Media and Communication
2026 • Volume 14 • Article 12460
https://doi.org/10.17645/mac.12460

cogitatio

ARTICLE

Open Access Journal

Piloting the Integration of AI-Driven Detection Methods to Counter Disinformation in Organizational Processes

Lucas Stampe and Christian Grimme

Department of Information Systems, University of Münster, Germany

Correspondence: Lucas Stampe (lucas.stampe@uni-muenster.de)

Submitted: 27 March 2026 **Accepted:** 8 May 2026 **Published:** 22 July 2026

Issue: This article is part of the issue "The Role of AI for Counter Speech: Detection, Intervention, and Risks" edited by Lena Frischlich (University of Southern Denmark – Odense), Cathy Buerger (Dangerous Speech Project), and Magdalena Obermaier (LMU Munich), fully open access at https://doi.org/10.17645/mac.i500

Abstract

With the rise of generative AI, the increasing threat of automatically generated uncivil content (including misinformation for information warfare up to cyber-bullying purposes) makes the protection of open online discourse even more pressing than before. In the implementation of measures for countering these threats, the different intervention objectives of stakeholders, their workflows, and IT support need to be considered. Stakeholders include online social network moderators, journalists, fact-checkers, social listeners, as well as authorities and organizations with safety- and security-related tasks. Given the sheer volume of online social network content, automated or community-based solutions are required to detect (automated) misinformation. However, detection solutions are predominantly message-focused and target end-users, leaving experts without systematic, large-scale perspectives on coordinated disinformation campaigns to guide countermeasures. To bridge this gap, we conduct an expert-centered study on the integration of detection methods proposed by the research community. Our contributions are threefold: (a) We adopt disinformation features from a previous study and draw connections to literature on detection methods; (b) semi-structured interviews yield vignettes that expose the spectrum of goals, constraints, tools, and decision-making processes employed by experts, informing requirements for method integration; (c) we design a demonstrator that showcases representative methods to uncover unexplored concepts, probe affordances and limitations in context, and evaluate conceptual fit during the interviews. Together, these steps bridge the gap between data- and AI-driven detection techniques from research and the practical needs of diverse stakeholders confronting targeted and large-scale disinformation.

Keywords

artificial intelligence; disinformation; disinformation detection; disinformation intervention; open-source intelligence; social media

1. Introduction

With the rise of social media in almost every aspect of life and work, there is a growing need for computer-assisted analysis of the data circulating on these platforms, particularly in the context of open-source intelligence (OSINT). OSINT is the collection and analysis of publicly available information from open sources such as (online) media, social networks, public records, and publications (Pastor-Galindo et al., 2020). Experts in this field deal with public security in general, but also with the protection of private companies (e.g., to prevent damage to reputation or products), individual citizens (such as the defense against cyberbullying, doxxing, harmful speech, etc.), or even entire nations (Pastor-Galindo et al., 2020). The scope and complexity of the analysis range from distinguishing true from false information to identifying artefacts, actors, intentions, and campaigns (Browne et al., 2024), often with the aim of triggering a legal response.

Based on analyses and observations, both researchers and societal stakeholders have already proposed a wide range of intervention strategies, including warnings about and corrections to misinformation, indicators for detecting breaches of norms (e.g., harmful speech), and for highlighting relevant content (Hartwig et al., 2024). However, these approaches are generally directed towards informing and protecting end-users. They usually do not, or only to a limited extent, address the information needs and scope for action of OSINT analysts or stakeholders relying on OSINT support in protecting against complex attacks. At the same time, these practitioners, researchers, and society must keep pace with the rapidly evolving forms of inauthentic behavior on online social networks (OSNs) to continue to ensure open public discourse on social media (Unger et al., 2025).

For example, the complexity and dynamics of disinformation campaigns are inherently difficult to detect (Quandt & Klapproth, 2023)—a challenge that can be addressed through technological advances such as (generative) AI, which is at the same time exacerbated by (generative) AI (Bontcheva et al., 2024; Das, 2025; Farooq & de Vreese, 2025; B. Grimme et al., 2023; Mylrea, 2025). Accordingly, a wide range of detection methods can be found in the literature (see, e.g., Alam et al., 2022; Tufchi et al., 2023). Developed methods are only rarely evaluated using standardized benchmarks; validation is usually carried out through demonstration using very specific case studies. In most cases, however, the evaluation of methods and countermeasures often takes place in isolation from the real-world operational context. Only a limited number of studies address actual process integration, and these cover only a narrow set of methods and characteristics (e.g., Cerone et al., 2020). Alternatively, they serve primarily as demonstration examples within research projects, which are only accessible to a limited extent once the projects have been completed (e.g., https://machine-vs-rage.net). Practitioners from different disciplines work with varying roles, objectives, workflows, and incentives, and often rely more on personal motivation than on institutional support to improve their processes (Unger et al., 2025). It therefore remains unclear to what extent existing detection methods meet the real needs of practitioners without a deeper understanding of their workflows and objectives.

To address this research gap, we are conducting a (vignette-based) pilot study in the context of combating disinformation on social media, in which we interview practitioners using case studies to gain initial qualitative insights into their approaches for detecting disinformation and the corresponding countermeasures. The aim is to derive insights for the development of computer-assisted methods that better support practitioners' work.

The interviews take place in two parts and include a method demonstrator in the form of an analysis dashboard. Its design is based on a literature review and is guided by a framework of disinformation characteristics derived from a previous expert study (Unger et al., 2025). The topic of disinformation detection has been chosen in this work as data validation constitutes an integral part of OSINT (Hwang et al., 2022), as it lies within the authors' (technical and methodological) research focus, whilst also having methodological overlaps with many areas of OSINT application—in the hope that the findings obtained may also have implications for these areas.

The remainder of this article is structured as follows: Section 2 provides the conceptual background, Section 3 links disinformation features from literature and practice, Section 4 outlines the methodology and demonstrator, and Section 5 presents the results, followed by key conclusions in Section 6.

2. Background

The concept of disinformation is commonly defined along the dimensions of truthfulness and intent (Wardle & Derakhshan, 2017), distinguishing "misinformation" (unintentional falsehood), "disinformation" (intentional falsehood), and "malinformation" (true content used maliciously). Automation and coordination elevate such content through links to campaigns (C. Grimme et al., 2022; Rice & Atkin, 2013), which can be understood as "dark participation" shaped by actors, motives, targets, audiences, and processes (Quandt, 2018). With the rise of generative AI, information warfare increasingly threatens open discourse on OSNs (Ferrara, 2023; Hartwig et al., 2024). While fact-checking identifies falsehoods, debunking contextualizes them based on assumptions about intent. Even true content may be disseminated deceptively in coordinated or automated ways. Accordingly, fact-checking providers increasingly offer forensic debunking, and content creators frequently expose others using such services. Prebunking, by contrast, provides counter-information pre-emptively, akin to vaccination (Lewandowsky & van der Linden, 2021).

Given the scale and diversity of social media content, automated and community-based detection approaches are essential (Shao et al., 2016). Platforms employ measures such as contextual links and "community notes," which have proven more effective than simple fact-checking (Drolsbach et al., 2024). However, generative AI complicates the detection of inauthentic behavior (C. Grimme et al., 2022). Traditionally, content has been analyzed as isolated artefacts or static networks (Alamleh et al., 2023; Ferrara et al., 2016; Gao et al., 2025), but recent work conceptualizes it as a dynamic data stream (Assenmacher et al., 2020; Cresci, 2020; B. Grimme et al., 2023). This shift highlights the need for combined methods to detect coordinated and automated behavior (Plikynas et al., 2025).

"OSINT" refers to analyzing publicly available data, mainly in digital spaces (Browne et al., 2024; Pastor-Galindo et al., 2020). Key challenges include data complexity, misinformation, source reliability, and ethical issues. In high-validity contexts, it is termed OSINT-V (Hwang et al., 2022). Although processes are not standardized, they typically involve data collection, processing, analysis, and reporting. Methods range from text and social media analysis to geospatial techniques (Hwang et al., 2022; Pastor-Galindo et al., 2020). "Social listening" has become increasingly important, particularly for managing reputational risks. OSINT actors pursue diverse goals and may act both as targets and agents (Unger et al., 2025).

The German Federal Ministry of Research, Technology and Space (Bundesministerium Forschung, Technologie und Raumfahrt, 2019) emphasizes the need for scientific literacy and dialogue between science,

policymakers, and the public. However, integrative studies bridging research and practice remain scarce. Existing research focuses largely on societal impacts rather than professional application. Delphi studies identify key challenges, including technological-regulatory gaps, malicious actors, mistrust, limited detection and debunking capabilities, restricted data access, opaque algorithms, and cross-platform dynamics (Kruger et al., 2024; Mahl et al., 2025). Suggested strategies include regulation, improved information literacy, and cross-sector collaboration. Hameleers and van der Goot (2025) show how expert references are used to support claims of veracity and can be exploited by mimicking established communication strategies. This aligns with findings that multiple perspectives are crucial for assessing truthfulness (Caramancion, 2023). Chystoforova and Reviglio (2025) highlight asymmetries between platforms and experts in governance contexts. Finally, Rojas Torrijos and Garrote Fuentes (2025) demonstrate that characteristics of disinformation can aid in assessing truthfulness.

3. Relating Scientific Advances of AI-Based Detection Methods to Disinformation Features From Practice

This study focuses on expert practice. Consequently, particular interest lies in the characteristics and thus, implicitly, the indicators and methods that experts use to identify disinformation and disinformation campaigns. As a starting point for the design of the expert study conducted in this work, we draw on the findings of a recent large-scale qualitative study that examined the perception of disinformation among experts from the fields of public administration, business, journalism, and politics (Unger et al., 2025). To this end, we categorize 22 features mentioned in the report by Unger et al. (2025) into six feature groups proposed by the same authors (see Table 1). In line with the authors' findings, the features in the "Content," "Language," and "Source" groups are primarily used by experts for the detection of disinformation artefacts. The features in the "Temporal," "Organizational," and "Dissemination" groups relate to the detection of disinformation campaigns.

Table 1. Disinformation features derived from expert interviews.

Disinformation	**Content features**	**Language features**	**Source features**
	Fabricated content Decontextualization Distortion of information Cherry-picking/exclusion Clickbait	Emotionalized language and radical statements Poor quality (including translation)	Known disinformation sources. Account features (followers, profile picture, etc.) Imitations of sites
Disinformation Campaign	**Temporal features**	**Organizational features**	**Dissemination features**
	Pattern of dissemination (long-term) High dissemination speed Immediacy	Recognizable strategy Organized structures for implementation Political motives Repetition Trolls, bots, fake accounts	Cross-media distribution Multi-modality Platform-specific adaptation Designed to fool mainstream media

Note: The categorization was confirmed to be reflective of the study results together with the original authors. Source: Adapted from Unger et al. (2025).

For the analysis carried out in this work, however, it is important to link the (expert-based) characteristics of disinformation and campaigns identified by Unger et al. (2025) with technical and algorithmic analysis approaches. The literature contains numerous taxonomies and frameworks for the analysis of big data and social media. These are not only highly heterogeneous but are also often tailored to specific contexts (Dasari & Kaluri, 2023; Hamzah & Vu, 2018; Hemmati et al., 2024; Rahman & Reza, 2022). Given our aim to identify a general alignment between experts' perspectives on disinformation and scientific findings from the literature, we draw on the following four fundamental (and well-established) methodological categories:

- Supervised learning: Training models such as support vector machines, random forests, (deep) neural networks, and others for classification or prediction.
- Unsupervised learning: Exploration and learning with little or no prior knowledge, e.g., through clustering or dimensionality reduction.
- Network analysis: Representation of structural features and dynamics of information propagation, whereby networks themselves can be used as features for other approaches.
- Large language models/vision-language models: Although training may be unsupervised and improvements monitored via "human-in-the-loop" procedures, the application is often detached from the concepts of unsupervised or supervised learning.

The traditional distinction between supervised and unsupervised learning is supplemented here by the perspective of the network level, which is particularly important in the context of social media as it can capture and analyze social connections and interactions. As generative models, as described in Section 2, have become highly relevant in content generation and evaluation, this level is introduced as a fourth methodological category (from a detection measure perspective).

In recent years, the detection of fake news based on the content of posts or news articles has become a widely researched field, in which deep learning methods are frequently employed (Capuano et al., 2023). For example, specific features such as "clickbait" are the subject of research in the field of machine learning, as is, e.g., the detection of fake accounts (Roy & Chahar, 2021; Zuhro & Rakhmawati, 2020). Network analysis offers a variety of approaches for detecting disinformation on social networks (Pacheco et al., 2021), including the consideration of different data types (such as content, activity, and identity) as well as data formats (such as text and images). Although network analysis is clearly well-suited to capturing the organizational characteristics of disinformation campaigns, temporal information can also be captured and modelled, for example through the timing of coordinated actions (Mannocci et al., 2024) or information diffusion (Michalski et al., 2020). Large language models and related derivatives such as vision-language models have now become an integral part of the literature on disinformation detection. These models naturally address features based on the use of language, such as emotions (Liu et al., 2024). Similarly, identifying attitudes towards various topics can provide insight into the political motives of different actors (Pangtey et al., 2025).

The search for conceptual alignments between expert characteristics and identification approaches found in the literature certainly warrants a study in its own right, beyond the scope of this work. Here, we assume with a high degree of certainty that many of the features used in practice are well represented in the literature. Given the multitude of methodological developments, we also assume that the integration of representative methods into an interactive dashboard will provide a good basis for discussion during our interviews and help to clarify whether and how these methods can support expert practice.

4. Research Approach

To test the integration of scientific advances into relevant tools for experts confronted with disinformation, this study analyses the circumstances under which experts struggle with disinformation. To take the experts' perspective into account in the interviews, characteristics of disinformation from the conceptual framework presented in Table 1 are drawn upon. On this basis, we implement analysis methods from the categories identified in Section 3 in an analysis dashboard, which serves as a method demonstrator during the interviews, acting both as a guide and as a catalyst for discussion and for gaining deeper insights into the experts' perspectives. Figure 1 summarizes our overall approach. In it, we describe the specific steps and our considerations regarding them.

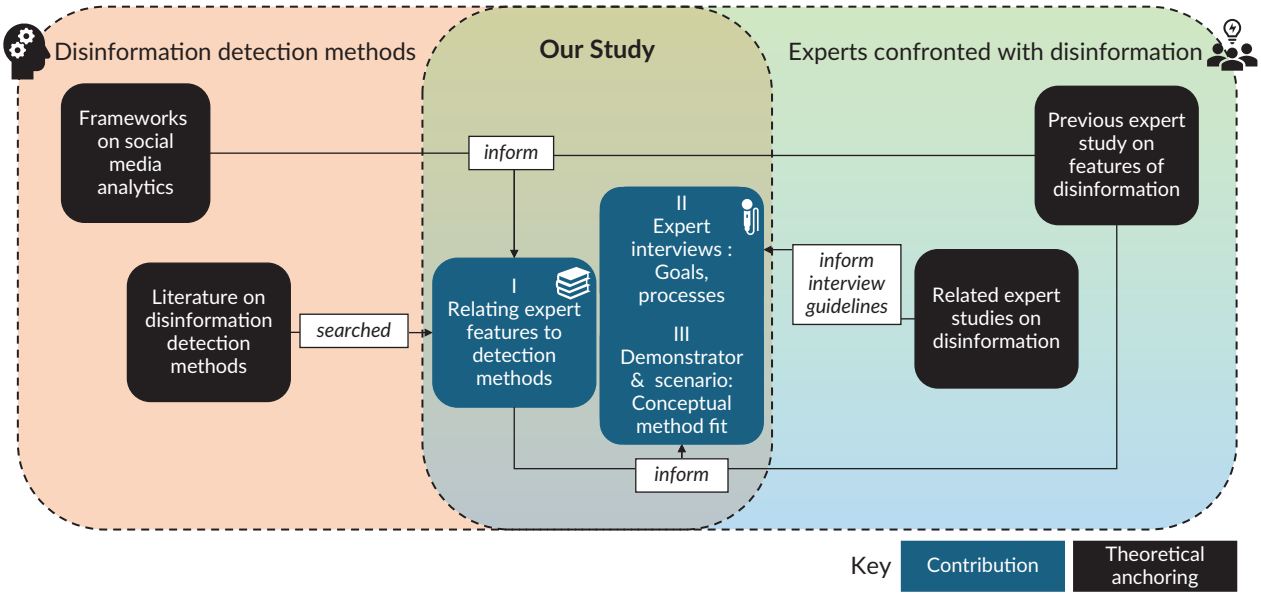

Figure 1. Overview of our research approach.

4.1. Expert Interviews

As a pilot study to align disinformation research with practice, we adopt a qualitative design (Flick, 2023) using case studies to capture comparable decision-making processes. Semi-structured expert interviews explore objectives, workflows, and constraints, supported by a methodological demonstrator (analysis dashboard) presenting a shared, real-data scenario. This follows the vignette approach of enabling contextualized interpretation of actions (Barter & Renold, 1999). Interviewees come from administration, business, internal security, the armed forces, and research institutions in German-speaking countries, including actors engaged in public rebuttal. All participants are listed in Table 2 under pseudonyms.

We structured the interviews and initial coding along four dimensions:

- Goals (e.g., general goals where disinformation might be an issue, goals relating to handling disinformation itself, etc.);
- Native workflows (e.g., process steps, OSNs used, websites consulted, assumptions made, people involved, tools used, disinformation features identified, etc.);

- Constraints (e.g., data requirements, policy or compliance concerns, training/ownership, tech, or media literacy, etc.);
- Conceptual fit (e.g., task-technology fit, processes integration, etc.).

Experts from fact-checking organizations were excluded because their goals and operational context differ. Fact-checkers verify circulating claims to inform the public, closely examining them and their effects on public consensus. By contrast, experts in safety- and security-related domains focus on threat management and vulnerability reduction, requiring a broader overview of current circumstances to quickly assess data validity.

Experts drew on current and past projects. Interviews followed Myers and Newman (2007) and lasted about one hour. After discussing existing practices, participants interacted with the demonstrator in a guided "think-aloud" session (van Someren et al., 1994). A final debriefing captured additional reflections and suggestions. Our guiding questions for the interviews can be found in the Supplementary File.

Table 2. Interviewed experts.

Expert	Application area	Expert role	Country
AF1	National armed forces	IT project manager	Germany
AF2	National armed forces	OSINT project manager	Switzerland
DR1	Media and disinformation research	Method development	Germany
DR2	Media and disinformation research	Method development	Germany
CM1	Crisis management	Crisis management expert	Austria
CM2	Crisis management	Crisis management researcher	Germany
FA1	Federal agency	OSINT IT support	Germany
FA2	Federal agency	OSINT IT support	Germany
PA1	Police authority	Researcher and university teacher	Germany
PA2	Police authority	OSINT analyst	Germany
TO1	Tech organization (non-profit)	OSINT tool provider	Germany
TO2	Tech organization	OSINT analyst	Germany
SL1	Social listening service provider	OSINT analyst	Germany
SL2	Social listening service provider	OSINT CEO	Germany

4.2. Method-Demonstrator

The interviews were conducted using a specifically implemented analysis dashboard designed to identify inauthentic behavior in social media data streams. It reflects the identified four literature-based categories (see Section 3) and integrates representative detection methods aligned with competencies described by Unger et al. (2025). The demonstrator includes components indicating coordination and automation, such as temporal content clustering, network analysis, detection of AI-generated text, and AI-supported summarization. To support initial exploration, basic elements such as descriptive statistics, aggregations, and time-series analyses were included, enabling a quick understanding of the dataset.

The demonstrator was used during the semi-structured expert interviews with a real dataset (X, formerly Twitter) on climate change, including noise and inconsistencies (Pohl et al., 2022). It comprises over 110,000

textual posts from almost 70,000 unique users across three days. Data were collected using hashtags related to climate change (denial) and key movements and actors. Although limited to 1% of the total data stream at any time, the dataset revealed patterns indicative of disinformation using our sample of computational methods. Participants could freely combine analytical components to reflect real-world workflows. Some methods were derived from prior research, though alternative specialized tools exist. To minimize bias, only basic views were initially visible, with further visualizations activated on demand. A detailed description of the demonstrator and its components is provided in the Supplementary File.

5. Results

5.1. Coding Scheme

To ensure a high degree of accuracy in coding, we distinguish between conceptual and implementation feedback to ensure that our design decisions do not compromise statements regarding the general concept of a method group. The method demonstrator was successfully used to illustrate further concepts, many of which mirrored the first half of the interview, but above all created a common ground for interviewers and interviewees to discuss the potential and challenges of computer-assisted support. To increase the transparency of our coding process, Figure 2 illustrates how our first-order concepts lead to higher-order concepts and ultimately to aggregated dimensions.

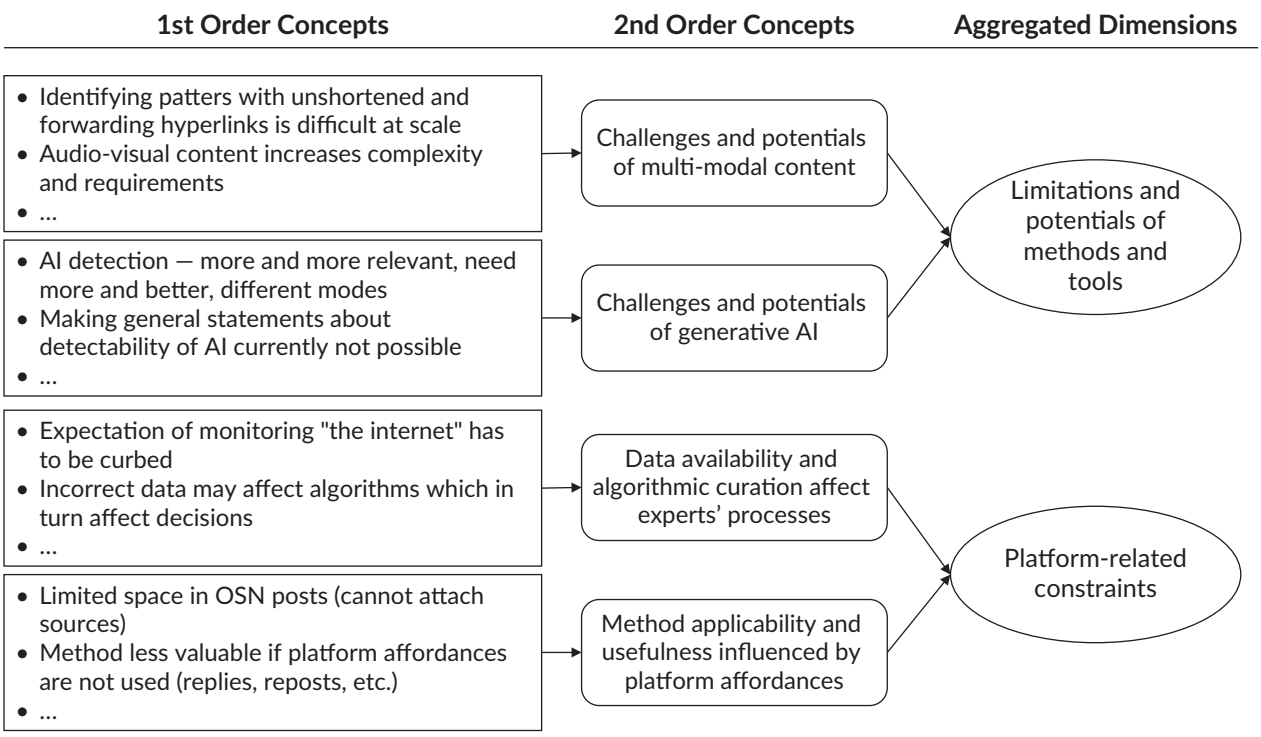

Figure 2. Exemplary codes and derived higher-order concepts. Source: Adapted from Corley and Gioia (2004).

As part of the analysis, overarching themes were developed to identify spectra conceptualizing the expert situation. In addition, the study aimed to map the conceptual suitability of detection methods, particularly regarding their conditions of application and potential integration hypotheses. Against the backdrop of an

exploratory research interest, reflexive thematic analysis according to Braun et al. (2023) was chosen, as it enables an inductive and flexible evaluation and is not based on the application of a fixed coding framework. This methodological openness is particularly suitable for exploratory studies, as coding and theme development can evolve iteratively over the course of the analysis. Data collection and analysis were conducted by a researcher who was working on an interdisciplinary project on the detection of disinformation, adhering to the interpretivist perspective of reflexive thematic analysis.

Although four initial dimensions (see Section 4.1) were incorporated into the coding, the analysis process revealed that the boundaries between them became increasingly blurred. For example, process-related tasks often also served the function of addressing specific constraints. These overlaps are reflected in the iterative coding as well as in the subsequent cross-dimensional analysis and confirm the decision to adopt an inductive approach. Nevertheless, the initial dimensions proved to be heuristically valuable for identifying overarching concepts and themes.

In the following, we present the themes resulting from our analysis. First, we address themes that deal with actors, roles, and processes in the application context of our respondents to outline the overarching context in which the interviewed experts operate. We then focus on themes that specifically concern computer-aided support and derive conclusions from them for the research and development of new methods as well as their integration into processes.

5.2. OSINT Actors, Roles, and Processes

5.2.1. Detection Goals Depend on Application Context and Lead to Different Role Allocations Among Actors

Across application contexts, we identified a spectrum of roles that can be categorized into three groups: receivers of open-source information required to fulfill various goals (OSINT stakeholders); analysts responsible for the collection and/or verification of such data (OSINT analysts); and actors supporting analysts in their work (OSINT supporters). This also functions as a form of division of power where the actors requiring OSINT data are not necessarily the same actors carrying out the analysis. Across the cases reported by the experts, many different "setups" of roles among involved actors were described. In these, OSINT stakeholders either require support or outsource the analysis, while supporters or analysts may be part of the same overarching organization (yet separated by divisions or geography) or belong to entirely different organizations.

OSINT supporters provide assistance through OSINT capabilities and know-how. There are generally specific positions within agencies and related institutions tasked with supporting different stakeholders. These stakeholders may include government bodies, agencies, ministries, politicians, police, public relations units, or researchers. Instead of conducting analyses themselves, supporters equip stakeholders with OSINT expertise and insights into how disinformation is created, spreads, and can be detected. This may also include providing data access, particularly for researchers (TO1).

OSINT analysts, in contrast, carry out the actual analyses and produce status reports and overviews for stakeholders such as private companies, journalists, or the armed forces. Their deliverables include, for example, status reports on threat situations for company locations, automated overviews of relevant topics

in OSNs, and regional assessments to evaluate military situations. Although not their primary objective, some experts reported that their work also involves debunking false assumptions to counter short-sighted or black-and-white thinking rooted in a polarized media landscape (CM1).

5.2.2. Processes Related to OSINT Activities Depend on Roles and Range From Inquiry-Driven to Recurring Tasks

Depending on the organization and its relationship to stakeholders, OSINT supporters and analysts receive inquiries through various entry points, ranging from business contacts to direct requests, including web-based "reporting portals." Before analysis begins, responsibilities must be determined. For example, the police must assess whether an inquiry is relevant to their area of operation and, if necessary, forward it to individuals with specific expertise related to the case. Initially, relevant data sources must be identified, and the data prepared for analysis. In the case of private companies, the interests of specific departments regarding social listening need to be clarified: "We have customers who require a more comprehensive monitoring of scenes propagating disinformation. And then we have customers who set it a lot more granularly" (SL1). The range of topics evolves over time and listening reports must be adjusted accordingly as they are iterated over several weeks. In contrast, the military may work closely with its own dedicated IT support division to define requirements for tools that will be used operationally.

An analysis generally begins with keyword searches to identify commonly occurring topics and actors relevant to the analysis goal. In most cases, the objective is to determine which individuals or groups posted first, to what extent this can be traced, and how information was disseminated to reveal connections or regional proximity among potentially malicious actors. During the process, analysts may decide to exchange information with other institutions, such as agencies, to the extent that this is permitted. Finally, a report is delivered, and potential follow-up queries may be received. These reports range from inquiry-driven outputs (primarily for public authorities) to regular reports issued daily to monthly (primarily in the private sector). Similarly, stakeholders within the armed forces may receive regular status reports, and agencies may forward identified actors and associated campaigns to the initiating body or other responsible parties (e.g., the Ministry of the Interior).

Figure 3 summarizes our findings pertaining to the roles different actors assume and their interactions.

5.3. Current State and Barriers of OSINT Practice, Processes, and IT-Supported Disinformation Detection

The integration of OSINT into organizational processes varies significantly across domains, particularly between public and private sectors. While public sector actors are generally familiar with OSINT workflows, process maturity and awareness in private companies often depend on prior exposure to disinformation incidents and are frequently limited to basic sensitization efforts. Identifying where disinformation fits within an organization, whether in security, communication, or elsewhere, remains a key challenge. At the same time, many analytical approaches, such as data aggregation, network analysis, and visualizations of activity over time, are already well known and widely used. These methods are often perceived as "pretty standard" and valued for their ability to structure and screen large datasets. However, current tools derive most of their value from data access and preprocessing rather than advanced analytical capabilities. As a

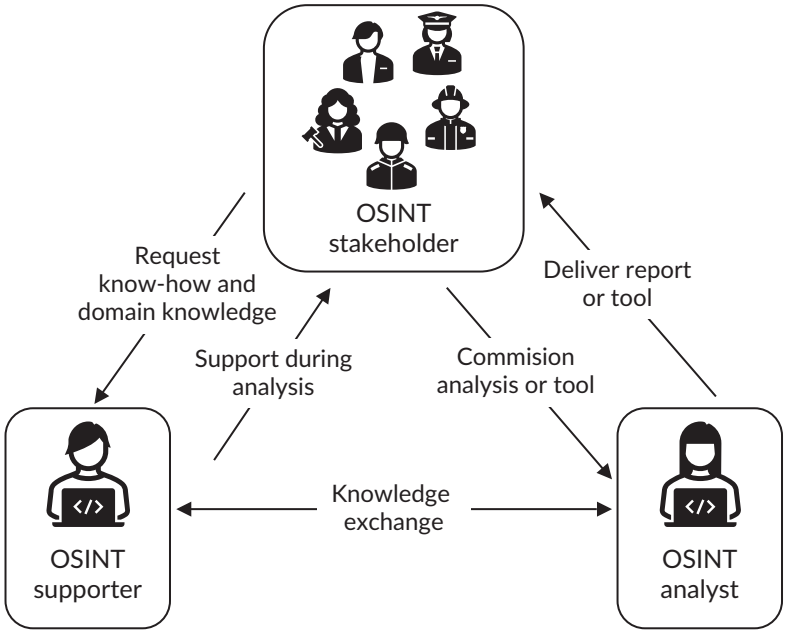

Figure 3. Abstract overview of roles and their interactions.

result, substantial resources are still devoted to data acquisition, management, and filtering, with multi-layered, multi-perspective analysis seen as essential but resource-intensive.

Practitioners' experiences and reports reflect these observations. Even in the initial stages of social listening processes, identifying relevant disinformation topics or attack vectors can be challenging: "Where is the topic located? In security? In communication? What's the interaction?" (SL2). While recent high-profile cases have increased awareness and interest in OSINT capabilities, organizations not yet affected by disinformation campaigns often show limited engagement. Combining different analytical perspectives was emphasized as crucial: "The standard, which analysts…need—multi-layered filtering. I can look at different aspects." (FA2). Despite this familiarity, more advanced approaches—such as clustering, prediction, or large language models-based analysis—are not yet widely adopted, partly due to resource constraints and limited operational integration. Barriers to adoption include the effort required to handle uncertainty in IT-supported analysis—"But if we want computers to ease our work a little, we will always have this risk" (FA2)—as well as structural challenges in transferring research into practice: "It often takes very long with [universities] financing projects" (TO1). Limited computational resources during development, narrow application scopes, and unclear cost-benefit ratios further hinder adoption: "The cost-benefit ratio now…is not really proportional" (TO1). Ultimately, the relevance of any tool depends on its alignment with practical needs: "[We are looking at] what the users need to be able to do, which tools they need, and which processes are required" (DR2).

We may conclude from these observations that method development should place greater emphasis on integration into existing workflows, particularly by supporting data acquisition, preprocessing, and multi-layered filtering as core tasks. While understanding of the need and knowledge of the usefulness of monitoring processes and methods is available in principle, operational and resource-related barriers exist. Tools and methods should, thus, ensure relevance to specific organizational needs and decision-making processes. Improving usability and reducing resource requirements—especially computational demands—are

critical for broader adoption. At the same time, bridging the gap between research and practice requires faster and more sustainable transfer mechanisms. Finally, tools should explicitly account for uncertainty and support analysts in managing it, rather than attempting to eliminate it. This can foster trust and effective use in real-world settings.

5.4. Potentials and Challenges of AI-Driven Disinformation Detection

5.4.1. Regulatory Constraints Affecting IT Support, and Constraints Caused by Unclear or Changing Responsibilities

Regulations and role-specific constraints strongly shape the application of computational methods during analysis. For public agencies in particular, a key issue is that legal requirements limit what is permissible. They often even restrict the use of advanced IT and AI despite their technical feasibility. As noted by participant (FA1), "Companies may do all that is not legally forbidden, agencies may only do what is legally permitted by law." These constraints include data protection regulations, platform terms of service, and strict requirements for data handling, such as where data is stored, processed, and how models are trained. Consequently, AI-based tools are often not used at all due to unresolved data concerns: "The question is always: how is it trained, where is the data located, where is the data processed?" (PA2). Limited access to OSN data, especially after research projects end, and restricted computational resources further hinder operational deployment. At the same time, organizations tend to prefer domestically developed or university-based solutions over foreign or commercial tools, which reflects concerns about trust and compliance.

In addition, responsibilities across actors such as police, military, intelligence services, and private organizations are often unclear, overlapping, or subject to change. Determining jurisdiction can be difficult from the outset: "At what point is this still a police matter, and at what point does it cross into tactical military operations?" (AF2). These ambiguities may persist throughout the analytical process and can require cases to be reassigned or discontinued. For example, an investigation may need to be handed over once it becomes clear that a campaign originates domestically rather than abroad. Tensions also arise when analytical units are asked to evaluate the impact of decisions they supported, which may conflict with their defined roles: "But we don't want to analyze this for the press officer…this is not our task" (FA2). Overall, what is technically possible often exceeds what is legally or organizationally permitted or wanted.

These findings highlight that method development must explicitly account for regulatory and organizational realities. Transparent data handling, clear documentation, and options for local deployment to address data sovereignty concerns are essential aspects in the regulatory context. At the same time, flexibility regarding shifting responsibilities (e.g., by enabling traceability, modular use, and smooth handover by actors) needs to address assigned duties and jurisdiction. To achieve this and to facilitate the transition from prototypes to operational systems, sustainable data access, adequate resources, and integration into actual processes and workflows are important infrastructural issues.

5.4.2. Platform Affordances and Policies Influence Processes, Including Choice of Detection Methods

Platform-specific policies and affordances significantly shape how OSN data is accessed, interpreted, and analyzed. A central reason for this is limited availability of data due to dynamically changing platform

policies, formats, and access conditions. As a result, data collection must be continuously adapted, making the expectations of comprehensive monitoring unrealistic. At the same time, excluding platforms due to access difficulties is often not considered acceptable, as it may create critical blind spots. Beyond access, algorithmic curation mechanisms such as recommender systems pre-select content in largely opaque ways, influencing what data is available for analysis in the first place. This can introduce biases into analytical models and decision-making. Such feedback loops are particularly problematic in sensitive domains, where biased data can reinforce existing patterns. Overall, experts operate under conditions where data availability, quality, and visibility are externally shaped and only partially controllable. In addition, platform affordances strongly influence the applicability and usefulness of analytical methods. Different OSNs provide distinct interaction mechanisms—such as hashtags, mentions, replies, or reposts—which define how information spreads and can be analyzed. These features enable certain types of analysis while limiting others, and they vary considerably across platforms. Consequently, methods often need to be adapted to platforms, making cross-platform analyses more difficult.

These challenges are reflected in the interviewees' experiences. For example, one participant emphasized the limits of data access, "I had to keep on explaining that we cannot 'read the internet'" (AF1), while another stressed that omitting platforms is not an option, "If it's not allowed, then we have blinders on" (SL2). The influence of algorithmic curation and resulting biases was highlighted in the context of predictive policing: "So we would monitor a certain area more, and in turn find more violations. It's a vicious cycle" (PA1). Platform-specific constraints also affect method choice, as illustrated by the limited applicability of text-based analysis: "That sort of content analysis based on text…is not going to get me far on TikTok" (FA2). Similarly, certain analytical approaches depend on platform features, which are not always available: "I see this [type of] analysis often on X, but outside of X very rarely… because it is difficult to establish a connection….For example, on Telegram, users are often anonymous" (AF1). Moreover, narratives may differ significantly across platforms—"Specifically on Telegram, it is quite challenging when texts become very long…and you may find that there is no real equivalent on other platforms" (DR1)—reflecting varying user demographics and contexts. Even content creation constraints can impact reliability, as noted: "Information I can include in an Insta-message is limited….If I now include the list of references…I have no more space left for the content" (PA1).

These findings imply that useful methods and analysis tools need to handle incomplete, biased, and dynamically changing data. At the same time, they must be informative about data provenance and algorithmic influence. Methods should be adaptable to different platform affordances by abstracting common user interaction patterns to allow cross-platform analysis. Moreover, practitioners often need support in identifying and mitigating biases introduced by platform curation. At the same time, analyses are always snapshots: It should be clearly communicated to users that a complete view of the data is unrealistic and that a comprehensive analysis is neither technically nor practically feasible.

5.4.3. Analysis Approaches Range From Problem-Centric to Feature-Centric

Experts' analytical approaches can be positioned along a spectrum ranging from problem-centric to feature-centric strategies. On one end, analysts focus on concrete threat scenarios and specific investigative goals, often applying deductive reasoning to identify malicious actors or activities: "Our central question is always: What threats are emerging?" (SL2). On the other end, more abstract approaches rely on known

disinformation features, such as linguistic cues, account characteristics, or patterns of coordinated behavior. This often involves searching for specific indicators such as fake accounts, deepfakes, coordinated campaigns, or known disinformation sources. At the same time, however, participants highlight the limitations of generalized frameworks: "Regarding [detection] goals…you will get a completely different answer from someone working for the police, in cyber defense, or from the military….Something like this could have higher explanatory power the closer it is to the phenomenon or goals" (AF2). Analysis approaches vary significantly depending on the application domain, detection goals, and perceived threats, with experts often combining bottom-up exploration with selective use of known indicators. Moreover, some features are not exclusive to disinformation or have become less reliable as adversaries adapt their techniques. In practice, experts also adapt to evolving adversarial strategies, such as the use of shortened links, constantly changing domains, or increasingly sophisticated fake websites that mimic legitimate media. These dynamics further challenge purely feature-based detection and require continuous adjustment of analytical approaches.

This suggests that method development must go beyond purely feature-based analysis. Instead, a flexible, context-sensitive analysis should be enabled. This means that tools should allow for both deductive and problem-oriented investigations and offer experts the possibility of shifting focus by selecting and combining methods/heuristics appropriate to the situation. Ultimately, effective tool support must reflect the diversity of expert practices and offer flexibility, rather than prescribing a single analytical paradigm.

5.4.4. Limitations and Potentials of Methods and Tools

The effectiveness of methods and tools is shaped by both technical limitations and emerging opportunities, particularly regarding multi-modal content and generative AI. A key challenge lies in the increasing importance of audio-visual data, which raises complexity and requires new analytical capabilities. Humor and memes further complicate detection, especially from ethical and political perspectives. At the same time, multi-modal analysis offers new opportunities, for example enabling the "multi-level placement" of corrective messages (e.g., counter speech) across different formats. Another challenge concerns the identification of patterns in hyperlink structures, as shortened and chained links are often used to obfuscate target websites, making large-scale analysis difficult. Detecting AI-generated content is becoming more and more important, yet also more difficult as deepfakes improve. Distinguishing real from artificial content is still difficult, and general statements about detectability are seen as currently impossible. At the same time, AI offers new capabilities, such as interactively exploring data, identifying semantic connections between features, or combining multiple methods to detect coordinated manipulation. However, concerns remain regarding adversarial attacks, data integrity, and the lack of preparedness of many systems to handle such risks.

This is also reflected in practitioners' experiences. AI-based summaries were seen as useful for gaining a "quick and dirty" overview (FA1), yet problematic in contexts where decisions must be traceable. At the same time, interactive AI support was highlighted as valuable for uncovering relevant aspects that might otherwise be missed (AF1) and for understanding relationships between features and narratives (PA1). Again, the relevance of specific methods strongly depends on the application context. For example, data aggregation techniques such as hashtag analysis were considered of limited value in security contexts: "The top topics or hashtags in most cases are not the deciding factor. In a pinch, the needle in the haystack is sufficient, if detected early enough" (SL2). In contrast, identifying specific actors or linking content across platforms was seen as

more relevant. Temporal analyses can help associate activity patterns with geographic regions: "This is very interesting, to look at how it maps to the working hours of different countries" (AF2). Network analysis was highlighted as particularly useful for identifying key actors, understanding campaign dynamics, and supporting communication with stakeholders: "Having support like this is generally received very well" (SL2).

We conclude from the above that method development must integrate multi-modality and account for the growing complexity of content. At the same time, it needs to remain robust against obfuscation strategies. AI-based tools should balance automation with transparency and traceability, especially in high-stakes contexts. Combining multiple approaches and perspectives is preferable over relying on single methods, while tools must be tailored to specific application domains, avoiding overly generic or overly specialized solutions. The support of data export, cross-platform analysis, and flexible integration into workflows can enhance practical usefulness, while avoiding overreliance on dashboards that may create a false sense of security.

5.4.5. Automation vs. Manual Verification

The use of AI in practice is often restricted to supporting tasks such as sorting, filtering, and pre-qualifying large datasets, instead of making autonomous decisions. Verification processes vary widely, ranging from fully manual assessments to software-supported approaches such as forensic tools for detecting image manipulation. The level of verification effort depends on contextual factors, including the source or sender and their potential intent. As one participant noted, "From each side, an intention must be assumed. Because of that, I must contextualize information and put more or less energy into the verification" (FA2). While AI can assist in structuring data (e.g., by summarizing posts, extracting text via optical character recognition, generating captions, transcribing audio, or identifying relevant topics), it is primarily used to make data searchable and filterable. Additional applications include anonymizing non-public individuals and filtering user-generated content in line with "privacy by design" principles (AF1). Despite these capabilities, the credibility of information must be assessed independently of its source, and final judgments remain with human analysts.

Experts' experiences reflect these practices. AI is commonly used to flag or pre-select relevant content, yet analysts always validate all outputs of automated data screening. At the same time, both human and automated processes are prone to error. Manual screening of channels is described as resource-intensive and error-prone, while automated matching across datasets may lead to incorrect conclusions if not carefully validated. In some cases, detecting AI-generated content is considered extremely challenging: "Right now, my only question is: (a), can I even associate this [post] with the event, and (b), was this AI-generated? It's an absolute nightmare" (AF2). Participants also stressed the need to contextualize OSN data relative to other sources: "You can't just treat an X-post or a Facebook-post as equal to analysts' estimations or police reports. You need to weigh the data" (SL1). Although AI summaries can save time and provide a "quick and dirty" (FA1) overview, they often lack traceability, which limits their use in contexts requiring legally robust evidence. Similarly, clustering methods or automated analyses may be perceived as too vague or untransparent: "Since this cluster is a bit cluttered due to the semantic relatedness" (TO2).

These findings highlight that AI-supported tools should prioritize human-in-the-loop approaches, focusing on pre-structuring data rather than replacing analytical judgment. Systems should enhance traceability by clearly linking outputs to underlying data. This is valued more highly than abstract explainability.

Transparency about models, adjustable sensitivity, and options to select specific AI components can increase trust. Interactive functionalities for exploring, filtering, and cross-checking results are of particular importance. Overall, effective integration requires balancing automation with human expertise, ensuring that AI augments rather than replaces critical reasoning in operational contexts.

5.4.6. Fusion of Information and Data

A common trend across application areas is the fusion of OSINT data with additional, context-specific sources to enhance analysis. The primary goal of this integration is to supplement, cross-check, and increase the reliability of available information, thereby minimizing falsifiability. While this practice is not unique to OSINT, it reflects a broader methodological standard of combining independent sources. Data fusion includes linking OSINT data with geographic information systems, proprietary databases, sensors, or trustworthy public sources such as official statistics and domain-specific studies. It also involves integrating multiple modalities (e.g., text, audio, video, geolocation) and correlating information across platforms. In some cases, organizations extend OSINT with their own historical data or external datasets, which can provide a competitive advantage over generic AI systems. At the same time, approaches such as AI-based detection were seen as useful to further strengthen the reliability of public data.

These practices are reflected in experts' accounts. As one participant emphasized, "You perform a metadata evaluation and a plausibility check, asking 'is there a second source?'" (CM2). Another highlighted the added value of proprietary data: "This is way better than the baseline data that ChatGPT and Gemini and so on are built on" (SL2). Practical examples include referencing official statistics—"I work relatively often with criminal statistics…from public and reliable websites like the federal offices of criminal investigation" (PA1)—or combining OSNs data with additional contextual information. However, participants also pointed to limitations. Understanding complex phenomena may require deep domain knowledge that cannot be replaced by data alone: "The central element is…system—and complexity-theory based…only those who know the whole know the details" (CM1). Yet, data fusion introduces its own challenges. Combining multiple sources can obscure the origin of information and make it more difficult to detect inconsistencies or falsified content. As more data and methods are layered, errors may propagate: "I have new findings and build on them, then an error cascades right through to the end" (AF2). Additionally, valuable proprietary data may remain inaccessible due to confidentiality concerns: "At the same time, it is understandable that they may not want to hand them out" (SL2).

These observed aspects suggest that tools should support flexible integration of heterogeneous data sources while maintaining transparency about data provenance. Mechanisms to trace origins, identify inconsistencies, and prevent error propagation are critical. At the same time, systems should allow selective inclusion of relevant data and support cross-platform and multi-modal analysis. Finally, integrating expert knowledge alongside data-driven methods remains essential, as data fusion alone cannot fully capture complex, context-dependent phenomena.

5.5. Discussion and Conclusion

Interviewing experts from different fields in private and public domains uncovered how different actors assume different roles, and in which way these influence analysis processes. Each application context has its own actors and corresponding roles, shaping constraints and analysis approaches. We provide an overview

of all themes and 2nd-order codes in Figure 4 as a reference. Following our above argument, these are shown as context- and role-related above, and process- and IT-focused below.

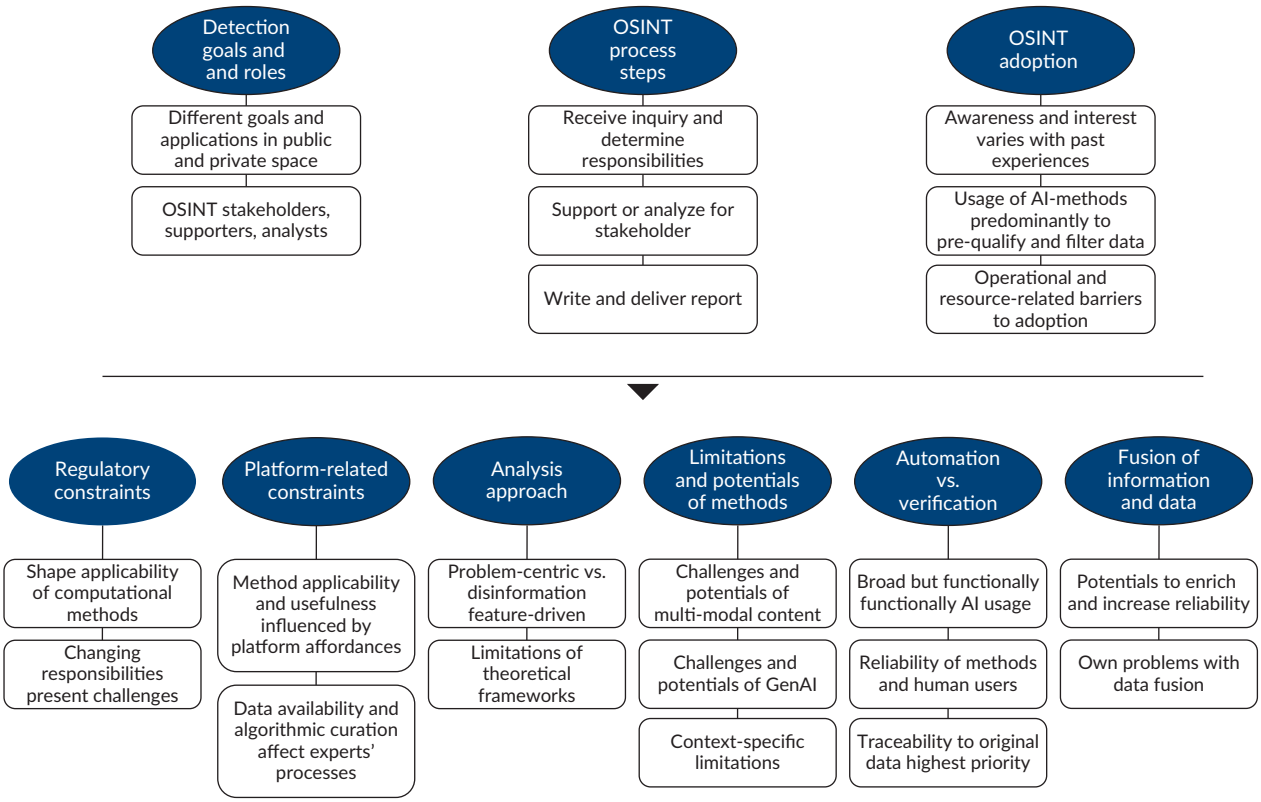

Figure 4. Overview of themes and 2nd order codes.

We identify a tension between limited data availability and the need for advances in cross-platform research, particularly in Sections 5.4.2., 5.4.4., and 5.4.6. Many OSINT tool providers integrate data from vendors, allowing their customers to partially overcome these constraints. Consequently, data access remains a core issue for researchers. On the other hand, OSINT analysts report a lack of human resources for analysis. Our results therefore encourage closer cooperation between OSINT tool developers and analysts, building on prior work suggesting research avenues (Assenmacher et al., 2022; Bruns, 2021). Some requirements may not apply to foundational or even applied computational research and are better addressed at the organizational R&D level. This includes local deployment options to address data sovereignty concerns and smooth handover across jurisdictions as in Section 5.4.1., and sensitivity adjustments calibrated to specific analysts' workflows as in Section 5.4.5. Still, applied and design-oriented research should account for the organizational, technical, and regulatory conditions under which methods must ultimately operate.

Overall, the expert interviews show that the success and applicability of methods depend less on their performance in isolation than on their integration into organizational, technical, and regulatory contexts, as well as their ability to specifically complement human judgment. To guide method development, we propose adhering to the following principles (forming the acronym YMCA):

- Yoke to wisdom: The incorporation of human expertise remains central. Automated procedures should support analytical processes, not replace them.

- Multi-method integration: Effective systems rely on a combination of multiple methods and the integration of multimodal data to ensure resilience against deception strategies.
- Context specificity: The diversity of disinformation phenomena requires flexible, context-sensitive approaches that combine both problem-centered and feature-based analyses.
- Accountability: Methods must make uncertainties visible, disclose data sources, and communicate the limitations of analyses. Traceability and interactive control options are crucial for trust and effective use.

Through our focus on experts from security-related areas, our results are limited to that domain. Despite the sample size, we are confident in the results as the final dimensions were core issues in almost every interview. In addition to the themes identified in our analysis, we believe that the technical integration of disinformation detection methods into existing tools would constitute a minor issue. Such tools often already include different modules to be used as seen fit by OSINT analysts; implications for those not directly involved in the analysis are more difficult to derive. Rather than a reluctance to use AI due to lack of expertise or resistance to modify existing workflows, we observe a willingness to better understand a potentially useful indicator to support validation for countermeasures, including public rebuttals.

Finally, the method demonstrator does not allow an evaluation of the selected methods' effectiveness. Rather, our research illuminates whether the goal of a detection approach, as exemplified by the corresponding methods in the demonstrator, is relevant to these experts. There is a plethora of developments for each method included in our demonstrator aiming to detect disinformation and disinformation campaigns along different features (see the Table in the Supplementary File). Our results highlight various additional aspects that need to be considered, and therefore encourage more applied research and prototyping instead of further research to improve detection performance. Our derived principles may guide the adaptation of computational methods to specific application areas by asking how exactly these principles apply to a specific field.

Acknowledgments

The authors would like to thank the reviewers for their constructive feedback and Raquel Silva from Cogitatio Press for the great support.

Funding

Publication of this article in open access was made possible through the institutional membership agreement between the University of Münster and Cogitatio Press.

Conflict of Interests

The authors declare no conflict of interests.

LLMs Disclosure

We acknowledge the use of ChatGPT to improve the readability of the manuscript, DeepL as a dictionary, and ChatGPT for aiding in the coding of the demonstrator.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Da San Martino, G., Shaar, S., Firooz, H., & Nakov, P. (2022). A survey on multimodal disinformation detection. In N. Calzolari & C.-R. Huang (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6625–6643). International Committee on Computational Linguistics.

Alamleh, H., Al Qahtani, A., & El Said, A. (2023). Distinguishing human-written and ChatGPT-generated text using machine learning. In *Systems & information engineering design symposium* (pp. 154–158). IEEE. https://doi.org/10.1109/sieds58326.2023.10137767

Assenmacher, D., Clever, L., Pohl, J. S., Trautmann, H., & Grimme, C. (2020). A two-phase framework for detecting manipulation campaigns in social media. In G. Meiselwitz (Eds.), *International conference on human–computer interaction* (pp. 201–214). Springer.

Assenmacher, D., Weber, D., Preuss, M., Calero Valdez, A., Bradshaw, A., Ross, B., Cresci, S., Trautmann, H., Neumann, F., & Grimme, C. (2022). Benchmarking crisis in social media analytics: A solution for the data-sharing problem. *Social Science Computer Review*, *40*(6), 1496–1522. https://doi.org/10.1177/08944393211012268

Barter, C., & Renold, E. (1999). The use of vignettes in qualitative research. *Social Research Update*, *25*(9), 1–6. https://sru.soc.surrey.ac.uk/SRU25.html

Bontcheva, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dutkiewicz, L., Krack, N., Teyssou, D., Nucci, F. S., Spangenberg, J., Srba, I., Aichroth, P., Cuccovillo, L., & Verdoliva, L. (2024). *Generative AI and disinformation: Recent advances, challenges, and opportunities*. Fraunhofer-Publica. https://publica.fraunhofer.de/entities/publication/c3c82b4f-d189-482e-a98c-fc29f5a264d5

Braun, V., Clarke, V., Hayfield, N., Davey, L., & Jenkinson, E. (2023). Doing reflexive thematic analysis. In S. Bager-Charleson & A. McBeath (Eds.), *Supporting research in counselling and psychotherapy: Qualitative, quantitative, and mixed methods research* (pp. 19–38). Springer.

Browne, T. O., Abedin, M., & Chowdhury, M. J. M. (2024). A systematic review on research utilising artificial intelligence for open source intelligence (OSINT) applications. *International Journal of Information Security*, *23*(4), 2911–2938. https://opal.latrobe.edu.au/ndownloader/files/48342541

Bruns, A. (2021). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. In S. Walker, D. Mercea, & M. Bastos (Eds.), *Disinformation and data lockdown on social platforms* (pp. 14–36). Routledge. https://doi.org/10.4324/9781003206972-2

Bundesministerium Forschung, Technologie und Raumfahrt. (2019). *Grundsatzpapier des Bundesministeriums für Bildung und Forschung zur Wissenschaftskommunikation*. Bundesministerium für Bildung und Forschung. https://www.bmftr.bund.de/SharedDocs/Publikationen/DE/1/24784_Grundsatzpapier_zur_Wissenschaftskommunikation.html

Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content-based fake news detection with machine and deep learning: A systematic review. *Neurocomputing*, *530*, 91–103. https://doi.org/10.1016/j.neucom.2023.02.005

Caramancion, K. M. (2023, July). An interdisciplinary perspective on Mis/Disinformation control. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1–6). IEEE.

Cerone, A., Naghizade, E., Scholer, F., Mallal, D., Skelton, R., & Spina, D. (2020). Watch'n'Check: Towards a social media monitoring tool to assist fact-checking experts. In G. Webb, Z. Zhang, V. S. Tseng, G. Williams, M. Vlachos, & L. Cao (Eds.), *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 607–613). IEEE. https://doi.org/10.1109/iceccme57830.2023.10253252

Chystoforova, K., & Reviglio, U. (2025). Framing the role of experts in platform governance: Negotiating the code of practice on disinformation as a case study. *Internet Policy Review*, *14*(1), 1–28. https://doi.org/10.14763/2025.1.1823

Corley, K. G., & Gioia, D. A. (2004). Identity ambiguity and change in the wake of a corporate spin-off. *Administrative science quarterly*, *49*(2), 173–208. https://doi.org/10.2307/4131471

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, *63*(10), 72–83. https://doi.org/10.1145/3409116

Das, R. (2025). *Cyberterrorism: The rise of misinformation and disinformation*. CRC Press.

Dasari, S., & Kaluri, R. (2023). Big data analytics, processing models, taxonomy of tools, V's, and challenges: State-of-art review and future implications. *Wireless Communications and Mobile Computing*, *2023*(1), Article 3976302. https://doi.org/10.1155/2023/3976302

Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS nexus*, *3*(7), Article pgae217. https://doi.org/10.1093/pnasnexus/pgae217

Farooq, A., & de Vreese, C. (2025). Generative AI and disinformation: Introduction. *International Journal of Communication*, *19*, 3559–3576. https://ijoc.org/index.php/ijoc/article/view/26088

Ferrara, E. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday*, *28*(6). https://doi.org/10.5210/fm.v28i6.13185

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Flick, U. (2023). An introduction to qualitative research. Sage. https://doi.org/10.4135/9781036231712

Gao, M., Duan, Q., Liu, B., Xiao, Y., Wang, X., & Chen, Y. (2025). Higher-order information matters: A representation learning approach for social bot detection. In M. Echa, C. Park, & N. Park (Eds.), *Proceedings of the 34th ACM international conference on Information and Knowledge Management* (pp. 675–685). ACM. https://doi.org/10.1145/3746252.3761162

Grimme, B., Pohl, J., Winkelmann, H., Stampe, L., & Grimme, C. (2023). Lost in transformation: Rediscovering LLM-generated campaigns in social media. In D. Ceolin, T. Caselli, & A. Tulin (Eds.), *Multidisciplinary international symposium on disinformation in open online media* (pp. 72–87). Springer. https://doi.org/10.1007/978-3-031-47896-3_6

Grimme, C., Pohl, J., Cresci, S., Lüling, R., & Preuss, M. (2022). New automation for social bots: From trivial behavior to AI-powered communication. In F. Spezzano, A. Amaral, D. Ceolin, D. Fazio, & E. Serra (Eds.), *Multidisciplinary international symposium on disinformation in open online media* (pp. 79–99). Springer. https://doi.org/10.1007/978-3-031-18253-2_6

Hameleers, M., & van der Goot, E. (2025). Look at what the real facts and experts say! The use of expert references and objectivity claims in disinformation: A qualitative exploration and typology. *Journalism*, *26*(7), 1469–1487. https://doi.org/10.1177/14648849241257383

Hamzah, M., & Vu, T. T. (2018). A taxonomy of twitter data analytics techniques. In *Proceedings of the 32nd IBIMA Conference, Seville, Spain* (pp. 15–16). Springer CCIS. https://ibima.org/conference/32nd-ibima-conference

Hartwig, K., Doell, F., & Reuter, C. (2024). The landscape of user-centered misinformation interventions-a systematic literature review. *ACM Computing Surveys*, *56*(11), Article 292. https://doi.org/10.1145/3674724

Hemmati, A., Arzanagh, H. M., & Rahmani, A. M. (2024). A taxonomy and survey of big data in social media. *Concurrency and Computation: Practice and Experience*, *36*(1), Article e7875. https://doi.org/10.1002/cpe.7875

Hwang, Y. W., Lee, I. Y., Kim, H., Lee, H., & Kim, D. (2022). Current status and security trend of OSINT. *Wireless Communications and Mobile Computing*, *2022*(1), Article 1290129. https://doi.org/10.1155/2022/1290129

Kruger, A., Saletta, M., Ahmad, A., & Howe, P. (2024). Structured expert elicitation on disinformation, misinformation, and malign influence: Barriers, strategies, and opportunities. *Harvard Kennedy School Misinformation Review*, *5*(7). https://doi.org/10.37016/mr-2020-169

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2024). Emotion detection for misinformation: A review. *Information Fusion*, *107*, Article 102300. https://doi.org/10.1016/j.inffus.2024.102300

Mahl, D., Kessler, S. H., Schäfer, M. S., Jobin, A., Zeng, J., & Georgi, F. (2025). Conspiracy theories and misinformation in digital media: An international expert assessment of challenges, trends, and interventions. *Communications*, *51*(1), 233–258. https://doi.org/10.1515/commun-2024-0123

Mannocci, L., Mazza, M., Monreale, A., Tesconi, M., & Cresci, S. (2024). *Detection and characterization of coordinated online behavior: A survey*. arXiv. https://arxiv.org/abs/2408.01257

Michalski, R., Jankowski, J., & Bródka, P. (2020). Effective influence spreading in temporal networks with sequential seeding. *IEEE Access*, *8*, 151208–151218. https://doi.org/10.1109/ACCESS.2020.3016913

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and organization*, *17*(1), 2–26. https://doi.org/10.1016/j.infoandorg.2006.11.001

Mylrea, M. (2025). The generative AI weapon of mass destruction: Evolving disinformation threats, vulnerabilities, and mitigation frameworks. In W. Lawless, R. Mittu, D. Sofge, & H. Fouad (Eds.), *Interdependent human–machine teams* (pp. 315–347). Academic Press. https://doi.org/10.1016/b978-0-443-29246-0.00007-9

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and case studies. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*(1), 455–466. https://doi.org/10.1609/icwsm.v15i1.18075

Pangtey, L., Bhatnagar, A., Bansal, S., Dar, S. S., & Kumar, N. (2025). *Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions*. arXiv. https://doi.org/10.48550/arXiv.2505.08464

Pastor-Galindo, J., Nespoli, P., Mármol, F. G., & Pérez, G. M. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. *IEEE Access*, *8*, 10282–10304. https://doi.org/10.1109/access.2020.2965257

Plikynas, D., Rizgelienė, I., & Korvel, G. (2025). Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning. *IEEE Access*, *13*, 17583–17629. https://doi.org/10.1109/access.2025.3530688

Pohl, J. S., Assenmacher, D., Seiler, M. V., Trautmann, H., & Grimme, C. (2022). Artificial social media campaign creation for benchmarking and challenging detection approaches. In P. O. S Vaz de Melo, W. Jeng, & C. Buntain (Eds.), *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence. https://doi.org/10.36190/2022.91

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48. https://doi.org/10.17645/mac.v6i4.1519

Quandt, T., & Klapproth, J. (2023). Dark participation: A critical overview. In M. Powers (Ed.), *Oxford research encyclopedia of communication*. Oxford Academic. https://doi.org/10.1093/acrefore/9780190228613.013.1155

Rahman, M. S., & Reza, H. (2022). A systematic review towards big data analytics in social media. *Big Data Mining and Analytics*, *5*(3), 228-244. https://doi.org/10.26599/bdma.2022.9020009

Rice, R. E., & Atkin, C. K. (2013). *Public communication campaigns*. Sage.

Rojas Torrijos, J. L., & Garrote Fuentes, Á. (2025). The factuality of news on Twitter according to digital qualified audiences: Expectations, perceptions, and divergences with journalism considerations. *Journalism and Media*, *6*(1), Article 3. https://doi.org/10.3390/journalmedia6010003

Roy, P. K., & Chahar, S. (2021). Fake profile detection on social networking websites: A comprehensive review. *IEEE Transactions on Artificial Intelligence*, *1*(3), 271–285. https://doi.org/10.1109/tai.2021.3064901

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In J. Bourdeau, J. A. Hendler, & R. N'kambou (Eds.), *WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 745–750). ACM. https://doi.org/10.1145/2872518.2890098

Tufchi, S., Yadav, A., & Ahmed, T. (2023). A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, *12*(2), Article 28. https://doi.org/10.1007/s13735-023-00296-3

Unger, S., Klapproth, J., Boberg, S., Bösch, M., Vief, N., Stöcker, C., & Quandt, T. (2025). Features of disinformation: An expert interview study on the perception of disinformation among political, governmental, media and business elites in Germany. *Journal of Elections, Public Opinion and Parties*, *35*(3), 472–494. https://doi.org/10.1080/17457289.2025.2514199

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: A practical approach to modelling cognitive processes*. AcademicPress.

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27). Council of Europe. https://www.firstdraftnews.org/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-de%CC%81sinformation-1.pdf

Zuhro, N. A., & Rakhmawati, N. A. (2020). Clickbait detection: A literature review of the methods used. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, *6*(1), 1–10. https://journal.unipdu.ac.id/index.php/register/article/view/1561

About the Authors

Lucas Stampe is completing his PhD in information systems at the University of Münster, Germany. His research interests center on the development of quantitative computational methods designed to make sense of the complex phenomena inherent in online media, aiming to understand the dynamics that shape and sustain open public discourse.

Christian Grimme is an associate professor at the University of Münster, Germany, heading the Computational Social Science and Systems Analysis group. His research specializes in social media analytics, online propaganda, and algorithms. His methodological focus is on the development of hybrid algorithms as well as on graph-based and multi-objective optimization.