how and when they wanted to protect themselves from harm. The Guardian programme demonstrates a safer web is achievable. With good design and people-centred technology, social media can be a less frightening place. It can be restored to what it was always intended to be a place for connection, for creativity, and for meaningful expression where everyone, and especially women, can feel protected and respected.

7.2 How it meets the objective

The objective There was one the project was easy to identify at the beginning. The purpose of it was to build a tool that empowers women to feel safer and more at ease sharing their lives on social media, particularly Instagram. Looking back on the objectives set out at the start, Guardian is delivering in spades. The second big aim was to identify abusive, disturbing, or harassing messages as they come in, not after users have already seen them. By coupling an adapted DistilBERT model with real-time monitoring through a Chrome extension, the much lauded Guardian emerged out of the shadow. It scours the content of messages in a matter of seconds and indicates the potential level of risk before a user receives the message. The more than the usual blocked message checks was another important aim. Guardian labels messages into a range of categories, from safe to extreme harassment. It's a bit more nuanced of an approach for users to understand the type of experiences they're actually having rather than seeing everything under one umbrella. Privacy was a concern from the beginning, and this requirement is adequately fulfilled. Unlike conventional systems, Guardian does not transmit or leak the content of messages to any server. All the processes, such as text processing, classification, and decision making, are performed locally on the user's device. This keeps everything under lock and key.

User control and autonomy was also important. It is simple for users to toggle Guardian on or off, and to re-examine or unhide any messages that were

flagged, on their terms. It's all with their permission, and their knowledge.

The two strikes policy was a success in dealing with long-term harassment. automatically hide that user's future messages from you as well and send you an alert. This eliminates the burden on users having to keep an eye on or report offenders who repeatedly break the rules.

Now emotional well-being was what motivated this entire project. Guardian enhances this by limiting the amount of traumatic content being viewed and engaging in discussions in a more secure atmosphere. As Guardian runs silently in the background, women will be able to enjoy handling their Instagram accounts with greater peace of mind, more confidence and freedom.

8. FUTURE SCOPE

8.1 What Improvements can be made

Guardian already provides robust protection from online abuse, but a few tweaks could make the system more practical and user-friendly in the future.

One important factor that needs to be improved is the language support. At the moment the model performs best in English, but harassment is increasingly expressed in regional languages, slang, and mixed language communication, particularly on social media. With multilingual training data, Guardian would be able to help a wider range of users belonging to a variety of communities and cultures.

Further enhancement in terms of classification accuracy to consider emotional tone, context and prior conversation is also possible. This would help it identify even more subtle forms of harassment, sarcasm, threats dressed up as compliments, and patterns of abuse that might repeat over time.

Further the UI can also be tweaked to allow more customization. For example, users could define their own sensitivity levels or choose content categories that are most relevant to them. A dashboard for flagged messages, strike counts, and prevalent harassment types could potentially increase user awareness and control.

Increased support for platforms is another useful direction. Guardian is for now only built for Instagram web messaging, but abuse takes place across many digital platforms. More such developments could also bring the real time protection support to WhatsApp Web, Facebook Messenger, X and even email providers.

Safety escalation features could be added as well. For example, alerting users ahead of time that they are responding to a known harasser, or displaying quick links to support services, mental health resources, or platform reporting tools in the event that the harassment intensifies.

Users who want to contribute anonymized data to improve the AI model could opt in to a secure cloud mode. This would enable classification performance to increase as a function of time and at the same time keep user privacy intact.

8.2 Future Practical Relevance

Guardian will be more necessary as digital communication changes. But online harassment is increasingly a rite of passage for many women. Traditional reporting systems are typically slow, stressful and reactive. Only when users are harmed can anything be done. Guardian overcomes that by taking action ahead of damage. Prevention is Now a Big Part of Online Safety.

Going forward, Guardian could be a game changer for the way social platforms treat user well-being. Features such as real-time message protection, AI-based severity detection, and long-term tracking of harasser behaviour point to a model of social media that could be taken up globally by companies. It could help change policy and pave the way for new standards in responsible platform design.

In addition to individuals, Guardian could also serve communities, schools and organizations. Institutions interested in providing a safer online space for their members particularly younger women or individuals new to the world of social media could incorporate this instrument into their digital safety curriculums. This could enhance education and accountability in respectful communication.

Kinetics also saw potential for collaboration with law enforcement and cyber safety officials. While Guardian keeps everything private by default, users could choose to share records of severe harassment in secure formats to help report and stop repeat offenders. This could make legal action more accessible for those who need it.

REFERENCES

- 1 Rajan, Benson. "Harassment and abuse of Indian women on dating apps: a narrative review of literature on technology-facilitated violence against women and dating app use." *Humanities and Social Sciences Communications* 12.1 (2025): 55.
- 2 Balabantaray, Subhra Rajat, Mausumi Mishra, and Upananda Pani. "a sociological study of cybercrimes against women in india: deciphering the causes and evaluating the impact on the victims." *international journal of asia-pacific studies* 19.1 (2023).
- 3 Alauddin Middy, Bimal Mandal (2021). *Cyberviolence Against Women: Exploring Patterns of online Gender Based Harassment* (2021)
- 4 Bhat, Rashid Manzoor, and Peer Amir Ahmad. "Social Media and the Cyber Crimes Against Women-A Study." *Journal of Image Processing and Intelligent Remote Sensing (JIPIRS) ISSN* (2022): 2815-0953.
- 5 Negi Advocate, Dr Chitranjali. "An overview of worldwide cyberbullying and cyberviolence against women, teenagers, LGBTQ on social media: Facebook, Instagram, Telegram, WhatsApp, Snapchat, YouTube, LinkedIn and Twitter." *Boston College International and Comparative Law Review, Forthcoming* (2023).
- 6 Galdi, Silvia, and Francesca Guizzo. "Media-induced sexual harassment: The routes from sexually objectifying media to sexual harassment." *Sex Roles* 84.11 (2021): 645-669.
- 7 Burke Winkelmann, Sloane, et al. "Exploring cyber harassment among women who use social media." *Universal journal of public health* 3.5 (2015): 194.
- 8 Qian Zhang, Qinxuan Chen, Xinrong Gu (2024). *Trolling, Cyberstalking, Body-shaming, Slut-shaming – A Study on Online Abuses of Social Media* (Zhang, 10)
- 9 Mochamad Iqbal Jatmiko, Muh. Syukron, Yesi Mekarsari (2020). *Covid-19, Harassment and Social Media: A Study of Gender - Based Violence Facilitated by Technology During the Pandemic* (Jatmiko, 2024)
- 10 Kejsi Take, Victoria Zhong, Chris Geeng, Emmi Bevenssee, Damon McCoy, Rachel Greenstadt (2024). *Stoking the Flames: Understanding Escalation in an Online Harassment Community* (Take 2024, 23)
- 11 Abarna, Sheeba, et al. "Identification of cyber harassment and intention of target users of socialmedia platforms." *Engineering applications of artificial intelligence* 115 (2022): 105283.
- 12 Barlett, Christopher P., et al. "Social media use and cyberbullying perpetration: A longitudinal analysis." *Violence and gender* 5.3 (2018): 191-197.
- 13 Talpur, Bandeh Ali, and Declan O'Sullivan. "Cyberbullying severity detection: A machine learning approach." *PloS one* 15.10 (2020): e0240924.
- 14 Ashraf, Noman, Arkaitz Zubiaga, and Alexander Gelbukh. "Abusive language detection in youtube comments leveraging replies as conversational context." *PeerJ Computer Science* 7 (2021): e742.
- 15 Herry, Emily, and Kelly Lynn Mulvey. "Gender-base cyberbullying: Understanding expected bystander behavior online." *Journal of Social Issues* 79.4 (2023): 1210-1230.

- 16 Sahu, Tamanna, and Ritu Raj. "The role of social media in shaping Women's paranoia about Harassment and Safety." *International Journal of Interdisciplinary Approaches in Psychology* 3.5 (2025): 1333-1344.
- 17 Nayayani, A. (2024). Women's Safety in Digital Space. *Indian Journal of Public Administration*, 70(3), 546-561
- 18 Chadha, Kalyani, et al. "Women's responses to online harassment." *International journal of communication* 14 (2020): 19-19.
- 19 Nadim, Marjan, and Audun Fladmoe. "Silencing women? Gender and online harassment." *Social Science Computer Review* 39.2 (2021): 245-258.
- 20 Ghowrui, Ahana, et al. "Effects of cyberbullying on women's mental health." *IJPR* 6.1 (2024): 25-29.
- 21 Lal, Disha, Udaya Kumar Giri, and Shrish Kumar Tiwari. "Virtual Vulnerability: Addressing Cyber Harassment against Women in India." *DS Journal of Cyber Security* 2.3 (2024): 1-14.
- 22 Dar, Showkat Ahmad, and Dolly Nagrath. "Are Women a Soft Target for Cyber Crime in India." *Journal of Information Technology and Computing* 3.1 (2022): 23-31.
- 23 B. Vijayalaxmi. (2020), *Cybercrime Against Women in India: A Critical Analysis*. (B, 2020)
- 24 Neha Gupta, Soyonika Gogoi (2025). *Cyberbullying And Gender: An Analysis Of Legal Protections For Women On Social Media* (Gupta, 2025)
- 25 Mythili, K. C., and K. Nagamani. "Safeguarding women in digital spaces: Legal responses to cyber harassment and objectification social media." *Development Policy Review* 43.5 (2025): e70039.
- 26 Nusrat Ali Rizvi (2025). *The Role of Law in Combatting Gender-Based Violence on Social Media and Online Platforms* (Rizvi, 2025)
- 27 Lazarus, Suleman, Mark Button, and Richard Kapend. "Exploring the value of feminist theory in understanding digital crimes: Gender and cybercrime types." *The Howard Journal of Crime and Justice* 61.3 (2022): 381-398.
- 28 Gabriel Grill, Jane IM, Sarita Schoenebeck, Marilyn Iriarte (2023). *Women's Perspectives on Harm and Justice after Online Harassment* (Grill, 2023)
- 29 Orusa Karim, Sumbl Ahmad Khanday (2026). *Cybercrime and Women-Online Harassment, Stalking, And Blackmail; A Sociological Analysis* (Karim, 2026)
- 30 Gallegos, Ada, et al. "Cyberbullying against women in digital environments examining manifestations, risk factors, and institutional responses." *Discover Psychology* (2026).