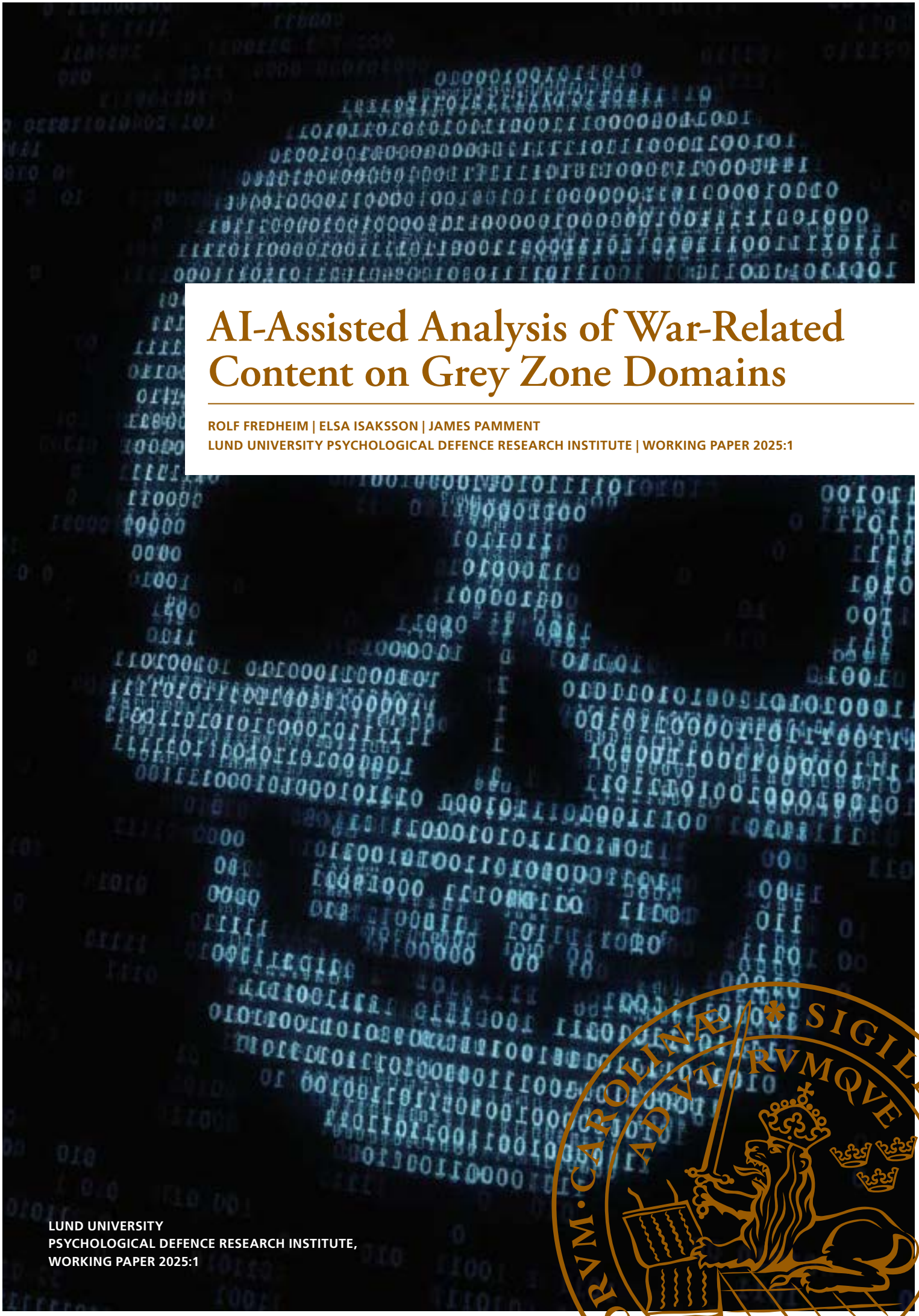
AI-Assisted Analysis of War-Related Content on Grey Zone Domains

Fredheim, Rolf; Isaksson, Elsa; Pamment, James

2025



AI-Assisted Analysis of War-Related Content on Grey Zone Domains

ROLF FREDHEIM | ELSA ISAKSSON | JAMES PAMMENT

LUND UNIVERSITY PSYCHOLOGICAL DEFENCE RESEARCH INSTITUTE | WORKING PAPER 2025:1

LUND UNIVERSITY
PSYCHOLOGICAL DEFENCE RESEARCH INSTITUTE,
WORKING PAPER 2025:1

AI-Assisted Analysis of War-Related Content on Grey Zone Domains

Dr. Rolf Fredheim
Psychological Defence Research Institute, Lund University

Elsa Isaksson
Psychological Defence Research Institute, Lund University

Dr. James Pamment
Psychological Defence Research Institute, Lund University

LUND
UNIVERSITY

Lund University Psychological Defence Research Institute
Working Paper 2025:1

About the Lund University Psychological Defence Research Institute

PDRI is an independent research institute based at the Department of Strategic Communication, Lund University. Its core funding is provided by the Swedish Psychological Defence Agency. In addition, PDRI receives funding from different sources to produce specific publications or develop research tracks.

About Psychological Defence Research Institute Working Papers

In this publication series, researchers connected to PDRI present short analyses or briefings on issues relevant to the public understanding of psychological defence. This includes work on the concept of psychological defence, its associated capabilities, the tactics, techniques and procedures used by threat actors, and the use of new technologies or new platforms to exert information influence.

Funding for this research has been provided by the AI Lund initiative and the Swedish Psychological Defence Agency.

Cover image generated by Google Gemini

Department of Communication

ISBN 978-91-8104-792-9 (print)

ISBN 978-91-8104-793-6 (electronic)

Working Paper 2025:1

Printed in Sweden by Media-Tryck, Lund University

Lund 2025

Table of Contents

Executive Summary

The Lund University Psychological Defence Research Institute conducts studies of under-researched platforms, such as our recent analysis of information influence operations in video games.[1] For the next study in this series, we are exploring methods of analysing so-called grey zone domains within information influence operations.

Grey zone domains are web platforms of dubious legality, where violent and shocking content (mostly videos and images) is hosted. The content is often aggregated from private messaging apps, member-only chat forums, social media, and the dark web. Once the content is hosted by one of these aggregator web platforms, it can be reshared hundreds or thousands of times on other digital media platforms, with each re-share posing a new challenge to that digital platform's content moderation algorithms. Grey zone domains may, therefore, have evolved into a crucial step in information influence infrastructure, particularly for enhancing the resilience of certain types of content-to-content moderation. At present, there is little scientific data to explain how much of a problem this infrastructure is.

This project focuses on two types of grey zone domains that sometimes aggregate content relevant to information influence operations. The first is gore sites, or aggregators of violent content, including murders, torture, suicides, terrorism, and accidents. We identified several sites of this kind, hosting tens of thousands of videos and images that have together been viewed tens, perhaps even hundreds of millions of times. The second is porn sites, or aggregators of sexually explicit content that sometimes involves gore or other signifiers of information influence. We identified a small number of porn sites that carry relevant content. In some cases, grey zone domains carry both types of content in the same aggregator, positioning violence and porn content in the same feed.

The relevance to information influence is perhaps most obvious in relation to disturbing content shared from the Russia-Ukraine and Israel-Palestine conflicts. We have found thousands of examples of grey zone websites carrying violent content from these conflicts, including graphic examples of torture, mutilation, and death from the wars, but also sexually explicit violent content tied to the conflicts. But to what extent is the spread of this content part of information influence operations? Can we identify specific users,

[1] Pamment, Falkheimer & Isaksson, *Malign Foreign Interference and Information Influence on Video Game Platforms: Understanding the Adversarial Playbook* (Stockholm: Swedish Psychological Defence Agency, 2023).

platforms, or countries using grey zone domains as part of their influence operations targeting EU member states?

The disturbing nature of the content aggregated by these websites makes researching them a challenge, and in general they are a poorly understood aspect of information influence infrastructure. We have therefore developed a method for researching grey zone domains that minimises the analyst's exposure to disturbing content, and that enables a closer analysis of the behaviours that can help to identify relevant content.

For this project, we developed an AI-assisted system to analyse sensitive content on these domains. The system consists of a web crawler to collect data from target sites, a content classifier using GPT-4o, and a customisable dashboard that enables complex filtering in order to hone in on the most relevant content while shielding human analysts from disturbing material. This working paper is a proof-of-concept for the system that primarily focuses on demonstrating the scale of the problem and the potential for classifier reliability. Further applications in the realms of counter-FIMI and law enforcement are feasible based on this prototype.

This pilot study focuses on one grey zone domain that aggregates both gore and pornographic content. Our analysis found substantial war-related content, with approximately 17% of classified posts related to ongoing conflicts. Individual posts were viewed hundreds of thousands of times, suggesting that some material hosted on this single grey zone domain went viral on other digital platforms.

The methodology demonstrates the potential for monitoring hard-to-access online spaces while minimising analyst exposure to harmful content. The results show the effectiveness of AI in reducing human exposure to traumatic content while maintaining analytical accuracy.

1 Introduction

Imagine that you are a Russian intelligence operative tasked with spreading pro-Russian, anti-Ukrainian content to international audiences. After almost four years of war, there is plentiful graphic footage that can be spun to show cruelty, bravery, cowardice or incompetence. Your challenge is not finding content but disseminating it. Western platforms like YouTube or Twitter (X) automatically remove or label gory footage. To varying degrees, 'friendlier' platforms like VK and Telegram offer a place for sharing war-related propaganda content, but not to mainstream EU audiences. There is, however, another option.

Certain web domains exist in a regulatory grey zone. They host material depicting illegal violence (murder, executions, sexual violence, mutilation and dismemberment, cannibalism, etc) not allowed on mainstream social media. On the surface, it would not be surprising if these web domains offered safe spaces for hosting war propaganda, either as part of an organised operation by a military seeking to shape perceptions of a conflict, or by internet fanboys who found a path of least resistance to raise advertising revenue or support their political convictions.

The scenario prompts the question: how would we know? These domains are 'unsafe for work' and only reach niche audiences, at least initially. Researchers and analysts would struggle to trawl through thousands of gory or pornographic images and videos from their work computers. AI systems designed to avoid violent and sexual material often fail to monitor these sites, and their existence is for most a taboo subject. Law enforcement of grey zone domains appears limited. Social media platforms bear little, sometimes no, responsibility for externally hosted content. In this pilot study we therefore present a novel methodology for analysing these unsafe spaces, using AI to support the analyst's task.

This working paper has two goals:

1. To conduct a pilot study that estimates the extent to which war-related content is shared on these platforms. This gives us a first reliable snapshot of the 'universe' of content that could be exploiting grey zone domains in support of information influence operations targeting the EU.

2. To develop and test a robust methodology and operational system for using AI to analyse the material shared on these domains. A successful pilot study of one grey zone domain will enable us to expand the analysis to a larger selection of relevant domains in a later study.

This working paper presents a proof-of-concept of an AI-assisted content classification system for assessing and categorising harmful content published on grey zone domains. The system should rapidly and precisely process diverse and sensitive online content while minimising exposure risks for human analysts. We seek a methodology to identify and analyse trends in online content within platforms that may escape traditional monitoring, particularly focusing on hard-to-access or ethically challenging content sources pertaining to ongoing conflicts and geopolitical events. The system is designed to alert stakeholders to new problem content relative to their specific areas of interest.

2 Methodology

2.1 Data Collection

We aimed to collect all posts from one publicly accessible grey zone domain, focusing on content aggregated on the sites during a 6-month period. During our data collection period, two major conflicts dominated global attention: the ongoing war in Ukraine and the conflict in Gaza.

An exploratory analysis showed stability in content distribution across categories since November 2023. Before this period, some content appeared to be missing, prompting us to restrict historical content to a 6-month observation window. This is because the domain restricted some content behind a paywall after a specific time window. We considered only freely available material without authentication since this content is more likely to be cross-posted.

This gave us an initial dataset of 8,300 posts. We configured our monitoring system to check for new content daily. Our pilot dataset therefore consists of 9,200 posts from 1 November 2023 to 10 July 2024.

We developed a bespoke web crawler to systematically crawl and collect data from the grey zone domain. For the pilot study, we adapted the scraping framework to a single domain but built it to use configuration files that would allow rapid and precise scaling to other sites. The scraper was designed to gather specific fields such as links to media, number of views, and titles.

The system is designed to offer continuous monitoring as well as historical data collection. It crawls each target website daily. It stores collected data in a database and sends new datapoints to the AI classification system for initial assessments. These assessments are pushed to a dashboard for visualisation, and a labelling interface for analyst validation.

2.2 AI Classification System

Our AI classification system uses GPT-4o, a large language model from OpenAI capable of processing text and image inputs. We chose this model because it can process sensitive material without rejecting inputs, unlike some alternatives that moderate both inputs and outputs. OpenAI and Google use different content moderation strategies. OpenAI moderates outputs, while Google moderates both inputs and outputs. Google's input moderation makes its models unsuitable for this project, as they would reject the sensitive

content we aim to analyse.[2] OpenAI's models, however, can process highly sensitive content as long as the schema doesn't require outputting potentially harmful material.

The classification process involves showing each post's title, description, and associated images or video screenshots to the model alongside our schema. The schema consists of category descriptions for Content Type, Target, Perpetrator, Relevance, and Tags.

The classification system analyses the combined text and visual inputs to generate classifications. Our system ensures the model's responses adhere to a consistent format compatible with the schema. The AI-generated classifications are a starting point for human annotators, who can review and adjust these initial assessments.

This AI-assisted approach allows for efficient processing of large volumes of potentially sensitive material while minimising direct exposure for human analysts. The model's ability to handle text and visual content enhances the accuracy of initial classifications, particularly for posts where the visual element is key to understanding the content's nature or relevance.

2.3 Classification Schema

The classification schema for the pilot study prioritises the surfacing of content relevant to military or security concerns.

The categories assessed were:

1. Content Type:
    - None: Content that doesn't fit into any other category
    - Violence: Depictions or descriptions of violent acts, Graphic depictions of bodily harm or death
    - Pornography: Sexually explicit content
    - Other: Content that doesn't fit the above categories but may be relevant
2. Target
    - One or more of: None, Women, Russia, Ukraine, Israel, Palestine, Sweden, NATO, Gangs, Other
3. Perpetrator
    - One or more of: None, Women, Russia, Ukraine, Israel, Palestine, Sweden, NATO, Gangs, Other
4. Relevance:
    - Binary classification (True/False) for relevance to military/terrorism purposes

[2] https://ai.google.dev/gemini-api/docs/safety-settings

- Based on the description: "This project is interested in content used for military/terrorist purposes. Material, however shocking or gruesome, with no military or security relevance, is irrelevant."

5. Tags:
    - Analysts could additionally apply one or more of these tags:
    - None, War related, Terrorism/extremism related, Race/ethnicity, Gender, Sexuality, Other

Data generated according to this schema allows analysts to examine intersections between content types and targets and identify patterns or trends.

2.4 Analyst Verification and Labelling Interface

Instead of looking at posts directly, data is pre-labelled in the open-source labelling software Label Studio. We configured it to display AI-generated classifications for review.

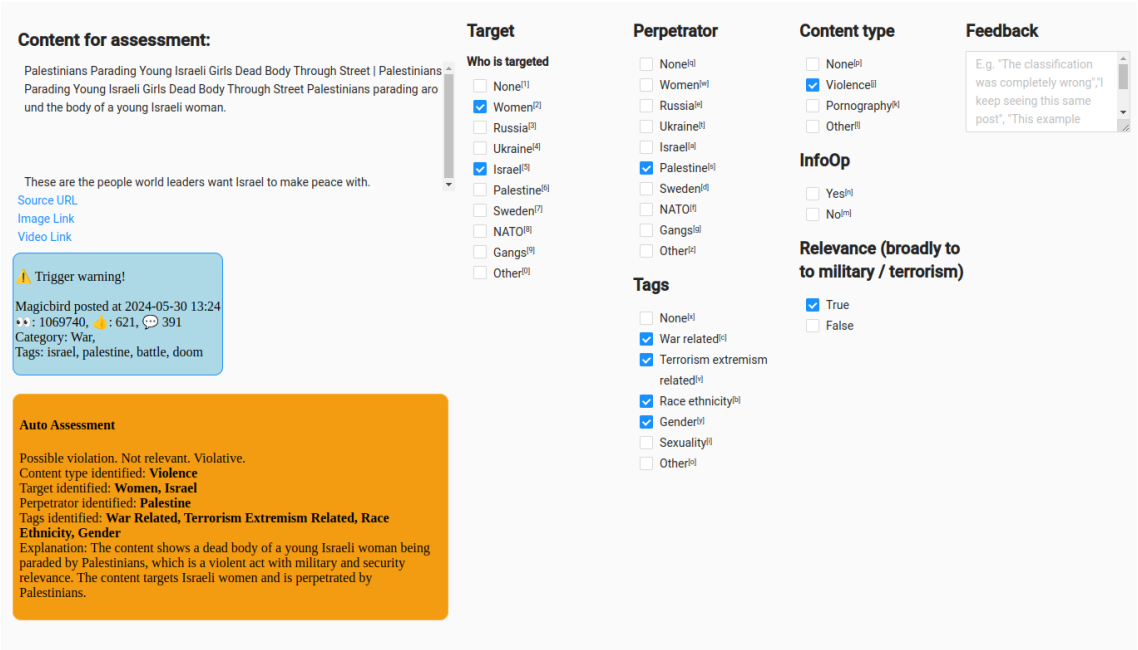


Figure 1: The Labeling interface showing the content display area, classification options, and AI-assisted pre-labelling.

The labelling interface, shown in Figure 1, shields analysts from disturbing content and enables accurate labelling. It features a prominent trigger warning at the top of the blue box, alerting analysts to potentially upsetting content. The interface provides embedded content links rather than displaying material inline, giving analysts control over which content they wish to view. The original post text is shown, which may suffice for analysts to understand the content's nature without viewing graphic images or videos. Metadata

such as posting details and engagement metrics provide context without exposing the analyst to the content.

A prominent orange box displays an automated assessment of the content, including potential violations, content type, target and perpetrator identification, relevant tags, and a detailed explanation. This AI assessment often enables analysts to label content accurately without directly viewing the original material, significantly reducing exposure to potentially traumatic imagery or descriptions.

The example in Figure 1 shows a data point describing disturbing scenes where Hamas paraded the dead body of a young Israeli woman through the streets. The AI assessment identifies the violent nature of the content, the involved parties, and the broader context of conflict and extremism. It classifies the content type as Violence, with Israel and women as the targets and Hamas as the perpetrator. The AI tags the content as War Related and Terrorism Extremism Related, recognising its relevance to ongoing conflicts. It also applies Gender and Race Ethnicity tags, acknowledging the specific targeting of an Israeli woman. This could potentially support analyses of ethnicity and gender using this methodology.

2.5 Data Visualisation and Monitoring

AI classification and human validation results were collated in a dashboard, allowing for quantitative analysis of data patterns and calculation of reliability measures. The dashboard, which utilises the open-source data visualisation tool Kibana, is connected to our database. We visualised content distribution across categories, trend analysis over time, and comparisons between AI and human classifications.

The dashboard allows us to:

1. Track newly classified content
2. Filter for specific date ranges, content types, and relevance scores
3. Monitor inter-annotator agreement and AI classifier performance

The dashboard also helps analysts identify trends and anomalies in the classified data and conduct ongoing monitoring and quantitative analysis.

2.6 Task handler

Data collection was automatic and continuous. To streamline the process, we used a task handler to initiate and monitor the web scraping. The task handler also triggers AI classification for new content, assigns labelling tasks to human annotators, and tracks system uptime and performance. It can raise alerts for system health and processed

content, such as detecting highly relevant content. This setup could alert relevant agencies to data points of interest.

The task handler coordinates all components. It initiates the nightly scraper, logs new content to the database, triggers GPT-4o classification for fresh posts, and dispatches items to Label Studio. It tracks performance, raises alerts for system health or high-priority content, and can run follow-up queries when items are deemed relevant.

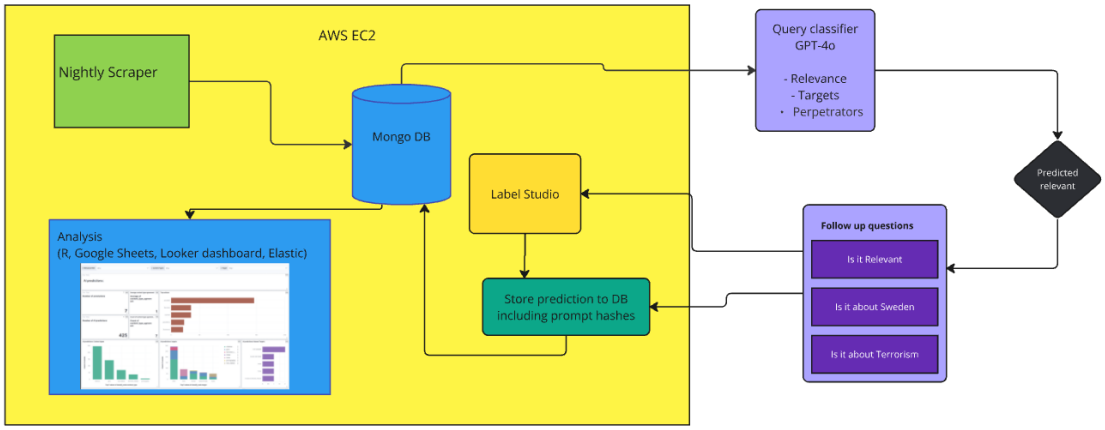


Figure 2: The AI-assisted data processing pipeline. The diagram shows how the nightly scraper, MongoDB, Label Studio, and GPT-4o classification interoperate. The scraper populates the database, which feeds content to GPT-4o for classification. Predicted results route back to the database and Label Studio for analyst review, with follow-up questions triggered as needed.

3 Results

3.1 War-Related Content

Within the dataset of 9,200 posts, the AI classifier identified 2,500 as war-related and about Russia, Ukraine, Israel or Palestine. This represents 27.2% of the total content. This substantial proportion indicates that war-related content forms a significant part of the material hosted on grey zone domains.

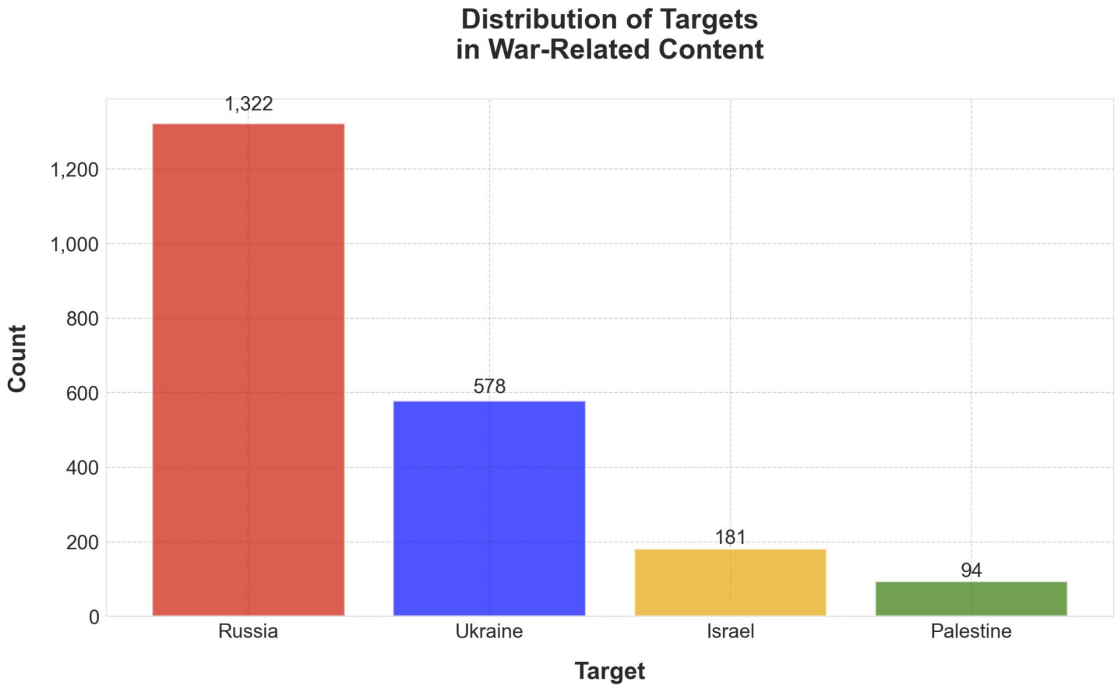


Figure 3: Distribution of Targets in War-Related Content. The graph depicts the distribution of perpetrators in war-related content classified as relevant by AI on grey zone domains. It focuses on posts where Russia, Ukraine, Palestine, or Israel is identified as either perpetrators or targets.

Russia appears as the target of violent acts in approximately 1,300 posts, more than double the next highest target. Ukraine is the second most common target, featured in almost 600 posts. Israel follows with almost 200 posts, while Palestine is the least targeted, appearing in roughly 100 posts. This distribution suggests a disproportionate focus on the Russia-Ukraine conflict compared to the Israel-Palestine conflict within the analysed content.

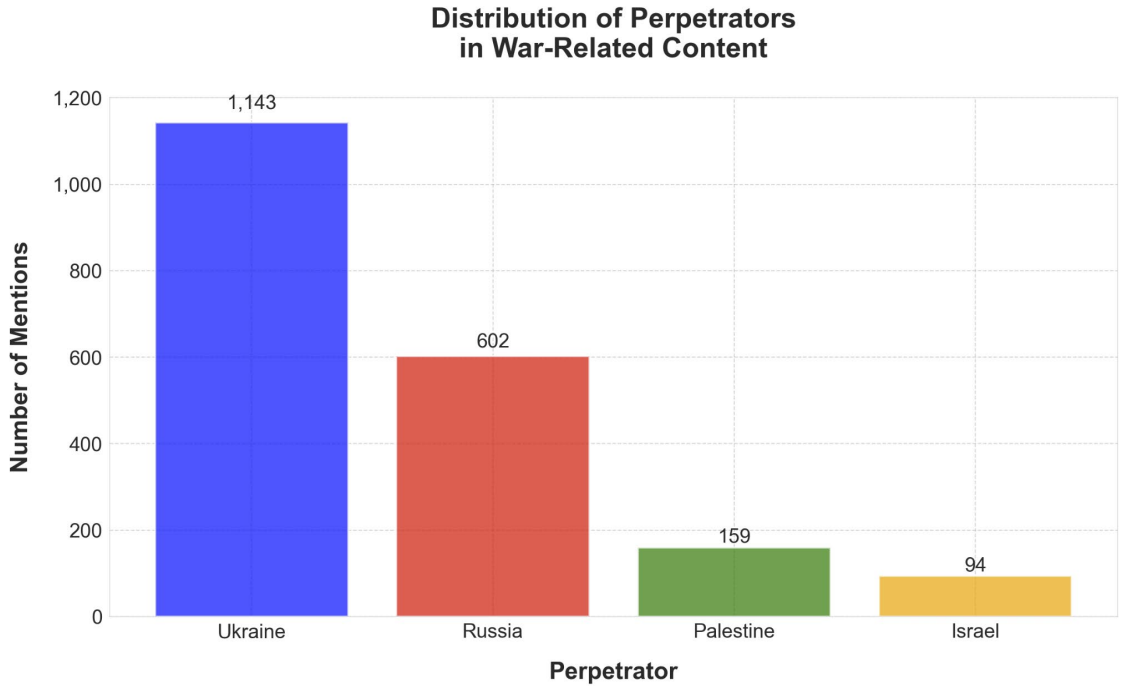


Figure 4: Distribution of Perpetrators in War-Related Content. The graph depicts the distribution of perpetrators in war-related content classified as relevant by AI on grey zone domains. It focuses on posts where Russia, Ukraine, Palestine, or Israel is identified as either perpetrators or targets.

Ukraine is most frequently identified as the perpetrator of violent acts, appearing in approximately 1,150 posts. This is nearly double the number of posts identifying Russia as the perpetrator. In the Israel-Palestine conflict, Palestine is more often portrayed as the perpetrator (159 posts) compared to Israel (94 posts).

This material is not just hosted but actively consumed. The most viewed post, with over 1 million views, shows Hamas parading the body of a young Israeli woman. Other highly viewed content includes mass shootings, footage released by ISIS showing the Moscow concert hall terrorist attack, and graphic war footage from Ukraine and Gaza. These posts often attract hundreds of thousands of views.

Grey zone domains act as repositories for content that can later be spread on mainstream platforms. Our analysis shows significant traffic to this material, indicating its continued spread onto mainstream platforms. For example, according to Semrush, a recording of the 2022 Buffalo supermarket shooting, which was live-streamed on Twitch but quickly removed, has 617 inbound links from 233 separate referring domains.

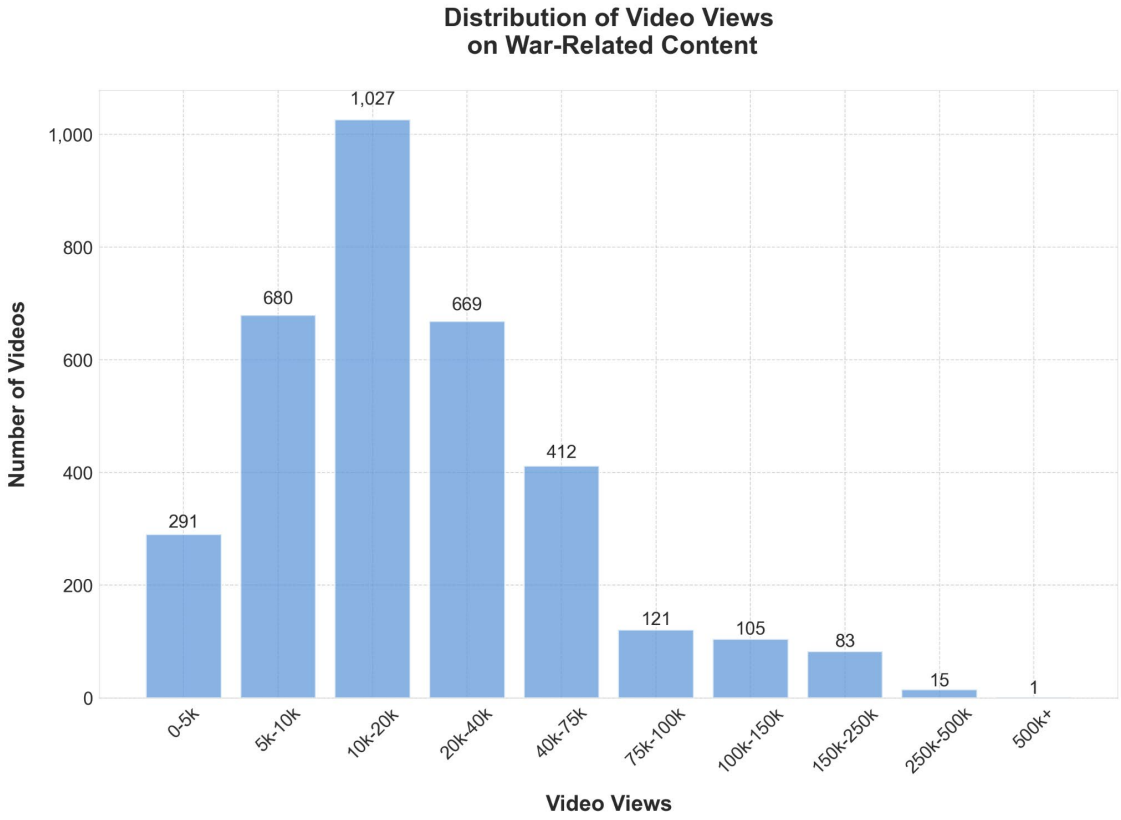


Figure 5: Distribution of view counts for war-related content.

The histogram in Figure 5 shows a skewed distribution of view counts, with a long tail extending into hundreds of thousands of views. Most posts receive modest view counts — the median is 18,900. However, a small number of posts attract large numbers of views, creating an extended tail to the right.

According to Semrush, in June 2024, the domain received 15.8 million visits from 3.5 million unique visitors, with an average of 6.27 pages viewed per visit and an average visit duration of 8 minutes and 47 seconds.

This distribution pattern strongly suggests viral spread. The majority of posts attract modest attention, likely from regular site visitors. However, the outliers with hundreds of thousands or even millions of views indicate content that has spread far beyond the site's usual audience. This points to external sharing and embedding of content, drawing significant inbound traffic from other sources. It is highly probable that parts of this material are being shared on more mainstream platforms that then link or direct users to this domain, perhaps for more complete or uncensored versions.

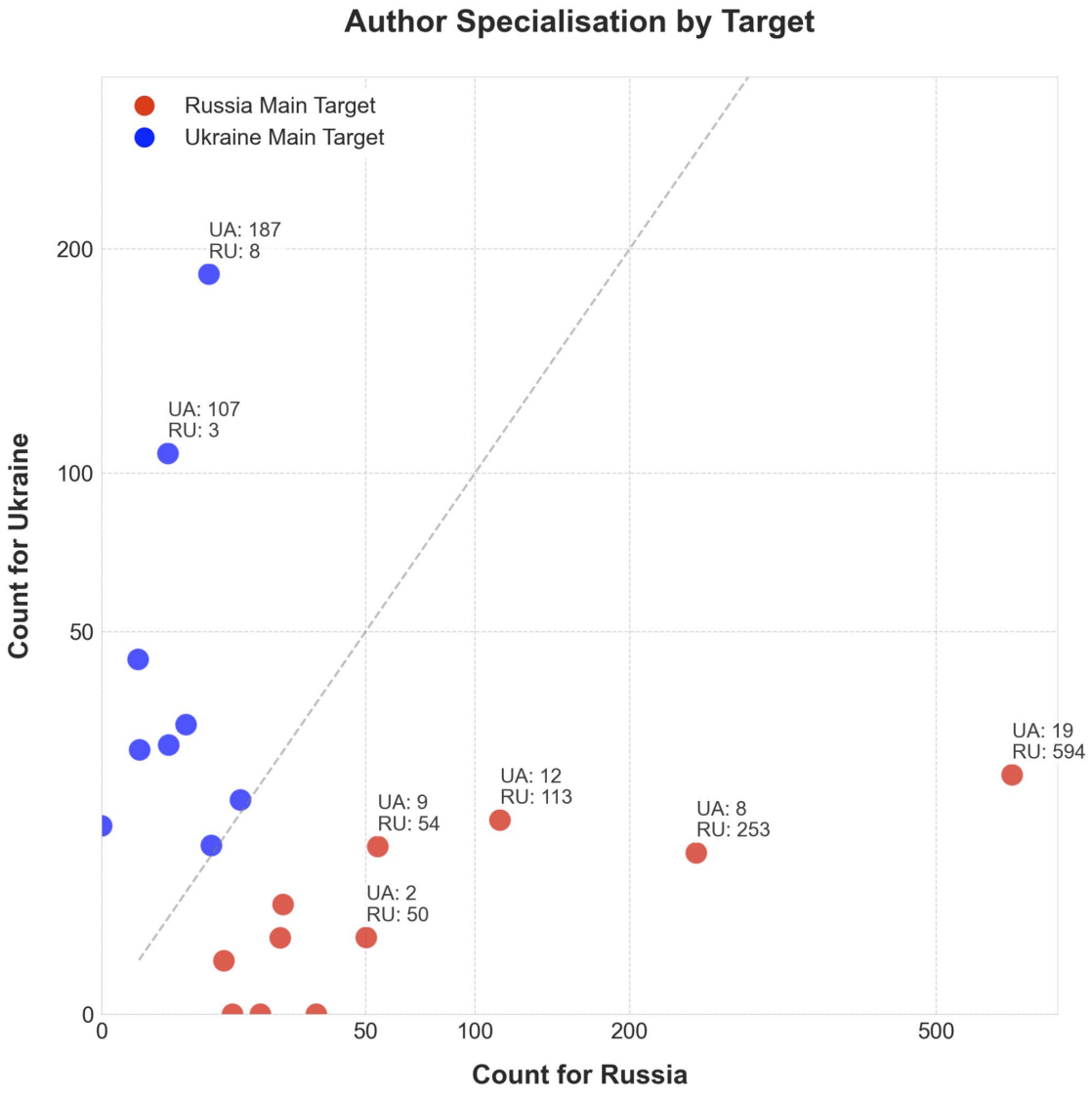


Figure 6: Graph showing focus area for users sharing violent content from the Russia-Ukraine war. It shows counts of posts targeting Russians (x-axis) versus Ukrainians (y-axis). Red points represent accounts predominantly posting content targeting Russians, while blue points indicate accounts focusing on content targeting Ukrainians. The counts show the number of uploads where Ukraine and Russia are identified as targets. All but two accounts are positioned away from the diagonal line, suggesting a strong bias towards one side. This uneven distribution highlights actors who are motivated to highlight the actions of one side or the other, rather than simply sharing gore.

The figure above shows that most users who upload violent content depicting incidents in the Russia-Ukraine war aren't indiscriminately sharing gore or violent footage; they are taking a side. Some accounts specialise in videos targeting Russians, others focus on the deaths of Ukrainians. The area above the dotted line contains accounts predominantly posting content where Ukrainians are targets and Russians are perpetrators. Conversely, the area below the line displays accounts focused on content where Russians are targeted by Ukrainian forces. If users were merely interested in sharing violent war footage, the points would cluster along the dotted line. However, only two accounts fall near this line, indicating indiscriminate sharing. Most other accounts show a clear bias, suggesting an

intent to emphasise the actions of one side. Motivations for this surely vary, but one possible explanation is that some of the actors are engaged in information influence operations.

The violent content depicting incidents in the Russia-Ukraine war often depicts torture, humiliation, or executions, frequently using derogatory language and celebratory framing. Some illustrative examples classified by our AI-assisted system as targeting Ukraine include a post titled *"Ukrainians get burned alive | They couldn't make it out in time and got baked,"* which depicts extreme violence against Ukrainian victims. A post titled "trench full of dead Ukrainians | *RU POV"* shows a trench full of dead Ukrainians. Another post, *"Ukrainian POW gets his ass kicked,"* shows violence against a captured Ukrainian soldier. Similarly, a sniper attack video titled *"RU Sniper Pair Get A Double Kill"* displays precision strikes against Ukrainian forces. Equally, there are also many examples of violent content targeting Russian forces. One post titled *"Ukraine Death Comp | Invaders Must Die! Slava Ukraine!"* features violent statements alongside images of destruction, explicitly calling for violence against Russian forces. Other examples include the post *"Dead Russian bodies | Unsuccessful Russian assault in Priyutne direction. 8 dead ziggers."* and the post *"Meat grinder for Russian invaders (graphic!)"*, in which the violent deaths of Russian soldiers are glorified, using terms like "Russian cockroach" and expressing enthusiasm for Ukrainian drone strikes.

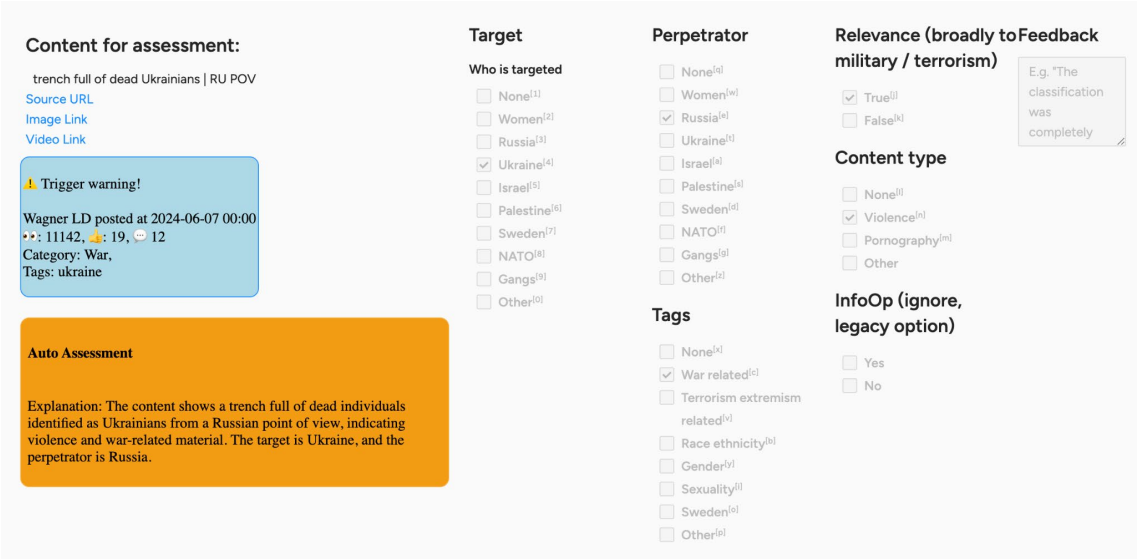


Figure 7: AI classification of post "trench full of dead Ukrainians | RU POV"

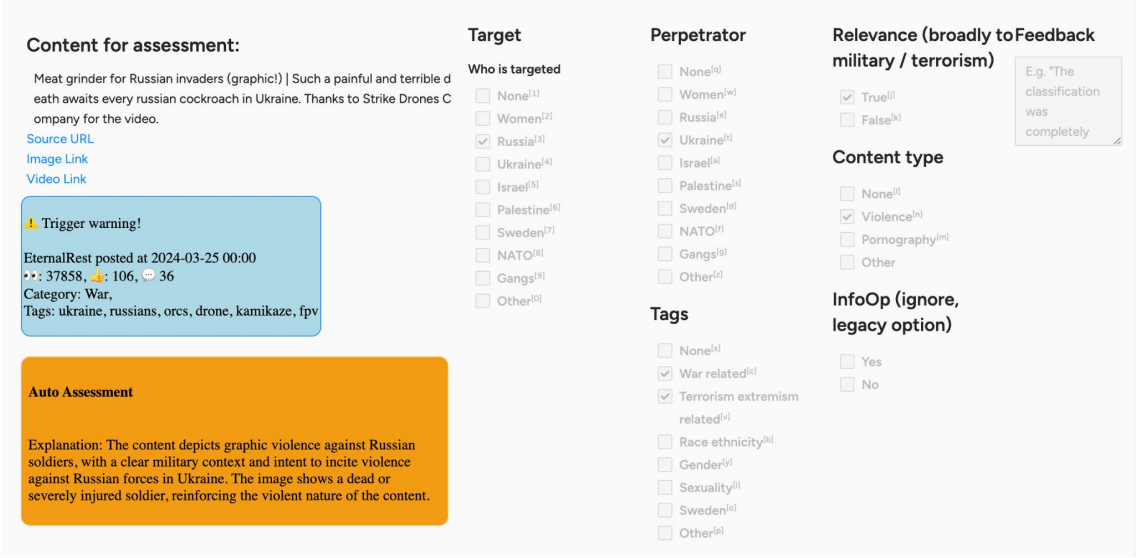


Figure 8: AI classification of post "Meat grinder for Russian invaders (graphic!)"

3.2 Precision Metrics

To evaluate the performance of our AI-assisted classification system, we conducted a double-blind review. In this process, two human annotators independently assessed the same set of data points that the AI had classified. Our objective was to determine how closely the AI's classifications aligned with human judgments across various categories: relevance, tags, perpetrator, and target.

In the relevance category, which assesses whether the content is pertinent to military or terrorist purposes, we observed a strong agreement between the human annotators. They agreed on 94.4% of the cases, with a Cohen's kappa coefficient of 0.87 and an F1 score of 0.94, indicating excellent consistency. When comparing the AI's classifications to those of the human annotators, the AI agreed with human annotator 1 in 75.9% of cases and with annotator 2 in 79.1% of cases.

A closer examination of the disagreements revealed that there were 32 instances where both human annotators agreed the content was not relevant, but the AI classified it as relevant. These are considered false positives. There were zero false negatives—cases where the AI failed to identify content that both human annotators agreed was relevant.

The majority of these false positives involved gang-related violence (examples: "Gang member executed by rival", "Drug dealer killed by police", "Members of the Zicuiran Cartel beheaded"). Such examples clearly have security and terrorism relevance if seen from a particular perspective, and clearer instructions would likely have prevented this systematic error.

Stripping out the gang-related posts we found the AI agreed with annotator 2 96% of the time and with annotator 1 94% of the time. This high level of agreement underscores the AI's reliability in accurately classifying relevant content when ambiguous categories are accounted for.

3.2.1 Perpetrator and Target Categories

In identifying perpetrators, the AI demonstrated high alignment with human judgments. The AI agreed with human annotator 1 in 89.5% of cases and with annotator 2 in 92.2% of cases, with Cohen's kappa coefficients of 0.83 and 0.86 respectively. These results indicate a strong ability of the AI to match human assessments in this category.

For the target category, the AI agreed with annotator 1 in 87.7% of cases and with annotator 2 in 76.3% of cases. The most common disagreements involved the AI identifying 'women' as targets in posts where human annotators did not. Despite these differences, the overall agreement rates suggest that the AI is effective in correctly identifying targets in most instances.

3.2.2 Overall

Agreement between the human annotators and the AI was strongest for content deemed relevant. By contrast, the human annotators disagreed more with each other over content they considered irrelevant. For example, one video featured a woman committing suicide. In this instance, one human annotator identified no target, while the other selected “women.” This is a discrepancy that could easily be resolved with a clearer guidebook.

Within content that both human annotators agreed was relevant, the AI aligned closely with their assessments. It matched their consensus on relevance in 119 out of 121 cases, identified the same perpetrator in 105 out of 107 cases, and agreed on the same target in 93 out of 96 cases. In those few examples where all three assessments differed, the AI tended to be more inclusive, marking content as relevant even when the human annotators did not.

For instance, in three videos - “half naked woman feels angry with her lover,” “female black cannibal feasting on raw meat,” and “thief nearly killed after trying to rob old lady” - the AI labelled “women” as the target, whereas both human annotators selected “none.” In these edge cases, the AI’s cautious approach is beneficial. Rather than risking the omission of potentially significant content, it effectively flags the material for human review if uncertain, reinforcing the reliability of the overall filtering process.

The Cohen's Kappa scores for human-AI agreement were virtually identical to those between the two human annotators. In some categories, the AI even outperformed human

annotation. A Kappa score of 0.8 is considered excellent,[3] and our scores were consistently higher. This demonstrates that AI-generated assessments can be relied upon to a high degree, provided that ongoing monitoring is conducted to adjust and mitigate any systematic errors or misunderstandings. Most errors were edge cases, where a post did not neatly fall into our schema.

[3] Landis and Koch (1977), https://www.jstor.org/stable/2529310

4 Conclusions

This pilot study demonstrates that grey zone domains serve as critical nodes in disseminating war-related and potentially propagandistic content. Our analysis revealed that approximately 27.2% of the content on the examined grey zone domain was related to ongoing conflicts, particularly the wars in Ukraine and Gaza.

These sites do not exist in isolation. The significant viewership of this material, with some posts reaching millions of views, highlights the potential for viral spread beyond these platforms. Users actively cross-share the material from grey zone domains to more accessible digital spaces which amplifies the reach and impact of the harmful content. This cross-platform dissemination makes it difficult for platforms to curb the spread of violent and propagandistic material once it has been seeded in these safer spaces. As our analysis shows substantial war-related content on these platforms, it indicates that they serve as secure and resilient hubs for hosting and circulating war propaganda.

The AI-assisted system we developed effectively minimised human analysts' exposure to harmful content while still maintaining high classification accuracy. The AI demonstrated strong alignment with the annotations made by human analysts, particularly in identifying relevant content, targets, and perpetrators. This approach provides a scalable and reliable method for monitoring hard-to-reach digital spaces where harmful or propagandistic material thrives. This capability can thus shield human analysts from harmful material and filter out the most harmful and violative content for further analysis.

Our findings underscore the importance of further investigating how grey zone domains function within broader information influence operations. These platforms do not operate in isolation; content hosted on them is often cross-shared to more mainstream platforms, amplifying its reach. This suggests that grey zone domains could be deliberately leveraged by state and non-state actors to bypass moderation efforts and disseminate war propaganda or extremist content.

Previous publications in the working paper series:

2024:4 Grahn, H. & Pamment, J., "Exploitation of psychological processes in information influence operations"

2024:3 Nygren, T. & Ecker, U.K.K., "Education as a countermeasure against disinformation"

2024:2 Daskalovski, Z. & Damjanovski, S., "Psychological Defence and Strategic Resilience: North Macedonia's response to hybrid threats and malign foreign influence and interference"

2024:1 Palmertz, B., Weissmann, M., Nilsson, N. & Engvall, J., "Building Resilience and Psychological Defence"

2023:1 Fredheim, R., Ahonen, A., & Pamment, J., "Denying Bucha: The Kremlin's influence tactics in the aftermath of the 2022 Bucha atrocity"

Printed by Media-Tryck, Lund 2024 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Lund University
Faculty of Social Sciences
Psychological Defence Research Institute
Working Paper 2025:1
ISBN 978-91-8104-792-9 (print)
ISBN 978-91-8104-793-6 (electronic)