When Words Become Warnings.
Assessing Threatening Communication in Online Spaces.

Lukas Lundmark[1], Lisa Kaati[1], James Silver[2], and Amendra Shrestha[1]

[1] Stockholm University, Stockholm, Sweden
firstname.lastname@dsv.su.se
[2] Boston University, Boston, USA
silverjm@bu.edu

Abstract. This paper presents a scalable framework for assessing threatening online communication using eight theoretically grounded warning indicators. These indicators are extracted through a hybrid approach that integrates dictionary-based methods, machine learning classifiers, and large language models. Our findings show a statistically significant correlation between the presence of these indicators and expert assessments of individuals deemed to pose a risk of targeted violence. This underscores the potential of the proposed indicators to support reliable and efficient threat assessment in digital environments.

1 Introduction

In recent years, digital environments have become critical arenas for the emergence of new offenders that commit targeted violent attacks. A growing number of targeted attacks - such as those in El Paso, Christchurch, Baerum, and Halle - have been preceded by digital communication that reveals ideological alignment, grievance-fueled narratives, and signals of intent. These cases illustrate how online spaces increasingly serve as both radicalization platforms and staging grounds for offenders on a pathway to violence.

Identifying violent offenders before they act remains a persistent challenge for law enforcement and security agencies. Detection is difficult not only because offenders often operate in isolation, but also because their motivations are diverse, complex, and highly individualized, making early intervention inherently uncertain. Threat assessment offers a promising approach for identifying individuals at risk of committing acts of violence and for preventing such attacks. To support this goal, various threat assessment protocols have been developed - some designed to evaluate the likelihood of repeated violence among known offenders, and others aimed at identifying individuals who may commit a violent act for the first time [11].

While much of the existing research on threat assessment has centered on offline contexts - where evaluators have direct access to individuals or detailed background information - there is growing interest in extending these methods to digital environments. The key challenge is determining how to reliably detect

behavioral or linguistic indicators in online communication that signal elevated risk and can inform early intervention strategies.

In threat assessment, warning behaviors for targeted or intended violence play an important role since they can be viewed as indicators of increasing or accelerating risk. Warning behaviors are described by [14] as any behavior that "precedes an act of targeted violence, is related to it, and may, in certain cases, predict it." Possibilities to identify certain warning behaviors in social media are not new, and have been examined in earlier work [4–7, 18].

2 Assessing Online Threats

Threat assessment is a structured, evidence-based process designed to identify, evaluate, and manage the risk that an individual will engage in targeted violence. In online environments, this process focuses on detecting behavioral and linguistic indicators that signal elevated risk before violence occurs.

Assessing threatening communication in online environments presents significant challenges. Digital messages often lack tone, context, and reliable identity cues, making it difficult to distinguish between genuine intent and rhetorical or provocative expression. These difficulties are further compounded by the vast volume of content and the dynamic nature of online language, where new terms, coded references, and evolving rhetorical styles emerge rapidly—complicating both detection and interpretation.

Drawing on prior research in behavioral and digital threat assessment [5, 11, 12, 18], we propose a set of eight theoretically grounded warning indicators that capture key dimensions of escalating risk in online communication. These indicators are rooted in the concept of warning behaviors and have been selected for both their empirical relevance and their practical detectability in digital text. This makes them suitable for integration into both manual and automated threat assessment workflows at scale.

The eight indicators reflect behavioral and linguistic patterns that, when observed in online environments, may point to a heightened risk of targeted violence. They include: Anger, Grievance, Othering, Leakage of intent to harm, Warrior mentality, Influence from previous offenders and attacks, Alignment with an extreme ideology, and Preoccupation. Together, these indicators offer a multidimensional framework through which analysts and computational tools can assess risk in written online communication. Each indicator is described below.

Anger
Anger is a core emotional indicator associated with elevated risk of targeted violence. It is commonly expressed through strong hostility, resentment, or frustration, often directed toward specific individuals, groups, or institutions, and in some cases linked to violent intent [18].

In written communication, anger can be identified through the frequent use of aggressive language, insults, threats, or violent rhetoric. Additional linguistic

markers include heightened emotional intensity, the use of excessive exclamation marks, capitalization for emphasis, and words or phrases explicitly conveying rage or frustration.

Grievance
Grievance is a central motivational factor in many cases of targeted violence, often arising from a perceived sense of being wronged or treated unfairly [1]. This perceived injustice can evolve into a strong emotional driver, fostering resentment, a desire for revenge, or a mission to restore perceived lost status. Grievances may be personal (e.g., social rejection, failure, humiliation) or political/ideological, and frequently contribute to the pathway toward violence.

In written communication, grievance may be expressed through persistent complaints about being wronged, treated unfairly, or denied something deserved. Indicators include mentions of revenge, justice, or the need to "correct" a perceived injustice. A key signal is the repetition of victimhood narratives, which may escalate in emotional intensity and frequency over time.

Othering
Othering refers to the linguistic and cognitive division between an in-group ("us") and an out-group ("them"), often serving to dehumanize, stereotype, or justify hostility toward the latter [10]. This distinction is psychologically significant: wars, prejudice, and discrimination are deeply rooted in this "us vs. them" framing [16]. The use of pronouns - especially frequent use of third-person plural forms such as 'they' or 'them' has been shown to be a reliable linguistic marker of negative identification with an out-group [16].

Research has demonstrated that third-person plural pronoun use is one of the strongest linguistic predictors of extremist rhetoric in online groups, including American Nazis and militant animal rights groups [15]. Prior work by Kaati et al. [7] further showed that violent lone offender manifestos contained significantly higher rates of such pronoun use compared to non-violent texts.

In written communication, othering is manifested through frequent use of third-person plural pronouns, dehumanizing language, stereotypes, and portrayals of the "them" group as dangerous, corrupt, or inferior. Such language often implies that violence or exclusion is justified, making othering a key indicator of escalating radicalization and risk for targeted violence.

Leakage of intent to harm
Leakage refers to the communication to a third-party of intent to harm a specific target, conveyed generally through written, spoken, or online statements (as opposed to threats which are made directly to a potential target) [14]. Such communication can be intentional or unintentional, and may vary in its level of specificity regarding the intended act of violence.

Empirical research shows that leakage is a common precursor in cases of targeted violence, including school shootings and attacks on public figures [14]. According to Meloy and O'Toole [14], public figure attacks are often preceded

by indirect, conditional, or bizarre and threatening communications aimed at individuals associated with the target, rather than direct threats to the primary target. Across various studies, the occurrence of pre-attack leakage has been observed in 46% to 67% of cases.

In written communication, leakage may appear as mentions of upcoming actions, boasts about potential violence, veiled threats, or references to past violent events used as inspiration. Statements suggesting planning, preparation, or justification for harm are strong indicators of leakage and signal a heightened need for intervention and monitoring.

Warrior mentality

The use of military terminology in threatening communication is often associated with the warning behavior known as identification, defined by Meloy et al. [12] as a behavioral pattern indicating a desire to become a "pseudo-commando", to adopt a warrior mentality, or to identify with prior attackers or agents of a cause. Such individuals increasingly frame themselves as actors engaged in a noble battle, justifying the use of structured violence.

In written communication, this manifests through references to combat, duty, or warfare outside of a formal military context. Other indicators include mentions of weapons, tactical gear, or training for violent action. The individual may explicitly or implicitly frame their actions as part of a larger ideological or grievance-driven battle. The presence of military terminology in such contexts signals an increasing identification with violent roles, and is therefore a key variable in threat assessment.

Influence of previous offenders and attacks

Identification is a form of warning behavior described by Meloy and Hoffmann [12], in which an individual expresses admiration or emotional connection with past perpetrators of violence - particularly those responsible for ideologically or grievance-driven attacks. This behavior signals both a psychological affinity with previous offenders and a potential intent to emulate or continue their actions.

In written communication, influence from previous offenders and attacks may be identified by explicit references to past perpetrators, expressions of admiration for their actions, replication of their language or manifestos.

Alignment with an extreme ideology

Alignment with an extreme ideology involves the adoption of rigid, absolutist belief systems that legitimize or promote the use of violence to achieve ideological, political, religious, or social objectives. These belief systems often frame violent action not merely as acceptable, but as necessary—or even morally justified—to advance the cause.

In written communication, alignment with an extreme ideology may be identified by the use of extremist terminology, slogans, or coded language linked to known ideologies or ideological justifications for violence.

Preoccupation
Preoccupation refers to a persistent and sustained focus of attention on a particular person, topic, cause, or grievance [13]. Within the context of threat assessment, it is considered an early warning behavior that may signal the development of an individual's emotional investment or intent toward a potential target of violence or ideological cause.

While many individuals demonstrate strong interest or engagement with particular subjects without progressing toward violence, preoccupation becomes a cause for concern when it begins to dominate the individual's cognitive and emotional landscape, narrow their worldview, and fuel a grievance-driven mindset [13]. In this state, the individual's thinking and communication often become obsessive, repetitive, and increasingly isolated from other life domains.

In written communication, preoccupation is reflected in repeated references to the same person, cause, ideology, or event. The language used may become progressively intense, polarized, or absolute, signaling the potential escalation toward violent ideation or targeted action.

3 Technologies for Detecting the Warning Indicators

To support a scalable assessment of warning indicators in large volumes of online text, we employed a hybrid extraction approach combining rule-based, supervised, and generative methods. Each of the eight warning indicators was extracted using techniques best suited to its linguistic and semantic characteristics. Our pipeline leverages three complementary strategies:

- Dictionary-based extraction - for indicators where explicit lexical cues are well-understood and can be captured through curated dictionaries.
- Supervised classification - for indicators where a trained model can learn complex patterns of expression beyond simple lexical matches.
- Generative large language model (LLM) extraction - for indicators that require deeper contextual understanding or interpretation beyond what is practical with rules or classifiers alone.

The extraction methods for each warning indicator are described below.

3.1 Anger, Othering, Grievance and Influence from previous offenders and attacks

For the indicators Anger, Othering, Grievance, and Influence from previous offenders and attacks, we employed a dictionary-based approach. Manually curated dictionaries were constructed based on prior research on linguistic markers of these phenomena [2, 17, 18], and were further adapted to cover the specific language observed in our datasets. Each text segment was scanned for matches to these dictionaries. If the relative frequency of matched terms exceeded a defined threshold, the corresponding warning indicator was marked as present. We used the same threshold as described in [6].

3.2 Alignment with an extreme ideology

The indicator Alignment with an extreme ideology was extracted using a set of supervised classifiers trained on different examples of extremist communication. The classifiers, based on a transformer architecture (RoBERTa), were fine-tuned on datasets and described [19]. Each classifier outputs a probability score for ideological alignment. Texts with scores above a defined confidence threshold (40%) were flagged as expressing alignment with an extreme ideology.

3.3 Leakage, Warrior mentality, and Preoccupation

The remaining three indicators - Leakage, Warrior mentality, and Preoccupation - require interpretation of broader semantic and behavioral context, which is difficult to capture using dictionaries or simple classifiers alone.

For these indicators, we We utilized a zero-shot (or few-shot) prompt-based extraction technique leveraging a large language model (LLM). Using a generative LLM (based on GPT-4), we designed prompts tailored to each indicator. The LLM was asked to evaluate whether a given text segment contained evidence of the relevant behavior or language pattern, based on definitions aligned with prior threat assessment literature. Outputs from the LLM were validated against a sample of manually annotated texts to ensure reliability and to calibrate extraction thresholds.

This hybrid extraction strategy allowed us to combine the high precision of dictionary-based methods, the flexibility and generalization of supervised models, and the deep contextual reasoning capabilities of LLMs. By tailoring extraction methods to the nature of each warning indicator, we aim to maximize both accuracy and scalability in the automated analysis of threatening communication.

4 The Presence of Warning Indicators in Threatening Communication

To evaluate the usefulness of the eight warning indicators in assessing threatening communication, we have created a dataset consisting of 95 texts, each authored by a unique user, containing only threatening communication.

The threatening communication that we use was extracted from an Incel forum - such forums are well-known to host threatening communication. In [9], the presence of violent language on the incel forum incels.is was examined. Almost 22% of the posts contained some form of expression of violent thought or intent. The forum where we have collected our data is called Blackpill. Blackpill is an independently operated platform moderated by self-identified incels [3]. A detailed description of the forum and its characteristic forms of threatening communication is provided in [8]. We collected all publicly available data from Blackpill, covering the period from the forum's launch on September 7, 2020, through January 21, 2024. The resulting dataset comprises 378,767 posts authored by 1,072 unique accounts.

Since our primary objective is to assess threatening communication we first needed to extract such content using a specialized classifier described in [8]. Applying this classifier, we identified a total of 20,993 posts authored by 501 individuals as containing threatening material.

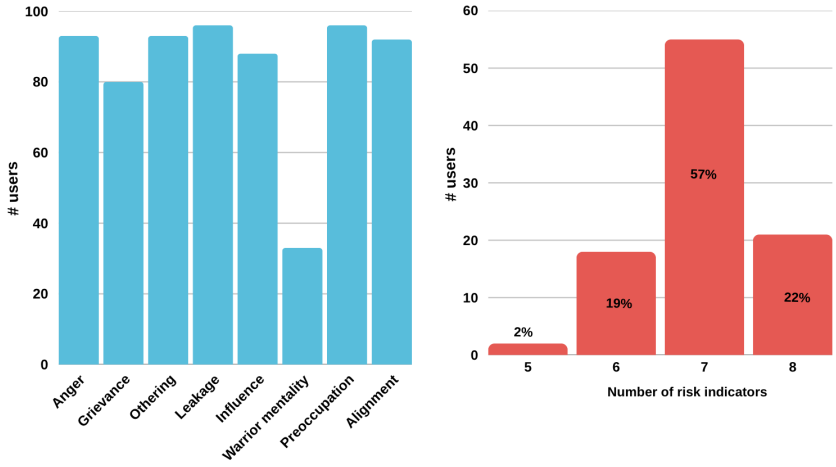

Fig. 1. Warning indicators for each individual present in the threatening communication (left) and the distribution of the number of warning indicators for each individual (right).

It is important to note that the dataset we created consists exclusively of threatening language and violent fantasies and should be seen as high-risk, concerning content.

We applied our hybrid detection method to determine which indicators were present in each text. The results, summarized in Figure 1 (left), reveal that:

- 100% of texts showed signs of Preoccupation and Leakage.
- 97% included Anger and Othering.
- 96% contained Alignment with extreme ideology.
- 92% indicated Influence from previous offenders.
- 83% included Grievance.
- 34% demonstrated a Warrior mentality.

The frequent occurrence of individual warning indicators in threatening online communication underscores the challenge of discerning whether users genuinely pose a threat or are merely engaging in violent rhetoric. When examining the number of indicators per individual, however, a more nuanced picture emerges. As illustrated in Figure 1 (right), 57% of users displayed seven indicators, while 22% exhibited all eight. The remainder showed six indicators (19%)

or five (2%). These findings indicate that not all eight warning indicators are consistently present in the threatening communications we analyzed.

5 Correlation Between Warning Indicators and Persons of Concern

To evaluate how effectively the eight warning indicators predict an elevated risk of violence, each text was independently assessed by 2-4 trained threat assessors, all with professional backgrounds in law enforcement or psychology. Each assessor reviewed a set of 25 texts each and was asked to judge whether the author of the text should be considered a person of concern warranting further investigation. The assessors were not provided with a formal threat assessment protocol and were instructed to rely solely on their professional judgment. Given that all texts in the dataset contained threatening or violent language, we anticipated that the majority of individuals would be classified as persons of concern. Our assessors judged 66 authors (69%) to be concerning, resulting in a dataset with a high concentration of individuals flagged for elevated risk. For our analysis, we used a majority vote approach, classifying an author as "concerning" if the text was flagged as such by at least two assessors. This resulted in 20 texts (21%) being classified as concerning under this criterion.

To examine the relationship between our experts judgments and the presence of warning indicators, we calculated the Pearson correlation coefficient between the majority vote labels assigned by the assessors (i.e., whether the individual was considered a person of concern) and the number of warning indicators automatically extracted from each text. The resulting correlation coefficient was $r = 0.306$, indicating a moderate positive correlation between the presence of warning indicators and the assessors' judgments. In other words, texts containing more warning indicators were more likely to be classified as concerning by the human assessors.

This correlation was statistically significant ($p = 0.0024$), indicating that the likelihood of observing this correlation by chance is very low (less than 0.25%). These results suggest that the eight warning indicators provide a meaningful and reliable framework for detecting signals of potential violence that aligns well expert human judgment.

6 Discussion

Our findings reveal a statistically significant correlation between the number of detected warning indicators and cases where experts identified an author as a person of concern. This suggests that the eight warning indicators are not only detectable through automated methods but also meaningful and relevant in professional threat assessment contexts. Although the correlation is moderate, it supports the potential integration of these indicators into triage systems and early detection workflows.

However, several limitations should be acknowledged. The data used in this study were drawn exclusively from a high-risk online forum known for violent and extremist content, which may limit the generalizability of the findings to broader or more diverse digital environments.

Moreover, while our hybrid extraction approach offers a balance between interpretability and scalability, it remains susceptible to false positives—particularly in cases involving sarcasm, coded language, or statements taken out of context.

Another potential limitation lies in the absence of a formalized threat assessment protocol for the human evaluators, which may have introduced variability in their judgments. Nonetheless, this choice reflects real-world conditions, where practitioners often rely on professional intuition and contextual judgment during early-stage evaluations.

Overall, the results underscore the value of combining theory-driven frameworks with computational tools in the domain of digital threat assessment. Future research should aim to validate these indicators across diverse platforms and examine the practical implications of such tools in operational threat management settings.

7 Conclusion and Directions for Future Work

This study explored how eight theoretically grounded warning indicators can be used to assess the potential for targeted violence in written online communication. Our results demonstrate that these indicators - anger, grievance, othering, leakage, warrior mentality, influence from previous offenders, alignment with an extreme ideology, and preoccupation can be detected through a combination of manual and automated methods. The presence of warning indicators aligns moderately with expert human judgment. Specifically, we found a statistically significant positive correlation between the number of warning indicators present in a text and the likelihood of it being classified as concerning by trained threat assessors.

These findings suggest that the eight warning indicators provide a meaningful and scalable framework for digital threat assessment and can support early identification of individuals on a potential pathway to violence in online spaces. Importantly, the indicators can be used both by human analysts and by automated systems to prioritize cases for deeper review.

Future work should expand validation of the warning indicators across diverse platforms and ideological contexts, explore how these indicators evolve over time, and integrate linguistic signals with behavioral data for more robust threat models. Improving automated extraction methods, especially using advanced language models, and conducting operational validation with law enforcement partners are also key priorities. These efforts aim to enhance the effectiveness of digital threat assessment in preventing targeted violence.

References

1. Clemmow, C., Gill, P., Bouhana, N., Silver, J., Horgan, J.: Disaggregating lone-actor grievance-fuelled violence: Comparing lone-actor terrorists and mass murderers. Terrorism and Political Violence pp. 1–26 (2020)
2. Cohen, K., Johansson, F., Kaati, L., Mork, J.C.: Detecting linguistic markers for radical violence in social media. Terrorism and Political Violence **26**, 246–256 (2014)
3. ICSR: The incel subculture (July 2024), https://doi.org/10.18742/pub01-189
4. Kaati, L., Shrestha, A., Sardella, T.: Identifying warning behaviors of violent lone offenders in written communication. In: ICDM workshop SoMeRis (2016)
5. Kaati, L., Shrestha, A., Akrami, N.: Predicting targeted violence from social media communication. In: 2022 IEEE/ACM ASONAM. pp. 383–390 (2022)
6. Kaati, L., Shrestha, A., Akrami, N.: General risk index: A measure for predicting violent behavior through written communication. In: 2023 IEEE International Conference on Big Data (BigData). pp. 4065–4070 (2023)
7. Kaati, L., Shrestha, A., Cohen, K.: Linguistic analysis of lone offender manifestos. In: International Conference on CyberCrime and Computer Forensics (ICCCF) (2016). https://doi.org/10.1109/ICCCF.2016.7740427
8. Lundmark, L., Kaati, L., Shrestha, A.: Visions of violence : Threatful communication in incel communities. In: 2024 IEEE International Conference on Big Data (BigData). pp. 2772–2778 (2024)
9. Matter, D., Schirmer, M., Grinberg, N., Pfeffer, J.: Investigating the increase of violent speech in incel communities with human-guided gpt-4 prompt iteration. Frontiers in Social Psychology **2** (2024)
10. McCauley, C., Moskalenko., S.: Mechanisms of political radicalization: Pathways toward terrorism. Terrorism and Political Violence **20**(3) (2008)
11. Meloy, J., Gill, P.: The lone-actor terrorist and the TRAP-18. Journal of Threat Assessment and Management **1**(3), 37–52 (2016)
12. Meloy, J., Hoffmann, J., Guldimann, A., James, D.: The role of warning behaviors in threat assessment: An exploration and suggested typology. Behavioral Sciences & the Law **30**(3), 256–279 (2012)
13. Meloy, J., Hoffmann, J., Guldimann, A., James, D.: Warning behaviors in threat assessment: An exploration and suggested typology. Behavioral Sciences and the Law **30**(3), 256–279 (2012)
14. Meloy, J., O'Toole, M.E.: The concept of leakage in threat assessment. Behavioral Sciences and the Law **29**, 513–527 (2011)
15. Pennebaker, J.W., Chung, C.K.: Computerized text analysis of al-Qaeda transcripts. In: Krippendorf, K., Bock, M.A. (eds.) The Content Analysis Reader. Sage, London, UK (2008)
16. Pennebaker, J.W., Chung, C.K.: Language and social dynamics. In: Technical Report 1318. University of Texas at Austin, Texas, USA (2012)
17. Shrestha, A., Kaati, L., Akrami, N.: PRAT - a tool for assessing risk in written communication. In: 2019 IEEE International Conference on Big Data (Big Data). pp. 4755–4762 (2019)
18. Shrestha, A., Akrami, N., Kaati, L.: Introducing digital-7 threat assessment of individuals in digital environments. In: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 720–726 (2020)
19. Shrestha, A., Kaati, L., Akrami, N.: Linguistic alignments: Detecting similarities in language use in written communication. In: 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). p. 619–623