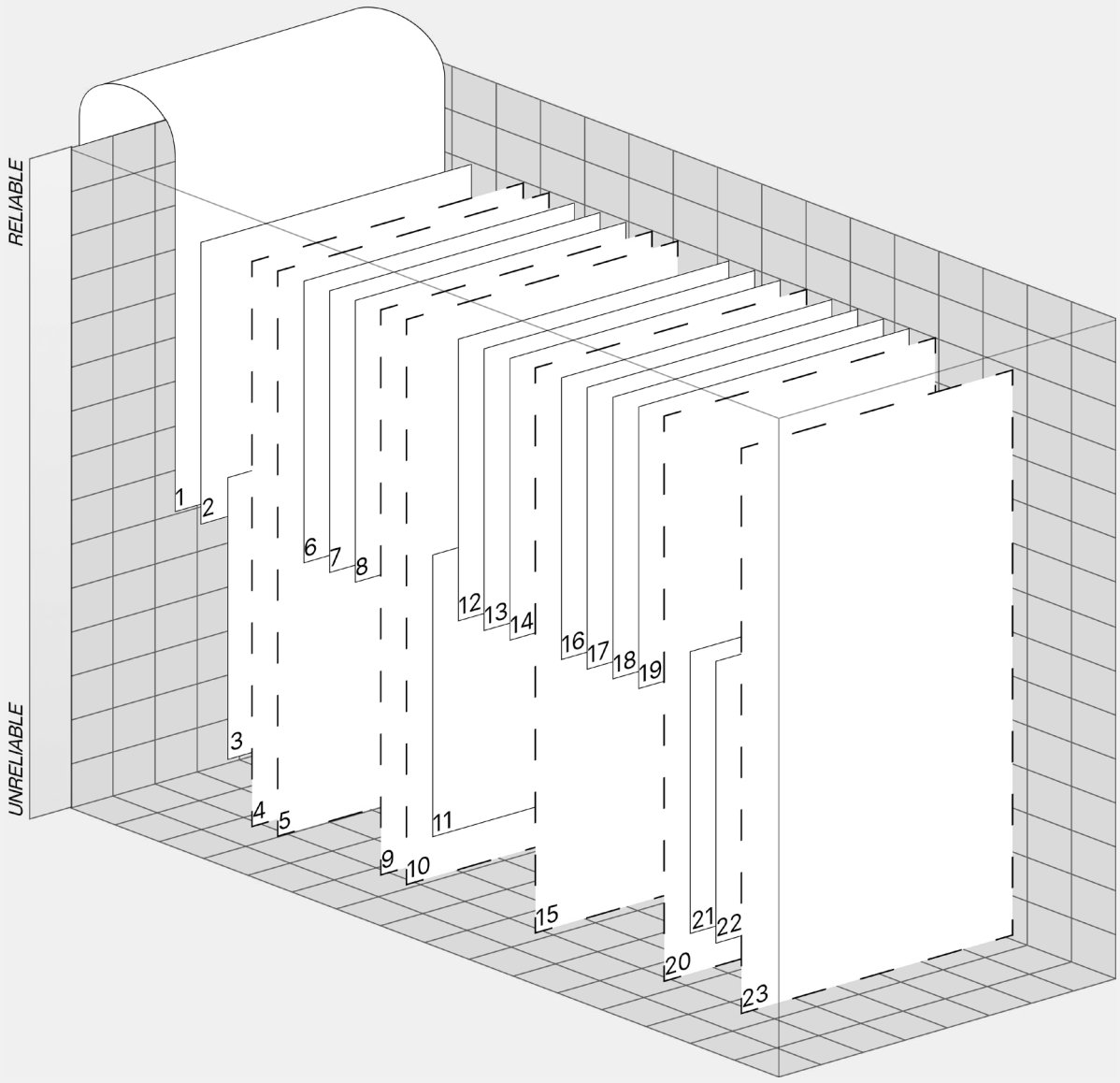
HANDBOOK ON THE USE OF THE AMSTERDAM MATRIX

FOR THE EVALUATION OF THE TRUSTWORTHINESS OF OSINT ON SOCIAL MEDIA PLATFORMS



CIDER
OSINT FOR UKRAINE

OSINT
FOR
UKRAINE

THE HANDBOOK

WHAT IS THIS

A guide to effectively assessing the trustworthiness of (self-declared) OSINT accounts on Twitter/X, including the use of the Amsterdam Matrix developed in collaborative work during the dmi-datasprint.

AUTHORS

Johanna Hiebl, Janthe Van Schaik, Fernando Tabárez Rienzi, Guillen Torres; Deniz Dirisu.

CONTRIBUTORS

The Matrix was initially developed by an international group of researchers and students in cooperation with OSINT for Ukraine including: Lonneke van der Velden, Cees van Spaendonck, Helge Moes, Winnie Lee, Johanna Hiebl, Yamine Mohamed, Vita van Lennep, Maartje Kral, Koen Bruning, Rutger Overstegen, and Viggo Rijswij), facilitated by the researchers Tomás Dodds, Guillén Torres and Joanna Sleigh in cooperation with OSINT for Ukraine (Deniz Dirisu, Daria Delawar).

DESIGNER

Tommaso Prinetti.

INDEX

EXECUTIVE SUMMARY

Within the last few years, Open Source Intelligence (OSINT) has emerged as a transformative tool for information gathering. There has been a dramatic increase in user-generated content on social media, including material explicitly intended to analyze and clarify current events.

However, the changing dynamics of social media, with access to Open Source Information (OSINF), introduce challenges in assessing the trustworthiness of publicly shared content, challenging its usefulness and role within investigations. The handbook on the use of the Amsterdam Matrix for the evaluation of OSINT on social media provides a comprehensive framework for systematically evaluating the reliability of OSINT-labeled posts published on Twitter/X.

Central to this handbook is the Amsterdam Matrix, a tool designed to assess tweet trustworthiness, which a group of scholars and OSINT practitioners and students developed through iterative coding. The project group qualitatively identified 23 parameters categorized by information type, argumentative qualities, textual communication style, and OSINT linguistic traits. Through the development parameters, the goal of the manual is to develop an approach in assessing trustworthiness in the field of OSINT.

The Matrix's application on Twitter/X demonstrates the potential to address challenges faced by practitioners, academics, and journalists to verify OSINT content. However, it's framework can potentially be applied to any other digital platform where OSINT is communicated to open audiences.

Developed during the Digital Methods Winter School in 2024, the Matrix and handbook highlight that evaluating OSINT content trustworthiness depends fundamentally on evidence-based criteria rather than stylistic elements.

EXECUTIVE SUMMARY

The Matrix was further developed into the handbook by OSINT For Ukraine (OFU), highlighting that the methodology could likely serve as a foundation for developing AI-driven tools to automate reliability evaluations, enhancing the speed of the OSINT verification and avoiding online misinformation.

INTRODUCTION

With the lightning speed of the news cycle, especially during conflicts and crises, many turn to social media platforms seeking clarity in the ongoing flux of information. Since the 2010s, Twitter/X in particular has emerged as a crucial platform for OSINT practitioners to share their analyses and insights on global events, conflicts, crimes, and corruptions by publishing and commenting on verified footage from war scenes (Foster & Valcartier, 2013).

This shift highlights the growing interest and demand for open-source research leveraging materials like satellite imagery, flight-tracking websites, and on-the-ground footage. With its open, accessible nature, OSINT invites anyone to participate, constructing a community that, lacking formal qualification requirements, remains mainly unstructured.

THE APPROACH OF CREATING THE AMSTERDAM MATRIX DURING THE DIGITAL METHODS WINTER SCHOOL 2024

During its annual Winter School 2024, the Digital Methods Initiative based at the University of Amsterdam explored these concerns within the context of digital investigations enhanced by artificial intelligence (AI). As AI reshapes the landscape of digital fact-finding and detection and may not only detect deepfakes but also generate them, the Winter School program focused on the impact of these technologies. The project focuses on the transformative impact of OSINT, providing non-state actors with the capabilities to engage in intelligence collection through the Identification of military targets and debunking of false content.

Overall, 21 groups addressed emerging digital investigative epistemologies, including fact-checking, debunking, source verification, and algorithmic auditing, to counter disruptions in the media landscape, from disinformation and content manipulation to ironic trolling and in-

fluence campaigns. Among them, one group tackled the complexity of the information ecosystem and increased the regularity of targeted malign influence operations connected with the growing importance of OSINT amateurs (cf. Joshi, 2023; Cochrane, 2022).

The Matrix encompasses 23 distinct categories to assess the trustworthiness of posts made by OSINT accounts on Twitter/X. By analyzing and iteratively categorizing a dataset of tweets, the project group aimed to define indicators that classify an OSINT tweet as trustworthy or not. To identify the reliability of OSINT content, the group discussed the methods to assess through the following questions:

"What is the type of information that is incorporated in the tweet? How are the argumentative techniques mobilized in the tweet? What is the style of communication that the author uses in the text of the tweet?"

This resulted in developing the following categories: (1) data, information, and sources; (2) argumentation quality; (3) style of communication; and (4) community. These categories were considered the main pillars of the Matrix, which were then split into 23 different weighted (1) reliable; (2) neutral; and (3) unreliable) variables that establish whether a tweet, thread, or account was deemed unreliable.

BROADER IMPLICATIONS AND PRACTICAL APPLICATIONS

The Matrix aims to systematically evaluate the trustworthiness of posts made by OSINT accounts on Twitter/X, offering a structured approach for practitioners, researchers, and platform users. In general, the proposed assessment with the Amsterdam Matrix exemplifies a solution-oriented approach to fill methodological gaps in fact-checking and source verification within journalism, research, and intelligence communities.

With the digital evolution and the development of OSINT as a new informational discipline, this manual provides a theoretical framework on how to evaluate a source's reliability and information credibility based on NATO´s Admiralty Code. In the next step, the methodological development of the Matrix is redrawn, extensively elaborating on the distinct 23 parameters presented in the application.

The annex provides further contextualization about the complexities of open source data and the differences between OSINT, OSINF, and OSINV. We advise readers to use detailed methodology and cheat-sheet whilst evaluating OSINT content, to assess the quality and trustworthiness. This can also be used as a guide when producing OSINT.

PRIVACY AND ETHICAL CHALLENGES OF OSINT

The use of publicly accessible information in producing OSINT raises critical issues of legality, ethics, and privacy. Although accessing, analyzing, and disseminating open data is legal, it can be misused by malicious actors to spread misinformation, sway public sentiment, or engage in detrimental actions, underscoring the need for caution in verifying OSINT. Equally important are the ethical dimensions surrounding the practice. OSINT practitioners must ensure their work serves legitimate purposes and avoids harm, adhering to principles of transparency, fairness, and accountability to maintain trust (OHCHR, 2022).

Privacy concerns are also essential in the ethical deployment of OSINT. Public data ranging from social media activity and public records to digital footprints can expose highly intricate profiles of individuals' habits, interests, and behaviors. While much of this information is voluntarily shared, it is often done so without a full understanding of the potential consequences.

Therefore, OSINT practitioners are required to judiciously weigh their investigative objectives against the imperative to respect individual privacy, ensuring adherence to legal frameworks and ethical standards. The Amsterdam Handbook and Matrix provide guidance and a framework to address these challenges, helping practitioners navigate the complexities of open data while maintaining integrity and credibility in their work.

THE THEORETICAL FOUNDATION: EVALUATING SOURCE RELIABILITY AND INFORMATION CREDIBILITY: NATO'S ADMIRALTY CODE

For open-source investigators, evaluating the reliability of sources and the credibility of the information gathered is critical to building rigorous methods to assess trustworthiness. This evaluation mitigates the risks associated with using low-quality or misleading information in investigations. NATO's information evaluation system, also known as the Admiralty Code, provides a structured framework for assessing both source reliability and information credibility in intelligence assessments.

The dual categories allow us to separate the trustworthiness from the plausibility of the information. Adapted from the British Admiralty system developed during World War II, it categorizes sources and information using an alphanumeric scale: source reliability is rated from A to F, with A being the highest score. In contrast, information credibility is rated from 1 to 6, with 1 being the highest score. This dual-category system separates the quality of the source from the quality of the information, allowing for a more nuanced evaluation.

TABLE 1. NATO SOURCE RELIABILITY AND CREDIBILITY	
SOURCE RELIABILITY	Source reliability refers to the trustworthiness of a source providing information. A reliable source delivers accuracy and dependable data but no source is flawless. All sources, even those perceived as reliable, should be critically evaluated on errors, regardless of whether the information is original, relayed, or beyond the source's usual expertise. Assessing reliability requires a subjective judgment, informed by the source's historical performance and built on consistent interaction over time. Developing a well-informed evaluation of a source demands both critical analysis and experience.
INFORMATION CREDIBILITY	Information credibility refers to the plausibility of the information within the context and involves assessing whether the information is believable, supported by corroborating evidence, and aligned with the existing body of knowledge. Relying solely on corroboration can introduce confirmation bias, necessitating the use of specific analytical techniques to mitigate this risk.

Figure 1. NATO Source Reliability and Credibility (NATO Standardization Office, 2025)

Maintaining high standards in evaluating sources and information is essential in safeguarding against misinformation and disinformation. By focusing on these criteria, investigators ensure their analyses are both accurate and trustworthy, ultimately supporting informed decision-making across various fields. Unlike automated processes, this evaluation requires careful manual analysis to uphold its integrity.

Regarding the Information Evaluation Rubric (Table 2, see next page), a source rated as 'A' is deemed completely reliable, while other credible sources have confirmed information rated as '1'. Conversely, a source rated 'F' indicates that its reliability cannot be judged, and information rated '6' means its truth cannot be determined. This structured approach offers a clear rubric for investigators to evaluate information systematically.

TABLE 2. NATO AJP-2 INFORMATION EVALUATION RUBRIC			
RELIABILITY OF THE SOURCE		CREDIBILITY OF THE INFORMATION	
A	Completely reliable	1	Confirmed by other sources
B	Usually reliable	2	Probably true
C	Fairly reliable	3	Possibly true
D	Not usually reliable	4	Doubtful
E	Unreliable	5	Improbable
F	Reliability cannot be judged	6	Truth cannot be judged

Figure 2. NATO AJP-2 Information Evaluation Rubric (NATO Standardization Office, Figure 2: 2025)

The Admiralty Code is adaptable and can be applied across various intelligence types, from human intelligence (HUMINT) to OSINT. As the global landscape evolves, involving the rise of social media platforms, our source verification processes must be adapted. By adapting the Admiralty dual-layered framework, the Amsterdam Matrix incorporates the systematic methodology to address the unique challenges of assessing the trustworthiness of OSINT, specifically on the Twitter/X platform.

METHODOLOGICAL APPROACH FOR DEVELOPING THE MATRIX

To get a clear scope of what constitutes OSINT practices on Twitter/X, a multi-step approach was employed to classify tweets and accounts that claim to produce OSINT and contribute to conversations about OSINT.

Initially, the group began with an explorative familiarization of the pre-collected database of tweets with the hashtag #OSINT provided and scraped with the 4CAT-Tool[1]. In contrast, others simultaneously examined tweets from OSINT accounts pre-identified by laying the foundation for identifying distinguished patterns to classify a tweet as trustworthy/non-trustworthy tweets.

1 4CAT is a tool used to process and analyse data from online social media platforms

Based on this exploratory practice, a manual content analysis was deemed to be the most suitable method to break down tweets into variables that signify trustworthiness or lack thereof. Therefore, the group dissected the anatomy of OSINT tweets by identifying nuanced variables within the tweet or account that signaled degrees of trustworthiness (as green flags, such as source links, expert knowledge and untrustworthiness ("red flags", such as the use of AI-created content, excessive hashtags).

These variables were then compiled into a structured framework with clear labels and explanatory classifications. Each variable could be sorted into one of these three categories:

1) Reliable/Trustworthy/Present, (2) Incomplete, and (3) Unreliable/Untrustworthy/Absent, iteratively refined through group discussion and sample testing. This framework was then tested and improved by the group collectively. They tested the parameters for one tweet to

ensure tht different individuals can use the Matrix to obtain consistent conclusions. If the majority of the coders gave the same ratings to the same post, it showed that the parameters allowed reliable identification.

In another step, every group member coded selected tweets individually, and discrepancies again were solved collaboratively through group discussion. Furthermore, to test and make final revisions, the refined Matrix was applied to 250 pre-selected tweets collected via double coding to verify the results. For the Matrix evaluation, 0 is incomplete, -1 is false, and 1 is true.

This demonstrated the Matrix's applicability and extension to evaluate Twitter/X accounts and resulted in the final Matrix.

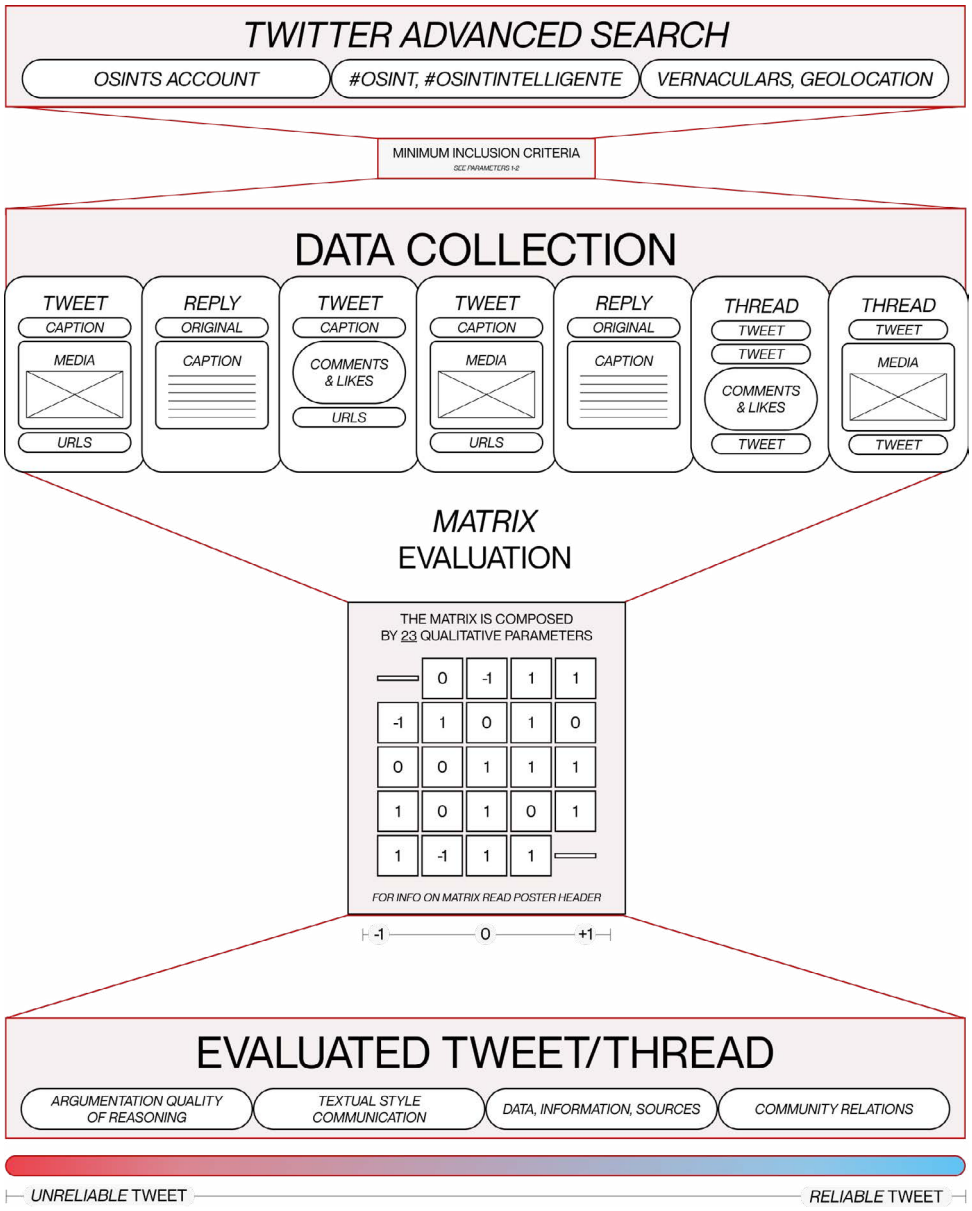


Figure 3. Matrix's Original Blueprint

THE MATRIX IN APPLICATION: EXAMPLES FROM TWITTER/X

Through iterative coding, the project group qualitatively derived 23 different parameters. Each parameter represents an indicator that suggests that an OSINT tweet is trustworthy/untrustworthy. The parameters focus on the categories of what type of information is included in a tweet, what kind of argumentative techniques are employed, the communicative style that is used in the textual part of a tweet, and the extent to which characteristics of a specific OSINT linguistic variety are present.

These variables are ordered but not equally weighted or unambiguously descriptive, so manual definitions had to be agreed upon. Each parameter signals whether a tweet aligns with credible OSINT practices.

Overall, the 23 distinct parameters can be divided into five distinct categories of analysis, which are inclusion criteria (1-2), data, information, and sources (2-7), the written argumentative quality (8-13), the style of written communication (14-19), and the community (15-23).

The parameters are visualized as follows: For each of the 23 individual parameters, an assessment is made as to whether they are present and thus contribute to the trustworthiness of a tweet. If the parameters are predominantly present, they are marked in blue, which indicates a high level of trustworthiness within a tweet. If certain parameters are not applicable or are even missing, they are marked in red. If it is not possible to make a clear assessment, the parameters are marked in gray.

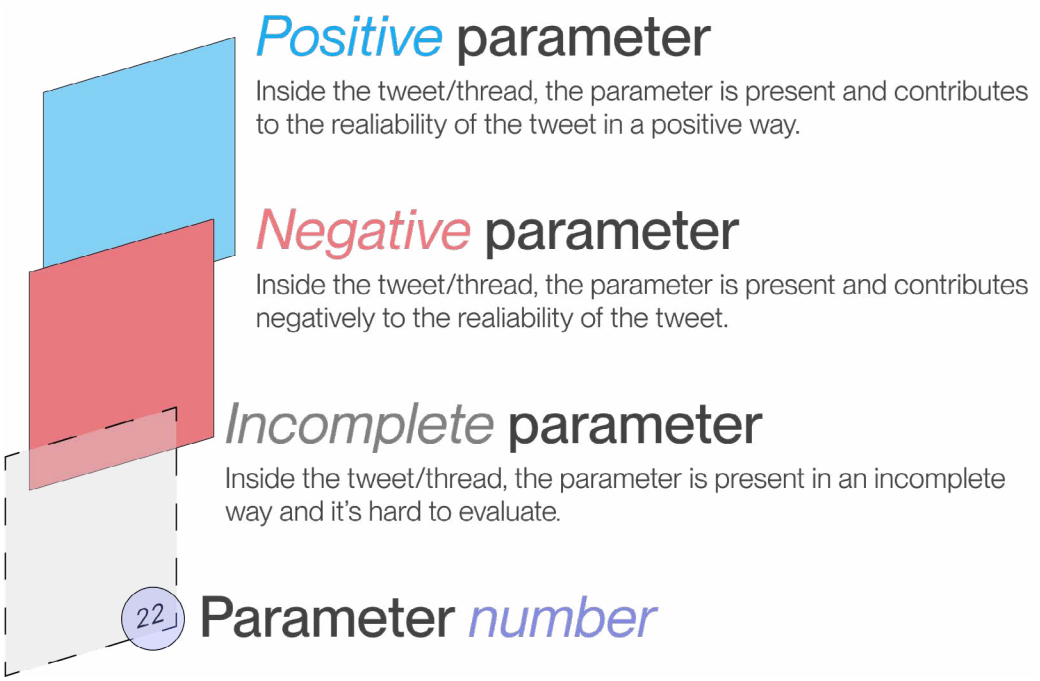


Figure 4. Classification of present parameters

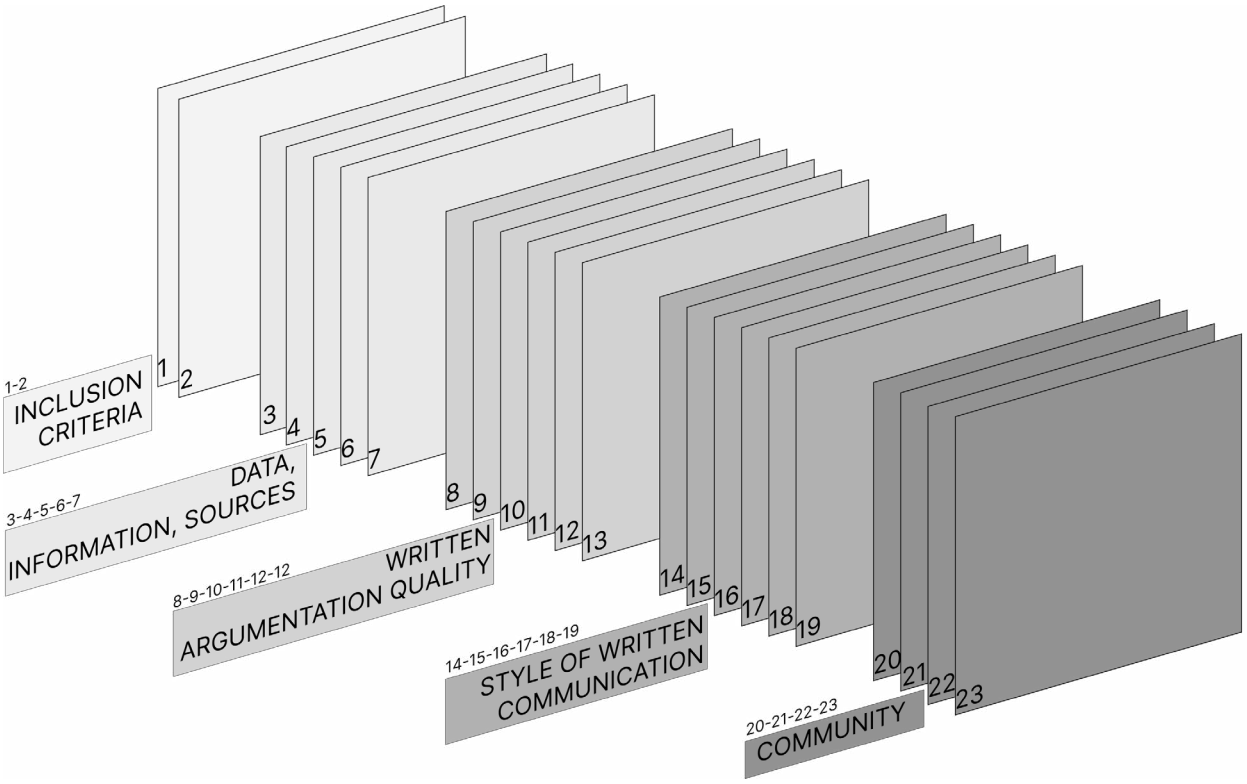


Figure 5. Clusters of distinct analysis parameters

The separation between the individual parameters indicates the division into distinct analysis categories (Inclusion criteria, data, information, and sources (2-7), the written argumentative quality (8-13), the style of written communication (14-19) and the community (15-23).

THE INCLUSION CRITERIA (1-2):

1 // ***PRESENCE OF RELEVANT DATA OR INFORMATION***
A tweet includes factual data and/or information related to the claims made (e.g., GPS location alongside a photo), establishing a basis for its trustworthiness.

2 // ***ADDS AN INSIGHT***
The tweet adds unique insights, such as verified details, that distinguish it from untrustworthy or purely speculative content.

DATA, INFORMATION AND SOURCE INTEGRITY (3-9):

3 // ***LINKS TO SOURCES***
A hyperlink (as a source) leads the viewer to a credible source of information supporting the data or claims made.

4 // ***PRESENCE OF CREDITS***
Referencing a source as an explicit statement clarifying the information's origin. The absence of a clear source suggests that the source is untrustworthy.

5 // ***ANNOTATED CONTENT***
Audio-visual media is annotated by an author to substantiate claims made about the data and/or information. Content that is not annotated or is distracting indicates untrustworthiness.

6 // ***MANIPULATION***
Deliberate manipulation of media to deceive leads to untrustworthiness.

7 // ***CREATING NOISE***
Excessive elements (such as emojis, too many hashtags, AI-generated elements, or watermarks) or other non-relevant content not supporting a claim could be distracting for the user and reduce trustworthiness.

8 // *SPECIALIZED KNOWLEDGE*

A tweet provides specific information that makes a claim stronger, such as geolocation or an identified weapon or vehicle as specialized military expertise. Such specific information suggests that the author of a tweet has expertise in a certain domain and strengthens trustworthiness.

9 // *DISCLOSURE OF THE METHOD*

The author outlines or references the method(s) used and how certain claims were constructed. A lack of transparency in communicating the method signals untrustworthiness.

WRITTEN ARGUMENTATIVE QUALITY (10-13):

10 // *LOGICAL ARGUMENTATION*

Outlined claims follow a logical sequence, and relevant data and/or information is provided. If there is no logic and coherence to the argument, this is an indicator of untrustworthiness. This does not include logical fallacies.

11 // *ACKNOWLEDGING INFORMATION GAPS*

Potential gaps in the argumentation, data, and/or information are acknowledged in trustworthy OSINT. If such gaps stay unacknowledged/unmentioned, this is an indicator of untrustworthiness.

12 // *RESULT IS REPLICABLE*

Trustworthy OSINT investigations provide enough detailed steps of information for a reader to replicate the findings and make the same claim(s). Insufficient details or other information that leads to different claims suggests lower trustworthiness.

13 // *SELF-AFFIRMATION AS ARGUMENT*

Over-emphasis on the author's credibility as a primary argument and means to justify or try to influence or persuade the reader to believe a certain argument/claim is not trustworthy, e.g., terms such as "trust me" suggest persuasion rather than evidence-based reasoning.

STYLE OF WRITTEN COMMUNICATION (14-19):

14 // *EXCESSIVE USE OF HASHTAGS*

Describes overusing hashtags or using such hashtags that do not relate to the presented content at all, which creates problematic effects on trustworthiness. Excessive hashtags suggest the intent to amplify rather than inform.

15 // *LOGICAL STRUCTURE*

Trustworthy OSINT structures into logically built, coherent, and well-structured sentences to present information so that the argument is clear to follow. If the argument is not presented that way, this is untrustworthy.

16 // *EMOTIVE LANGUAGE*

Using language that is appealing to the reader creates untrustworthiness because emotive language can hinder or distract from the argumentation.

17 // *VICTIM NARRATIVES*

A communication style falling into victimization language is untrustworthy. This regards, e.g., the suffering of children or oppression by colonialists etc., which skew perception and reduce objectivity.

18 // *HARMFUL LANGUAGE*

Language that incites violence on groups or individuals, any racist or sexist undertones, etc., undermines trustworthiness.

19 // *SARCASM*

Using a tone of irony or sarcastic language on a large scale distracts from the argument or building a conclusion and is particularly untrustworthy if this is the dominant tone of an OSINT post.

COMMUNITY VERNACULAR (20-23):

20 // *REFERENCING OSINT TOOLS/TECHNOLOGIES*

Trustworthy OSINT provides links or references to tools and technologies that other OSINTers can use for their investigation.

21 // *COMMUNITY VERNACULAR USAGE*

Trustworthy OSINT correctly uses informal linguistic standards, which, by referring to the language or variation spoken, suggests familiarity with the OSINT community that outsiders won't necessarily be able to relate to. This concerns an OSINT-related tone and is not related to specialized knowledge. Untrustworthy OSINT tweets lack or wrongly apply the vernacular.

22 // *TAGGING REPUTABLE THIRD PARTIES*

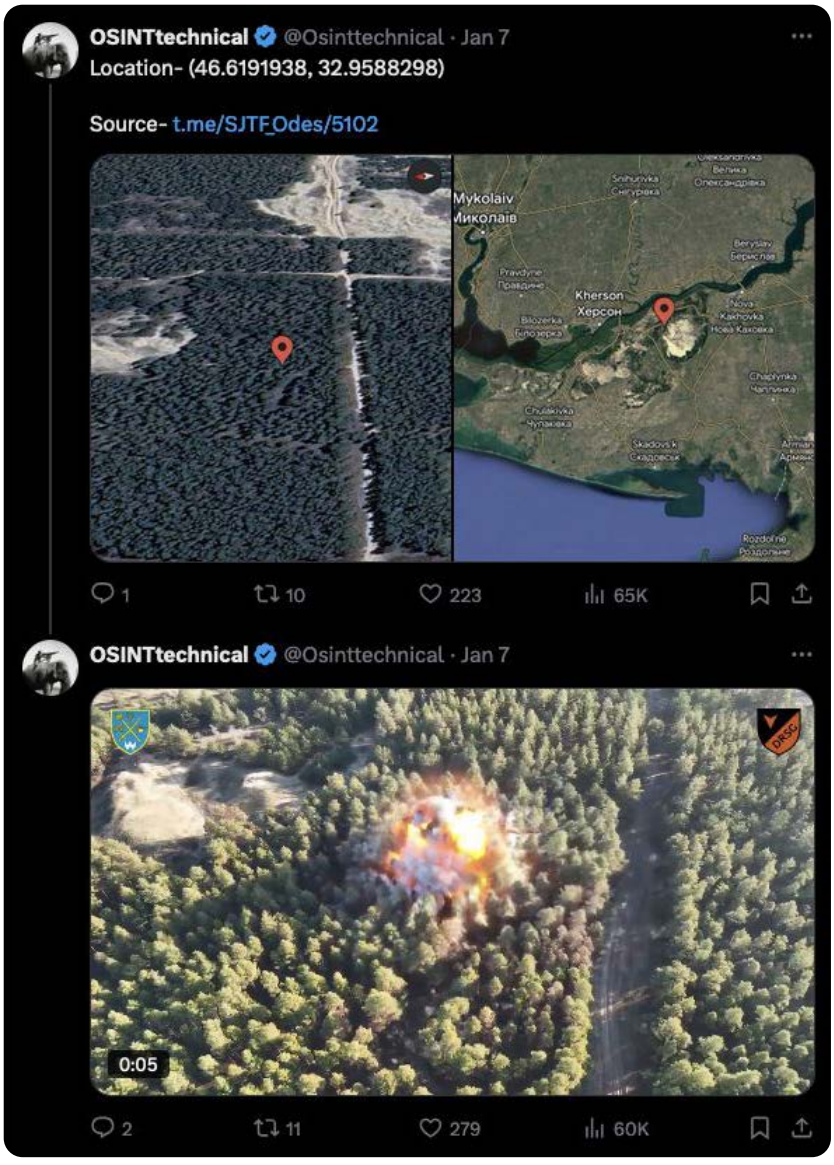
Tagging third parties on Twitter/X using a handle (@) creates trustworthiness, as a user opens up to scrutiny, which is a sign that information can be validated or reviewed. One example is tagging reputable sources, such as @Geoconfirmed.

23 // *ENGAGEMENT (LIKES, RETWEETS)*

Trustworthy tweets receive feedback through likes and retweets, which are not from the same author who published the tweet. High-quality engagement suggests trustworthiness.

EXEMPLARY CASE APPLYING THE MATRIX

To test the Matrix with its parameter categories, the group collaboratively engaged in a test run in which they analyzed the following thread on Twitter/X, which was posted by the account **@OSINTtechnical** in English on 7 January 2024.



https://twitter.com/technical/status/1743934358673649996

Figure 6. Exemplary tweet from OSINT technical's account

EXEMPLARY CASE APPLYING THE MATRIX

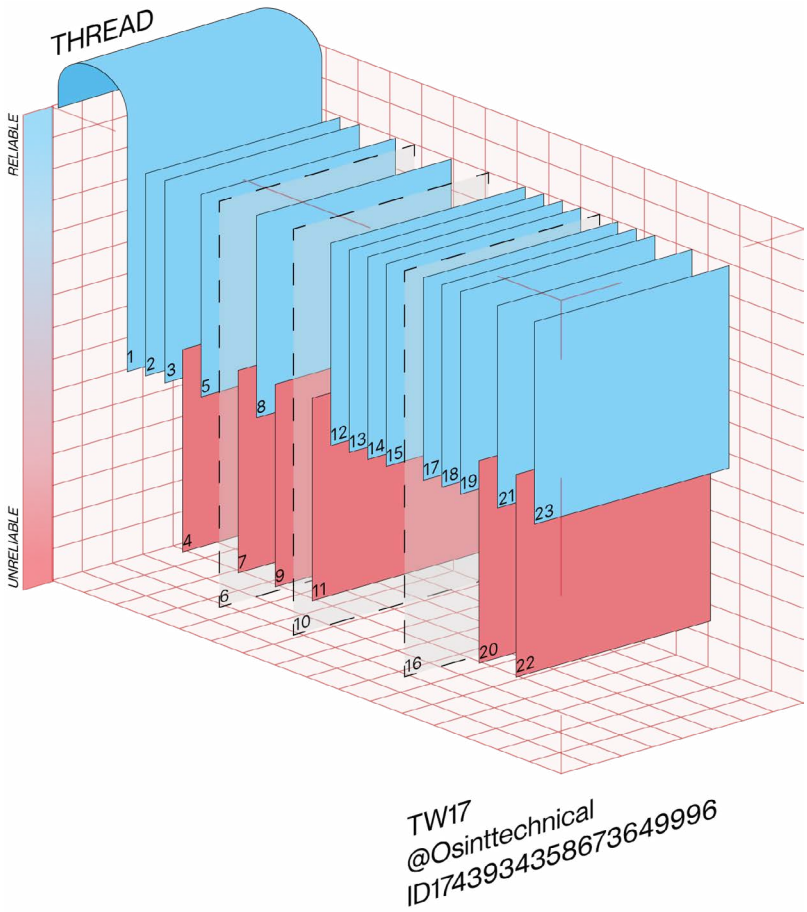


Figure 7. Visualization of applying the Matrix to the tweet from @OSINTtechnical

Regarding inclusion criteria, the group detected the presence of relevant information by the provision of geolocation coordinates and corresponding map visuals that establish a factual basis for a claim. This fulfills the criterion of including relevant information (1) and is therefore evaluated as trustworthy/blue. By showing annotated map coordinates and explosive visuals, the thread adds unique insights (2) to a potential military strike or incident beyond speculative information and is therefore evaluated as trustworthy/blue as well.

Regarding the second category of Data, Information, and Source Integrity, the thread includes a link to the source of the Telegram Channel of Add *(Defense Forces of Southern Ukraine / Сили оборони Півдня Україні)* and points to additional evidence, which enhances trustworthiness and is therefore evaluated as present/blue. Although the original appearance of the content is mentioned, there is no explicit mention of credits (3) to original data contributors or third-party verified content and is therefore evaluated as missing/red.

EXEMPLARY CASE APPLYING THE MATRIX

Furthermore, the map visuals are annotated (4), showing specific geolocation markers. This substantiates the claims visually and is therefore evaluated as present/blue. During collaborative coding, the group did not detect signs of deliberate manipulation (e.g., fabricated images or falsified data) evident in the content provided; therefore evaluated as not manipulated (5) and marked blue. The content of the thread is evaluated as direct and free of excessive elements like emojis or irrelevant hashtags, and the message appears focused and concise, not creating noise (7) and marked blue.

The usage of the map overlays and accurate coordinates, which demonstrates expertise and specialized knowledge (8) in geolocation analysis, is common in OSINT practices and therefore evaluated as present/blue. To finish the block on data integrity as a category, there is no detailed explanation of the methods used for geolocation of verification (9) disclosed. In this case, for instance, matching images with satellite views could have been used and is therefore evaluated as missing/red.

Regarding the third category of written argumentative quality, the tweet presents information in a logical sequence (10), mentioning the geolocation coordinates and depicting the map and related visuals to create coherence between the data points, and is therefore evaluated as present/blue. However, the tweet does not acknowledge potential gaps and limitations (12), such as the lack of independent verification, and is therefore evaluated as missing/red. Moreover, the exact methods for obtaining the insights and the final results are not explained, so the results don't seem replicable (13) to a reader.

Nevertheless, in an OSINT community vernacular, there might be a common understanding of how to verify the coordinates independently, and that's why this parameter was evaluated as partially present/grey. To conclude the category of the written argumentative quality, the example refrains from relying on self-affirmation as an argument (14) to create personal credibility but focuses on naming evidence. Therefore, this parameter is marked as absent/blue.

In the fourth category, the style of written communication is evaluated. For instance, in this thread, there is no excessive use of hashtags (14), and is therefore marked as absent/blue. The thread appears to be structured logically (15), drawing a clear connection between the coordinates, maps, and visuals, and is therefore evaluated as present/blue.

EXEMPLARY CASE APPLYING THE MATRIX

The language used in the post is neutral and avoids emotional appeals and is therefore evaluated in the absence of emotive language (16) and marked blue. The same accounts for the absence of narratives invoking victimization and emotional bias (17) and harmful or incendiary language (18) and is marked as absent/blue as well. Lastly, the tone remains factual and avoids irony or sarcasm (19).

In the fifth category, the parameters of belonging to a certain community vernacular of OSINT are evaluated. The tweet does not explicitly reference specific OSINT tools or technologies (20) used for analysis and is therefore marked as missing/red. The post example reflects standard OSINT practice, i.e., the usage of geolocation and satellite imagery, which represents the usage of a community vernacular (21) and is marked as present/blue. Neither third-party accounts nor entities are tagged for verification or additional insights (22), which is evaluated as missing/red. Lastly, the tweet thread has a significant number of likes, retweets, and comments, and therefore, the engagement metrics suggest credibility (23), which is evaluated as present/blue.

In summary, through the application of the Matrix, there were 16 parameters evaluated as present/blue with factual content, logical argumentation, and visual substantiation avoiding content manipulation and excessive emotional appeals. Nevertheless, with six parameters stated as missing/red, there can be some weaknesses detected, such as the lack of transparency regarding methods, the absence of distributed credits, or collaboration with the content-origin, which weakens the overall credibility.

This showcase of the application of the Amsterdam Matrix allowed us to systematically assess the trustworthiness of an OSINT post on Twitter/X and shed light on possible steps that would create a higher level of trustworthiness, such as disclosing the method used in detail and including a brief explanation of geolocation and verification techniques. Moreover, listing original datasets, maps, and tools being used could enhance replicability. Likewise, tagging third parties would illustrate engagement with reputable OSINT accounts. Lastly, the mention of any potential uncertainties or limitations in the presented post to highlight gaps in the investigation would heighten the level of trustworthiness.

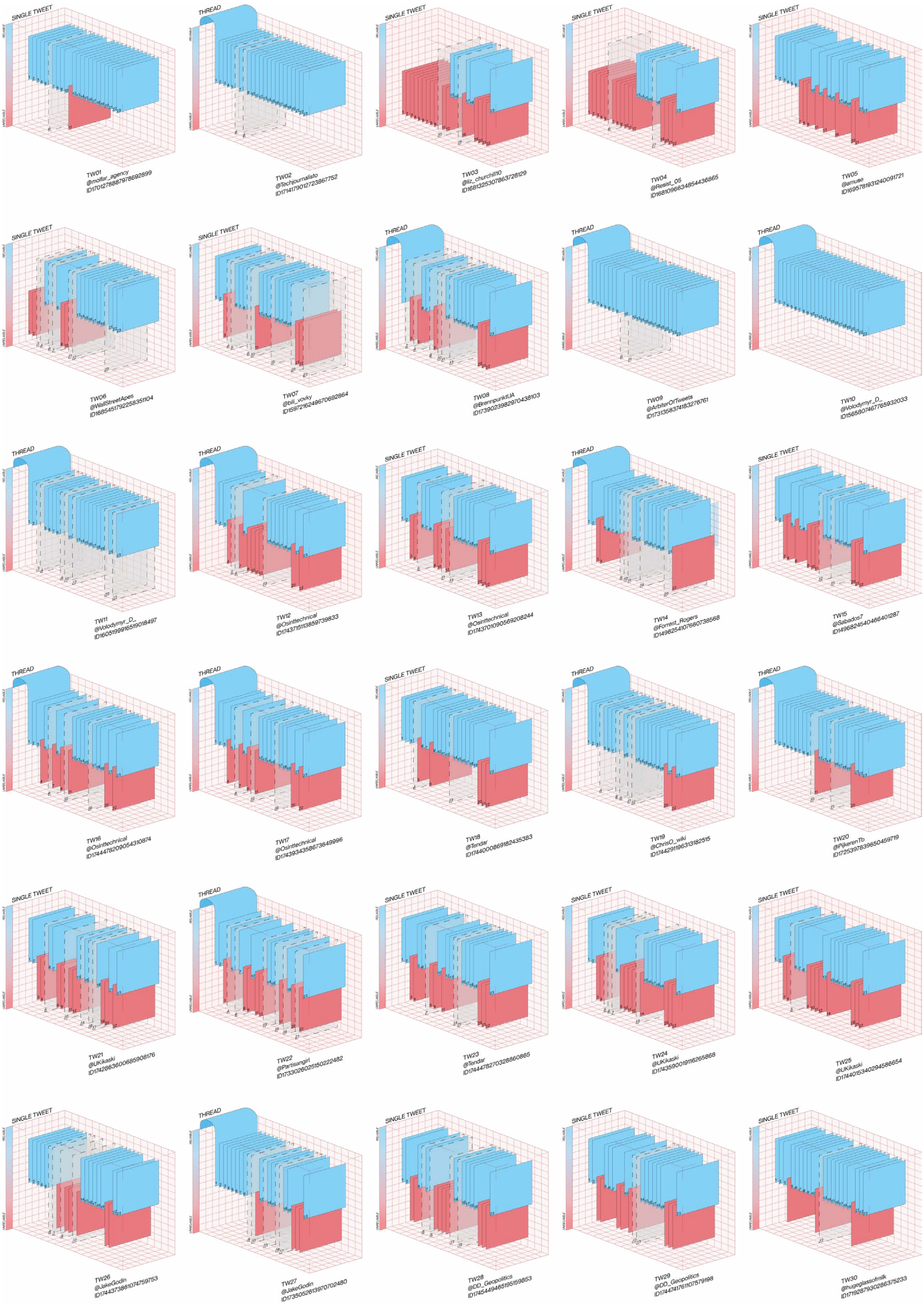


Figure 8. Original complete mosaic of all the matrixes

DISCUSSION

FINDINGS

The Amsterdam Matrix offers critical insights into the nuanced practice of Open Source Intelligence (OSINT) on Twitter/X, demonstrating the multifaceted challenges of evaluating the trustworthiness of OSINT tweets. Through assessing simplistic notions that credibility rests on a singular characteristic, the findings advocate for a multidimensional and systematic approach. The interconnected variables underpin credibility assessment in a dynamic online environment.

To be able to evaluate the complex nature of OSINT tweets effectively, a nuanced evaluation framework to ensure an accurate reliability assessment is required. This is underscored by the difficulty of validating, sometimes self-attributed OSINT content under the hashtag #OSINT On Twitter/X, especially with restrictive platform affordances that hinder external referencing or extensive discussions.

The Amsterdam Matrix was developed to address these challenges by offering a strategy for assessing reliable OSINT effectively and concisely, tailored to platform constraints. This highlights the importance of an adaptive approach from the OSINT community in adjusting their communicative practices to coherent and universal practices and, therefore, establishing an evaluation framework for OSINT practitioners, researchers, and analysts.

In the course of developing the Amsterdam Matrix to assess the trustworthiness of OSINT content on Twitter/X, an iterative and collaborative coding process was employed as the key methodology. This method not only emphasized the indispensability of a multifaceted approach for accurate reliability assessment but also proved essential for defining the parameters of effective OSINT practice. By continuously refining evaluation criteria and methodologies, the focus shifted to the adaptability required for credibility assessment, which elevates its importance in distinguishing trustworthy OSINT from misinformation. This process of manual content analysis illuminated the heterogeneous demands involved in assessing the trustworthiness of dis-

seminated OSINT tweets, offering critical insights that contribute to the development of a more robust credibility evaluation framework.

The investigation of OSINT tweets and threads began with the assumption that an elaborate communication style used in the posts correlates with a high level of trustworthiness. Interestingly, the study did not reveal a consistent correlation between the communication style in OSINT tweets and their trustworthiness. Whether formal or informal communication styles were used in the posts, they failed to indicate the credibility of the content.

Overall, this is only one example from the Matrix serving as a critical reminder that assessing trustworthiness in OSINT content relies solely on evidence-based approaches rather than on stylistic cues.

LIMITATIONS OF THE AMSTERDAM MATRIX

While the Amsterdam Matrix presents a systematic approach for evaluating the trustworthiness of OSINT posts on Twitter/X, it is essential to acknowledge its limitations. These limitations stem from the exclusive focus on Twitter/X as a platform-specific issue, manual content analysis, or subjectivity from coders as constraints regarding the scope and method. Moreover, the issues of anonymity inherent in OSINT practices, driven by operational and information security concerns, as well as personal security concerns of OSINTers, add another layer of complexity to our findings and determine what is a credible OSINT post.

SCOPE AND METHODOLOGICAL CONSTRAINTS

Although our initial application of the Amsterdam Matrix focused on Twitter/X, the framework itself is flexible and can be refined for use on a range of social media platforms such as Instagram and Facebook. By extending its scope beyond Twitter, the Matrix has the potential to enhance our understanding of OSINT across diverse sectors. Nonetheless, because this framework concentrated on a single platform, further validation on additional networks will be important for broadening the Matrix's generalizability.

Additionally, the manually compiled data collection consisted of approximately 250 posts that were qualitatively analyzed in depth by the whole group. As collaborative coding procedures are relatively time-consuming, the group stuck to this rather small sample. This is also the reason why larger sample sizes and automated methods were not yet integrated. Still, future studies could benefit thereof to be able to capture broader trends and enhance statistical validity

PLATFORM CONSTRAINTS COMMUNICATION STYLES AND CREDIBILITY

The coding and analyzing procedure also visualized platform-specific challenges that undermine the trustworthiness of OSINT content: Twitter/X itself hinders the production of reliable OSINT content by limiting the number of signs, formatting, or censorship of sensible content. The platform's inherent architecture, therefore, limits the provision of comprehensive context and authentication. As observed in the coded posts, there is a lack of threaded discussion or linking external sources, such as if footage stems from an external Telegram channel, which thereby worsens the credibility dilemma (Theocharis et al., 2023).

For instance, Twitter/X's verification system, which is symbolized by a blue tick as a purchasable checkmark, yet it defeats the purpose itself, as the label can be purchased by anyone. While verification could have been an indicator of authenticity, the commercialization of the verification label has made it insecure to misuse and enables less credible accounts to appear more authoritative. Therefore, even verified accounts may disseminate dubious information or misleading content, such as memes or jokes, underscoring the need for a more robust framework to evaluate trustworthiness.

Another example revealed by the Matrix is the lack of correlation between the communication style used in OSINT tweets and their trustworthiness. For instance, tweets using numerous hashtags, such as #OSINT, might be considered deceiving, suggesting that stylistic elements can sometimes mislead rather than inform. Conversely, the coded examples appeared more trustworthy if they were a thread rather than a single tweet, likely due to their allowance for a more structured argumentation and provision of further context to support claims made.

AI-GENERATED CONTENT AS AN EMERGING CHALLENGE

Lastly, the establishment of the Matrix also detected the increasing integration of artificial intelligence-generated content in OSINT posts as a red flag. AI-generated videos and images can be technologically advanced, present a demanding challenge to assess trustworthiness as their undue reliance or even misuse might lower the credibility of an OSINT post. Based on the analyzed posts, it became clear that, manipulated or AI-generated images can potentially mislead audiences and undermine the integrity of the information presented. These findings emphasize the importance of vigilance and critical and manual evaluation when encountering AI-generated content in OSINT tweets.

CONCLUSION

Despite the above-mentioned limitations, the Amsterdam Matrix and the qualitative, collaborative exploration of the data set provide notable takeaways about the challenges and opportunities associated with OSINT on Twitter/X and social media platforms more broadly. The rise of OSINT as a valuable method for information gathering has brought with it the necessity to carefully assess the credibility of content disseminated on social media platforms.

The Amsterdam Matrix creates a first approach in comprehensively assessing the reliability of content. The findings highlight the difficulties in determining the trustworthiness of OSINT content, revealing the need for a holistic and nuanced approach. Systematically and iteratively, critical indicators were identified into categories such as data, information, sources, argumentation quality, style of communication, and community, emphasizing the multi-sided factors that influence trustworthiness but also offering a structured tool for assessing trustworthiness in OSINT posts.

The Matrix reveals that the style of communication alone does not correlate with trustworthiness, challenging assumptions about the role of rhetoric in evaluating OSINT. Moreover, the research underscores the difficulty of presenting reliable OSINT practices on Twitter/X without resorting to threads or external links, emphasizing the importance of understanding the boundaries and practices within the OSINT community.

The Matrix presented in this research can serve as a practical framework for practitioners, researchers, and platform users. Beyond its immediate applications, the developed Matrix can serve as a guiding framework for various stakeholders, including journalists, fact-checkers, academics, and other information consumers.

As the digital landscape continues to evolve and the reliance on open-source information grows, having a structured and comprehensive tool becomes increasingly vital for responsible information consump-

tion. Journalists, in particular, can benefit from our Matrix as a guideline to enhance the rigor of their investigative processes by verifying claimed OSINT content.

With a systematic approach to evaluating OSINT content, fact-checkers can strengthen efforts to combat misinformation and disinformation. Academics who delve into social media dynamics and information dissemination can use the presented findings to inform their research and build upon the executed methodology, applying collaborative coding procedures.

The dynamic nature of social media and open-source information requires adaptive and innovative solutions. Furthermore, this could result in the establishment of clear guidelines for OSINT practitioners on the transparency of research, operational security when navigating the internet, and the ethical implications of sharing information that is intentionally or unintentionally nontrustworthy.

Looking ahead, the ultimate goal is to leverage our research to contribute to developing AI tools capable of automating the checking process. By training AI models on the identified reliability indicators identified within the Matrix, these tools could be empowered to distinguish between trustworthy and unreliable OSINT content and, therefore accelerate the verification process, enabling faster and more precise distribution of information.

Future directions of research can develop more comprehensive methodologies for assessing the reliability of OSINT content on other platforms, recognizing the interplay between communication style, platform constraints, and trustworthiness evaluation by building upon the current Matrix. Cross-platform analysis research could broaden the understanding of the evolution of source verification over social media by including platforms such as TikTok or Mastodon, their algorithmic behavior, and the algorithms. As OSINT content does not exclusively appear on Twitter/X anymore, this would allow a deeper understanding of how platform affordances influence content dissemination and the patterns of algorithmic influence.

The presented Amsterdam Matrix proves that establishing the reliability of OSINT in social media requires recognizing the fact that reliability is not determined by a single factor only. It demands a comprehensive and nuanced examination of all categories within the Matrix.

Nevertheless, it is essential to acknowledge that certain categories may carry more weight than others during this holistic analysis, underscoring the need for a balanced consideration of each factor's importance. The Matrix thus represents an important stepping stone toward developing a robust framework for systematically assessing the reliability and validity of OSINT content on platforms like Twitter/X, ultimately aiding in the verification process and mitigating the spread of misinformation.

REFERENCES AND BIBLIOGRAPHY

Block, Ludo. 2021.
"GDPR Essentials for OSINT Research" Blockint, July 28, 2021.
https://www.blockint.nl/methods/gdpr-essentials-for-osint-research/

Coulthart, Stephen, and Brian Nussbaum. 2022.
A definition of open source intelligence.
The Open Source Intelligence Laboratory.

Forrester, Bruce, and Defence R&D Canada Valcartier. 2013.
"Twitter as a Source for Actionable Intelligence".
Presented at the 18th Command and Control Research and Technology Symposium.

Gerashchenko, Anton (@Gerashchenko_en). 2024
"Photos appeared online that reportedly show the results of last night's missile attack on Belbek airfield in occupied Crimea. This looks like a radar from an S-400 complex. There has been unconfirmed information that a MiG-31K aircraft has been struck."
Twitter/X, May 15, 2024.
https://x.com/Gerashchenko_en/status/1790732932518216093

Heuer, Richards J., Jr. 2007.
Improving Intelligence Analysis with ACH. Pherson Associates.
February 21, 2007.
https://pherson.org/wp-content/uploads/2013/06/Improving-Intelligence-Analysis-with-ACH.pdf

Hribar, Gorazd, Iztok Podbregar, and Tjaša Ivanuša. 2014.
"OSINT: A 'Grey Zone'?"
International Journal of Intelligence and CounterIntelligence 27, no. 3: 529–549.
https://doi.org/10.1080/08850607.2014.900295

REFERENCES AND BIBLIOGRAPHY

Iversen, T. A. 2023.
"Autumn Approaches: Part 3" Reports by T. A. Iversen (newsletter), September 8, 2023.
https://thelookoutn.substack.com/p/autumn-approaches-part-3

Lee, Tristan, Kolina Koltai, and Giancarlo Fiorella. 2024.
"OSHIT: Seven Deadly Sins of Bad Open Source Research." Bellingcat, April 25, 2024.
https://www.bellingcat.com/resources/2024/04/25/oshit-seven-deadly-sins-of-bad-o pen-source-research/

Manjhu, Karan. 2023.
"Twitter is closing free access to its API starting Feb. 9". TechStory, February 3, 2023.
https://techstory.in/twitter-is-closing-free-access-to-its-api-starting-feb-9/

Meyer, Chris. 2021.
"The Intelligence Cycle: How to Process Information Like an Analyst." The Mind Collection, August 22, 2021.
https://themindcollection.com/intelligence-cycle-how-to-process-information-like-an- analyst/

North Atlantic Treaty Organization. 2016.
NATO Standard AJP-2: Allied Joint Doctrine for Intelligence, Counter-intelligence and Security. NATO.

Office of the Director of National Intelligence. 2022.
Intelligence Community Directive 203 Amendment: Analytic Standards. December 21, 2022.
https://www.dni.gov/files/documents/ICD/ICD-203_TA_Analytic_Standards_21_Dec_2022.pdf

Office of the United Nations High Commissioner for Human Rights (OHCHR). 2024.
Berkeley Protocol on Digital Open Source Investigations. Geneva: OHCHR.
https://www.ohchr.org/sites/default/files/2024-01/OHCHR_Berkeley-Protocol.pdf.

REFERENCES AND BIBLIOGRAPHY

Reuser, Arno. 2024.
“On the Difference Between OSINT, OSINF, Information, Intelligence.” February 8, 2024.
https://opensourceintelligence.biz/on-the-difference-between-osint-osinf-information

NATO Standardization Office 2025.
AJP-2.1: Allied Joint Doctrine for Intelligence Procedures. Edition B, Version 1. Ratification Draft 1. NATO UNCLASSIFIED.
Accessed April 5, 2025..
https://jadl.act.nato.int/ILIAS/data/testclient/lm_data/lm_152845/Linear/JISR04222102/sharedFiles/AJP21.pdf

Theocharis, Yannis, Shelley Boulianne, Karolina Koc-Michalska, and Bruce Bimber 2023.
“Platform Affordances and Political Participation: How Social Media Reshape Political Engagement.“ West European Politics 46, no. 4 (2023): 788–811
https://doi.org/10.1080/01402382.2022.2087410.

CHEATSHEET: THE MATRIX

INTRODUCTION

The Matrix presented in this handbook is a practical guide to assess the credibility of OSINT content on Twitter/X, offering a structured framework for practitioners, researchers, and platform users.

CLAIM

Sub-category	Description	Check
Are there claims being made?*	If there are no claims, the tweet should not be included. (1=yes or 0=no)	☐
Status of data and/or information	Data and/or information related to the claim(s) made are provided. For example, GPS location is given about a photo.	☐
Links to sources	A source (hyperlink) is provided that takes the viewer to the source of information and/or data.	☐
Provides credits	References a source (not a hyperlink but that clarifies the source of information and/or data) or provides a reference that is not complete (i.e, someone told me...)	☐

STYLE OF COMMUNICATION

Sub-category	Description	Check
Excessive use of hashtags	Use of hashtags that do not relate to the content, or that do not support the content or consumption	☐
Logical structure	The sentences and presentation of information is structured so that the argument is clear to follow.	☐
No overuse of emotive language	Emotive language appeals to the emotions of the reader, which hinders or distracts from the argumentation or the conclusions	☐
No victim literature	The communication style does not fall into victimization styles. For example, emphasizing the suffering of children or the oppression by colonialists, etc	☐
No harmful language	There is no language that incites violence in groups or individuals. Also, the language does not include any racist or sexist undertones, etc	☐
No overuse of sarcasm	There is no use of irony or sarcastic language that distracts from the conclusion or argument.	☐

COMMUNITY

Sub-category	Description	Check
Share details on OSINT tools/technologies	Provides links or references to tools and technologies that other OSINTers can use for investigations	☐
Uses community vernacular	Effective/correct use of community vernacular, which refers to the language or dialect spoken by a community, that another community won't necessarily understand.	☐
Tagging verified third parties (i.e., @GeoConfirmed)	Tagging third parties on Twitter/X using a handle (@). This is considered a reliable characteristic, as the user opens themselves up to scrutiny, which is a sign that the information can be validated/reviewed. For example, tagging reputable sources such as: @GeoConfirmed.	☐
Some engagement (likes, retweets)	The content received feedback through likes and retweets. The retweets and or likes are not simply from the same author	☐

ARGUMENTATION QUALITY

Sub-category	Description	Check
Adds an insight	An insight to the data and/or information is provided, one that is more than a repetition of the information. It adds new information that doesn't exist anymore. For example, a verification could be considered an insight.	☐
Adds specialised knowledge	Provides specific information such as geolocation, a weapon, or a vehicle, which suggests that the person has expertise in a specific domain.	☐
Method is communicated clearly	A method is provided and described, or referenced in a link.	☐
The argumentation is logical	The conclusion follows logically from the claims and the provision of data and/or information. There are no logical fallacies.	☐
Admits gaps in information	Acknowledgement of potential gaps in the argumentation, data, and or information	☐
Result is replicable	There is enough information for another person to redo the investigation and achieve the same claim	☐
Use of self-affirmation of argument	Over-emphasis on the author's credibility as a means to justify the argument or claim. For example using terms such as "I know so", "trust me", etc.	☐

DATA-INFORMATION SOURCES

Sub-category	Description	Check
Annotated content	Media (photo or video) is annotated by the author to support the claims made about the data and/or information.	☐
Manipulation	Potential signs of media manipulation to deceive.	☐
Noise (text, emoticons, media, AI, watermarks...)	Addition of elements or non-relevant content that could be distracting for the user or do not support the claim(s). For example, text, emoticons, media, AI, watermarks, etc.	☐

NAVIGATING THE COMPLEXITIES OF OPEN SOURCE DATA: DISTINGUISHING OSINT, OSINF AND OSINV

The Matrix is designed to address the diverse range of information quality encountered across open sources on the internet. The Amsterdam Handbook defines Open Source Intelligence (OSINT) as the collection and analysis of open-source information to produce actionable insights, including monitoring ongoing events, ensuring the credibility of information, or detecting emergent threats. Additionally, open-source information is "publicly available information that any member of the public can observe, purchase or request without requiring special legal status or unauthorized access" (Berkeley Protocol on Digital Open Source Investigations, 2022).

Traditionally, OSINT was considered a branch of military intelligence. However, with widespread social media, internet access, and technological advancements in the 2010s, its application has expanded significantly. OSINT has since become integral to investigative journalism, news reporting, and international justice, complementing conventional investigative methods. Built on the Berkley Protocol (2022) definition, OSINT is transforming, increasingly becoming the primary means of information gathering and investigation for a diverse array of actors, reflecting its growing importance and effectiveness in the digital age. In an era of information characterized by high volume, variety, and velocity, and where news breaks at lightning speed, many turn to open-source accounts and experts for clarity. This shift highlights the growing credibility and demand for open-source research — leveraging tools like satellite imagery, flight-tracking websites, and on-the-ground footage. It's accessible, public, and open to anyone. Yet, because of this, there is also room for the inexperienced and devious to add to the information pool. It is thus important to become acquainted with the different types of information and understand the principles by which an open-source investigator can advance trustworthy, accurate, and reproducible findings. Understanding the nuances between Open Source Intelligence (OSINT), Open Source Information

(OSINF), and Open Source Investigation (OSINV) aids in assessing the trustworthiness of OSINT accounts on Twitter/X. This is a higher-order classification, whereas our Matrix disaggregates components for closer examination. Therefore, this classification establishes the theoretical basics.

OPEN SOURCE INFORMATION (OSINF)

OSINF encompasses information available freely or commercially from various sources, including the Internet, physical media, and telecommunications. Examples include newspapers, broadcasts, and voicemails. OSINF serves to inform the general public and can be utilized by various entities. For example, an OSINF investigator might process combat footage to identify specific military units or weapons and their locations, as shown in Figure 9.

Figure 9. OSINF example by Anton Gerashchenko

1) However, such data lack contextualisation within a broader framework and, most importantly, lack actionable insights.

Within the intelligence production cycle (as demonstrated in Figure 10), OSINF covers the collection and processing stages, providing raw, uncontextualized data (Meyer, 2014). However, such data lacks the broader context and actionable insights that OSINT provides.

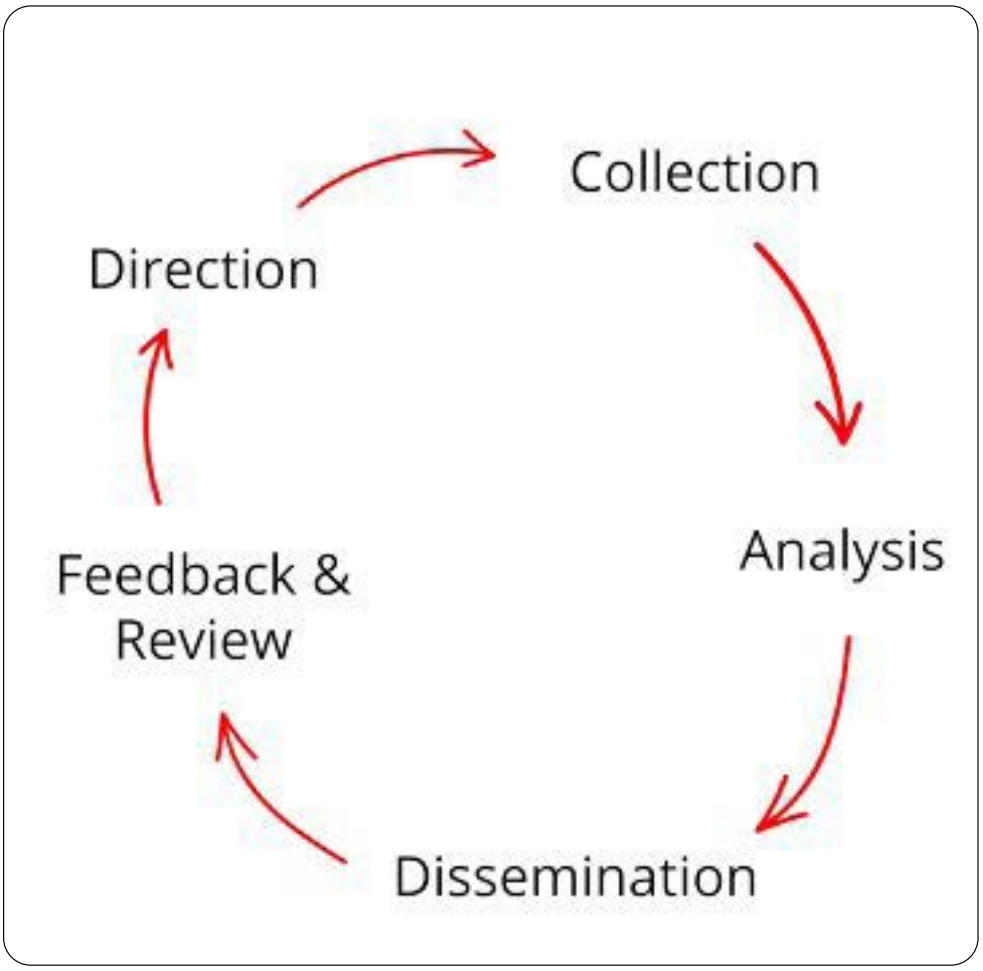


Figure 10. The intelligence production cycle (Meyer, 2014)

OSINT involves collecting various pieces of OSINF, gathering the main takeaways from each verified piece of information, and using that to craft an analytical hypothesis. In turn, multiple hypotheses are tested for robustness against each other, a process known as Analysis of Competing Hypotheses[2]. While OSINT and OSINF can share the same data sources and are related in their use of publicly available data, they are not synonymous; OSINT transforms OSINF into a coherent narrative that meets specific intelligence needs, aiding national security, law enforcement, and investigative journalism.

[2] See Heuer Jr. *Improving Intelligence Analysis with ACH*

OSINT carries on to the next stage of production: analyzing public data to generate intelligence that fulfills a pre-directed intelligence requirement (Figure 11). OSINT uses a variety of sources to provide a narrative analysis presenting a logical argument of synthesized evidence. OSINT has applications across diverse professional spheres, such as national security, law enforcement, core international crimes investigations, and investigative journalism, as well as aiding in strategic decision-making.

The main distinction between OSINF and OSINT lies in their purposes: OSINF informs, while OSINT facilitates strategic decision-making and investigative actions. OSINF includes social media posts and news reports, providing basic details for further analysis. OSINT synthesizes this information into actionable insights. While OSINF targets a broad audience, OSINT is tailored.

From Dublin to the US

A profile at myvisajobs.com provides a biography for Cherkasov's cover identity: he worked as a travel agent in Brazil before attending Trinity College Dublin as an undergraduate studying political science (2014-18), then getting his Master's at Johns Hopkins' SAIS programme (2018-20).

Victor Muller — Contact Victor Muller

Phone:	verified employer only
Email:	verified employer only
Street:	verified employer only
From:	member or employer only
Languages Spoken:	member or employer only

Career Profile

Degree:	Master's Degree
Career Level:	Entry Level
Occupation:	Life, Physical, and Social science
Target Title:	Analyst, Research Assistant
Target Locations:	22209 , Washington, DC , New York, NY
Skills:	Conflict; Reasearch; International Relations; Foreign Policy; Social Movements; Elections;

Screenshot of the MyVisaJobs profile for "Victor Muller" – the cover identity of Russian GRU asset Sergey Cherkasov.

In 2017, when starting his last year at Trinity College Dublin, "Victor" started a blog focused on geopolitics called "Politics of Us".

While at first glance the blog seems incredibly popular with thousands of comments on each article, in fact the vast majority of interactions appear to be spam with very little organic engagement. The ideology expressed in the blog would appear milquetoast for a Western political science student without any pro-Russian leanings – discussing methods of increasing democracy in developing nations, calling Putin a "cancer", and focusing on the importance of grassroots organisations fostering peace in Africa.

Victor Muller

I am a Johns Hopkins University MA student of the School of Advanced International Studies. I hope that my blog will contribute towards the

Figure 11. OSINT example by Bellingcat

OPEN SOURCE INVESTIGATION (OSINV)

Open Source Investigation (OSINV) arises from non-state actors using OSINT techniques for digital investigations. OSINV is driven by editorial considerations rather than official directives, representing civil society's effort to uncover truths that may be politically sensitive. This is a term that was created in response to the increasing use of OSINT by institutions outside of traditional intelligence agencies. It reflects the evolution of technology in enabling groups outside government to make use of open source data for discovering information on matters of human rights violations, abuses of power, and other ways of speaking truth to power. Rather than strategic considerations, OSINV is driven by a desire for justice against a perceived inability or negligence by governments to do what is morally correct.

OSINV encompasses diverse sources such as social media content, satellite imagery, video recordings, governmental records, and academic literature. The democratization of investigative tools has significantly empowered journalists, researchers, advocates, and advocacy organizations to illuminate urgent global challenges, including human rights abuses, wartime atrocities, corruption, disinformation strategies, and environmental degradation. By integrating disparate pieces of information, OSINV practitioners can uncover patterns and evidence that might otherwise remain obscured, thereby contesting prevailing narratives and holding powerful entities accountable.

A distinguishing feature of OSINV is its commitment to open data and its collaborative framework. Investigations typically engage a decentralized network of contributors from various backgrounds, utilizing collective knowledge and interdisciplinary methodologies to corroborate findings.

This crowdsourced methodology not only enhances the credibility of OSINV activities but also broadens its impact within a globalized and interconnected digital environment. Furthermore, the outcomes of these investigations are often disseminated publicly, promoting increased transparency and facilitating broader societal interaction with critical issues. In the digital era, OSINV has risen to prominence as a significant field, particularly as the advent of information technology has blurred the distinction between formal and informal investigatory processes. It serves as a clear illustration of how civil society can

function as a vigilant observer, holding governments, corporations, and other powerful actors responsible for their actions in a world that is increasingly intricate and saturated with data.

Having established definitions, it is also important to be aware of the practical and ethical challenges in open-source research. These include ensuring data is reliable, combating misinformation, and adhering to legal standards (particularly regarding data protection). Open source investigators must conduct due diligence: verification of OSINF is time-consuming and requires solid evidence of authenticity. Unethical practices, known as 'Grey OSINT', involve using OSINT techniques that may breach legal boundaries, highlighting the importance of strict adherence to professional standards (Hribar et al., 2014).

EXTENSIVE METHODOLOGY FOR DEVELOPING THE MATRIX

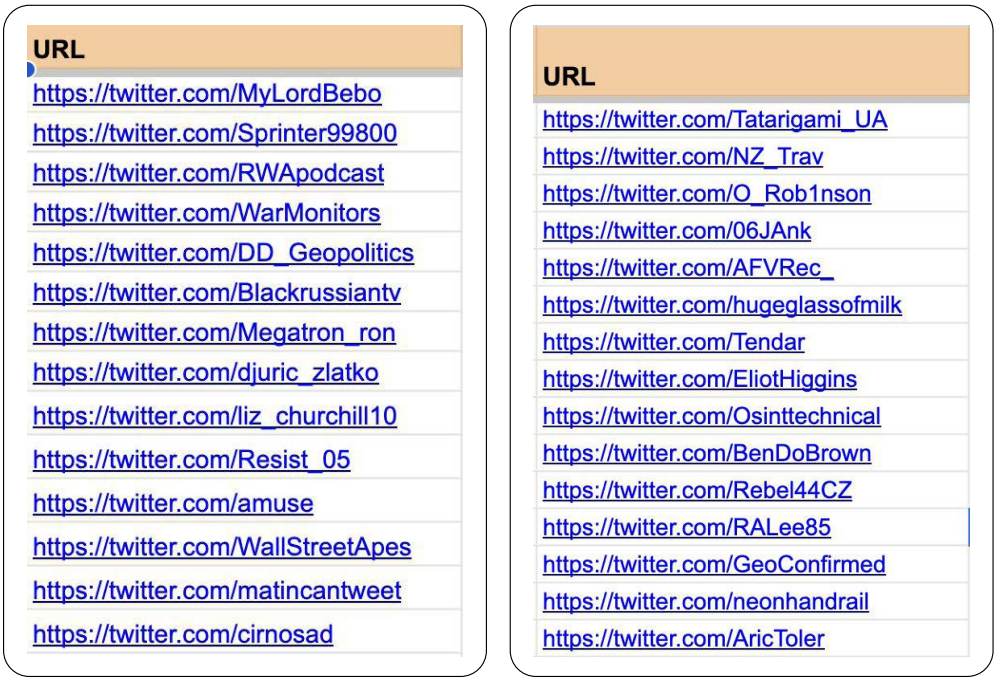
Building on the principles of source reliability and information credibility discussed in the previous section, this methodology focuses on adapting those evaluation frameworks to the specific challenges posed by social media platforms, particularly Twitter/X. While NATO's Admiralty Code provides a robust foundation for evaluating intelligence sources, the nature of open-source intelligence (OSINT) on social media, where information is often rapidly disseminated, unverified, and decentralized, necessitates a more specialized tool.

The Matrix we developed addresses this need by offering a systematic approach to evaluate the reliability of Twitter/X accounts and the credibility of their content. By applying the theoretical principles of source evaluation to the fast-paced, highly variable environment of OSINT on social media, this methodology ensures that investigators can reliably distinguish between trustworthy and untrustworthy sources in an ever-changing digital landscape. The following steps outline the systematic approach used to create and validate the Matrix, with the ultimate goal of providing a reliable framework for coding and analyzing tweets involved in OSINT practices.

A systematic, multi-step approach was employed to classify tweets and Twitter/X accounts engaged in OSINT practices.

STEP 1: DEFINING BOUNDARIES: FAMILIARISATION WITH OSINT TWEETS AND ACCOUNTS

In the initial step, the group began by familiarising themselves with OSINT tweets and accounts using two approaches: One subgroup explored a pre-collected dataset of tweets with the hashtag #OSINT to identify posts reporting on falsely reported events. The dataset "#OSINT" derives from 4CAT and can be accessed on request by contacting contact@osintforukraine.com. The source of the data is the Twitter API (v2) Search. 4CAT here queried the hashtag #OSINT on Twitter between 2020 and 2022, resulting in a dataset of around 1 million tweets. Simultaneously, another subgroup examined pre-selected OSINT accounts, pre-identified by OSINT for Ukraine either as problematic or trustworthy and investigated related accounts on Twitter/X.



URL
https://twitter.com/MyLordBebo
https://twitter.com/Sprinter99800
https://twitter.com/RWApodcast
https://twitter.com/WarMonitors
https://twitter.com/DD_Geopolitics
https://twitter.com/Blackrussiantv
https://twitter.com/Megatron_ron
https://twitter.com/djuric_zlatko
https://twitter.com/liz_churchill10
https://twitter.com/Resist_05
https://twitter.com/amuse
https://twitter.com/WallStreetApes
https://twitter.com/matincantweet
https://twitter.com/cirnosad

URL
https://twitter.com/Tatarigami_UA
https://twitter.com/NZ_Trav
https://twitter.com/O_Rob1nson
https://twitter.com/06JAnk
https://twitter.com/AFVRec_
https://twitter.com/hugeglassofmilk
https://twitter.com/Tendar
https://twitter.com/EliotHiggins
https://twitter.com/Osinttechnical
https://twitter.com/BenDoBrown
https://twitter.com/Rebel44CZ
https://twitter.com/RALee85
https://twitter.com/GeoConfirmed
https://twitter.com/neonhandrail
https://twitter.com/AricToler

Figure 12. Screenshot from proceedings of the Working Group OSINTer CENSUS during Data Sprint, January 2024

Both approaches (4CAT-derived data set and pre-selected accounts) involved filtering out tweets that did not mention specific events, as these were not useful for understanding the distinctions between good and bad OSINT accounts.

STEP 2: DISSECTING THE ANATOMY OF OSINT TWEETS

In the second step, the group analyzed OSINT tweets by identifying variables within the accounts' tweets that signaled degrees of trustworthiness/non-trustworthiness (e.g., use of AI images, excessive hashtags, and source links). Due to the closure of the Twitter API on 9 February 2023[3], this process involved iterative manual coding by the participants. Participants collaboratively defined and annotated variables that caught their attention, such as types of images, narratives, links, mentions, and hashtags. Later, individual notes were taken on how these variables were compared between true and false accounts.

3

Manjhu, K.,
Twitter is closing free access to its API starting Feb. 9

The following visualization illustrates the iterative process in data collection from Twitter/X and deriving categories for the Matrix, which were then evaluated by its 23 parameters and classified as unreliable or reliable tweets.

STEP 3: CREATING THE MATRIX

The variables were collectively identified through discussion and exploration of type images, narratives, hashtags, and mentions that called the attention of the group. These were compiled into a set of variables, such as hyperlinking / providing credits; Details on tools/ technologies; Distinction between assessment and opinion; Replicability; Audience Engagement; Emotive language, Sarcasm; Self-affirmation; Profile description; Spelling; AI-generated content; and manipulation of information. This set of variables was then input into ChatGPT to create a preliminary Matrix based on the above-mentioned variables with an explanation of how they can be 'red flags' or 'green flags, ' alongside a description of how they can be used to create truth or suggest untrustworthiness. The generated Matrix was refined and edited, whereby the hierarchy was evaluated, examples provided for each aspect, and iteratively tested on one sample, focusing on tweet threads that help create or verify narratives based on data, expertise, or assessment.

This process was initially carried out through discussions in the group and through the documentation of sticky notes and later digitized into the form of a working Matrix.

Aspect	Green Flags	Red Flags	Impact ranking	Example (Link)
Provision of information	Provides an assessment of data that is reasonable and insightful	Blurs fact and opinion	Adds an insight	An insight to the data and/or information is provided, one that is more than a repetition of the information. It adds new information that doesn't exist anymore. For example, a verification could be considered an insight.
Specialized knowledge	Shows depth of understanding in a field	Demonstrates lack of expertise or provides no new knowledge	High	
Hyperlinking and credits sources	Cites sources that are reputable	Provides no sources or links to unverified or biased sources	High	
Sources	Cites credible and diverse sources	No citations or reliance on questionable sources	High	

Figure 13. Identified "green flags| and "red flags" by the working group during the Digital Methods Winter School 2024

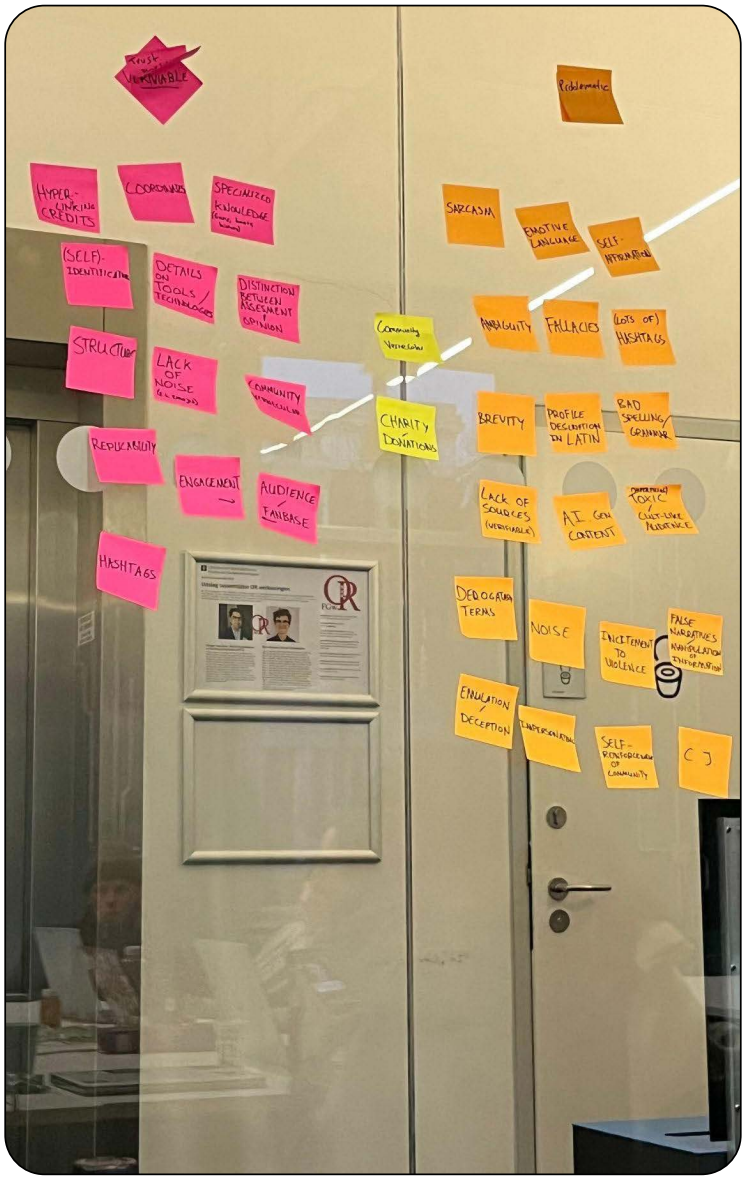


Figure 14. First classification as collaborative brainstorming during Winter School 2024 of what marks a good and bad OSINT post. Copyright: Johanna Hiebl

STEP 4: VALIDATION THROUGH INTERCODER RELIABILITY TESTING

The Matrix underwent validation through intercoder reliability testing, with discrepancies resolved through iterative group discussion. Each group member individually coded one tweet by parameters of (1) Reliable/Trustworthy/Present; (2) Incomplete; and 3 (Unreliable/Untrustworthy/Absent). Concretely, this resulted in the rating of a single parameter rated as “1” making the tweet point towards trustworthiness; if a single parameter is rated as “3”, it makes the tweet point to untrustworthy. By comparing the individual coding, discrepancies were identified, and the classification was defined where necessary. The intercoder reliability was checked twice, using different Twitter threads each time.

STEP 5: APPLICATION OF THE MATRIX TO CODE A DATASET

The refined Matrix was then applied to code a dataset of OSINT tweets collected through snowball sampling, resulting in a dataset of 250 tweets from OSINT accounts. Multiple coders reapplied the updated Matrix and analyzed the tweets, with double coding performed to verify the results.

STEP 6: ANALYSIS OF OSINT ACCOUNTS

Finally, accounts actively producing OSINT tweets were analyzed based on the identified variables (e.g., username, Twitter ID, profile description, picture, verified profile status, tags, hashtags, external links, institutional affiliation, engagement metrics, bio details, and email address). This analysis was conducted on a sample of 250 tweets from the previous dataset from OSINT accounts.