

HAVEN-A Behavioral-Aware AI Framework for Digital Harassment Detection

Dyuthi R¹, Ida Jane M², Hemalatha K³, R.A. Mabel Rose⁴

^{1,2,3}Student, Computer Science And Business Systems, Panimalar Engineering College

⁴Assistant Professor, Computer Science And Business Systems, Panimalar Engineering College

Abstract:

In today's modern technology evolving area, there is an increase in the use of social media, message applications, and location-based technologies. These mentioned technologies create new opportunities for global communication. But at the same time, the mentioned technologies also pave way for online harassment, cyberstalking, and intimate people via online. There were several measures that were taken to prevent these crimes, namely keyword filtering, content moderation, blocking, and user reporting. The mentioned mechanisms are efficient in identifying individual harmful messages, but they cannot be used to represent the progression of harassment across multiple interactions. Therefore, this paper proposes HAVEN [Harassment Awareness and Virtual Early-warning Network], an AI-based framework designed to identify potential escalation in online harassment by combining the message analysis with behavioral pattern analysis. This framework incorporates Natural-Language Processing [NLP], toxicity detection, threat detection, sentiment analysis, and many more to generate an interpretable risk assessment that supports user awareness along with evidence preservation. We have considered essential design requirements such as privacy protection, uncertainty management, and human oversight to make this framework more efficient. As HAVEN is currently a proposed system, the paper presents its conceptual architecture, intended workflow, and a future evaluation methodology. It aims to provide a foundation for developing proactive and privacy-conscious digital tools.

Keywords: Cyberstalking, Cybersecurity, Digital Harassment, Natural-Language Processing, NLP, Behavioural Analysis, Risk Assessment, Early Warning Systems

1 INTRODUCTION

In today's modern life, digital communication has become an integral part where social networking platforms, instant messaging applications, online communities, and location-based services are used to enable communication and share information comfortably. However, they are misused to facilitate cyberstalking, repeated unwanted communication, online harassment, threats, impersonation, and mainly image-based abuse. Technology-facilitated violence is not necessarily restricted to a single platform or interaction, but they might get followed by contact through alternative accounts, intimidating language, attempts to obtain private information or efforts to determine a person's location. At the same time, the significance of an interaction may depend on its relationship with previous interactions rather than on its content alone. UN Women has highlighted the diverse forms of technology-facilitated violence and the importance of improving the evidence and responses used to address it [UN Women 2023].

The safety mechanisms which are in use provide important protections. Keyword filters can identify predefined expressions; content moderation can flag or remove harmful material. Blocking and reporting mechanisms allow users to seek assistance. Cybersecurity research has explored the use of textual analysis and machine learning to identify behaviors associated with online stalking and online harassment [Frommholz et al., 2016].

Nevertheless, identifying potentially harmful communication remains a complex task. A threatening message may contain explicit language while persistent harassment may involve seemingly ordinary messages, and they become concerning due to their frequency, context, or relationship to earlier interactions. Consequently, evaluating individual messages may not adequately represent the broader behavioral pattern. Research and human rights reporting also demonstrate the technology-facilitated abuse that extends beyond hostile text, and they include stalking, impersonation, doxxing, manipulated images, and the non-consensual distribution of intimate material.

Human Rights Watch [2021] documented the consequences of digital sex crimes in South Korea, including continuing harm to survivors and difficulties in obtaining effective remedies. Similarly, Equality Now and Breakthrough [2025] documented survivor experiences of technology-facilitated gender-based violence in India, including several forms of digital abuse and barriers to effective support. These concerns motivate the need for digital safety approaches that consider both the content of communications and the behavioral context in which they may occur. AI may insist on organizing large volumes of information and identifying patterns that deserve further review. However, automated analysis must not be treated as definitive proof of criminal intent or as a replacement of survival judgement, professional investigation, or appropriate human intervention. To address this need, we propose HAVEN [Harassment Awareness and Virtual Early-warning Network].

The main contributions of this paper are,

- A conceptual architecture for combining message-level analysis with behavioral indicators across interactions.
- A proposed risk assessment approach that considers content, behavioral, contextual information, and quality of evidence.
- A structured workflow for organizing relevant digital evidence for review.
- A future evaluation plan addressing detection of performance, fairness, privacy, and responsible use.

The remainder of the paper reviews relevant research, describes the proposed framework, discusses real-world contexts that motivate its development, and outlines expected outcome and further evaluation.

2 RELATED WORKS

2.1 Cyberstalking and textual analysis

Cyberstalking involves the use of digital technologies to engage in persistent unwanted contact, monitoring, intimidation, or similar behaviour. Spitzberg and Hoobler [2002] examined cyberstalking related to interpersonal communication and the misuse of communication technologies. Their work provides a conceptual basis for understanding how digital tools can facilitate persistent unwanted behaviour. Frommholz et al, (2016) investigated textual analysis and machine learning in context of cybersecurity attacks, i.e. cyberstalking detection. Their work is relevant to the proposed use of computational methods for examining communication via digital. It also highlights the relationship between identifying potentially harmful behaviour and gathering digital evidence. These research directions motivate the content-analysis component of HAVEN. However, HAVEN proposes an additional emphasis on behavioural patterns across available directions. Its purpose is to explore whether combining content-level indicators with repeated –contact and escalated signals could support more contextual assessment.

2.2 Automated identification of abusive content

Datasets and machine learning approaches are designed to identify offensive language, personal attacks, and abusive behaviour.

Davidson et al. (2017) examined automated hate-speech detection and the problem of distinguishing hate speech from offensive language. This distinction is important because offensive wording doesn't always mean targeted harassment, and harmful behaviour may not always contain explicit insults. Founta et al. (2018) investigated large-scale crowdsourcing and the characterization of abusive behaviour on Twitter. Wulczyn et. Al (2017) studied automated identification of personal attacks in online discussions. Together these works provide correct background on data annotation, classification of harmful content, and the challenges associated with automated detection. For HAVEN, these studies motivate the proposed use of natural language processing, toxicity-related classification, and threat related signals.

2.3 Technology-facilitated violence and behavioral context

Technology-featured violence involves multiple digital channels and forms of abuse. UN Women (2023) reviewed evidence and collection of datum related to technology-facilitated violence against women. This report highlights the importance of considering how such violence is documented and understood, rather than treating digital harm as a narrow-text classification problem. The report by Equality and Now (2025) presents survivor narratives and legal responses relating to technology-featured gender-based violence in India. The reported experiences involve stalking, doxxing, image-based abuse, and the non-consensual distribution of intimate images. This context is relevant to design safety systems that accounts that account for variety of ways in which digital technologies may be misused. These sources motivate HAVEN's interest in combining multiple indicators and organizing relevant evidence. However, they do not establish that the proposed framework can detect every form of technology-facilitated abuse. The system's proposed capabilities would need to be accessed against clearly defined tasks and suitable data.

2.4 Research Gap and Motivation

The literature above describes and provides foundations for textual cyberstalking detection, abusive-content classification, and understanding technology-featured violence. However, a proposed early work must also address the relationship between the individual messages, repeated behaviour, changes over time, and the quality of available evidence. HAVEN is proposed as a conceptual framework for exploring this combination. The design objective is to consider content-level signals together with

behavioural patterns and evidence organization, while preserving human oversight. This is a design motivation rather than a claim that existing systems universally lack these capabilities. Establishing the framework's practical value would require a systematic review of compatible systems and empirical evaluation.

3 Proposed HAVEN Framework

3.1 System Overview

HAVEN is proposed as an AI-assisted digital safety framework that analyses available digital interactions to identify patterns potentially associated with escalating harassment. The intended workflow consists of data intake, preprocessing, content analysis, behaviour pattern analysis, risk assessment and evidence organization. The framework is intended to support a user or authorized potentially relevant information. It is not designed to automatically declare a person guilty, establish criminal intent, or guarantee that a future incident will occur. The proposed system can be represented through the following stages:



Figure 1: Conceptual architecture of the HAVEN framework.

The sequence represents a conceptual architecture. The specified data sources, integrations, and technical components would need to be finalized during implementation.

3.2 System Architecture

The architecture consists of six principal modules.

- 1. Input and Data Management Modules:** Receives digital interaction records that the user is authorized to provide, or that are otherwise lawfully available. Depending on the eventual implementation, inputs could include text messages, timestamps, interaction identifiers, and relevant user reports. Access to private accounts or communications must not be assumed without explicit authorization.
- 2. Preprocessing Module:** Cleans and organizes the available data. Possible operations include removing duplicate records, normalizing text, identifying language, preserving timestamps, and

structuring interactions into a consistent format. Original records should be retained separately when needed for evidence of integrity.

3. **AI-based content-analysis module:** Examines message content for signals such as abusive language, explicit threats, coercive wording and potentially harmful sentiment. These outputs are indicators for further assessment, not independent proof of harassment.
4. **Behavioural-pattern analysis module:** Examines temporal and interaction-related features such as repeated unwanted contact, changes in message frequency, persistence after a request to stop, or recurrence across available communication records.
5. **Risk-assessment module:** Combines relevant content and behavioural indicators with contextual information and evidence-quality considerations to produce an interpretable assessment for review.
6. **Alert and evidence-organization module:** Presents relevant interaction records, timestamps, and detected indicators, and the reasons an item was flagged. The intended purpose is to help a reviewer understand the basis of an alert and locate supporting information.

3.3 Input data and processing

The quality of automated analysis depends on the relevance and reliability of its input data. For this reason, HAVEN proposes a preprocessing stage before classification and behavioral analysis. The input representation may lawfully include message text, timestamp, sender, and recipient identifiers represented using privacy-preserving identifiers, user reports, and interaction history. The framework collects only information which should be necessary for its stated purpose. Preprocessing may include text normalization, language identification, duplicate detection, timestamp standardization, and the organization of interactions into chronological sequences. Multilingual communication and regional expressions require particular attention because a model trained on one language or cultural context may misinterpret another. The framework should also distinguish missing data from evidence that a behaviour did not occur. For example, the absence of messages from one platform does not establish that no contact occurred elsewhere. HAVEN's assessment should therefore reflect the limits of the available data.

3.4 AI-based content analysis

The content-analysis module is proposed to examine individual messages for several types of signals.

- Toxicity and abusive language detection
- Threat-related detection
- Sentiment-related analysis
- Contextual interpretation

The proposed use of these signals is informed by previous work on abusive-content classification and automated identification of personal attacks (Davidson et al., 2017; Founta et al., 2018; Wulczyn et al., 2017).

A central design consideration is the distinction between a model-generated signal and a verified conclusion. Sarcasm, quotations, reclaimed language, discussion of abuse, and legitimate disagreement may produce misleading classifications. Similarly, a message that contains no explicit abusive terms may still be concerning in its wider context. The framework therefore proposes passing content-analysis outputs to the risk-assessment stage rather than using any single classification as a final decision.

3.5 Behavioral analysis pattern

Behavioural pattern analysis is the central proposed feature of HAVEN. It aims to examine how available interactions relate to one another over time.

Potential indicators include:

- Repeated messages or contact attempts directed at the same individual.
- Continued contact after a clear request to stop, where that request is available in the records.
- Increasing frequency of contact over a defined period.
- Repeated threatening or coercive content across multiple interactions
- Recurrence of related behaviour across available channels, where the relationship between records can be established reliably.

These indicators are proposed features, not a validated list of predictors. Their definitions and usefulness must be tested before they are used to generate operational warnings. Context is essential. Frequent communication between consenting participants should not automatically be interpreted as harassment. The system should account for available consent-related information, legitimate reasons for communication, missing records, and the possibility of incorrect identity matching. Behavioural indicators should therefore be interpreted in combination and presented with appropriate uncertainty.

3.6 Risk Assessment and Early Warning

HAVEN proposes a risk-assessment stage that combines information from the preceding modules. A conceptual representation is:

$$R_t = f(C_t, B_t, E_t, Q_t)$$

where,

- R_t represents the estimated risk at time (t).
- C_t represents content-related indicators.
- B_t represents behavioral patterns.
- E_t represents relevant contextual information.
- Q_t represents data quality and uncertainty considerations.
- $f()$ represents the function combining these inputs.

This expression describes the proposed structure of the assessment; it is **not a trained model, validated mathematical formula, or measured risk score**. The function could eventually be implemented using a rule-based method, a trained model, or a carefully evaluated combination of both. The purpose of the risk-assessment stage is to support prioritization for human review. An alert could identify the factors that contributed to it, such as repeated contact, a threat-related classification, or a change in interaction frequency. The system should also indicate uncertainty and avoid presenting a risk category as a prediction with established accuracy until that claim is supported by evaluation.

3.7 Evidence Organization

A digital safety tool may be more useful when it helps a user or authorized reviewer locate relevant records instead of producing only a classification label. Frommholz et al. (2016) discuss the importance of digital evidence in the context of cyberstalking detection. HAVEN therefore proposes an evidence-organization module that can group potentially relevant interactions chronologically and associate each flagged item with the reason it was identified. Depending on the eventual implementation, an evidence record may contain a timestamp, source identifier, original content where appropriate, the relevant model-generated indicator, and a record of any subsequent review. Evidence handling must preserve privacy and avoid altering original records without documentation. The framework should not claim that

its summaries are legally admissible or sufficient to establish an offence. Evidence requirements vary by jurisdiction and context, and any use in a formal investigation would require appropriate procedures and professional review.

4 REAL WORLD CONTEXT AND CASE STUDIES

4.1 The Nth room case in South Korea

The Nth Room case in South Korea became a widely reported example of technology-facilitated sexual exploitation involving digital communication channels. Human Rights Watch (2021) documented digital sex crimes in South Korea, including non-consensual intimate imagery, the use of digital technologies to facilitate abuse, and the substantial difficulties survivors may face in obtaining effective remedies. The South Korean government also announced stronger measures against digital sex crimes in 2020 (Ministry of Gender Equality and Family, Republic of Korea, 2020).

This context illustrates that digital abuse can involve more than a single offensive message. Harm may be connected to repeated actions, coercion, the distribution of private material, and the use of digital platforms to reach or exploit victims. Such cases also demonstrate that technical detection is only one part of a broader response involving survivor support, investigation, platform responsibilities, and legal protections.

4.2 Lessons for the proposed framework

The Nth Room case is included as contextual motivation, not as a test of HAVEN. It does not establish that HAVEN would have identified the perpetrators, prevented the abuse, or reduced the harm. For the proposed framework, the broader lesson is the importance of considering the context and progression of harmful activity, where reliable and authorized data is available.

Content analysis may help identify certain threatening or coercive communications, while behavioral analysis may help organize repeated interactions for review. However, not every form of digital exploitation is visible in text messages, and HAVEN cannot be assumed to detect activity occurring in inaccessible or encrypted spaces.

4.3 Technology-facilitated gender-based violence in India

The report, *Experiencing Technology-Facilitated Gender-Based Violence in India: Survivor Narratives and Legal Responses*, which was published by Equality Now and Breakthrough (2025), documents survivor experiences involving forms of digital abuse such as stalking, doxxing, image-based abuse, and non-consensual distribution of intimate images. It also discusses challenges involving legal responses and access to support.

This report provides an Indian context for understanding the range of harms that a digital safety framework may need to consider. It is a survivor-experience and legal-response report, not an evaluation of HAVEN or evidence that a particular AI architecture would successfully detect these incidents.

4.4 Implications for HAVEN

The reports discussed in this section motivate three design considerations. First, digital abuse can take multiple forms, so text classification alone cannot cover every relevant risk. Second, records may be incomplete or distributed across different services, making reliable interpretation difficult. Third, technical warnings must be connected to appropriate human review and support rather than treated as a complete response to abuse.

Accordingly, HAVEN is proposed as a supporting framework for organizing available information and identifying potentially concerning patterns. Its practical usefulness would depend on the types of data it

can lawfully access, the reliability of its models, the needs of intended users, and the safeguards established for handling sensitive information.

5 EXPECTED OUTCOMES AND DISCUSSION

5.1 Expected Outcomes

Because HAVEN is proposed rather than evaluated, the following are anticipated outcomes that require future validation.

- **Context-aware identification:** Combining content indicators with behavioural patterns may provide a broader basis for review than examining isolated messages alone.
- **Prioritization of interactions:** An interpretable assessment may help users or authorized reviewers locate potentially urgent records within a large collection of communications.
- **Structured evidence review:** Chronological organization and explanations for flagged items may make it easier to examine the records associated with a warning.
- **Support for Early Intervention:** If future testing demonstrates that the framework identifies useful indicators reliably and in time, it may support earlier consideration of appropriate safety measures. This remains a possibility, not an established outcome.

5.2 Potential Advantages

The principal design advantage of HAVEN is its proposed combination of content-level and behavioral signals. This may help represent repeated patterns that an isolated message classifier cannot describe on its own. The framework also proposes separating model outputs from final decisions and providing information about why an interaction was flagged. However, additional components do not automatically guarantee better detection. A behavioural model could introduce new errors, while combining multiple uncertain signals could amplify bias. Any advantage claimed must therefore be established through comparative evaluation against suitable baselines.

5.3 Challenges and Limitations

Several challenges must be addressed before HAVEN can be implemented or deployed.

- **False Positives and False Negatives:** Legitimate communication may be incorrectly flagged, while subtle or context-dependent abuse may be missed.
- **Language and Cultural Variation:** Slang, code-switching, sarcasm, regional expressions, and differences between languages may reduce classification reliability.
- **Incomplete Information:** A system may not have access to all interactions relevant to an incident. Missing records can distort the apparent pattern.
- **Data Availability and Annotation:** Suitable datasets may be limited, inconsistently labelled, or unrepresentative of the intended users and languages.
- **Identity and Context Errors:** Incorrectly linking accounts or misinterpreting relationships could produce harmful conclusions.
- **Operational Constraints:** Real-time analysis, secure data storage, model maintenance, and integration with existing platforms may require substantial engineering and governance work.

These limitations mean that HAVEN should initially be treated as a research proposal requiring controlled development and validation.

5.4 Privacy and Ethical Considerations

HAVEN concerns highly sensitive communications and potentially vulnerable users. Privacy and

responsible use must therefore be treated as core design requirements rather than later additions. The NIST AI Risk Management Framework provides general guidance for identifying and managing AI-related risks and supporting trustworthy AI practices (Tabassi, 2023).

The proposed framework should follow data minimization, purpose limitation, access control, secure storage, and clearly defined retention policies. Data collection should be restricted to information that is authorized and necessary for the stated purpose. Users should be informed about how data is analysed and how alerts may be used.

The system should not automatically label an individual as a criminal, initiate punitive action solely from a model output, or make high-impact decisions without appropriate review. Where feasible, alerts should provide understandable reasons, communicate uncertainty, and allow correction or challenge. Evaluation should also examine differences in error rates across languages and relevant user groups.

Finally, an early-warning system must consider the safety of the person receiving an alert. An alert or notification could create additional risk in some circumstances, particularly where an abuser has access to a victim's device or accounts. The interface and deployment process must therefore be designed around user control, safe access, and context-sensitive support.

6 Future Evaluation Plan

6.1 Dataset Preparation

The first stage of future work is to define the detection task precisely and prepare suitable evaluation data. Potential data sources could include appropriately licensed public datasets, ethically collected research data, or consent-based examples. Private conversations should not be collected or processed without a valid legal and ethical basis. The dataset should include clearly defined labels and annotation guidelines. Where behavioural analysis is evaluated, records should preserve sufficient temporal information to assess sequences rather than only individual messages.

The evaluation should distinguish between message-level classification, interaction-level assessment, and sequence-level identification of potential escalation. Training, validation, and test data should be separated carefully to reduce data leakage. For example, messages from the same conversation or user should not be distributed across splits in a way that allows the model to exploit near-duplicate or identity-specific information.

6.2 Dataset Preparation

If a suitable dataset and implementation become available, the framework could be evaluated using precision, recall, F1-score, and a confusion matrix. Precision would measure the proportion of flagged cases that are positive according to the evaluation labels, while recall would measure the proportion of labelled positive cases identified by the system. F1-score would summarize the balance between precision and recall.

Because an early-warning system may be intended to identify serious cases, evaluation should also consider the consequences of false negatives and false positives. Depending on the task, additional measures could include calibration, detection delay, performance across languages, and the proportion of alerts that human reviewers judge useful. No numerical target or achieved score is asserted in this proposal. Appropriate thresholds would need to be selected based on the use case, the quality of available data, and the consequences of errors.

6.3 Comparative Evaluation

A future study could compare at least three approaches:

- A content-only baseline that classifies individual messages.
- A behavioural-only baseline that uses defined interaction patterns.
- A combined approach that uses content and behavioural signals.

This comparison would help establish whether the combined approach contributes measurable value over simpler alternatives. Ablation studies could remove individual components to examine their contribution. Human review could also be used to assess the interpretability and practical usefulness of the resulting alerts.

The study should report dataset composition, annotation procedures, evaluation settings, uncertainty, and limitations. Performance on on dataset should not be presented as proof of effectiveness across all platforms, languages, or forms of harassment.

6.4 Future Development

Future development may involve implementing the individual modules, building a secure prototype, and evaluating the workflow with appropriate oversight. Further work should investigate privacy-preserving processing, transparent explanations, multilingual support, fairness testing, and safe alert delivery.

Before deployment, the framework would require security testing, user-centred evaluation, and review of applicable legal and ethical obligations. Any proposed integration with third-party platforms would also depend on platform permissions, technical access, and relevant policies.

7 CONCLUSIONS

This paper has proposed HAVEN, an AI-based framework intended to support the early identification of potentially escalating digital harassment through content analysis, behavioural pattern analysis, risk assessment, and evidence organization. The proposal is motivated by research on cyberstalking detection, automated abusive-content classification, and reports documenting technology-facilitated abuse.

The central idea is to consider available interactions in context rather than relying exclusively on isolated message classifications. A conceptual risk-assessment function and modular system architecture have been presented to explain how the proposed components could work together. The paper has also discussed anticipated benefits, technical limitations, privacy safeguards, and a future evaluation plan.

HAVEN has not been presented as an implemented or experimentally validated system. Its effectiveness remains to be established through careful dataset preparation, comparison with baseline methods, fairness assessment, and human-centred evaluation. Ultimately, the framework should be developed as a support mechanism that assists appropriate review while respecting privacy, preserving user autonomy, and avoiding unsupported conclusions about individuals.

REFERENCES

1. Davidson, T., Warmley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), 512–515.
2. Equality Now and Breakthrough (2025). *Experiencing Technology-Facilitated Gender-Based Violence in India: Survivor Narratives and Legal Responses*. Published 4 November 2025.
3. Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of Twitter

- abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 491–500.
4. Frommholz, I., al-Khateeb, H. M., Potthast, M., Ghasem, Z., Shukla, M., and Short, E. (2016). On textual analysis and machine learning for cyberstalking detection. *Datenbank-Spektrum*, 16, 127–135.
 5. Human Rights Watch (2021). *“My Life Is Not Your Porn”: Digital Sex Crimes in South Korea*.
 6. Ministry of Gender Equality and Family, Republic of Korea (2020). *Harsher Punishment and Stronger Protection against Digital Sex Crimes*. 6 May 2020.
 7. Spitzberg, B. H., and Hoobler, G. (2002). Cyberstalking and the technologies of interpersonal terrorism. *New Media & Society*, 4(1), 71–92.
 8. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, NIST AI 100-1.
 9. UN Women (2023). *Technology-Facilitated Violence Against Women: Taking Stock of Evidence and Data Collection*.
 10. Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex Machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.