vera.ai: VERification Assisted by Artificial Intelligence

D4.2 - Coordinated sharing behaviour detection and disinformation campaign modelling methods

Project Title	vera.ai
Contract No.	101070093
Instrument	HORIZON-RIA
Thematic Priority	CL4-2021-HUMAN-01-27
Start of Project	15 September 2022
Duration	36 months

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.

Deliverable title	Coordinated sharing behaviour detection and disinformation campaign modelling methods
Deliverable number	D4.2
Deliverable version	V1.0
Previous version(s)	N/A
Contractual Date of delivery	14.07.2025
Actual Date of delivery	14.07.2025
Nature of deliverable	Report
Dissemination level	Public
Partner Responsible	UNIURB
Author(s)	Fabio Giglietto, Giada Marino, Anwesha Chakraborty, Nicola Righetti (UNIURB), Charis Bouchlis (ATC), Dimitris Karageorgiou (CERTH), Inès Gentil, Francesco Poldi (EUDL), Milica Gerhardt (IDMT), Martin Hyben (KiniT), Andrey Tagarev (ONTO), Xingyi Song, Mali Jin (USFD)
Reviewer(s)	Denis Teyssou (AFP), Miguel Colom, Mehdi Boussâa, Yanhao Li (ENS)
EC Project Officer	Peter Friess

Abstract	D4.2 presents methodologies for detecting coordinated sharing behavior and modeling disinformation campaigns, developed within Tasks T4.2-T4.3 of WP4. The research addresses the growing sophistication of disinformation operations through innovative network science methods and spatio-temporal narrative analysis techniques.
Keywords	Coordinated Inauthentic Behavior, Disinformation Detection, Social Media Analysis, Network Science, Multimodal Analysis, Spatio-Temporal Narratives

Copyright

Revision History

Version	Date	Modified by	Comments
V0.1	29/01/2025	Giada Marino (UNIURB)	Draft ToC
V0.2	01/06/2025	Giada Marino (UNIURB)	Initial draft
V0.3	16/06/2025	Giada Marino (UNIURB), Ines Gentil (EUDL), Francesco Poldi (EUDL)	Added Content Sections
V0.4	20/06/2025	Andrey Tagarev (ONTO), Mali Jin (USFD), Martin Hyben (Kinit)	Added Content Sections
V0.5	25/06/2025	Fabio Giglietto (UNIURB), Giada Marino (UNIURB)	Added content Sections
V0.6	28/06/2025	Fabio Giglietto (UNIURB)	Added Executive summary, user need alignment and KPI
V0.7	04/07/2025	Denis Teyssou (AFP), Miguel Colom (ENS), Mehdi Boussâa (ENS), Yanhao Li (ENS), Marina Gardella (ENS)	Feedback and comments
V0.8	07/07/2025	Fabio Giglietto, Giada Marino, Anwesha Chakraborty (UNIURB)	Addressed comments
V0.9	07/07/2025	Ines Gentil (EUDL), Francesco Poldi (EUDL), Mali Jin (USFD)	Final feedback and comments
V1.0	14/07/2025	Fabio Giglietto, Giada Marino, Anwesha Chakraborty (UNIURB), Olga Papadopoulou, Symeon Papadopoulos (CERTH)	Deliverable sent to EC

Glossary

Abbreviation	Meaning
DoA	Description of Action
WP	Work Package
CIB	Coordinated Inauthentic Behaviour
CSB	Coordinated Sharing Behaviour
CSDS	Coordinated Sharing Detection Service
DBKF	Database of Known Fakes

Table of Contents

Index of Tables

Index of Figures

Executive Summary

Since its inception, Work Package 4 of the vera.ai project has focused on developing tools and frameworks to detect and analyze coordinated disinformation campaigns across platforms. Recognizing the limitations of single-platform monitoring, the work prioritized a multimodal, cross-platform approach capable of adapting to evolving online dynamics.

This deliverable outlines progress made in Tasks T4.2 and T4.3, which focused respectively on developing detection methodologies and translating them into operational services. A central outcome was the development of CooRTweet, a coordination detection system designed to operate independently of specific platform APIs. By analyzing diverse media formats—including URLs, images, text, and audio—CooRTweet enables detection of network-wide coordinated behavior. In early testing using ground-truth datasets, CooRTweet achieved a 92% detection rate, outperforming baseline methods.

Complementing this technical development, the team proposed a four-branch framework for identifying Coordinated Inauthentic Behavior (CIB). This framework incorporates multiple dimensions—coordination, authenticity, source, and impact—to support a more comprehensive analysis of online operations. It was applied in various case studies, including Operation Overload and disinformation campaigns on TikTok, generating likelihood scores for coordination and helping to contextualize the findings.

Advances were also made in developing narrative discovery methodologies, particularly in tracking the emergence and evolution of disinformation themes across time and regions. By combining reinforcement-learning-based clustering techniques with document association algorithms, the project generated structured representations of evolving narratives. These methods yielded high coherence and clustering performance, and were embedded into visual tools for exploration and analysis.

These methodological contributions were translated into a set of services for operational use. A web-based interface, the Coordinated Sharing Detection Service (CSDS), provides real-time insights into detected coordination patterns. The VeraAI Alert System, a global-scale monitoring infrastructure for coordinated behavior detection, scaled from a small pilot to tracking over 10,000 coordinated links and more than 2,100 new accounts. A dedicated TikTok monitoring initiative processed over 1.2 million posts, identifying thousands of potentially coordinated behaviors despite platform access limitations.

The project also encountered challenges—most notably, the deprecation of Meta's CrowdTangle tool—which highlighted the risks of relying on unstable APIs and reinforced the importance of building platform-agnostic systems. Continued issues with research data access underscore the importance of regulatory frameworks, such as the Digital Services Act, in supporting transparency and accountability in platform research.

Overall, the work provides a foundation for scalable, cross-platform monitoring of disinformation, balancing automated detection with human analysis. It offers both methodological innovations and practical tools for researchers, journalists, and fact-checkers, supporting ongoing efforts to understand and mitigate the impact of coordinated disinformation in digital public spheres.

1 Introduction

This deliverable, *D4.2 Coordinated Sharing Behaviour Detection and Disinformation Campaign Modelling Methods*, presents the methodologies, tools, and validation results developed in Tasks T4.2 and T4.3 of WP4, **Analysis of Complex Disinformation Phenomena**. Task T4.2 focuses on the spatio-temporal analysis of disinformation campaigns and narratives, while Task T4.3 addresses the detection of coordinated sharing behaviour using network science methods.

The report introduces a set of methodological innovations designed to overcome current limitations in the detection and analysis of disinformation. These contributions serve the broader objective of WP4: enabling the systematic and scalable investigation of complex information manipulation phenomena across platforms and media types.

A key outcome is the development of **CooRTweet**, a **platform-independent engine for detecting coordinated sharing activity**. Unlike earlier tools, CooRTweet adopts a generalized architecture that supports cross-platform analysis and accommodates diverse content types, including text, URLs, images, and audio. It is available both as an R package for researchers and as a web-based interface tailored for practitioners such as journalists and fact-checkers. Preliminary validation using ground-truth datasets shows that the tool achieves 92% accuracy in identifying coordinated networks—significantly outperforming baseline approaches.

Complementing this, the deliverable introduces a **systematic framework for detecting Coordinated Inauthentic Behaviour (CIB)**. This four-branch detection tree evaluates coordination, authenticity, source characteristics, and potential impact, offering a structured, interpretable approach to identifying and categorizing CIB. The framework has been applied in case studies involving Operation Overload, Russian influence operations on TikTok, and QAnon-related activity.

In parallel, the methodology expands beyond traditional text-based detection by **integrating multimodal and cross-platform analysis techniques**. These include visual similarity detection, audio provenance analysis, and cross-modal validation, allowing for the detection of sophisticated, multi-platform campaigns.

The workflow also includes a **spatio-temporal narrative analysis** aimed at uncovering coordinated disinformation campaigns. This component explores how disinformation narratives evolve over time and across geographic locations, analyzing themes, topics, and contextual adaptation strategies.

1.1 Dynamic Defense System Workflow

The tools and methods developed in this deliverable are integrated into a broader, dynamic workflow designed to improve responsiveness to emerging disinformation threats. This system, illustrated in Figure 1 combines automated detection, expert validation, and continuous updating. It operates in near real-time and is adaptable to changes in platform policies and technological environments.

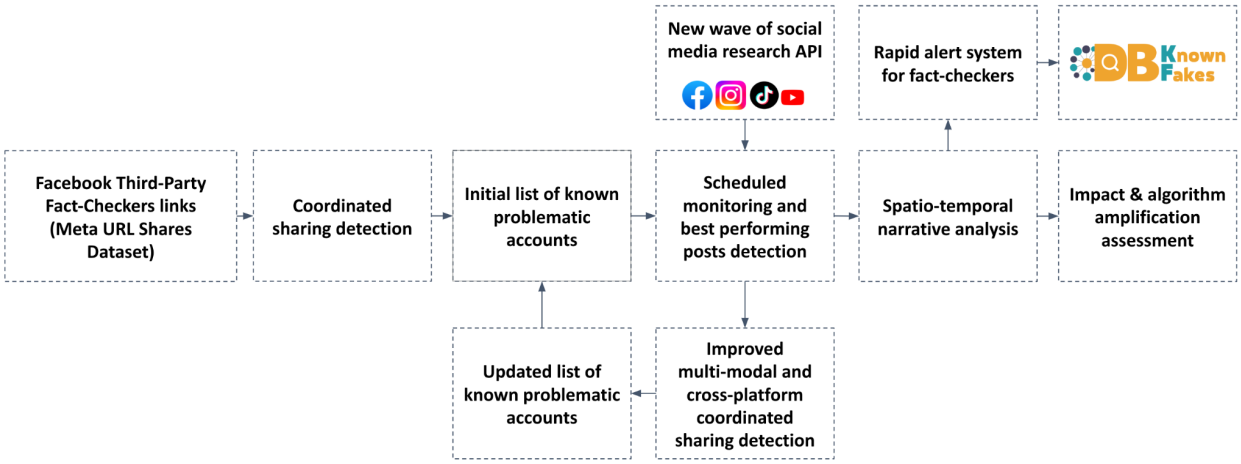


Figure 1 Multi-platform disinformation detection workflow combining coordinated sharing detection, continuous monitoring, and rapid alert systems.

The framework identifies initial problematic accounts, analyzes coordinated sharing patterns, and monitors scheduled posts to enable improved multimodal detection capabilities. Spatio-temporal narrative analysis assesses impact and algorithm amplification, while rapid alerts notify fact-checkers through RSS and a dedicated Slack App. Dynamic feedback loops continuously update the problematic accounts list, enhancing detection accuracy over time.

Several operational implementations demonstrate the effectiveness and applicability of the developed tools:

- The **VeraAI Alert System** monitored global disinformation from October 2023 to August 2024, scaling from an initial 1,225 accounts to over 10,000 identified coordinated links and more than 2,100 additional accounts. It revealed a variety of techniques, including the use of public groups, engagement baiting, political propaganda, and AI-generated content.
- In the **TikTok Coordinated Detection Project**, over 1.26 million posts were analyzed across 136 days, leading to the identification of more than 2,500 coordinated posts involving 8,574 account instances and over 2,200 distinct coordination networks.
- The **CIB Detection Tree** has been applied to content surfaced by these systems, identifying AI-generated campaigns such as the spread of false claims about Pope Francis's health, which quickly gained substantial engagement.
- A **Temporal Analysis Visualization Tool** has been developed to support the exploration of disinformation narrative dynamics. This web-based interface allows users to upload datasets and topic modeling results, generating interactive stream graphs and document relationship maps. The tool supports dynamic filtering by keyword, date, and topic hierarchy, offering interpretable insights into how disinformation narratives evolve and cluster over time.

Another crucial focus of this deliverable is the shifting landscape of platform data access. The deprecation of Meta's CrowdTangle in August 2024 has significantly disrupted research workflows. In response, this deliverable underscores the importance of platform-agnostic tools and highlights the role of regulatory

frameworks, such as the **Digital Services Act**, in ensuring continued access to data for public-interest research.

1.2 Deliverable Structure

The remainder of the deliverable is structured according to the components of the detection workflow:

- **Section 2** outlines user needs, performance indicators, and the research context that guided methodological development.
- **Section 3** details the methods for detecting coordinated sharing behaviour, including the CIB Detection Tree, the CooRTweet tool, and multimodal techniques.
- **Section 4** presents the spatio-temporal narrative analysis methods, supporting the identification and tracking of disinformation narratives across time and regions.

In summary, this deliverable presents a robust set of tools and frameworks that advance current capabilities for disinformation detection and analysis. Its main contributions include:

- A platform-independent coordination detection engine with multimodal support;
- A structured, scalable framework for identifying and evaluating Coordinated Inauthentic Behaviour;
- Operational implementations demonstrating real-world applicability;
- A temporal analysis tool for interactive exploration of disinformation narratives;
- A reflection on data access challenges and the need for sustainable, open research infrastructure.

By integrating automated detection with expert interpretation, the workflow outlined here enables a flexible and scalable approach to identifying coordinated behaviour and complex disinformation campaigns—contributing to broader efforts to safeguard digital information ecosystems.

2 Overview of User Needs, KPIs, and Research Contributions

The methodologies and implementations documented in this deliverable emerged from a comprehensive user-centered design process that placed the needs of media professionals, fact-checkers, researchers, and civil society organizations at the center of our research and development efforts. As disinformation operations have evolved to become increasingly sophisticated—employing AI-generated content, cross-platform coordination, and adaptive evasion strategies—the tools available to combat these threats have struggled to keep pace with this rapid evolution (DiResta & Goldstein, 2024).

Tasks T4.2 (Spatio-temporal analysis of disinformation campaigns and narratives) and T4.3 (Tackling coordinated sharing behaviour with network science methods) were conceived as part of the dynamic defense system workflow outlined in Section 1, designed to address critical capability gaps identified through consultation with end-users across the information verification ecosystem. This consultation process revealed a fundamental disconnect between the technical capabilities of existing detection tools and the operational realities faced by practitioners working to identify and respond to coordinated inauthentic behavior in real-world scenarios.

The vera.ai consortium's approach to addressing these challenges was grounded in the recognition that effective disinformation detection requires more than technical sophistication—it demands tools that are accessible, interpretable, and actionable for non-technical users while maintaining the analytical rigor necessary for complex coordination analysis. This dual imperative shaped our development of platform-independent methodologies that could transcend the limitations of platform's API-dependent tools while providing the transparency and explainability essential for building trust in automated detection systems.

Through iterative engagement with practitioners, including data journalists participating in our user testing sessions and OSINT researchers contributing to our validation processes, the vera.ai consortium identified and detailed in D2.1 AI against Disinformation: Use Cases and Requirements two primary categories of requirements that guided our research priorities. These include technical tool requirements addressing functional capabilities and performance needs, and design and usability considerations focusing on accessibility, interpretability, and workflow integration. These requirements directly informed the development of CooRTweet's universal detection architecture[1] and GUI[2], the VeraAI Alert System[3] and TikTok Coordinated Detection Project[4] real-time monitoring capabilities, and the comprehensive spatio-temporal analysis tools documented throughout this deliverable.

The successful achievement of our key performance indicators demonstrates not only the technical effectiveness of our approaches but also their practical utility in addressing the evolving challenges of information integrity in digital environments. This section provides a detailed analysis of how these user-

[1] https://cran.r-project.org/web/packages/CooRTweet/index.html
[2] https://coortweet.lab.atc.gr/
[3] Now discontinued after the CrowdTangle replacement with the Meta Content Library (MCL).
[4] https://fabiogiglietto.github.io/tiktok_csbn/tt_viz.html

driven requirements shaped our methodological innovations and examines the extent to which our implementations have successfully addressed the identified gaps in current detection capabilities.

2.1 Tool Technical Requirements

Platform Expansion and Cross-Platform Analysis: In D2.1 AI against Disinformation: Use Cases and Requirements[5] users expressed an urgent need to extend social network analysis capabilities beyond commonly investigated platforms like Facebook and Twitter to include established but overlooked platforms such as TikTok, Telegram, and YouTube. The research responded by developing CooRTweet as a platform-agnostic detection engine, successfully implementing coordinated behavior detection on TikTok through the TikTok Research API, and creating blueprints that can be adapted to any platform regardless of API availability or restrictions.

Real-Time Narrative Monitoring: Media professionals required tools to monitor evolving disinformation narratives across both mainstream commercial platforms and fringe networks. The VeraAI Alert System addressed this need by providing continuous monitoring capabilities that expanded from 1,225 initial accounts to identify over 10,000 coordinated links, enabling real-time detection of emerging threats and narrative evolution. Additionally, the spatio-temporal analysis methods described in Section 4 provide narrative discovery and temporal tracking capabilities that analyze how campaigns evolve across time and geographic regions.

Multilingual Detection Capabilities: Users emphasized the need to detect information operations and campaigns that operate across language barriers, particularly those using "seeding" keywords to initiate narratives in multiple languages. In the context of the works done for the spatio-temporal narrative analysis tool, the research developed multilingual clustering systems through ONTO's implementation and KInIT's narrative detection pipeline, enabling cross-linguistic coordination detection and narrative tracking across diverse linguistic contexts.

Multimodal Analysis: Users required capabilities to detect coordinated behavior across diverse content types beyond traditional text-based approaches, including images, videos, and audio content. The modal-agnostic architecture of CooRTweet addresses this need through integrated multimodal detection capabilities, demonstrated by the visual similarity methods[6] for cross-platform coordinated sharing detection (section 3.2.2) and audio provenance analysis techniques for coordinated sharing detection (section 3.2.3).

Speed and Efficiency in Analysis: Given the fast-paced nature of news cycles, users required tools that could accelerate the analysis of news articles and social media content without sacrificing accuracy. The automated detection capabilities of CooRTweet, combined with AI-powered content labeling using GPT-4o, significantly reduced the time required to identify and characterize coordinated networks.

Attribution and Actor Identification: Fact-checkers and journalists needed capabilities to recognize and determine individuals or groups intentionally spreading false or misleading information. The CIB Detection Tree framework and source assessment methodologies provided structured approaches to attribution while maintaining compliance with privacy standards.

[5] This deliverable is marked as sensitive / restricted and can't therefore be publicly distributed or linked.

[6] These methods are comprehensively described in D4.1 Cross-lingual and multimodal near-duplicate search methods.

2.2 Design and Usability Considerations

Transparency and Explainability: Users consistently emphasized the need for human verification and trust in automated results, requiring clear understanding of how tools reach their conclusions. The research addressed this through the development of visual assessment methodologies for CIB campaigns, comprehensive indicator frameworks, and interactive visualizations that allow users to examine the evidence underlying detection results.

Accessible Interface Design: The complexity of coordination detection was identified as a significant barrier, leading to demands for simpler dashboard solutions and better user interface design. The CooRTweet web interface underwent user testing and iterative improvement based on feedback from data journalists, resulting in streamlined workflows and enhanced interpretability.

Cross-Platform Operability: Users require seamless analysis capabilities across multiple social media platforms within a single workflow. CooRTweet's universal architecture directly addressed this need by enabling analysis of coordination patterns regardless of platform origin, supporting both single-platform and multi-platform investigations.

Personalization and Targeted Analysis: Media professionals required tools to focus investigations on specific events, actors, or time periods. The spatio-temporal analysis tools developed by USFD, ONTO and KInIT enable users to examine trends at specific points in time and filter results based on relevant topics, named entities, or geographic regions.

Alert and Notification Systems: Users required proactive notifications to identify emerging stories worth analyzing rather than relying solely on reactive investigation. The VeraAI Alert System's RSS Slack channel and automated detection pipeline provided timely alerts about high-performing content and newly identified coordinated networks.

2.3 Key Performance Indicators (KPIs)

Network-based Coordinated Sharing Behaviour Analysis

Agreement between Results of Coordinated Detection and Expert Ratings: The validation of CooRTweet using the South Korean National Intelligence Service (NIS) dataset achieved 92% accuracy in detecting known coordinated accounts, significantly outperforming existing methodologies. This validation against ground truth data demonstrates strong alignment between automated detection results and expert human assessment of coordinated behavior.

The CIB Detection Tree framework applied to three major case studies (Operation Overload, Russian TikTok influence operations, and QAnon's "Save the Children" campaign) achieved quantified CIB likelihood scores ranging from 32% to 64%, providing structured assessment metrics that enable comparison across different campaign types and validation of detection methodologies.

Disinformation Alerting Mechanism

Successful Use of Social Network Analysis and Alerting Mechanism: The research achieved substantial scale in identifying coordinated behavior across multiple platforms and timeframes, with detection capabilities exceeding the 50,000 accounts/incidents successfully suggested to media professionals threshold across multiple metrics:

Coordinated Content Detection: The VeraAI Alert System identified over 400,000 posts in a single AI-generated disinformation campaign exploiting Pope Francis health concerns within one week, demonstrating the system's capability to detect large-scale coordinated operations. The TikTok Coordinated Detection Project processed over 1.26 million TikTok posts, identifying 2,521 coordinated posts across 8,574 account instances and 2,248 distinct coordination networks, showcasing the methodology's scalability to video-first platforms.

Network Expansion and Discovery: The VeraAI Alert System demonstrated systematic network expansion from 1,225 initial monitored accounts to identification of 2,126 additional coordinated accounts and 10,681 coordinated links, proving the system's capability to surface previously unknown threats through iterative network analysis.

The alert mechanism's effectiveness was validated through the detection of multiple categories of problematic behavior, including exploited large groups, casino engagement bait schemes, political propaganda networks, and AI-generated disinformation campaigns. The system maintained continuous monitoring capabilities until Meta's CrowdTangle deprecation forced operational changes in August 2024. This established a proven framework for large-scale coordination detection that significantly exceeded performance targets. The expert opinion in Annex II details the challenges the team faced while adapting the Vera AI alerts to the Meta Content Library & API.

2.4 Overview of Research Contributions

This section provides an overview of the primary research contributions from the project, which directly address the challenge of increasingly sophisticated disinformation campaigns. Our work has focused on developing and validating novel methodologies that are both platform-independent and multi-modal, capable of detecting coordinated behavior across various content types. These efforts have resulted in scalable, real-world applications and a range of tangible outputs. The following subsections detail these innovations, their practical impact, and the resulting publications and tools.

2.4.1 Addressing Evolving Disinformation Campaign Sophistication

To counter sophisticated disinformation that uses AI, multiple platforms, and various media formats, the project introduced several key advancements:

Platform-Independent Detection Architecture: The development of CooRTweet represents a paradigm shift from platform-specific tools to universal detection engines capable of operating across any social media environment. This innovation directly addresses the challenge of platform API restrictions and ensures longevity of detection capabilities.

Multimodal Coordination Detection: The integration of visual similarity analysis, audio provenance detection, and cross-modal validation techniques extends coordination detection beyond traditional text-based approaches, enabling identification of sophisticated campaigns that employ diverse content types to evade detection.

Comprehensive CIB Assessment Framework: The CIB Detection Tree provides the first systematic framework for evaluating coordinated inauthentic behavior across four critical dimensions: coordination, authenticity, source attribution, and impact assessment.

2.4.2 Practical Impact and Scalability

Our research has demonstrated real-world effectiveness across diverse threat scenarios, from small-scale coordination clusters to large-scale information campaigns. The TikTok implementation successfully processed over 1.26 million posts, while the VeraAI Alert System maintained continuous monitoring capabilities until platform API restrictions forced operational changes.

The platform-independent architecture and multi-modal detection capabilities position the developed methodologies to address increasingly sophisticated disinformation campaigns employing AI-generated content, cross-platform coordination, and adaptive evasion strategies. The integration of automated detection with human expertise creates a scalable framework for maintaining information integrity in evolving digital environments.

2.5 Partners List of Outputs & Publications

During the reporting period, significant progress was achieved in both research and development. Fourteen research outputs were produced from T4.2 and T4.3 activities under the vera.ai project, including scientific reports, publications, tools, and organized conferences. These outputs are organized by tasks in the list below.

Outputs Related to Task 4.2 [Section 4]

1. Mu, Y., Bai, P., Bontcheva, K., & Song, X. (2024). Addressing topic granularity and hallucination in large language models for topic modelling. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2405.00611
2. Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 10160-10171.
3. Ontotext. (2023). Database of Known Fakes. https://dbkf.ontotext.com/

Outputs Related to Task 4.3 [Section 3]

4. Bontcheva, K., Scarton, C., Zareie, A., Giglietto, F., Marino, G., Righetti, N., Angus, D., FitzGerald, K., Graham, T., Zhu, G., Cuccovillo, L., Gerhardt, M., Karageorgiou, D., Papadopoulou, O., Papadopoulos, S., Tagarev, A., & Rossi, L. (2024, October 29). Coordinated Sharing Behavior Detection Conference. University of Sheffield, Sheffield, UK.
5. Gerhardt, M., Cuccovillo, L., & Aichroth, P. (2023). Advancing audio phylogeny: A neural network approach for transformation detection. IEEE International Workshop on Information Forensics and Security (WIFS), 1–6.
6. Gerhardt, M., Cuccovillo, L., & Aichroth, P. (2024). Audio provenance analysis in heterogeneous media sets, in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, Washington, USA, 2024, pp. 4387–4396.

7. Giglietto, F. (2024). tiktok_csbn: TikTok coordinated account live map. Github. https://github.com/fabiogiglietto/tiktok_csbn
8. Giglietto, F., Graham, T., & Righetti, N. (in press). Navigating coordination and inauthentic behaviour: Challenges and innovations in social media detection. In Routledge Companion to Social Media and Politics. Routledge.
9. Giglietto, F., Marino, G., Mincigrucci, R., & Stanziano, A. (2023). A Workflow to Detect, Monitor and Update Lists of Coordinated Social Media Accounts Across Time: The Case of 2022 Italian Election. Social Media + Society.
10. Marino, G., Giglietto, F., Chakraborty, A., Terenzi, M., & Olaniran, S. (forthcoming). Throwing Spaghetti, Seeing What Sticks: Iterative Deception in Digital Strategic Information Operations. AoIR2024, Fluminense Federal University.
11. Marino, G., & Giglietto, F. (2024). Integrating Large Language Models in political discourse studies on social media: Challenges of validating an LLMs-in-the-loop pipeline. In Sociologica (Vol. 18, Issue 2, pp. 87–107). Sociologica. https://doi.org/10.6092/ISSN.1971-8853/19524
12. Righetti, N., & Balluff, P. (2025). CooRTweet: A generalized R software for coordinated network detection. Computational Communication Research, 7(1), 1. https://doi.org/10.5117/ccr2025.1.7.righ
13. Romero-Vicente, A. (2025). Visual assessment of CIB in disinformation campaigns. EU Disinfo Lab. https://disinfo.eu/visual-assessment-of-cib-in-disinformation-campaigns/
14. Rossi, L., Giglietto, F., & Marino, G. (2023). Cracking Open the European Newsfeed. Journal of Quantitative Description: Digital Media, 3. https://doi.org/10.51685/jqd.2023.020

3 Tackling Coordinated Sharing Behaviour with Network Science Methods

This section presents the methodological foundations and practical implementations for detecting coordinated sharing behavior through advanced network science approaches, representing one of the core pillars of the vera.ai project's dynamic defense system workflow (see Section 1.1 Dynamic Defense System workflow). Building upon the user needs analysis and key performance indicators outlined in Section 2, this chapter details how the project has developed and deployed cutting-edge detection methodologies that address the evolving sophistication of modern disinformation campaigns operating across multiple platforms and content modalities.

The methodologies documented in this section directly respond to the critical challenge identified throughout the vera.ai project: the need for platform-independent, multi-modal coordination detection capabilities that can adapt to rapidly changing technological landscapes and platform policies. As demonstrated by the deprecation of CrowdTangle by Meta in August 2024 and the resulting disruption of CooRnet (Giglietto et al., 2020a) and the VeraAI Alert System, the prescient development of platform-agnostic approaches has proven essential for maintaining operational effectiveness in an environment of increasing API restrictions and policy changes.

At the core of this section is the presentation of CooRTweet, a platform-independent coordinated sharing detection engine that transcends the limitations of previous tools tied to specific platform APIs. Unlike earlier solutions such as CooRnet, which specialized in Facebook and Instagram link-sharing analysis, CooRTweet's generalized architecture enables unprecedented analytical flexibility across any social media environment, supporting single-platform and multi-platform investigations, mono-modal and multi-modal content analysis, and coordination detection around any "uniquely identifiable content." This methodological innovation directly addresses the dynamic defense system requirements established in Section 2, providing the foundational infrastructure for continuous monitoring and adaptive detection capabilities.

The section systematically builds from theoretical foundations to practical implementations, beginning with a comprehensive analysis of the evolution and scope of coordinated behavior research. This background establishes the academic and operational context that has shaped current understanding of coordination as "near-simultaneous sharing" by stable groups of accounts, while highlighting emerging challenges in multimodal and cross-platform detection that the vera.ai project specifically addresses. The discussion incorporates insights from the Coordinated Sharing Behavior Detection Conference held at the University of Sheffield in October 2024, which brought together leading international experts and formally recognized the field's remarkable growth from 200 published papers and 45 preprints between 2019 and 2024.

Central to the methodological contributions is the CIB Detection Tree framework developed by EU DisinfoLab, which provides the first systematic approach to evaluating Coordinated Inauthentic Behavior across four critical dimensions: coordination assessment, authenticity assessment, source assessment, and impact assessment. This framework addresses the fragmented nature of CIB definitions across digital platforms and the absence of formalised regulatory definitions, offering a unified analytical lens that has

been successfully applied to major case studies including Operation Overload (discovered in June 2024), Russian TikTok influence operations (flagged in December 2023), and QAnon campaigns (spotted in October 2017), yielding quantified CIB likelihood scores ranging from 32% to 64%. Beyond their analytical value, these case studies emphasise the importance of developing structured analytical techniques for retrospective evaluation and comparison of influence operations: an important objective for strengthening long-term response capabilities.

The section's methodological innovations extend beyond traditional text-based detection through the integration of multimodal analysis capabilities. These include visual similarity detection leveraging the Near Duplicate Detection Service, audio provenance analysis for tracing content transformation and reuse across platforms, and cross-modal validation techniques that detect contradictions between audio and textual content as indicators of deliberate manipulation. This comprehensive approach enables the detection of sophisticated campaigns that employ diverse content types to evade traditional detection methods, addressing the limitations of single-modality approaches highlighted in contemporary research.

Implementation and operational validation form a crucial component of this section, demonstrating the real-world effectiveness and scalability of the developed methodologies. The VeraAI Alert System operated at global scale from October 2023 to August 2024, expanding from 1,225 initial monitored accounts to identify over 10,000 coordinated links and 2,126 additional accounts, uncovering diverse operational categories including exploited large groups, casino engagement bait schemes, political propaganda networks, and AI-generated disinformation campaigns. The TikTok Coordinated Detection Project processed over 1.26 million posts across 136 monitoring days, identifying 2,521 coordinated posts and 2,248 distinct coordination networks, demonstrating successful adaptation to video-first platforms.

The section also addresses critical challenges and adaptations necessitated by platform policy changes, particularly the impact of CrowdTangle's deprecation on coordination detection research. Through comprehensive analysis of Meta's successor tools, the research reveals fundamental limitations in current platform transparency initiatives while demonstrating the foresight of developing platform-independent methodologies. This analysis contributes essential insights for regulatory frameworks like the Digital Services Act and highlights the critical importance of sustainable research infrastructure for maintaining information integrity.

Throughout this section, the integration of user-centered design principles ensures that sophisticated detection capabilities remain accessible to practitioners. The CooRTweet web interface, developed through iterative user testing with data journalists and OSINT researchers, exemplifies how complex network science methodologies can be translated into intuitive tools that support real-world investigations. The comprehensive validation against ground truth datasets, including the South Korean National Intelligence Service dataset where CooRTweet achieved 92% accuracy, demonstrates both technical effectiveness and practical utility.

The methodologies presented in this section establish a robust foundation for next-generation disinformation detection and analysis, positioning the vera.ai project to address increasingly sophisticated information manipulation campaigns employing AI-generated content, cross-platform coordination, and adaptive evasion strategies. The successful integration of automated detection with human expertise creates a scalable model for combating coordinated inauthentic behavior while preserving the nuanced

analysis required for complex disinformation campaigns, ensuring both technical precision and practical utility for maintaining information integrity in evolving digital environments.

3.1 The Evolution and Scope of Coordinated Behavior Research

Coordinated behavior is a core feature of online communication, particularly in the context of information campaigns. As Keller et al. (2019) note, coordinated messaging is inherent to any effort aimed at influencing public discourse. Similarly, Giglietto et al. (2020b) describe coordination as a defining characteristic of user participation in digital spaces. This understanding has evolved significantly since the early recognition of social media's role in collective action, from the Arab Spring to the Occupy Wall Street movement, which established foundations for understanding how digital platforms afford large-scale mobilization, coordinated action, and synchronized communication.

The academic study of coordinated behavior on social media has experienced remarkable growth, particularly following Meta's introduction of the term "Coordinated Inauthentic Behavior" (CIB) in 2018 (Gleicher, 2018). This growth was formally recognized at the Coordinated Sharing Behavior Detection Conference[7] held at the University of Sheffield on October 29, 2024, organized in the context of the vera.ai project, which brought together leading international experts to discuss state-of-the-art methodologies and challenges in the field. A full report on this conference is available on the project website at Insights and Report from the Coordinated Sharing Behavior Detection Conference[8].

As highlighted at the Coordinated Sharing Behavior Detection conference, recent years have witnessed a significant uptick in the study of coordinated communication networks across various social media platforms. The conference emphasized that these networks aim to influence and manipulate audiences on social media platforms and are often linked to problematic behaviors, including spreading misinformation and conducting state-sponsored information operations (Giglietto et al., 2020b; Keller et al., 2019; Starbird et al., 2019).

A Scopus database search for journal articles including the term 'coordinated inauthentic behaviour' shows that there are 200 published papers and 45 preprints between 2019 and 2024. The number of papers published that mention CIB almost doubled from 2022 to 2023, demonstrating the field's rapid expansion and increasing academic recognition.

3.1.1 Methods, Detection Approaches and Challenges

Recent research in computational social science has increasingly focused on detecting such coordination on social media platforms (Cinus et al., 2025). Studies have identified coordinated networks operating in a quasi-synchronous manner to disseminate specific content and narratives. These networks have been observed in national contexts such as Italy (Giglietto et al., 2020b), Germany (Righetti et al., 2022), Australia (Graham et al., 2020), Nigeria (Giglietto et al., 2022), South Korea (Keller et al., 2019), Brazil and France (Gruzd et al., 2022), as well as in transnational campaigns (Righetti, 2025; Righetti et al., 2025).

[7] https://sites.google.com/uniurb.it/csbdetectionconf
[8] https://www.veraai.eu/posts/coordinated-sharing-behavior-detection-conference-2024

Many of these operations are linked to right-wing populist movements, authoritarian regimes (Kulichkina et al., 2024), or economic incentives (Terenzi, 2025).

Core Methodological Frameworks

At the operational level, coordinated behavior is primarily understood through the lens of synchronicity—specifically, the near-simultaneous sharing of content by a stable group of accounts, identified as indicative of orchestrated operations (Giglietto et al., 2020b; Graham et al., 2020). The Coordinated Sharing Behavior Detection conference reinforced this understanding while highlighting emerging challenges in multimodal and cross-platform detection.

The most common methodology in social and communication sciences involves two main indicators:

- Posting Synchronicity: Emphasized as the core element for identifying coordinated actions across networks
- Repetition: Serving as a secondary criterion to refine the application of synchronicity criteria

Giglietto and colleagues conceptualize coordination as "near-simultaneous sharing," while Graham and colleagues highlight cases where accounts share identical content within short timeframes. The repetition of synchronized actions over time enhances the identification of coordinated networks, as they are expected to be relatively stable social media structures that repeatedly exhibit synchronized behavior, as opposed to isolated instances of synchronicity that may happen by chance.

Emerging Challenges and Research Frontiers

The Coordinated Sharing Behavior Detection conference brought together leading scholars in the field of coordination detection, shedding light on new research frontiers and pressing methodological challenges. A key limitation identified across presentations was the over-reliance on single-modality and single-platform analyses. As highlighted in the conference rationale, empirical research still tends to focus on isolated signals—such as URL dissemination, hashtags, or textual content—making it difficult to detect deeper coordination patterns at the level of issues or narratives. This gap is especially urgent in light of the growing use of generative AI, which enables the creation of diverse yet semantically aligned content.

Multimodal and Cross-Platform Detection

Several presentations addressed the need for multimodal and cross-platform approaches:

- Dynamic and Temporal Analysis: Stefano Cresci advocated for longitudinal methods, using sliding time windows to capture evolving patterns in networked behavior. This temporal lens helps distinguish between ephemeral coordination and sustained, strategic activity.
- Embedding-Based Methods: Daniel Thiele and Miriam Milzner proposed an embedding-based framework capable of identifying multimodal coordination. Their method accounts for the variability introduced by LLMs and generative AI, which challenge traditional similarity-based detection.
- Visual Content and Computer Vision: Daniel Angus introduced multimodal analysis frameworks that integrate computer vision and natural language processing to detect coordinated inauthentic behavior on visually intensive platforms.

- Cross-Platform Link Sharing: Jakob Kristensen demonstrated that URLs function as universal identifiers across platforms, allowing the detection of coordinated link-sharing behavior even in heterogeneous platform environments.

Non-Western and Authoritarian Contexts

A key contribution of the conference was the emphasis on linguistic diversity and geopolitical variation in coordination practices:

- Sociolinguistic Patterns: Raquel Recuero analyzed Russian influence operations on Brazilian Telegram, uncovering sociolects, signs of poor translation, and shared themes across channels as indicators of cross-channel coordination.
- Protest and Repression in Authoritarian Regimes: Aytalina Kulichkina explored how coordination facilitates both dissent and repression in Russia and China, illustrating its dual function in contested political environments.
- Content Moderation Gaps: Felipe Bonow Soares examined disinformation within Brazilian buy-and-sell Facebook groups, pointing to Meta's uneven enforcement of moderation policies in Portuguese-language contexts.

The Role of Generative AI

Generative AI emerged as a transformative force in coordination practices. Multiple speakers highlighted how AI-generated content—tailored, unique, and semantically consistent—renders traditional detection methods based on identical or highly similar content obsolete. As Luca Rossi emphasized in his conference address, temporal dimensions of coordination are increasingly vital. He introduced the metaphor of "pillar" (continuous) versus "pebble" (punctual) coordination to conceptualize different patterns of orchestrated behavior over time.

Access and Data Infrastructure

The deprecation of CrowdTangle APIs by Meta in August 2024 was identified as a turning point, symbolizing the broader shift toward restrictive data environments. The transition to "digital clean rooms" and proprietary data access models undermines researchers' ability to monitor coordination in real-time and raises critical concerns about the future of platform accountability.

In sum, the Coordinated Sharing Behavior Detection conference marked a methodological and conceptual evolution in the study of coordination. The field is moving beyond threshold-based and monomodal detection, embracing more complex, ethically aware, and platform-independent frameworks. Still, participants repeatedly stressed that advancing detection must go hand in hand with preserving legitimate forms of online collective action.

Methodological Innovation and Tool Development

The conference showcased a range of methodological advances aimed at overcoming existing detection limitations:

- Anomaly-Based Detection: Ahmad Zareie proposed a hierarchical anomaly-based model that focuses on behavioral outliers rather than content similarity. This approach addresses key limitations of traditional threshold-based models, especially in complex, evolving contexts.
- Knowledge Graphs: Jennifer Stromer-Galley illustrated how structured relationships between actors, topics, and content can be used to detect coordinated inauthentic behavior. Her work on

the 2024 U.S. Presidential Election demonstrated how knowledge graphs can enhance scalability and explanatory power in detection systems.

- Cross-Platform Tool Development: Beyond individual methods, several presentations emphasized the importance of building tools that operate across platforms and content types, reinforcing the move toward integrated, scalable solutions.

3.1.2 Platform Policies and Governance Frameworks

Although most social media platforms avoid explicitly referencing "Coordinated Inauthentic Behavior" (CIB) in their guidelines, many implicitly rely on notions of "authenticity" and "coordination" to distinguish suspicious from legitimate activity. This strategic ambiguity was the subject of Timothy Graham's keynote, "The 'Inauthenticity' Paradox: How Platforms Profit From and Shape Coordinated Inauthentic Behaviour."

Graham's analysis demonstrated how platforms selectively operationalize authenticity to serve business interests. Through case studies of cryptocurrency communities and Russian propaganda on X (formerly Twitter), he showed that enforcement of CIB policies is often inconsistent, allowing profitable inauthentic behavior to persist while publicly signaling commitment to integrity. This dynamic underscores the blurred boundary between policy enforcement and platform monetization strategies, raising concerns about accountability and transparency.

3.1.3 Ethical Considerations and Power Dynamics

Ethical concerns featured prominently across the conference. Presenters repeatedly emphasized that detection tools are not neutral instruments. As Magelinski et al. (2022) argue, most research to date has paid limited attention to the normative assumptions underpinning detection—particularly the binary classification of coordination as "good" or "bad."

The Coordinated Sharing Behavior Detection conference advanced this discussion by emphasizing the power implications of coordination detection. There is a growing risk that such tools may disproportionately target marginalized communities, reinforce dominant narratives, or shift power from the public sphere to tech platforms, law enforcement, and governments. The development of detection methodologies, therefore, must be coupled with ethical reflexivity, ensuring that efforts to combat manipulation do not suppress legitimate forms of political engagement or dissent.

3.2 Methodologies

This section introduces methodological advancements for detecting Coordinated Inauthentic Behavior and coordinated sharing in online environments. It first outlines the CIB Detection Tree, a theoretical framework developed to guide structured, multi-dimensional analysis of deceptive coordination. The section then presents practical detection pipelines, including the modality-agnostic CooRTweet, visual similarity analysis and audio provenance techniques, which operationalize and extend the framework to diverse platforms, formats and manipulation strategies.

3.2.1 A Theoretical Framework: CIB Detection Tree

The Coordinated Inauthentic Behaviour (CIB) Detection Tree is a methodological framework developed by EU DisinfoLab within the scope of the vera.ai project, aimed at supporting the identification and analysis of online influence operations that rely on deceptive coordination, falsified identities, and the manipulated amplification of content. As CIB becomes increasingly central to the dynamics of disinformation and influence campaigns, the need for a coherent and shared approach to its detection has grown more urgent. Despite broad acknowledgment of CIB across digital platforms, its definition remains fragmented, often shaped by the internal policies of individual services. At the same time, the European Union has yet to formalise a regulatory definition, even as recognition of CIB's impact intensifies in politically and socially sensitive contexts such as elections or conflicts. In response to these gaps, EU DisinfoLab has revisited and refined its original 2021 detection tree, offering a consolidated framework that aligns investigative logic with the demands of a rapidly changing information environment. The revised theoretical framework incorporates an updated toolkit, a clarified conceptual structure, and strategic reflections on the evolving role of artificial intelligence, which increasingly contributes to both the execution and detection of inauthentic activity. It seeks to address emerging challenges linked to AI-driven content generation and automation, while also equipping analysts, civil society, and the broader defender community with a unified lens and practical resources to diagnose and respond to complex CIB campaigns.

Detection Tree branches

The four branches of the detection tree are intended to function as interconnected layers of analysis that together build a diagnostic perspective of suspected CIB campaigns.

1. The Coordination Assessment Branch focuses on whether the activity is planned and executed by a network working toward a shared deceptive goal.
2. The Authenticity Assessment Branch explores whether the involved actors, behaviours, and content are misrepresented or fabricated.
3. The Source Assessment Branch works to identify the originators behind the operation, including their motivations and technical infrastructure.
4. The Impact Assessment Branch evaluates the reach and effectiveness of the campaign, with particular attention to amplification and influence outcomes.

These branches are designed to be applied jointly, creating a dynamic and holistic framework that accounts for both structural organisation and operational consequences of CIB.

Coordination Assessment

The Coordination Assessment Branch investigates whether the observed activity is the result of organised collaboration among agents, regardless of whether this collaboration is directed by humans or automated systems. Coordination in this context refers to the systematic organisation of actions that aim to achieve a misleading objective, typically through the manipulation of content flow and timing. It is explicitly recommended not to equate automation with coordination, as coordination can be purely human-led and does not require automation to be deceptive. In fact, the growing ability of AI to generate diverse and

individualised content, including profile images, further complicates the detection of coordination based solely on surface-level observation.

Analytical methods within this branch include behavioural analysis, which looks for signs such as synchronised posting, account creation patterns, and spikes in content volume; graph network analysis, which identifies clusters of closely interacting accounts and shared engagement behaviours; content analysis, which examines the repetition of narratives, shared media assets, and one-topic content production; identity and visual analysis, which explore the reuse of profile images or suspiciously uniform design features; metadata analysis, which checks for multiple accounts operating from identical IP addresses or configurations; and automation analysis, which attempts to spot systematic syndication patterns.

Authenticity Assessment

Following the identification of coordinated behaviour, the next step is to evaluate whether the actors and materials involved are authentic. This branch focuses on detecting deception through falsified identities, artificial engagement, and misleading content. Inauthenticity, in this framework, refers to the deliberate misrepresentation of account identity, content origin, or intent. Accounts may be real or fake, but what matters is whether their behaviour is genuine or contrived to create a misleading impression.

This branch applies several layers of analysis. Behavioural indicators include accounts showing sudden changes in activity, masked messaging, or artificially created interactions. Network indicators reflect abnormal engagement patterns or artificially boosted interactions. Identity analysis investigates impersonation, lack of personalisation, and use of randomly generated names. Content-language analysis considers factors such as unnatural phrasing, mistranslations, and use of fabricated or fantastical media. Visual analysis includes signs of image manipulation or reuse across multiple profiles. Metadata analysis points to patterns such as activity linked to VPNs or bot infrastructure. Automation analysis flags signs of bot behaviour and mass content generation. AI technologies exacerbate these challenges by enabling realistic fake personas and generating engagement at scale, but they also offer avenues for enhanced detection through pattern recognition and large-scale data analysis.

Source Assessment

The third step involves tracing the origin of the campaign. The Source Assessment Branch addresses the central question of attribution, recognising that the core of CIB activity is to obscure who the operators are and what their goals might be. Because of this, identifying the source requires the integration of multiple forms of evidence, including behavioural footprints, technical forensics, and contextual content clues. Attribution is often achieved through a combination of open-source intelligence, metadata tracing, and investigative synthesis.

The analysis includes identifying key amplifiers and first movers within a network, which often helps uncover the campaign's launch dynamics. Content analysis helps associate pieces of information with recognisable actors or known strategies. Identity analysis investigates recurring registration patterns or associations with known entities. Visual analysis looks for recurring background or environmental features that might reveal a common source. Metadata analysis includes tracking IP addresses, API requests,

Autonomous System Numbers, and links back to original content sources. While AI can complicate this process by masking traces or generating misleading evidence, it also contributes tools that help reverse engineer digital fingerprints and recognise reused technical infrastructure, as highlighted by OpenAI's recent investigations into influence operations (e.g., Nimmo et al., 2025).

Impact Assessment

Finally, the Impact Assessment Branch seeks to understand the scale and influence of the campaign. While coordination and authenticity provide evidence of manipulation, impact assessment tells us how much damage or influence has occurred. This involves looking beyond direct interactions to consider the broader reach, such as views, shares, and shifts in discourse. It considers both the effectiveness of amplification and the sophistication of content targeting.

Within this branch, behavioural analysis focuses on the roles of secondary accounts that help distribute the core messages. Network analysis examines how the content spreads across different platforms, identifies the involvement of public figures or influencers, and tracks dissemination tactics like sentiment hijacking or hashtag flooding. Content analysis measures polarisation and audience segmentation. AI further influences this space by making content production more efficient, enhancing targeting capabilities, and increasing automation. Yet the same AI tools also support defenders by enabling tracking and quantifying content reach and influence through algorithmic analysis.

The CIB Detection Toolkit

To complement the conceptual framework of the detection tree, the report presents a CIB detection toolkit that consolidates a broad range of analytical instruments tailored to each of the four branches. These tools enable both technical and behavioural investigations, supporting the identification of coordination patterns, the evaluation of authenticity signals, the tracing of campaign origins, and the measurement of impact. The selection includes network visualisation platforms, metadata extractors, forensic media analyzers, and social media monitoring solutions, each offering distinct functionalities suited to different investigative tasks. Tools such as Gephi, Cytoscape, and NodeXL assist in mapping coordination structures, while resources like the Verification Plugin (a tool developed in previous EU funded projects and enhanced with new functionalities in the context of vera.ai) and reverse image search engines help assess the authenticity of content. Source attribution is supported through technical forensics platforms like Maltego, WHOIS, and IP trace databases, while content virality and performance are tracked through tools like CrowdTangle (now deprecated), NewsWhip, and BuzzSumo. Many of these instruments also integrate AI-powered features, further enhancing their capacity to detect anomalies, synthesise patterns, and automate parts of the investigative process. Although the toolkit is inevitably limited by the evolving nature of CIB techniques and the discontinuation of some platforms, it serves as a practical foundation for defenders seeking to apply the detection tree across different cases and environments, and as a prompt for further innovation in the field.

Analytical Outcomes

As Coordinated Inauthentic Behaviour continues to evolve alongside technological advancements and the transformation of digital platforms, it becomes increasingly important for detection and mitigation

approaches to adapt accordingly. This toolkit contributes to a collective understanding of CIB by proposing a refined conceptual framework, an updated set of investigative tools, and a strategic reflection on the influence of artificial intelligence. This approach underscores the importance of a comprehensive, symptom-oriented methodology that is capable of identifying coordinated and manipulative behaviours, regardless of whether they are carried out by genuine or fabricated actors, and whether the execution relies on manual control or automation. EU DisinfoLab remains engaged with the broader defender community in an ongoing effort to strengthen detection capabilities and promote convergence around shared definitions and investigative standards.

3.2.2 CooRTweet: a Modality/Platform Agnostic Coordinated Detection Engine

The development of **CooRTweet**, an R-based engine for coordinated sharing detection, marks a significant methodological advancement in the study of inauthentic behavior on social media platforms. While previous tools in this domain have provided targeted approaches—such as *CooRnet* for Facebook/Instagram (Giglietto et al., 2020a) and the *Coordination Network Toolkit* for Twitter (Graham & QUT Digital Observatory, 2020)—CooRTweet stands apart due to its **conceptual generalization and architectural flexibility**. Rather than anchoring detection methods to specific platforms or data formats, CooRTweet is built upon a minimalist and abstract definition of coordination. This design choice carries far-reaching implications: by decoupling coordination from content type and platform constraints, CooRTweet becomes adaptable across diverse cases, from single-platform and single-modal analyses (e.g., retweets on Twitter) to more complex multimodal, cross-platform networks involving mixed media, hashtags, and other metadata.

This flexibility is not merely a technical asset but a **strategic response to growing challenges in digital research**, including shrinking data access due to API restrictions and platform policy changes. With the discontinuation of the CrowdTangle API and limited access to the Twitter/X API, tools heavily reliant on platform-specific infrastructure are increasingly difficult to maintain. CooRTweet addresses this structural limitation head-on by adopting a platform-independent approach. It can ingest any dataset conforming to a generic actor-object-time format, effectively **shifting the locus of power from the platform to the researcher**.

Moreover, CooRTweet redefines what constitutes a "coordinated object." While prior tools often focused on singular types of objects—such as URLs in the case of CooRnet—CooRTweet supports a wide variety of shared elements, including images, video descriptions, hashtags, and even combinations thereof. This generalized model is particularly important for contemporary studies of online coordination, where behaviors often span multiple modalities and reflect evolving manipulation strategies.

Central to CooRTweet's detection methodology are the principles of **synchronicity** and **repetition**, which serve as indicators of orchestrated behavior. Coordination, in this framework, is not inferred from content alone but from the **temporal and behavioral patterns** that emerge when multiple actors share content in close succession. This notion of behavioral convergence around shared media, rather than mere message similarity, offers a more robust signal for detecting inauthentic influence operations.

Validation: Performance Testing with Simulated Data

CooRTweet's architecture was rigorously validated through both synthetic and real-world testing (Righetti & Balluff, 2025). In simulated environments, the engine demonstrated linear scalability with dataset size and manageable resource consumption even in the face of increasing object complexity. These simulations confirm the tool's suitability for large-scale analyses, a key requirement for tracking contemporary information operations that unfold across massive datasets.

Righetti and Balluff (2025) ran 1000 simulations where they systematically varied the simulation parameters to generate diverse datasets. Figure 2 shows that the detect_groups() function's memory usage increases linearly with the size of the dataset, but quadratically with the number of unique objects. The generate_coordinated_network() function requires much less memory and also scales linearly.

The CooRTweet coordinated sharing detection engine underwent validation using the South Korean National Intelligence Service (NIS) dataset from Keller et al. (2019), which represents one of the few available ground truth datasets for coordinated behavior research. The authors of the study (Righetti & Balluff, 2025) generated two separate co-retweet (the retweeting of a third-party message within a short period of time) and co-tweet (where multiple accounts post the same message within a short time frame) networks. As illustrated in Table 1, for co-tweeting, CooRTweet's performance was similar to that of Keller et al. (2019), detecting 18% of the NIS accounts. This rate does not reflect a limitation of the method but rather the limited use of co-tweeting for coordination within the NIS network, as acknowledged by Keller et al. (2019). In the co-retweet network, however, CooRTweet outperformed Keller et al. (2019), detecting 84% of the NIS accounts using the same thresholds defined by Keller et al. (vs. 80%) and 91% with CooRTweet's default threshold (the median). Additionally, Righetti & Balluff (2025) demonstrate that CooRTweet's unique ability to combine networks coordinating across different modalities—in this case, tweets and retweets—enabled the detection of 737 NIS accounts (92%), improving overall performance.

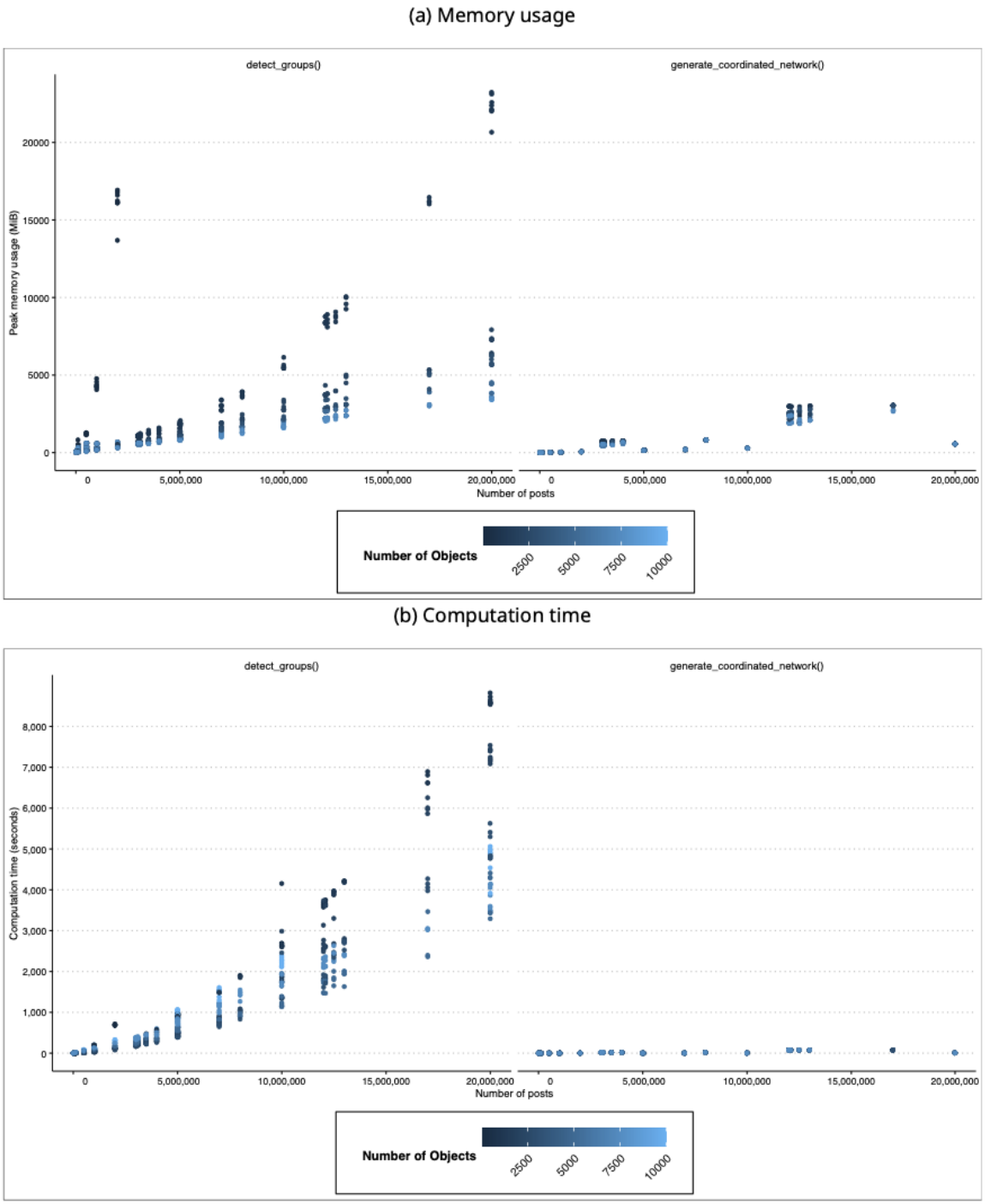


Figure 2 Benchmark results (Righetti & Balluff, 2025, p. 17).

Table 1 Results of coordination detection on NIS dataset (Righetti & Balluff, 2025, p. 16)

Metric	Keller et al. (2019)		Our Proposed Method		
	Co-Retweet	Co-Tweet	Co-Retweet	Co-Tweet	Combined
NIS accounts in network	642	144	731 (670)	139 (146)	737 (687)
Accuracy %	80%	18%	91% (84%)	17% (18%)	92% (86%)
Minimum Edge weight	5	2	2 (5)	3 (2)	2 (5)
Vertices	3,626	2,001	22,984 (4,350)	1,579 (2,335)	24,507 (5,042)
Edges	31,755	38,035	85,557 (39,321)	36,109 (40,062)	120,958 (68,280)
Components	325	362	2,778 (372)	257 (442)	3,025 (481)
Core suspects	440	662	458 (457)	381 (678)	638 (939)
NIS accounts among suspects	85%	-	89.9% (82.1%)	16.2% (17.9%)	90.8% (84.3%)

Note. The first row shows the total number of NIS accounts found in the coordination network, where the ground truth consists of 801 accounts. Keller et al. (2019)'s minimum edge weight thresholds correspond to the 97th percentile for co-retweets and the 40th percentile for co-tweets. We report results within brackets for CooRTweet using these thresholds, and those outside the parentheses using the default median threshold. The last row reports the testing results for suspect marking, which is the mean proportion of NIS accounts retrieved from the withheld test set (over 20 random runs).

CooRTweet Methodological Innovations

Traditional approaches to detecting coordinated behavior on social media platforms have relied on rigid threshold systems that categorize accounts as either coordinated or not coordinated based on predetermined criteria. This binary classification fails to capture the nuanced spectrum of coordination that exists in real-world social media environments. The CooRTweet coordinated sharing detection engine addresses this limitation by preserving data on accounts and connections regardless of whether they meet traditional coordination thresholds, using an innovative tagging methodology that assigns binary indicators (0/1) to network edges. This approach acknowledges that coordination exists on a continuum and enables researchers to examine networks of highly coordinated accounts while maintaining visibility into peripheral actors who may exhibit sporadic or emerging coordinated behaviors.

The detection engine provides analytical flexibility through multiple subgraph options and specialized utility functions, enabling researchers to explore coordination phenomena across various temporal scales and network configurations. This flexibility is crucial because coordinated behavior manifests differently depending on context, platform, and objectives. Some coordination campaigns involve rapid, synchronized sharing within minutes, while others unfold over days or weeks with subtle timing patterns.

The CooRTweet engine represents the first R tool designed with universal applicability for analyzing coordinated networks regardless of platform or content type, addressing the critical challenge of increasing fragmentation and instability of platform-specific analysis tools. Its platform independence is particularly valuable given the rapidly evolving landscape of social media platforms and their APIs, where researchers face frequent disruptions due to sudden changes, policy modifications, or platform closures.

By operating independently of specific platform architectures, the engine provides a stable foundation for longitudinal research projects and comparative studies across multiple platforms.

3.2.3 Visual Similarity for Cross-Platform Coordinated Sharing Detection

Visual media is one of the most engaging types of content shared by online social media platforms' users. Coordinating sharing of videos can be employed to amplify some otherwise unpopular narratives by exploiting the recommendation algorithms of the online platforms. Yet, automatically detecting whether some videos are shared in a coordinated manner requires the ability to answer whether two videos are related to the same event. Copies of the same video circulating online, segments of a longer video reshared by users, as well as response videos, are all examples of such similar videos. Yet, naive computational approaches, that cannot semantically analyze visual content, are unable to quantify the degree of similarity among different videos. To counter this issue and infuse CooRTweet the ability to operate directly on the visual modality, we build upon the visual similarity capabilities of the Near-Duplicate Detection (NDD) service that was previously developed under T4.1 of the WP4.

The NDD service employs the Distill-and-Select architecture (Kordopatis-Zilos et al., 2022) to efficiently search in large collections for videos that semantically match. On top of it, Self-Supervised Video Similarity Learning (Kordopatis-Zilos et al., 2023) enables the service to detect similar videos at three granularities, ranging from exact copies of video segments, to temporally aligned videos that depict the same incident, to ones only semantically related to the same event. Also, the underlying ViSiL model (Kordopatis-Zilos et al., 2019) allows searching both among arbitrary video segments, as well as on arbitrary spatial positions of a video, providing a visual similarity mechanism that is robust against several post-processing operations that could be applied to a video when being reshared across multiple accounts and social media platforms. Such robustness is particularly useful in cross-platform analysis, as different platforms can impose some required video editing operations, due to their support of different video formats, e.g. when sharing wide resolution videos from YouTube as short vertical videos in TikTok. These capabilities of the NDD service are exposed through a REST API, which we use to integrate it under our multimedia coordinated sharing detection pipeline.

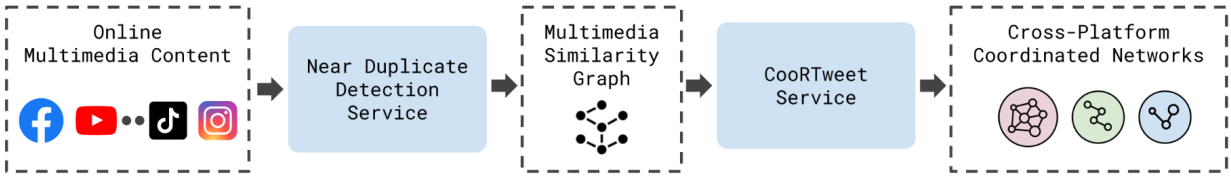


Figure 3 Pipeline for cross-platform multimedia coordinated sharing behavior detection

To perform cross-platform multimedia coordinated sharing detection we built a pipeline that, given the URLs of videos shared on popular online social media platforms, is capable of identifying accounts that incorporate some degree of automation among their sharings. In particular, starting from streams of public videos provided by platforms' APIs, like the Meta Content Library API and the TikTok API, we query the REST API of the NDD service to index them under a single collection. This collection has the ability to grow in real time and, due to the architecture of NDD service, to arbitrary numbers of videos. Then, for

each indexed video, we query again the NDD service's API to find previously indexed videos that semantically match to the given one and dynamically populate a multimedia similarity graph. Afterwards, we cluster this graph according to its connected components. The clustered multimedia similarity graph, as well the time of sharing of each video post, the platform identifier for the sharing account and the identifier of the video post are passed to the Coordinated Detection Service powered by CooRTweet. Finally, CooRTweet processes this data and generates a graph report of the accounts that exhibit patterns of coordinated sharing behavior. An overview of this pipeline is depicted in Figure 3.

Table 2 Statistics of clustered multimedia similarity graphs generated by the Near-Duplicate Detection Service. The YouTube, TikTok, Facebook and Instagram columns present the statistics for single-platform analysis; the right-most column presents the statistic

	YouTube	TikTok	Facebook	Instagram	Cross-Platform
#videos analyzed	1187	5940	338	1173	8638
Similar videos per video - AVG	1.16	24.29	0.42	0.84	17.19
Similar videos per video - STD	3.19	49.96	1.02	2.68	42.88
Similar videos for video - MAX	34	277	6	27	277
#video clusters	879	1586	284	948	3512

Using our pipeline for cross-platform multimedia coordinated sharing detection, we performed a case study by employing videos from four popular online social media platforms, i.e. YouTube, TikTok, Facebook and Instagram, aligning in time with the 2025 German federal election. In total, we processed 8638 videos, where 1187 originated from YouTube, 5940 from TikTok, 338 from Facebook and 1173 from Instagram. The different numbers of videos across different platforms can be attributed to the different rate of posting of video content on each platform. We used the NDD service to find for each considered video its near-duplicate ones under both a single-platform and a cross-platform setup. In the former case videos originating from each platform are considered separately, using separate collections of the NDD service, while in the later one, all videos are considered under a single media collection. Table 2 presents the average number of similar videos per video, as well as the size of the biggest cluster and the standard deviation in the number of videos per cluster. Also, the total number of clusters, i.e. groups of videos that do not include visually similar content, is displayed. As we see, different platforms exhibit different rates in resharing of original content, with TikTok dominating the number of resharings of the same content. Overall, in the cross-platform scenario we found a ratio of 17.19 near-duplicate videos for each original video shared online.

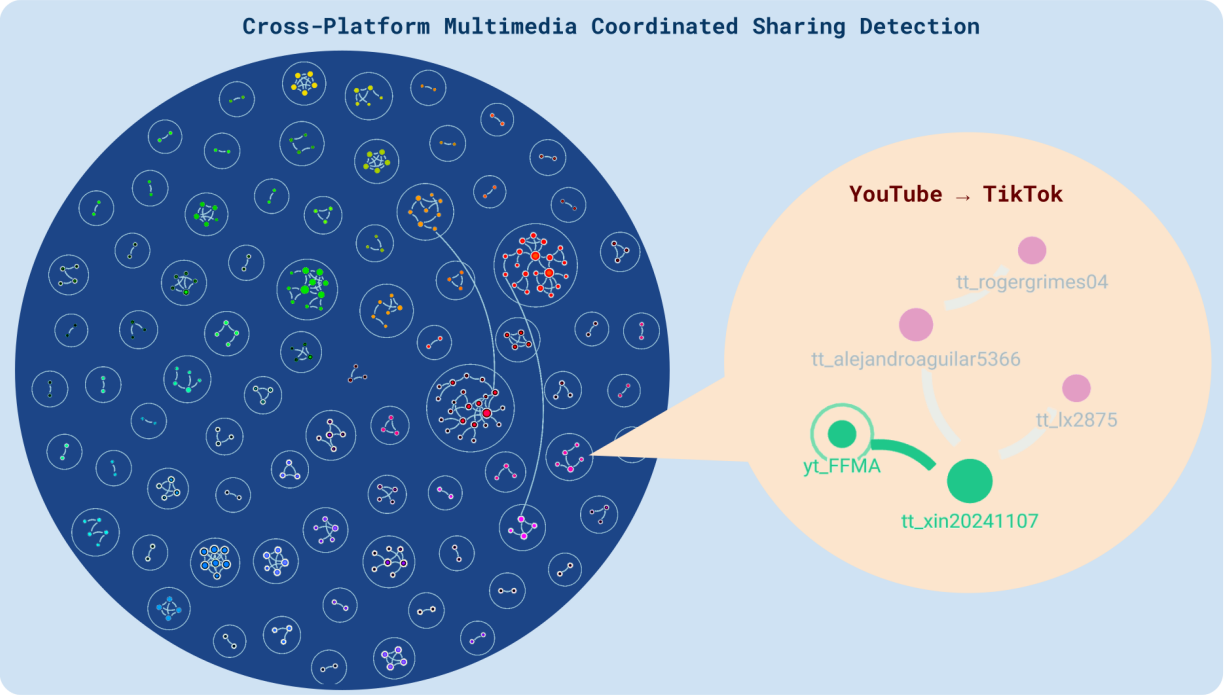


Figure 4 Cross-platform analysis of multimedia coordinated sharing networks. Each node of the graph represents a user account on an online social media platform, considering YouTube, TikTok, Facebook and Instagram. Edges represent a detection of coordinated sharing behavior among the corresponding accounts. On the left, each graph component represents a group of accounts exhibiting coordinated behavior throughout their online sharing of multimedia. On the right, a network of accounts with coordinated sharing behavior from YouTube and TikTok is presented. The analysis was performed using the Coordinated Detection Service powered by CooRTweet in conjunction with the Near Duplicate Detection service of vera.ai.

Finally, we employed the Coordinated Detection Service powered by CooRTweet[9] to analyze the clustered multimedia similarity graph generated under the cross-platform setup. We present in Figure 4 the resulting network of accounts that operate on the four considered platforms and exhibit some degree of automation in their sharings. In total, we identified 78 groups of accounts that exhibit patterns of coordination in their sharing of multimedia content. The largest network was found to incorporate 22 user accounts. Also, a single network of coordinated accounts can span several social media platforms. Point in case, the zoomed-in network at the right of Figure 4 depicts a pattern of coordinated sharing behavior among TikTok and YouTube accounts. The integration of the NDD and Coordinated Detection Service powered by CooRTweets under a single pipeline provide a powerful tool to detect these coordination patterns in the visual modality.

3.2.4 Audio Provenance Analysis in Coordinated Sharing Detection

Introduction and Methodology

Audio Provenance Analysis is a methodology developed to trace the origin, transformation, and reuse of audio content, particularly in large and heterogeneous datasets that often lack metadata or content descriptions. Its main goal is to convert a diverse collection of audio files into a directed acyclic provenance graph, which maps how audio segments have been reused, altered, and disseminated.

[9] https://coortweet.lab.atc.gr/auth/login

As illustrated in Figure 5, the process begins with **provenance clustering**, which uses partial audio matching (Maksimovic et al., 2021) to detect and localize reused segments between audio files. This step identifies near-duplicate files—those that are perceptually identical despite transformations—and partial duplicates, where only certain segments are reused. Unlike conventional audio matching techniques, this method supports the detection of multiple reused segments between pairs of files, allowing for a fine-grained understanding of content similarity.

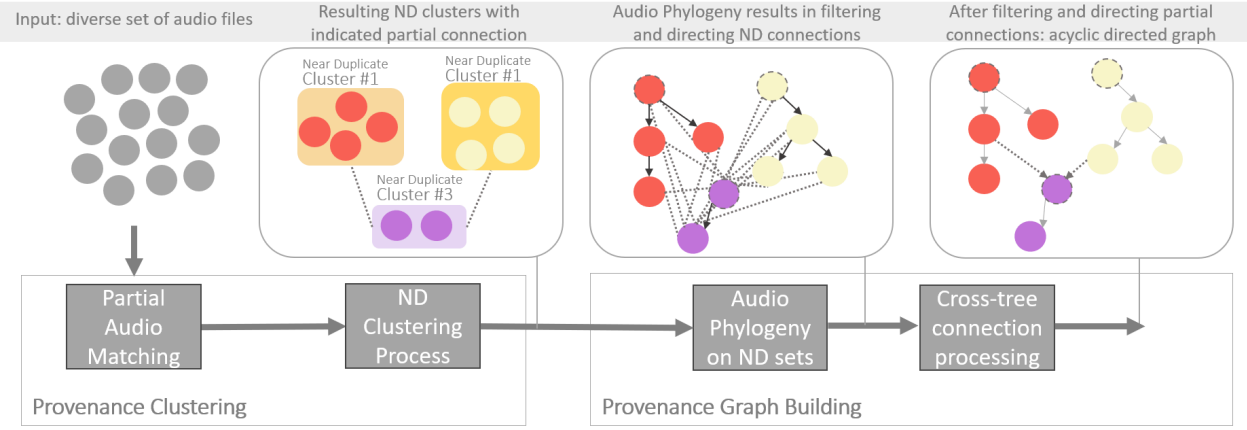


Figure 5 Workflow for Audio Provenance Analysis.

Once clusters are formed, the methodology proceeds to **provenance graph construction**. Within each cluster of near-duplicates, an audio phylogeny method (Gerhardt et al., 2023) is applied to infer directional relationships and build transformation trees. These trees help identify original source files and map how derivative files were derived from them through content-preserving transformations. To account for cases where audio compositions draw segments from multiple sources, cross-tree connection analysis is conducted. This step identifies when a file in one tree contributes a segment to a file in another, determining the most likely donor file and pruning redundant connections to maintain a clean, interpretable structure.

The final provenance graph encapsulates the dataset's transformation and reuse history. This structured representation is useful for verifying content authenticity, attributing sources, and identifying manipulation or misinformation. Introduced by Gerhardt et al. (2024), this methodology is the first of its kind for audio and adapts principles from image and video forensics to the specific challenges of the audio domain.

The underlying components—partial audio matching and audio phylogeny—were detailed in Deliverable 4.1. The framework was formally evaluated in two controlled scenarios, as documented in Gerhardt et al. (2024). In what follows, we extend that evaluation to the context of coordinated behavior detection.

Coordinated Behavior Detection via Audio Similarity

Beyond identifying individual cases of media manipulation, we evaluated audio provenance analysis as a tool for supporting the detection of coordinated sharing behavior on social media. Coordinated behavior detection, as defined by Giglietto et al. (2022), typically follows a two-step process: first, collecting data that links user accounts to the content they share and the timestamps of those shares; second, analyzing this data to detect patterns of synchronized or structured activity across multiple accounts. The second

step often relies on the CoorTweet tool (Righetti & Balluff, 2025), which clusters content based on a selected similarity indicator.

In the referenced study (Giglietto, 2024), the similarity indicator used was the presence of identical video descriptions across posts on TikTok. This proved effective in detecting coordinated behavior within that specific platform. However, it focuses on metadata rather than media content and assumes that actors will use the same textual descriptions—an assumption that may not hold across platforms or tactics. To overcome this limitation, we assessed audio provenance analysis as a content-based similarity metric, capable of detecting coordination signals independent of metadata.

The dataset used in this evaluation consisted of 513 TikTok accounts previously flagged for suspicious or potentially coordinated activity. These accounts were monitored continuously, and daily snapshots of their shared content were archived. We selected the snapshot from January 8, 2025, for our experiment.

We used two types of similarity indicators to detect coordination in this dataset: (a) identical video descriptions, replicating the method from Giglietto et al. (2024), and (b) clusters of near-duplicate and partial matches generated through audio provenance analysis. The description-based detection served as a proxy for ground truth, against which we evaluated the audio-based approach.

This snapshot included metadata for 4,034 videos, of which 3,993 were still accessible at the time of analysis. These videos were used as input for audio provenance analysis, which identified 1,222 near-duplicate clusters and 1,272 partial clusters. The resulting forest of acyclic graphs clearly illustrated content relationships, with root nodes representing original audio sources and directed edges tracing transformations or segment reuse (see Figure 6).

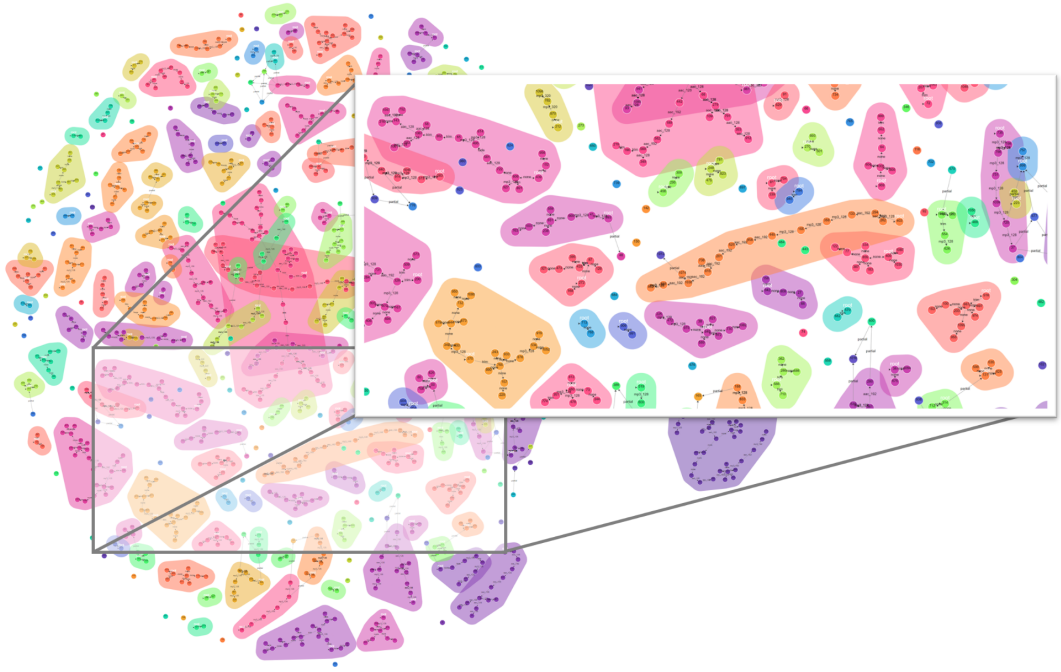


Figure 6 TikTok experiment: Final acyclic graph as output of audio provenance analysis. Single nodes represent individual TikTok videos. When nodes of the same color represent a set of near duplicates, these are illustrated within a bubble. Near duplicates are connected in a directed phylogeny tree with identified roots. Connections between ND trees or individual nodes indicate one or more partial matching segments.

To assess the value of audio-based clustering for coordinated behavior detection, we ran three configurations through CoorTweet using a 120-second time window: (a) description-based similarity, (b) audio-based similarity from ND and partial clusters, and (c) a combination of both. The results are visualized in Figure 7.

From the graphs in Figure 7, we observe that networks based on audio similarity (center) closely resemble those based on identical descriptions (left), though they feature slightly fewer nodes and edges. This difference likely stems from audio transformations such as overlays or time-stretching, which can obscure similarity detection. However, the combined approach (right) yields the most comprehensive network, revealing additional links that would be missed when relying on textual metadata alone.

These results indicate that audio provenance analysis is not only a viable content-based similarity metric but also a valuable complement to metadata-based methods. It can uncover coordination signals even when textual or metadata cues are altered, omitted, or inconsistent.

Beyond serving as a similarity indicator, audio provenance analysis offers two additional capabilities in the context of coordination detection. First, once a coordinated network is flagged as malicious, the provenance graph enables tracing how content propagated through the network. This helps identify root nodes or key influencers responsible for initiating or amplifying disinformation (see Figure 6).

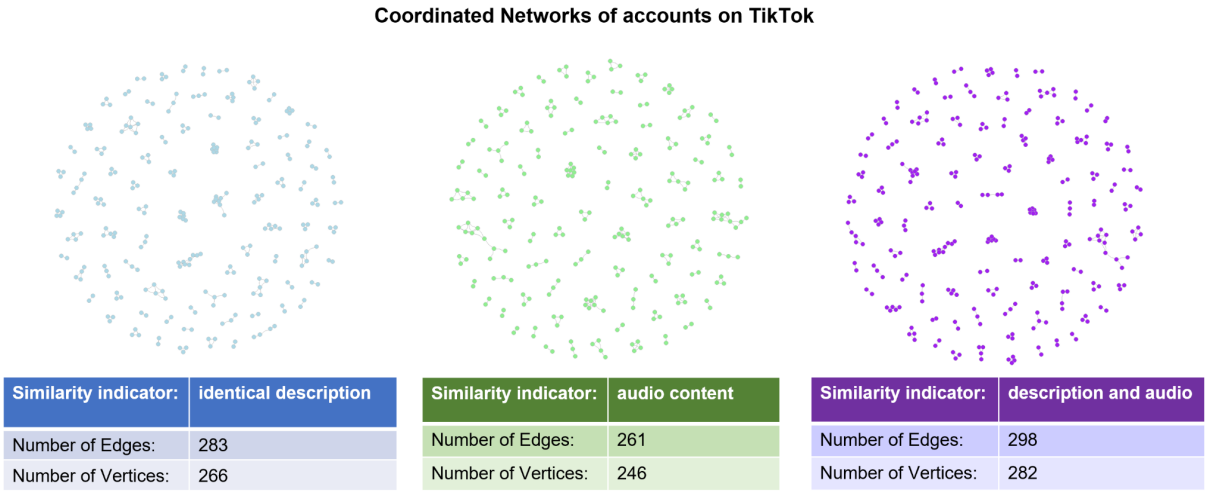


Figure 7 TikTok experiment: visualizing CooRTweet graph of detected coordinated accounts via: (left) identical titles, (middle) audio ND and partial duplicates and (right) combination of identical titles and audio similarity indication. Every node presents one user account on TikTok.

Second, combining audio provenance analysis with text-based decontextualization detection introduces a powerful cross-modal approach. When content reuse is confirmed through audio analysis, but the information inferred from the audio contradicts the message conveyed by the accompanying text, this discrepancy becomes a strong indicator of manipulation or coordinated inauthentic behavior (Romero-Vicente, 2025). Such cross-modal inconsistency may point to deliberate intent to mislead, especially in environments where actors attempt to mask disinformation under seemingly authentic multimedia content. Our approach for the audio-text decontextualization detection is described in detail in deliverable D3.3.

3.3 Implementations

In the following subsections, we present the implementations of each of the methodologies that we developed over the course of the project for this specific WP as described in the previous section.

3.3.1 Assessment of CIB disinformation campaigns

As disinformation and influence operations continue to grow in complexity, identifying signs of Coordinated Inauthentic Behaviour (CIB) remains a critical priority for researchers, journalists and online platforms. These operations do not simply distort facts; they manipulate public discourse, target democratic institutions, and erode public trust. The increasing use of cross-platform dissemination, AI-generated content, and coordinated behavioural patterns makes detection more challenging, but also more urgent.

Built on the detection framework developed by EU DisinfoLab in previous work and more extensively described in a report titled *Visual assessment of Coordinated Inauthentic Behaviour in disinformation campaigns*[10]. This section introduces a visual methodology to evaluate and interpret real-world campaigns that exhibit CIB characteristics. It applies a structured set of fifty generic indicators, designed to identify coordination, inauthenticity, source manipulation, and impact, to three recent case studies. Each indicator is used to assess whether specific features of a campaign are present or not, resulting in a quantified probability of CIB. By visually mapping disinformation activities against this standardised indicator set, the report aims to make CIB dynamics more accessible and understandable, and to provide a practical tool that supports media professionals, fact-checkers, researchers, and policymakers in identifying and contextualizing CIB operations.

The methodology translates complex analytical findings into intuitive visual outputs, enabling readers to grasp both the scope and nature of manipulation. Through this combination of structured analysis and accessible presentation, the report enhances the ability to monitor, understand, and communicate how CIB manifests across different types of disinformation campaigns.

Methodology and Presentation of Results

This structured methodology assesses whether disinformation campaigns exhibit signs of Coordinated Inauthentic Behaviour (CIB). This assessment is based on a set of fifty generic indicators developed by EU DisinfoLab, designed to capture a wide range of behavioural, structural, and content-based signals associated with CIB. These indicators are grouped into analytical categories including coordination, authenticity, source, distribution and impact, and are supported by additional dimensions: content design, metadata indicators, identity signals, visuals, behavioural patterns, network interactions and automation means.

Each case study is evaluated by checking whether these indicators are present. Some indicators may not be assessed due to a lack of data, irrelevance to the case, or non-inclusion in the investigation. All indicators carry equal weight. The final score is calculated by determining the percentage of indicators

[10] https://www.veraai.eu/posts/report-visual-assessment-of-cib-in-disinfo-campaigns.

found resulting in a CIB likelihood score ranging from 0 to 100. For example, if a campaign shows 40 out of 50 indicators, it receives a score of 80 percent, indicating a high probability of CIB activity.

To ensure accessibility and clarity, the results are presented using a combination of visual and tabular formats. Four color-coded gauges, Coordination, Authenticity, Source, Distribution & Impact, summarise in a single Final Assessment gauge the strength and prevalence of CIB features across each campaign. These gauges use a traffic light logic: red indicates a low likelihood of CIB (below 25 percent), yellow indicates a medium likelihood (between 25 and 75 percent), and green indicates a high likelihood (above 75 percent). Supporting these visuals are detailed indicator tables that outline exactly which CIB signals were present in each case. This layered approach allows readers to quickly understand overall findings while also enabling deeper analysis of the specific mechanisms and patterns that define each campaign.

Case Study Assessments

1. Operation Overload

This case study[11] covers a large-scale, multi-platform disinformation campaign revealed by CheckFirst and Reset.tech targeting European media and fact-checkers. The operation used social media content to manipulate newsrooms, urging them to verify misleading information and thus distracting them from authentic fact-checking tasks. The campaign aimed to create the illusion of widespread legitimacy and engagement by spreading identical content across languages and platforms.

The assessment reveals a medium-high likelihood of CIB. Strong signals were detected in coordination, authenticity, and distribution, supported by the use of AI-generated visuals, identical textual content, and synchronised behavioural patterns. The final CIB score is 64%, suggesting significant evidence of coordinated and inauthentic tactics in the operation (Figure 8).

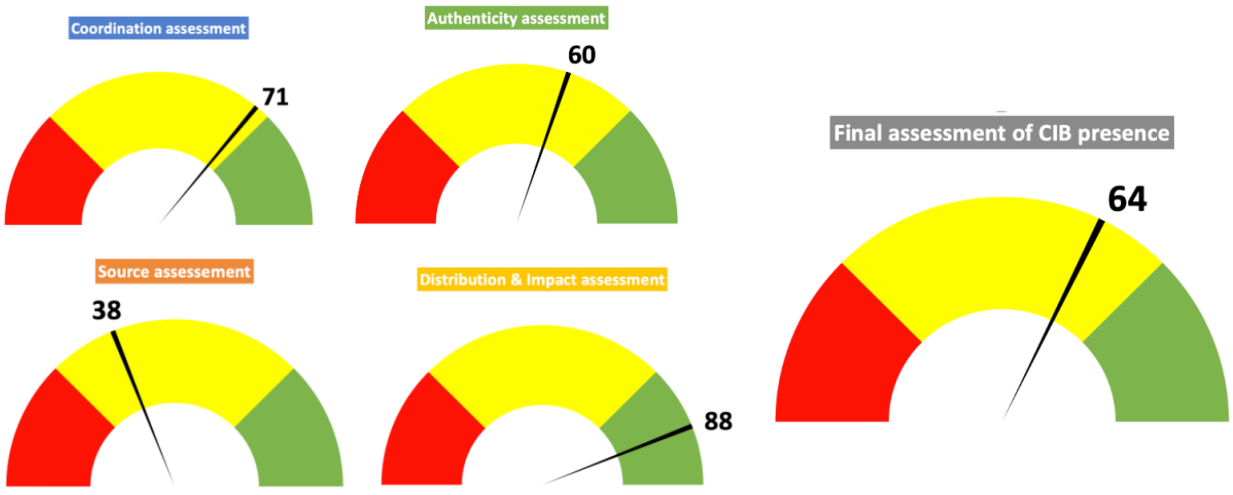


Figure 8 Visual representation of the CIB assessment showing medium-high likelihood, especially in the distribution, coordination and authenticity.

[11] https://checkfirst.network/operation-overload-how-pro-russian-actors-flood-newsrooms-with-fake-content-and-seek-to-divert-their-efforts/

2. Russian TikTok Influence Operation Targeting Ukrainian Minister

The second case[12] focuses on a massive influence campaign uncovered by DFRLab and BBC Verify, which targeted former Ukrainian Defence Minister Oleksii Reznikov. Over 12,800 TikTok accounts were involved in spreading false corruption allegations, often with AI-generated audio in various languages. The campaign aimed to undermine the Ukrainian government's image and reduce Western support during the ongoing conflict with Russia.

The campaign exhibited very high levels of activity in distribution and coordination. Numerous accounts posted similar content across languages, profiles displayed limited authenticity, and automation patterns suggested the use of bots. State media and fringe outlets helped amplify the narratives. Like the first case, this operation scored a 64% final likelihood of CIB, with particularly high scores in coordination and authenticity. The graphical representation is displayed in Figure 9.

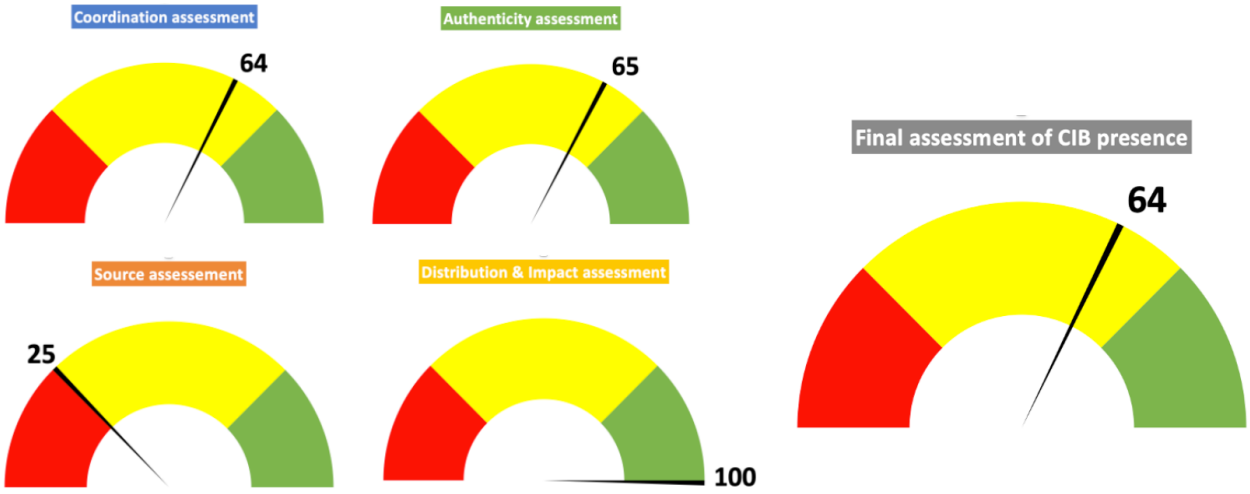


Figure 9 Visual representation of the CIB assessment showing medium-high likelihood, especially in distribution, coordination and authenticity.

3. QAnon's "Save the Children" Campaign

The third case[13] study examines how the QAnon movement hijacked the legitimate #SaveTheChildren hashtag during the summer of 2020 to disseminate conspiracy theories. By framing their false narratives within a humanitarian cause, they were able to mislead audiences and exploit public concern around child trafficking. This allowed the campaign to gain traction and introduce unfounded beliefs—such as the idea that elites extract adrenochrome from children—into mainstream discourse.

Despite the widespread impact and viral reach of this campaign, the evaluation found fewer indicators of coordination or inauthentic identity structures. Most content dissemination appeared organic rather than

12 https://dfrlab.org/2023/12/14/massive-russian-influence-operation-targeted-former-ukrainian-defense-minister-on-tiktok/

13 https://www.disinfo.eu/publications/disinformation-glossary-150-terms-to-understand-the-information-disorder/

bot-driven, and technical signals of centralised manipulation were limited. Consequently, the final CIB assessment was lower at 32%, suggesting a medium-low likelihood of coordinated inauthentic behaviour. However, high scores in distribution and impact highlight the campaign's social effectiveness, even if the coordination was less detectable (Figure 10).

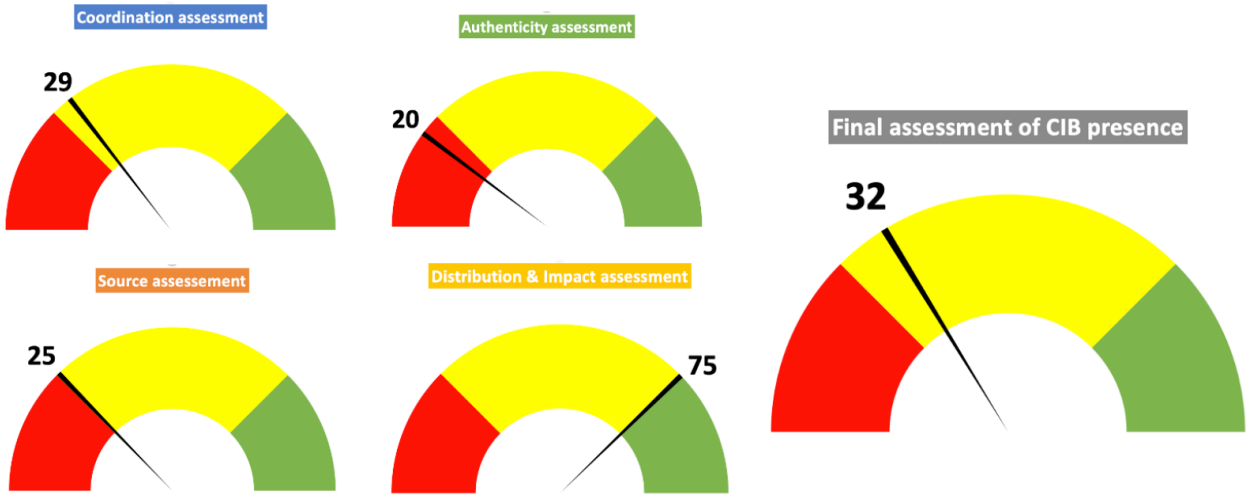


Figure 10 Visual representation of the CIB assessment showing medium-low likelihood despite high distribution and impact.

Summary of Insights

This approach has demonstrated how a structured, indicator-based methodology can effectively reveal the operational dynamics of disinformation campaigns that exhibit signs of Coordinated Inauthentic Behaviour. By applying a consistent framework across three diverse case studies, the analysis has shown how CIB can manifest through different layers of content replication, behavioural synchronisation, identity manipulation, and cross-platform coordination. The visual representation of these indicators not only clarifies complex patterns but also enables quicker and more intuitive interpretation of disinformation tactics. Importantly, the findings underscore that even campaigns with low technical coordination can still achieve high impact through emotional narratives and strategic hijacking of public discourse. As disinformation techniques continue to evolve, this visual assessment model offers a practical, scalable, and transparent tool for analysts, journalists, and civil society actors working to detect and mitigate manipulation in the digital information space.

3.3.2 Coordinated Sharing Behavior Detection Service Powered by CooRTweet

The CooRTweet Coordinated Sharing Detection Service was developed through a collaborative effort between the Athens Technology Centre (ATC) and researchers at the University of Urbino. This tool was designed to support data journalists, fact-checkers, researchers, and related practitioners in identifying and analyzing patterns of coordinated sharing across social media platforms. The current version of the

graphical user interface is publicly accessible[14]. A tutorial is presented in Annex III of the present document that guides potential users to effectively make use of this GUI.

At its core, the tool provides a structured workflow enabling users to upload their own data—often collected via social media APIs or content libraries—and receive a dynamic, visual representation of coordinated activity. Once the data is uploaded, CooRTweet processes account IDs, timestamps, and shared content fields. The system then generates an interactive network graph where nodes represent accounts and edges represent instances of near-simultaneous sharing. Users can examine this graph to identify clusters of likely coordination. Additional summary tables offer sortable and filterable views of accounts, clusters, and content. These outputs are downloadable in standard formats for further use.

The tool is composed of three main components: a frontend React application, a Node.js middleware, and a Python backend, all supported by a MongoDB database. The backend processes CSV file uploads and passes them to the CooRTweet library, which performs the core algorithm to generate the network graph data. To protect sensitive information such as API tokens, the Node.js middleware acts as a security layer. This layer can also be extended in the future to handle additional responsibilities like user authentication or request validation.

The frontend presents the output of the CooRTweet algorithm in a user-friendly interface using the reagraph library to visualize the coordinated networks. Because generating the graph is computationally intensive and may take some time, the system employs a polling mechanism to avoid request timeouts and long-lived network connections. When a user initiates a request, a task entry is created in the database. The frontend then polls the backend at regular intervals to check the status of this task. Once the algorithm completes execution, the task status is updated in the database, and the frontend retrieves and displays the final results.

The tool was built to support high-level use cases, including research into political campaigns by identifying networks that amplify specific narratives; disinformation detection by surfacing coordinated groups or bots promoting false claims; and monitoring of influence campaigns through the analysis of how content is amplified across platforms. With its integrated visualisation and export capabilities, CooRTweet helps users quickly make sense of complex information ecosystems.

Following the development of the first working version, a closed beta testing session was held on 18 February 2025, involving approximately ten participants, primarily investigative journalists and open-source intelligence (OSINT) researchers. The feedback from this session highlighted several opportunities for improvement. Participants suggested that the homepage could better communicate who the tool is intended for and what level of technical expertise is needed. It was also noted that the front-page text was overly dense, and that guidance on CSV formatting would be more effective if placed directly within the upload screen.

Feedback on the network graph interface pointed out the need to fix a rendering bug where some clusters appeared without visual outlines. Testers also recommended enabling real-time updates to the graph when parameters are modified, eliminating the need to reprocess the data after each change.

[14] https://coortweet.lab.atc.gr/auth/login

Additionally, it was suggested that node sizing be based on activity levels (i.e., volume of shared content) rather than node degree, to better reflect user behavior.

In terms of summary data, there was a call to revise the terminology used in the tables for clarity—specifically, replacing terms like "unique accounts" with simpler alternatives such as "accounts." Participants also wanted more intuitive table interactions, including the ability to click on a cluster in the graph and see the related entries in the tables, and vice versa. There was a consensus that the “average edge weight” column in one of the summary tables was confusing and should be removed. Further requests included a clearer download option for extracted data and improved documentation, such as legends explaining visual features like node size and edge thickness, as well as better guidance on how to interpret and manipulate the summary tables.

Between February and May, these suggestions were implemented into an updated version of the CooRTweet UI. The homepage was streamlined to clearly state the target user groups and provide a simple overview of the workflow. Table navigation was enhanced with clickable elements that link graph clusters and table rows, and vice versa, with a reset mechanism in place to exit filtered views. Simplified terminology replaced more technical terms, and the average edge weight column was removed. Finally, visual legends were introduced in the homepage to guide non-technical users to read the generated graphs meaningfully.

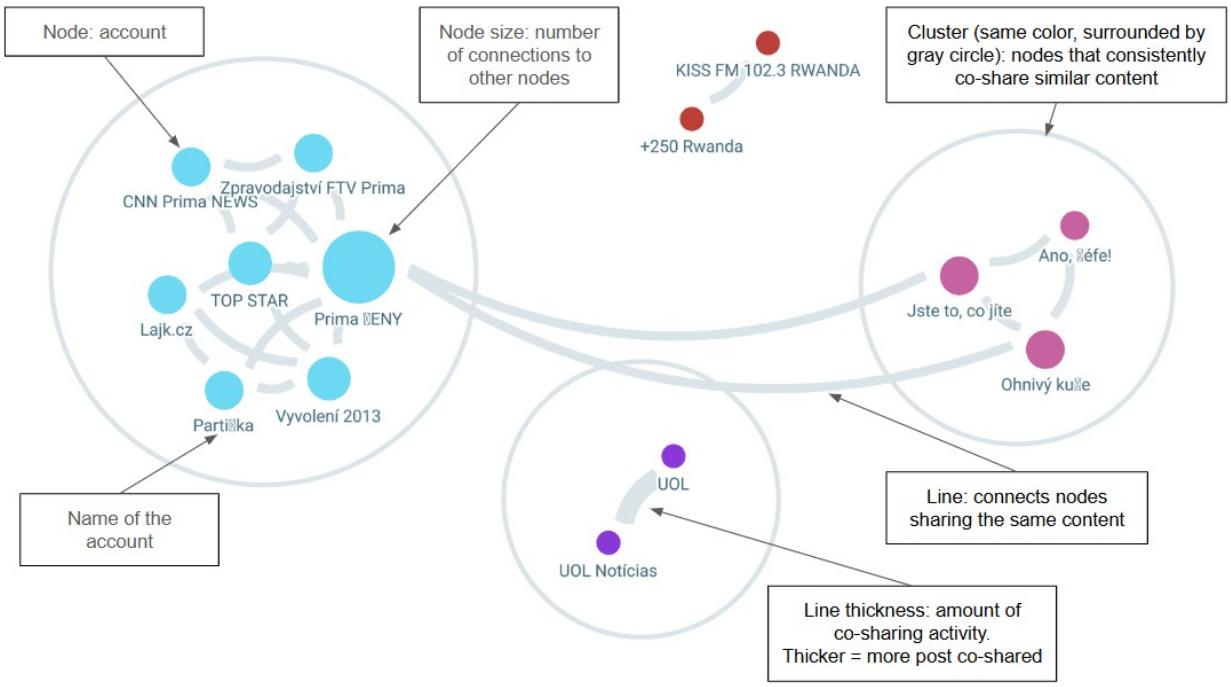


Figure 11 Illustration carried on the landing page of the CooRTweet Coordinated Sharing Detection Service, explaining how to read the graphs generated by CooRTweet CSDS GUI.

To assess the effectiveness of these improvements, a second user evaluation session was held on 5 June 2025, with a smaller group of seven participants, mostly composed of data journalists. Feedback from this

session was generally positive. Participants reported that the tool was easy to understand and use. The visual legends and clearer table labels improved interpretability, allowing users to better understand the significance of detected clusters and interactions.

To further refine the tool, the Urbino team has launched a public usability survey to gather broader feedback on the current interface and user experience. The results of this ongoing research will help guide the next phase of development, with particular attention to features such as cross-platform analysis, improved accessibility options, and expanded automation for larger-scale monitoring.

While the combination of CooRtweet's abstract architecture and the Coordinated Sharing Detection Service's user-friendly GUI significantly broadens and simplifies access to coordinated detection analysis across different platforms for non-technical users, data access remains a major unresolved challenge. To streamline the bring-your-own-data approach adopted, the team developed the Coordinated Sharing Detection Service pre-processor[15] that transform CSV data from Meta Content Library, TikTok Research API, BlueSky (via Communalytic), YouTube Data Tools, and Telegram into the standard CSDS format.

In conclusion, thanks to the Coordinated Sharing Detection Service, CooRTweet has evolved into a highly practical and accessible tool for detecting coordinated behavior on social media. By responding to user feedback with targeted improvements, the team has made it easier for journalists and researchers to uncover and interpret complex networks of influence and disinformation. With ongoing evaluation and development, the tool is well-positioned to become a vital asset in the broader fight against coordinated sharing behaviour, influence operations and manipulation in the digital public sphere.

3.3.3 VeraAI Alert System

The VeraAI Alert System represents a comprehensive implementation of the coordinated behavior detection workflow designed to operate at global scale. This system builds upon the foundational methodology first tested in the context of the 2022 Italian elections (Giglietto et al., 2023) and subsequently expanded to monitor coordinated influence operations worldwide. The system operates through a sophisticated multi-step process that systematically detects, analyzes, and updates lists of coordinated social media accounts, transforming static network analysis approaches into a dynamic, adaptive methodology capable of tracking evolving coordination patterns over time.

Unlike conventional detection systems that rely on one-time data collection, the VeraAI Alert System utilizes an iterative feedback loop that enables continuous surveillance and real-time adaptation to changing coordination strategies. This approach addresses the dynamic nature of coordinated operations on social media, where networks continuously evolve their tactics, platforms, and messaging to evade detection while maintaining operational effectiveness.

Global Implementation and Scale (October 2023 - August 2024)

The global deployment of the VeraAI Alert System began with a seed list of 1,225 Facebook accounts comprising both Pages and public groups. These accounts were strategically selected based on their documented history of problematic content dissemination: each had repeatedly shared at least four URLs

[15] This client-side helper tool is available at https://fabiogiglietto.github.io/csds-preprocessor/.

from a dataset of 36,091 web pages flagged as false by Meta's third-party fact-checking partners between 2017 and 2022 (Messing et al., 2020). This baseline ensured that the initial sample represented verified vectors of misinformation, providing a robust foundation for network expansion.

During its 10-month operational period, the system demonstrated remarkable effectiveness in uncovering the hidden architecture of coordinated influence operations. The system identified:

- **7,068 coordinated posts** exhibiting synchronized sharing patterns;
- **10,681 unique coordinated links** representing the content infrastructure of influence operations;
- **2,126 new accounts** not present in the original seed list, demonstrating the system's capacity for organic network discovery;
- **17 distinct networks** organized across thematic and geographic clusters.

Networks discovered by the system represented diverse operational objectives spanning anti-vaccine rhetoric, migration-related discourse, and geographically targeted campaigns across Southeast Asia, Eastern Europe, and Latin America. The discovery of these new accounts represents a 173% expansion beyond the original monitoring scope, highlighting the workflow's effectiveness in mapping the evolving landscape of coordinated inauthentic behavior.

The VeraAI Alert System leveraged CrowdTangle's API infrastructure to implement its core detection components through three interconnected subsystems, each optimized for specific aspects of coordination detection. These subsystems are detailed below.

Posts Alert Sub-System

This component focused on identifying high-performing content that exhibited statistical anomalies indicative of artificial amplification. The system collected up to 100 over-performing posts from both initial account lists and newly discovered accounts within 6-hour monitoring windows. Advanced engagement metrics analysis included:

- **Comment-to-share ratio analysis** to identify posts with unusual engagement patterns;
- **Combined performance scoring** integrating multiple engagement indicators;
- **Historical baseline comparison** against archived performance data;
- **Statistical outlier detection** using red-flag scoring systems.

Posts exhibiting significant deviations from expected engagement patterns were automatically flagged and delivered through Slack notifications and Google Sheets integration for expert review.

Coordinated Links Alert Sub-System

This specialized component extended the analysis to examine coordinated link-sharing behavior across the monitored network. The system processed URLs extracted from high-performing posts, implementing sophisticated deduplication and analysis workflows:

- **URL cleaning and grouping** by expanded forms to identify coordination despite link shorteners

- **Chronological tracking** to determine earliest appearance dates for coordination timeline analysis
- **CooRnet iteration processing** to detect temporal synchrony in link sharing
- **AI-powered network labeling** using GPT-4's natural language processing capabilities to generate descriptive labels for coordinated networks

The system prioritized up to ten coordinated URLs based on engagement metrics for deeper analysis, enabling focused investigation of the most impactful coordination campaigns.

Accounts Update Sub-System

The most sophisticated component maintained and dynamically updated the master account lists used throughout the detection process. This sub-system implemented advanced behavioral clustering techniques:

- **Coordinated Image Text Sharing Behavior (CITSB) detection** using OCR analysis of shared visual content;
- **Coordinated Message Sharing Behavior (CMSB) analysis** identifying accounts sharing identical or near-identical text content;
- **Network expansion algorithms** comparing newly detected coordinated accounts against existing lists;
- **Priority ranking systems** for incorporating the most active newly discovered accounts into future monitoring cycles.

This feedback mechanism transformed the system into a self-improving detection infrastructure that expanded its surveillance network based on observed coordination patterns.

Operational Discoveries and Network Characterization

The comprehensive Facebook network analysis conducted through the VeraAI Alert System revealed large sophisticated coordinated influence operations, identifying **14,832 Facebook accounts** organized into **207 distinct coordinated communities**. This network mapping represents a 443% expansion beyond the initial monitoring scope, uncovering a complex multi-layered coordination infrastructure that spans continents, languages, and operational objectives.

The full list of all the discovered networks is available in Annex I[16].

The analysis revealed a globally distributed coordination ecosystem with distinct regional specializations. Table 3 documents Latin America commanding nearly half of all accounts (49%) focused on political movements and progressive activism. Southeast Asia follows with 19% of accounts targeting commercial operations and e-commerce exploitation, while Africa represents 17% through entertainment networks

[16] The featured case studies presented in this section provide illustrative examples of the diverse coordination patterns identified through the VeraAI Alert System. For the sake of space and to maintain the focus of this deliverable on methodological aspects, detailed case studies of specific communities and their operational characteristics will be extensively discussed in deliverable D4.3.

and religious communities. Europe deploys nationalist movements and pro-Russian narratives across 11% of accounts, North America handles conservative political messaging with just 3%, and multi-regional operations coordinate cross-border activities with only 1% of total accounts despite having the most communities. Distinct regional specializations are observable in these organized online influence operations, with Latin America achieving the greatest reach and multi-regional efforts focusing on coordination rather than scale.

Table 3 Geographic Distribution of Coordinated Communities.

Region	Communities	Accounts	%	Primary Focus
Latin America	45	7,234	49	Political movements, progressive activism
Southeast Asia	38	2,891	19	Commercial operations, e-commerce exploitation
Africa	12	2,456	17	Entertainment networks, religious communities
Europe	31	1,567	11	Nationalist movements, pro-Russian narratives
North America	18	384	3	Conservative political messaging
Multi-regional	56	142	1	Cross-border coordination
Total	**207**	**14,832**	**100**	

Through AI-powered community detection and labeling, the analysis identified four primary coordination typologies, each representing distinct operational models and strategic objectives. Table 4 shows that political influence operations constitute the primary focus of these influence operations, followed by commercial exploitation, with smaller but notable presences in entertainment and gambling sectors. Political Networks dominate at 62% of all accounts (9,147) through systematic influence operations targeting democratic contexts. Commercial/Marketplace operations represent the second-largest category at 23% (3,421 accounts), employing geographic targeting and e-commerce exploitation strategies. Entertainment/Religious networks account for 9% (1,323 accounts) by exploiting authentic community structures, while Gambling/Casino operations, despite having only 5 communities, control 6% of accounts (941) through psychological manipulation and AI-generated content.

Table 4 Coordination Typologies by Scale and Scope

Typology	Communities	Accounts	%	Key Characteristics
Political Networks	86	9,147	62	Systematic influence operations across democratic contexts
Commercial/Marketplace	67	3,421	23	Geographic targeting, e-commerce exploitation
Entertainment/Religious	49	1,323	9	Authentic community exploitation
Gambling/Casino	5	941	6	Psychological manipulation, AI-generated content

Political coordinated networks constitute the dominant category, with Latin American operations representing the largest single operational cluster (Table 5). These networks demonstrate sophisticated understanding of algorithmic amplification strategies, cultural localization techniques, and synchronized content dissemination patterns across multiple linguistic contexts.

Table 5 Major Political Coordination Clusters.

Community Id	Accounts	Description	Regional Focus
1	1,842	Pro-AMLO, Morena, and Fourth Transformation (4T) supporters	Mexico
2	1,133	Brazilian political groups: pro-Bolsonaro, pro-Lula dynamics	Brazil
6	889	Pro-Gustavo Petro, Colombia Humana, anti-Uribe activism	Colombia
8	804	Argentina Kirchnerist and Peronist political support	Argentina
5	213	Czech and Slovak nationalist, anti-EU, pro-Russia groups	Central Europe

Among the other cases, we conducted a more in-depth investigation of community 2 (Table 5). We collected 15 million posts published by these accounts between 2021 and 2023. The analysis revealed unexpected cross-cutting communication patterns contradicting traditional echo chamber theories. Following Lula's 2022 electoral victory, pro-Bolsonaro networks showed increased rather than decreased mobilization, with 45% of highest-interaction posts involving systematic infiltration by opposing political actors (Marino et al., 2025).

Concerning Central Europe, a coordinated network of 15 Facebook groups with 683,790 combined followers generated 5,917 posts featuring three dominant propaganda frames: glorified leadership, military strength, and historical exceptionalism. The operation employed format-specific coordination strategies exploiting Facebook's algorithmic preferences for engagement-rich content.

Commercial coordinated networks (Table 6) demonstrate sophisticated geographic targeting and psychological manipulation techniques. Gambling networks, despite comprising only five communities, show exceptional concentration and employ manufactured trust signals, urgency conditioning, and AI-generated visual content.

A significant operational discovery involved coordinated campaigns promoting online casinos and gaming through engagement bait tactics, organized across multiple specialized communities within the network. The largest gambling coordination cluster, Community 100, with 795 accounts, focused on "Online casino freeplay groups featuring Juwa, Orion Stars, Fire Kirin," while Community 120, with 107 accounts, operated "Brazilian online casino and betting platforms with signup bonuses" (Giglietto et al., forthcoming).

Another case documented that malicious actors systematically infiltrated existing large, poorly moderated online communities to spread problematic content. Community 45 (1,665 accounts) covering African entertainment and marketplace networks and Community 21 (728 accounts) encompassing Latin American interest groups exemplify sophisticated organic amplification through authentic community co-optation.

Table 6 Commercial and Gambling Network Characteristics.

Network Type	Communities	Accounts	Monthly Posts	Key Tactics
Southeast Asian Marketplaces	28	1,890	~8,000	E-commerce ecosystem exploitation
Latin American Commerce	18	1,045	~6,000	Regional marketplace infiltration
Online Casino Platforms	2	902	~10,000	Psychological manipulation, fabricated receipts

Network Type	Communities	Accounts	Monthly Posts	Key Tactics
African Commerce	8	312	~2,500	Concentrated commercial manipulation

Technological and Operational Sophistication

The network analysis revealed clear evidence of **systematic automation and AI integration** across coordination clusters, indicating evolution from simple bot networks to sophisticated organizational structures. Automation and AI integration are detailed in Table 7.

Table 7 Sophistication Indicators Across Network Types

Indicator Category	Evidence	Scale
Scale Indicators	Average community size of 72 accounts (range: 2-1,842)	207 operational units
Organizational Complexity	Multi-layered structures beyond bot networks	Cross-community coordination
AI Integration	Exponential posting increases post-AI adoption	Systematic synthetic content
Geographic Targeting	Cultural localization and demographic focus	Multi-lingual operations
Platform Exploitation	Algorithmic amplification strategies	Cross-platform coordination

Technical Infrastructure and Alert Delivery

The VeraAI Alert System implemented a comprehensive notification infrastructure designed to support real-time response by fact-checkers and media watchdogs. The system delivered alerts through multiple channels to ensure rapid detection and response to coordinated inauthentic behavior.

RSS Slack Integration

A dedicated Slackbot (Figure 12) provided real-time notifications to verified fact-checking organizations and media professionals through an integrated RSS feed. This system enabled immediate notification of suspected coordination campaigns, delivering structured data that included account lists, coordination

evidence, and preliminary analysis. The integration was designed to seamlessly fit into existing fact-checking workflows and verification processes, allowing organizations to respond quickly without disrupting their established operational procedures.

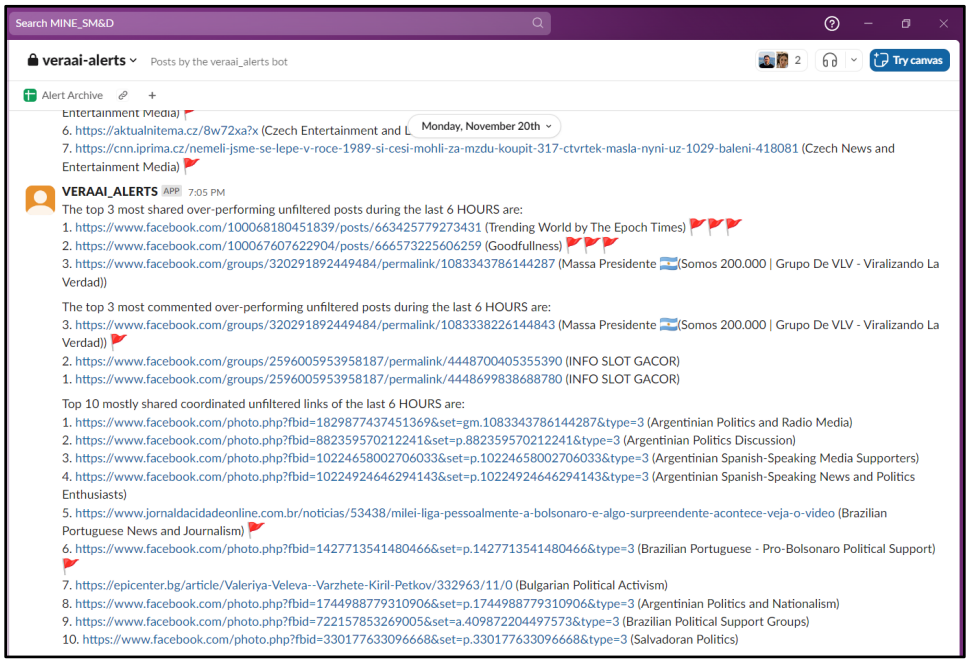


Figure 12 Slack channel of the RSS VeraAI Alert System.

Automated Data Archival

The system maintained comprehensive archives of detected coordination patterns to support both immediate response and long-term analysis. These archives enabled historical trend analysis across different campaign types, preserved evidence for investigative journalism and academic research, and provided longitudinal tracking of network evolution and adaptation strategies. This archival capability ensured that patterns could be identified over time and that evidence remained accessible for future investigation and verification efforts.

Platform Policy Impacts and Future Challenges

The operational effectiveness of the VeraAI Alert System was abruptly terminated on August 14, 2024, due to Meta's deprecation of CrowdTangle. This disruption highlighted critical dependencies in the misinformation research ecosystem and exposed fundamental limitations in current platform transparency initiatives. The transition from CrowdTangle to Meta's successor tools revealed insurmountable barriers[17] that fundamentally altered the landscape for coordinated behavior detection research.

[17] See Annex II for further details.

Meta Content Library (MCL) Limitations

A comprehensive assessment of adapting the workflow to Meta's Content Library (MCL) revealed critical technical infrastructure constraints that prevented effective implementation. The MCL's proprietary ID system cannot be used on Meta platforms, creating an unbridgeable gap between research data and live content verification. Without content URLs, fact-checkers cannot access posts in their original context, while the absence of automated monitoring capabilities in both Meta's Research Platform and SOMAR VDE eliminated the scheduled, real-time detection essential to the original workflow.

The analytical capabilities of the new system proved equally limiting. The absence of CrowdTangle's overperforming score requires multiple API calls to calculate manually, significantly reducing efficiency. Restricted search functionality for exact text matches beyond five words impedes coordination detection, while reduced functionality for analyzing visual content shared across networks further constrains investigative capabilities. These technical limitations transform what was once an automated process into one requiring teams of researchers and time investments extending from hours to days.

Current Operational Status and Challenges

The digital clean room isolation of Meta's research environments prevents real-time alert system implementation, which was essential for timely notification to stakeholders. Strict data deletion policies and export restrictions disrupt the continuity needed for tracking coordinated behaviors over time, while manual labor requirements have replaced automated processes entirely. This represents a substantial reduction in the research community's capacity to detect and respond to emerging coordination campaigns, fundamentally altering the operational landscape from an API-driven, real-time monitoring system to a manually intensive, retrospective analysis environment.

Regulatory and Policy Implications

The challenges encountered in adapting the VeraAI Alert System to post-CrowdTangle environments highlight critical gaps in current platform governance frameworks. The system's inability to function effectively within Meta's new research infrastructure raises important questions about platform accountability and the adequacy of current transparency measures for supporting independent research. The balance between user privacy and legitimate public interest research needs remains unresolved, while the role of regulatory frameworks like the Digital Services Act in ensuring meaningful research access continues to evolve.

These limitations underscore the urgent need for consistent, sustainable research infrastructure across major platforms and highlight the broader implications for industry standards in supporting academic and journalistic investigation of coordinated inauthentic behavior. The transition has effectively demonstrated how changes in platform policy can significantly impact the research community's ability to monitor and respond to emerging threats in the information ecosystem.

Future Adaptations and Methodological Evolution

Despite the substantial operational challenges introduced by platform policy changes, the core methodology of the VeraAI Alert System remains valid and adaptable to evolving research

environments[18]. Future iterations must emphasize resilient, platform-independent approaches that can withstand shifts in corporate policy while maintaining effectiveness in detecting coordinated inauthentic behavior.

Development of detection workflows that are resilient to a limited and unreliable researchers' APIs environment represents a critical adaptation strategy. In the short term, this approach necessitates implementing hybrid methodologies that combine responsible public interest scraping techniques with cross-platform data integration capabilities. This automatized, yet ethically-grounded, methodology allows for the continuity of research while the long-term infrastructure is built.

Looking forward, future systems must be architected around the principles of a new, regulated era of transparency. This moves beyond relying on voluntary platform cooperation. Instead, it envisions a system where next-generation researcher APIs—mandated by legal frameworks like Europe's Digital Services Act—feed coordination detection algorithms and other scheduled monitoring analysis. This entire process would be supported by a shared, distributed infrastructure and governed by a common ethics and privacy framework, two of the essential pillars for sustainable, independent analysis. Such an architecture ensures that research can operate across multiple platforms simultaneously, providing the comprehensive coverage needed for effective coordination detection without being subject to the policy whims of any single platform.

Collaborative Research Frameworks

Development of institutional partnerships that pool resources and expertise offers a pathway to overcome individual research limitations. Multi-institutional cleanroom access sharing can distribute costs and technical burdens while expanding analytical capabilities. Collaborative annotation and verification systems can leverage distributed expertise to improve detection accuracy, while distributed monitoring networks across academic institutions can provide comprehensive coverage that individual research teams cannot achieve. Standardized methodology frameworks for cross-institutional research ensure consistency and reproducibility while enabling large-scale collaborative investigations.

These adaptations collectively represent a fundamental evolution in coordination detection methodology, moving from platform-dependent monitoring to resilient, distributed approaches that can maintain effectiveness despite corporate policy constraints. The integration of advanced AI capabilities with collaborative research frameworks offers the potential to not only restore but also enhance the capabilities that existed under previous platform transparency regimes.

Impact Assessment and Research Contributions

The VeraAI Alert System has demonstrated that coordinated influence operations can be effectively detected, tracked, and analyzed at a global scale through structured, iterative monitoring approaches. The system's discoveries have contributed essential insights to the understanding of:

[18] See 3.3.4. TikTok Coordinated Detection Project for an implementation of the same workflow to TikTok.

- **Network Evolution Dynamics**: How coordination networks adapt and expand over time
- **Cross-Platform Operation Strategies**: The methods employed by influence operations to maintain effectiveness across different platform environments
- **AI Integration in Disinformation**: The emerging role of generative artificial intelligence in scaling and sophisticating influence operations
- **Platform Governance Challenges**: The limitations of current transparency frameworks in supporting effective detection and response

The methodology established by the VeraAI Alert System provides a foundational framework for future coordination detection research, demonstrating both the potential for systematic, large-scale monitoring and the critical importance of sustainable, platform-independent research infrastructure in maintaining information integrity across evolving digital environments.

3.3.4 TikTok Coordinated Detection

Following successful implementations on Facebook and Instagram, the TikTok Coordinated Detection pipeline represents a significant methodological advancement in identifying coordinated behavior across social media platforms. This system integrates scalable data retrieval mechanisms with sophisticated automated detection algorithms based on temporal and structural behavioral criteria, establishing a comprehensive framework for continuous monitoring of coordination patterns on TikTok.

Data Collection Strategy

The detection pipeline leverages traktok, a specialized R package that interfaces with the TikTok Research API to systematically retrieve metadata and video-level content from public user profiles. Operating on a daily cycle, the system queries all accounts within the monitored pool and downloads their most recent videos, creating a dynamic dataset that evolves as new accounts are identified and integrated based on behavioral similarity to existing entities.

The initial monitoring pool was strategically assembled through two complementary approaches:

Reactive Seeding: This approach responds to emergent content flows and trending behaviors. The initial analysis focused on posts tagged with #moscow, which emerged following the terrorist attack in Moscow. This entry point facilitated the detection of a preliminary cohort of accounts engaged in synchronized sharing activity, establishing the foundation for broader investigations into networked behaviors across social media environments.

External Intelligence Integration: The system incorporates vetted intelligence from trusted research partners within established consortia. For instance, on April 19, 2024, a comprehensive list of 513 potentially problematic accounts was provided by a partner within the vera.ai consortium. To ensure analytical rigor and minimize noise, only accounts appearing multiple times across intelligence sources were incorporated into the monitoring pool.

Account selection follows a systematic protocol where entities are included either through direct listing or by demonstrating temporally synchronized posting patterns that indicate coordinated behavior.

Detection Methodology

The core detection component utilizes CooRTweet, specifically adapted for TikTok data structures, to identify clusters of accounts exhibiting coordinated behavior patterns. The coordination detection operates under precisely defined parameters: time_window = 120 seconds, min_participation = 2 posts, edge_weight = 0.9. These parameters define the coordination threshold: posts must be published within 120 seconds of each other (time_window), by at least two different accounts (min_participation), and must share at least 90% similarity (edge_weight) to be considered coordinated.

For each daily processing batch, the system constructs comprehensive network graphs where nodes represent individual accounts and edges denote pairs of accounts sharing identical or nearly identical content within the specified coordination timeframe. These graphs undergo structural analysis to extract meaningful components (clusters), with advanced descriptive metrics including degree centrality, network density, and clustering coefficients computed for both individual nodes and entire subgraphs.

Account and Network Characterization

The system employs a multi-dimensional characterization framework for identified accounts, incorporating both structural and engagement-based features:

Account-Level Metrics:

- Number of coordinated posts and participation frequency
- Temporal patterns and regularity of sharing behavior
- Network position (degree and centrality within coordination clusters)
- Engagement metrics (average views, likes, comments, and shares per video)

Cluster-Level Analysis: Networks are profiled comprehensively by size, density, internal cohesion, and content characteristics. These metrics enable sophisticated assessment of whether coordination appears organic, spam-driven, or indicative of strategic manipulation campaigns.

Preliminary analytical findings reveal the coexistence of small, tightly synchronized dyads alongside larger, looser coalitions that maintain thematic coherence over extended periods. Notably, several clusters demonstrate coordinated reposting of specific storytelling formats—including Reddit-style narratives and viral anecdotal content—suggesting deliberate attempts to simulate authentic content virality and organic user engagement.

Content Labeling and Qualitative Analysis

To provide comprehensive qualitative insights into detected cluster behaviors, each coordinated network undergoes systematic annotation using AI-generated content labels. These labels are produced through an LLM-based (OpenAI's GPTo-mini in the current implementation) content summarization procedure applied to the complete set of posts within each cluster, designed to capture recurring narrative themes, distinctive stylistic features, and specific calls to action. This interpretive layer supports advanced downstream analysis of coordination intent, content genre classification, and strategic manipulation identification.

System Performance and Scalability

The pipeline has demonstrated robust operational performance, successfully processing thousands of TikTok videos from a monitoring pool currently exceeding 4,800 accounts. Despite occasional API-related

challenges, including rate limits, content removal, and account access restrictions, the system maintains consistent data collection cycles while continuously identifying new coordinated entities. The iterative architectural design ensures that the monitored ecosystem continuously adapts to evolving behavioral patterns and emerging content strategies.

The system's scalability has been validated through sustained operation across diverse coordination scenarios, from small-scale synchronized posting to large-scale information campaigns, demonstrating its effectiveness across the full spectrum of coordinated behavior manifestations.

Operational Results and Performance Metrics

Over the monitoring period from December 2024 through June 2025, the TikTok Coordinated Detection pipeline has demonstrated significant operational success and detection capabilities (detailed in Table 8). Across 136 active monitoring days, the system processed over **1.26 million TikTok posts**, identifying **2,521** coordinated posts representing a coordination rate of 0.2% of all analyzed content. This detection rate, while seemingly modest, represents a substantial volume of coordinated activity given the scale of content processing, with an average of 18.5 coordinated posts detected daily and peak coordination events reaching 63 posts in a single day.

The system's account-level detection capabilities have proven particularly robust, identifying **8,574** coordinated account instances across the monitoring period, averaging 63 coordinated accounts per day. Peak coordination events involved up to 396 accounts participating in synchronized activities, demonstrating the system's capacity to detect both small-scale coordinated clusters and large-scale coordination campaigns. The pipeline successfully identified **2,248 distinct coordination networks**, with an average of 16.5 networks detected daily and peak detection events identifying up to 94 networks simultaneously.

System reliability maintained a 66% success rate across the monitoring period, with 136 successful data collection cycles and 70 failed attempts primarily due to API limitations and access restrictions. The reported 66% success rate pertains to data retrieval only, and does not account for downstream processing or coordination detection accuracy. This performance demonstrates the system's resilience in maintaining consistent monitoring capabilities despite technical challenges inherent to large-scale social media data collection and often aggravated by the bugs afflicting early versions of the research APIs made available by platforms to external researchers (Pearson et al., 2025).

The project is still ongoing. All detection results are made publicly available and published on a daily basis through an interactive web interface[19], ensuring transparent access to coordination findings and supporting broader research community efforts in understanding platform dynamics and coordinated behavior patterns.

[19] https://fabiogiglietto.github.io/tiktok_csbn/

Table 8 Key Metrics from TikTok Coordination Detection Pipeline.

Metric	Value
Monitoring Period	Dec 2024 – Jun 2025
Total Posts Processed	1,260,000+
Coordinated Posts Detected	2,521 (0.2% of total)
Average Coordinated Posts/Day	18.5
Peak Coordinated Posts/Day	63
Distinct Coordination Networks	2,248
Average Coordinated Network/Day	16.5
Peak Coordinated Networks/Day	94
Peak Coordinated Accounts/Day	396

Methodological Innovation and Applications

This implementation demonstrates that coordinated behavior detection on TikTok can be operationalized through the strategic integration of automated data collection, structured network analysis, and intelligent content summarization. By maintaining continuously updated account pools and generating detailed network-level analytical outputs, the system makes substantial contributions to the documentation and interpretation of coordination patterns as they evolve in real-time.

The methodology represents a significant step forward in the ongoing effort to enhance platform dynamics observability and accountability, providing researchers, platform operators, and policymakers with sophisticated tools for understanding coordinated information campaigns and their impacts on digital discourse.

Preliminary observations

The TikTok Coordinated Detection pipeline identified several distinct patterns of coordinated behavior, with networks ranging from small dyadic partnerships to larger coalitions engaged in synchronized

content distribution. The detected networks exhibited sophisticated coordination strategies, including the systematic use of TikTok's duet feature to amplify identical content across multiple accounts. These networks employed both temporal coordination—posting within the 120-second detection window—and structural coordination through duet chains that created interconnected content webs. Some networks demonstrated a clear thematic focus, with some clusters coordinating around political narratives (both pro-Trump and anti-Trump content, as shown in Figure 13), while the majority engaged in mere amplification activities of repurposed content (e.g., Reddit viral posts).

Notably, the analysis revealed that numerous networks exploited TikTok's duet functionality as a primary coordination mechanism, allowing accounts to share identical video content while maintaining the appearance of organic user engagement. This duet-based coordination strategy aligns with findings from recent research by Luceri et al. (2025), who identified similar coordinated exploitation of platform-specific features for inauthentic amplification campaigns. The detected networks often combined duet coordination with temporal synchronization, creating multi-layered coordination patterns that were more sophisticated than simple simultaneous posting behaviors observed on other platforms.

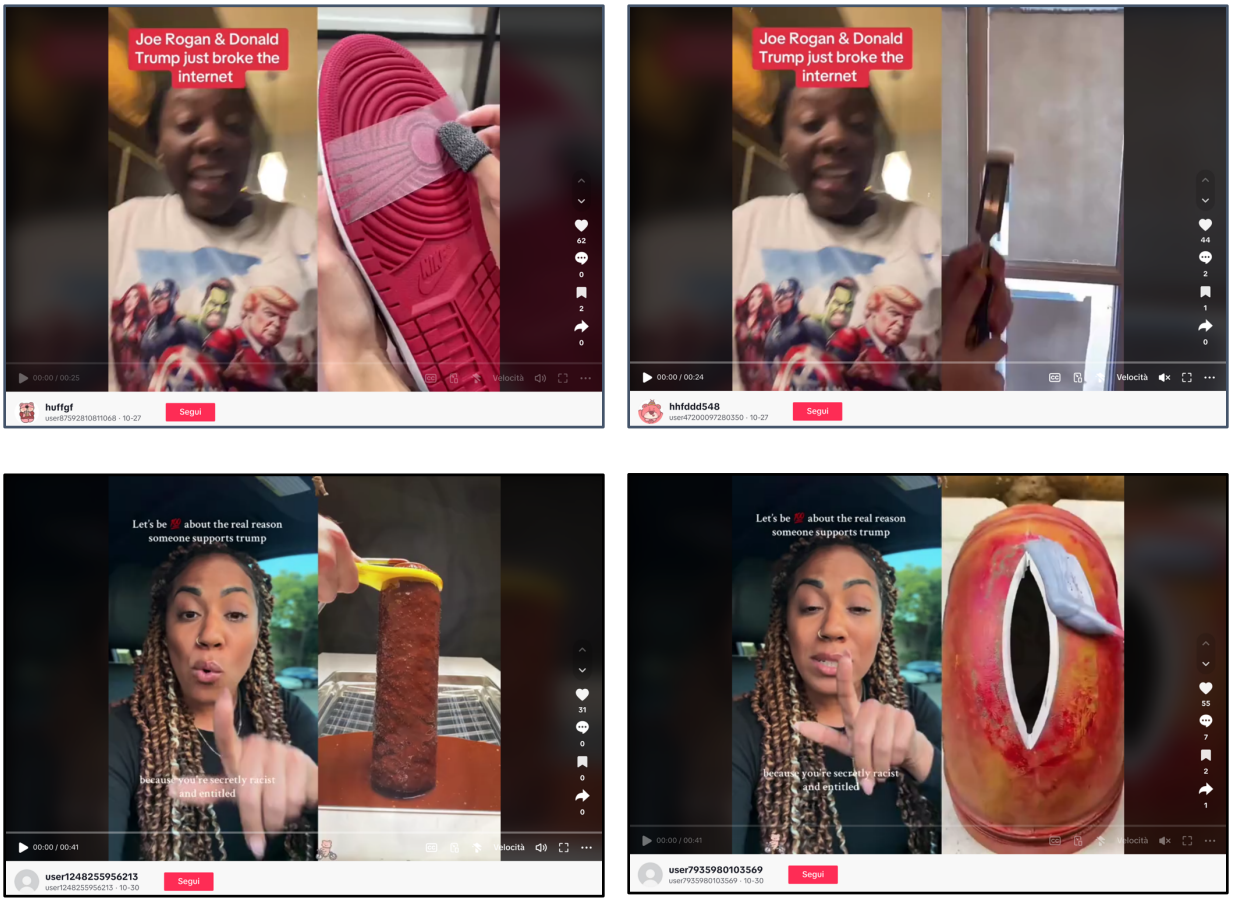


Figure 13 Example of pro and anti-Trump coordinated content. The same content (left part of the splitted screen video) is posted at about the same time by multiple accounts and accompanied by completely unrelated content (on the right part of the video) with a purely attention-grabbing role.

The identified networks demonstrated varying levels of operational sophistication, from basic content replication to complex narrative coordination involving complementary messaging strategies designed to dominate specific topic discussions or trending hashtags during politically sensitive periods.

Future Directions

The successful deployment of this detection pipeline establishes a foundation for expanded coordination analysis across additional platforms and behavioral modalities. The system's modular architecture supports integration with broader coordination analysis infrastructures, enabling cross-platform coordination detection and comprehensive mapping of multi-platform information campaigns. This methodological framework contributes to the emerging field of computational social science focused on understanding and mitigating coordinated inauthentic behavior across digital platforms.

4 Spatio-temporal analysis of disinformation campaigns and narratives

Harmful and misleading narratives are central tools in disinformation strategies, influence campaigns, and broader foreign information manipulation and interference (FIMI) operations. These narratives are routinely employed to sway electoral outcomes, deepen social polarisation, and erode public trust in democratic institutions at both national and EU levels (Roselle et al., 2014). One of the key challenges in countering them lies in Europe's cultural and linguistic diversity, which gives rise to country-specific narrative ecosystems shaped by distinct political contexts. At the same time, some misleading narratives transcend national boundaries and circulate widely across Europe. Their transnational reach makes them particularly difficult to monitor and analyse, due to the wide variety of languages, platforms, and media environments involved. These narratives often exploit societal vulnerabilities—such as distrust in electoral processes, economic insecurity, or contentious issues like gender, religion, immigration, vaccination, and climate change—and frequently become central themes in disinformation campaigns.

Despite widespread recognition of the need to address these narratives, there is still no shared and operational definition of what constitutes a "harmful narrative." This is partly because such narratives are not always entirely false. Within this broader category, disinformation narratives are more clearly defined. The European Digital Media Observatory (EDMO), for example, defines a disinformation narrative as a coherent message that emerges from a consistent set of content and can be demonstrated as false through fact-checking (Panizio, 2024).

To respond effectively, there is an urgent need for scalable methods to detect recurring narrative patterns and identify similarities across content. These capabilities are crucial for uncovering coordinated behaviours, tracing how narratives spread across platforms, and understanding their reach and influence on different audiences.

An equally important, though often overlooked, challenge is the lack of longitudinal analysis—studies that track disinformation narratives and campaigns over extended periods. Such analyses are essential for understanding how narratives evolve in response to political changes, media attention, or platform interventions. Addressing this gap requires the development of open-source tools that support both long-term and cross-platform monitoring, enabling researchers and practitioners to trace how narratives adapt and unfold across diverse digital spaces and in connection with real-world events.

Task 4.2 aims to respond to this need by developing tools specifically designed to detect narratives along with their spatial and longitudinal characteristics.

4.1 Background

This section documents the progress of the research outcomes achieved within the vera.ai project related to the spatio-temporal analysis of disinformation campaigns and narratives carried out in T4.2.

The main goal of task T4.2 is to support media professionals and analysts by enabling automatic detection and visualization of how disinformation campaigns and narratives emerge, evolve and spread across time and geographic space. The task focuses on developing AI-based methods to identify the temporal

dynamics and spatial distribution of disinformation narratives, uncovering coordinated behavior and tracking the evolution of themes and rhetoric across regions and languages. This perspective provides insights into identifying patterns, origins and escalation paths of harmful content and adds value for early warning and contextual understanding of disinformation operations.

Throughout the task, we have researched and implemented various methods for modeling the temporal and geographic aspects of disinformation. Our work focused on narrative clustering and tracking over time, hierarchical topic modeling and name entity extraction. We evaluated these methods on different datasets, showing their ability to capture emerging disinformation trends, identify peaks of activity and detect shifts in narrative focus.

In addition, we developed a visualization tool capable of performing spatio-temporal analysis of online content. This tool allows users to visualize the spread of narratives over time, monitor evolving disinformation topics and detect potential coordinated campaigns.

Finally, the insights and technical experience gained in this task contributed to broader cross-task integration and future planning. Our spatio-temporal analysis approaches provide a more nuanced and contextualized understanding of disinformation activity compared to static or content-only approaches. The ability to observe temporal and spatial patterns enhances the interpretability and actionability of disinformation detection, making the tools especially valuable to investigative journalists and analysts.

In the following subsections, we provide more in-depth information on our results related to the proposed methodologies, implementation and evaluation.

4.2 Methodology

The methodology we developed focuses on providing tools that enable end users to conduct large-scale spatial and temporal analysis. The pipeline includes **narrative clustering**, which involves extracting central claims from articles and applying named entity recognition to identify entities mentioned in the documents, including spatial and temporal information. Using this information, a clustering method organizes articles published in similar locations and timeframes as narratives for further analysis. Additionally, a **narrative evolution analysis tool** offers relatedness classification to link relevant articles across the temporal dimension and provides topic modeling functionality to analyse the evolution of topics over time. Finally, a **chatbot interface** allows users to interact with the extracted features from the above tools, including narrative clustering, locations, dates, and topics.

4.2.1 Narrative Evolution Analysis

Topic modelling can be used efficiently for narrative theme analysis and is an important feature for document-relation classification. One of the main limitations of existing topic models is their lack of interpretability and the inability to control topic granularity. We propose a large language model based topic-modelling approach to address these limitations and make the method more suitable for narrative analysis. The approach involves training a large language model with reinforcement learning (Mu, Bai, et al., 2024) to generate topics (Mu, Dong, et al., 2024). Each topic is also represented by a word embedding from the LLM. We apply a hierarchical clustering algorithm - HDBSCAN (Campello et al., 2013) to cluster

the topics and build a topic hierarchy, enabling users to choose the desired level of granularity for their analysis.

Understanding the development of disinformation narratives over time and across regions requires identifying documents that refer to the same event or topic, despite variations in language, timing or location. To facilitate this, we aim to predict pairwise relationships between documents in order to group together those that are topically or semantically related. By identifying such relationships at scale, we can construct coherent storylines that trace how a narrative unfolds over time and reveal its origin, spread, and transformation. These storylines provide a structured view of disinformation dynamics, offering valuable context for both automated systems and human analysts. Figure 14 illustrates an ideal structure of storylines based on document relationships:

Figure 14 The ideal structure of storyline-based temporal analysis of disinformation narratives. This structure not only indicates which documents are related, but also organizes them along a timeline, with arrows pointing from earlier to later documents. Documents covering the same topic (e.g., Typhoon Haikui) are grouped as a main branch, while those covering related topics (e.g., Typhoon Higos) appear as sub-branches. This approach helps users gain a clearer understanding of how related narratives evolve over time.

In this study, we develop a document pair relationship classifier inspired by the event extraction algorithm proposed by Liu et al. (2020), which employs a two-layer architecture to construct document relationship graphs for long texts. The first layer constructs a keyword co-occurrence graph using community detection and associates documents to communities based on their keyword occurrence. The second layer further extracts document subsets from these communities and builds document relationship graphs using a pre-trained document pair relationship classifier. We adopt only the classifier component to ensure compatibility with both short and long texts.

We construct the following features for each document pair: Jaccard similarity of keyword sets, cosine similarity of Term Frequency-Inverse Document Frequency (TF-IDF) features, cosine similarity of Latent Dirichlet allocation (LDA) features, cosine similarity of embeddings using a sentence-transformers model[20], and Jaccard similarity of name entity sets. Originally, Liu et al. (2020) trained a Support Vector Machine (SVM) model (Hearst et al., July-Aug 1998) to predict the document pair relationship using an additional set of 5000 documents (not publicly available). We replace the model with a Logistic Regression (LR) model for time efficiency. We train an LR model on a publicly available dataset from Liu et al., 2020, where each document is labeled with a story ID. The model takes the above features as input and outputs

[20] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

a predictive score for each document pair. We then construct a document relationship graph by grouping all related documents, defining pairs with a predictive score > 0.5 as related.

4.2.2 Multilingual Narrative Discovery Pipeline

The KInIT's narrative detection pipeline accepts input in three flexible formats to accommodate different user needs and levels of computational complexity. First, users can perform clustering over an existing corpus of check-worthy claims—either from the Meta Content Library (a static snapshot) or the dynamically updated MultiClaim dataset powered by KInIT's MONANT system. This setup enables the exploration of pre-computed narrative hierarchies at scale, with interactive filtering based on relevant topics or named entities. Second, the pipeline supports direct processing of raw text provided by the user, from which check-worthy claims are extracted using the Central Claim Extraction service developed by KInIT. These claims are then clustered into narrative hierarchies. While this approach enables on-demand narrative analysis, it may be computationally intensive and lacks the broader context provided by a reference corpus. Third, and most relevant for practical use, the system supports a hybrid mode where user-supplied text is analyzed to extract new check-worthy claims that are integrated into the existing corpus-based hierarchy. This allows users to situate novel claims within the broader narrative landscape. Named entity recognition is applied to each extracted claim to support targeted filtering and thematic exploration.

On top of this flexible input layer, our pipeline applies multilingual contrastive embeddings to encode the semantic similarity of claims and their named entities across languages and narrative styles. Retrieved or extracted claims are embedded using a fine-tuned multilingual encoder, and a similarity graph is constructed via k-nearest-neighbors based on cosine distance. The core of the clustering process is a two-tiered hybrid pipeline. At the first level, BERTopic is used to generate an initial segmentation of the corpus into broad, semantically coherent clusters, each annotated with representative keywords and topic summaries. This provides an interpretable and efficient overview of high-level narrative structures.

To capture finer-grained distinctions, a conditional hierarchical sub-clustering procedure is applied selectively to large BERTopic clusters—those exceeding a predefined size threshold (default: 50 claims). Within these clusters, we first apply the Louvain algorithm to identify macro-narratives by detecting communities in the semantic similarity graph. Subsequently, HDBSCAN is used within each Louvain cluster to uncover micro-narratives that reflect distinct rhetorical framings or sub-themes. Small clusters below the threshold are preserved in their original BERTopic form, ensuring scalability and computational efficiency without sacrificing resolution. Each cluster and sub-cluster is automatically labeled and summarized using a multilingual large language model to produce interpretable narrative descriptions. The resulting narrative hierarchy is visualized through the DataMapPlot framework, supporting interactive exploration and filtering by named entities or topics. The pipeline also supports human-in-the-loop refinement, enabling analysts to re-label, merge, or split clusters as needed. These annotations are fed back into the embedding model to continuously improve future iterations. This modular and adaptive architecture supports both robust offline analysis and real-time integration of novel content into an evolving narrative map.

The pipeline consists of several stages described by the diagram in Figure 15.

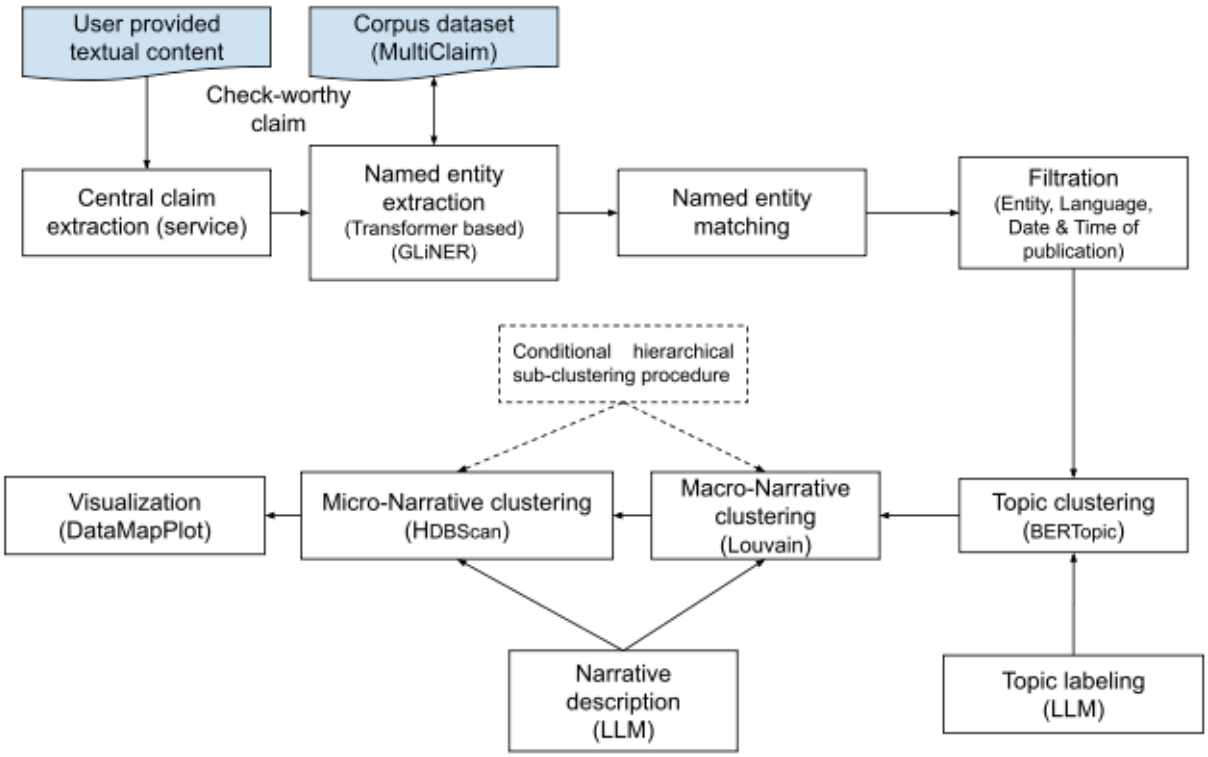


Figure 15 KInIT Narrative detection pipeline diagram consisting of two types of input: a) user-provided text in combination with the Central Claim extraction service, and b) corpus dataset. The process consists of preprocessing and filtration of the input and a multi-step clustering using the BERTopic and HDBScan in combination with Louvain clustering. All the cluster descriptions are generated using the LLM based on the named entities (BERTopic) or directly from the claims (DBScan).

4.2.3 Multilingual Claim Clustering and Disinformation Narrative Uncovering

The ONTO team has developed a system designed for the multilingual clustering of disinformation claims. It facilitates the discovery and analysis of cross-national and multilingual disinformation narratives by uncovering complex cases of disinformation content reuse.

The system leverages data from the vera.ai user-facing tool Database of Known Fakes (DBKF), incorporating the metadata and enrichments from it into the clustering process to improve the discovery, description, and organisation of the narrative elements it discovers.

Easier access to the insights available in the clustering output is a key feature. The ONTO team developed a chatbot interface to the DBKF, which provides a more intuitive access to the data that can also act as an incremental introduction to its advanced features.

System Overview

The clustering system was initially developed to detect debunks of virtually identical claims. There is both a huge variety in the disinformation covered by debunks globally, but also significant overlap in the key concepts addressed by debunks. To that end the system is designed to produce large specific claim clusters that discuss very closely tied claims.

The system employs a sophisticated clustering algorithm, specifically chosen for its ability to handle large datasets and multilingual data. When applied to nearly 200,000 claims in the DBKF, about half of the claims are unique and do not get clustered, meaning only one debunk addresses the specific claim. Of the remaining claims, most fall into small clusters of 2-6 claims, with only about 1% of clusters containing 10 or more claims. The average cluster size is 2.7, reflecting the system's precision in grouping closely related claims.

In addition to processing debunks, the system is capable of clustering generic content such as social media posts. In a test export of 19,000 TikTok and Meta posts, 75% of all posts were combined into larger clusters, with an average size of 5 posts per cluster. This result is expected, as social media posts tend to be more repetitive and less unique. Combining these two experiments demonstrates the system's versatility in quickly processing large quantities of data and tying social media posts to known instances of disinformation claims and debunks.

Prerequisites

The system requires access to both structured and unstructured data, including raw text, metadata, and additional enrichments. It uses the data collected in the DBKF ingestion process to tag clusters with key concepts, authors, and events and to construct timelines of narrative development. These enrichments are most effective when presented as a knowledge graph, enabling the exploration of linked clusters.

Clustering

The cluster building process comprises four main steps (Figure 16).

First, the process begins with document chunking and embedding calculation. This step ensures that all documents are divided into appropriately sized segments, and embedding vectors are computed for each segment using paraphrase-mpnet-base-v2[21]. These embeddings capture the semantic information of the text, facilitating accurate clustering.

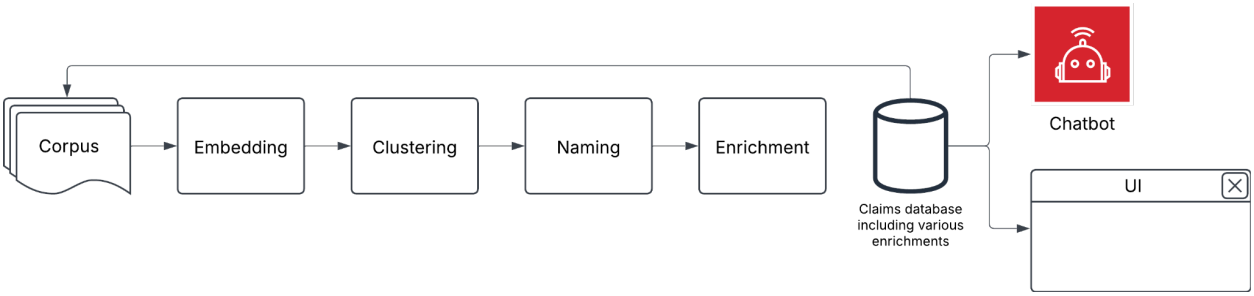


Figure 16 Cluster building process which consists of (i) chunking the texts and calculating embeddings for each chunks, (ii) cluster creation for documents on a corpus level, (iii) creating brief descriptive cluster names, (iv) enrichment with metadata such as key concepts, and (v) ingestion into the claims database where it can be accessed by both the UI and chatbot.

The second step involves cluster creation. Utilizing a GPU-enabled version of the HDBSCAN algorithm (cuml[22]), this step generates clusters and provides probabilistic assignments of each document chunk to a cluster. Documents are either assigned to a specific cluster, if a sufficiently good match is found without competing alternatives, or remain in their own individual clusters.

[21] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2
[22] https://github.com/rapidsai/cuml

Next, clusters are named using a local Small Language Model (Phi-4-mini-instruct[23]). This model employs a specialized prompt and specific parameters to produce concise and clear cluster names, enhancing the interpretability and usability of the clustering results.

Finally, cluster metadata is calculated based on the Knowledge Graph. This metadata includes key concepts, authors, languages, and the start and end dates of cluster activity. Such information enriches the clusters, providing valuable context and insights for further analysis.

Streaming Data

The current cluster building process does not inherently support streaming data. However, as a compromise, the system can generate embeddings for new documents and attempt to incorporate them into existing clusters or initiate new clusters for them. This approach remains valid as long as the volume of newly ingested documents is significantly smaller than the initially clustered documents.

Should the system ingest large corpora or experience a substantial increase in new documents over time, a complete rebuild of the clusters becomes necessary. This ensures the integrity and accuracy of the clustering process.

Accessibility

The results of the clustering process are made accessible through the Database of Known Fakes (DBKF) interface, offering users two primary methods to explore and interact with the data.

Firstly, users can utilise the cluster search and cluster view screens (Figure 17). These interfaces allow for the discovery of clusters through a powerful faceted full-text search, enabling efficient navigation and filtering of the clustered data based on various criteria.

[23] https://huggingface.co/microsoft/Phi-4-mini-instruct

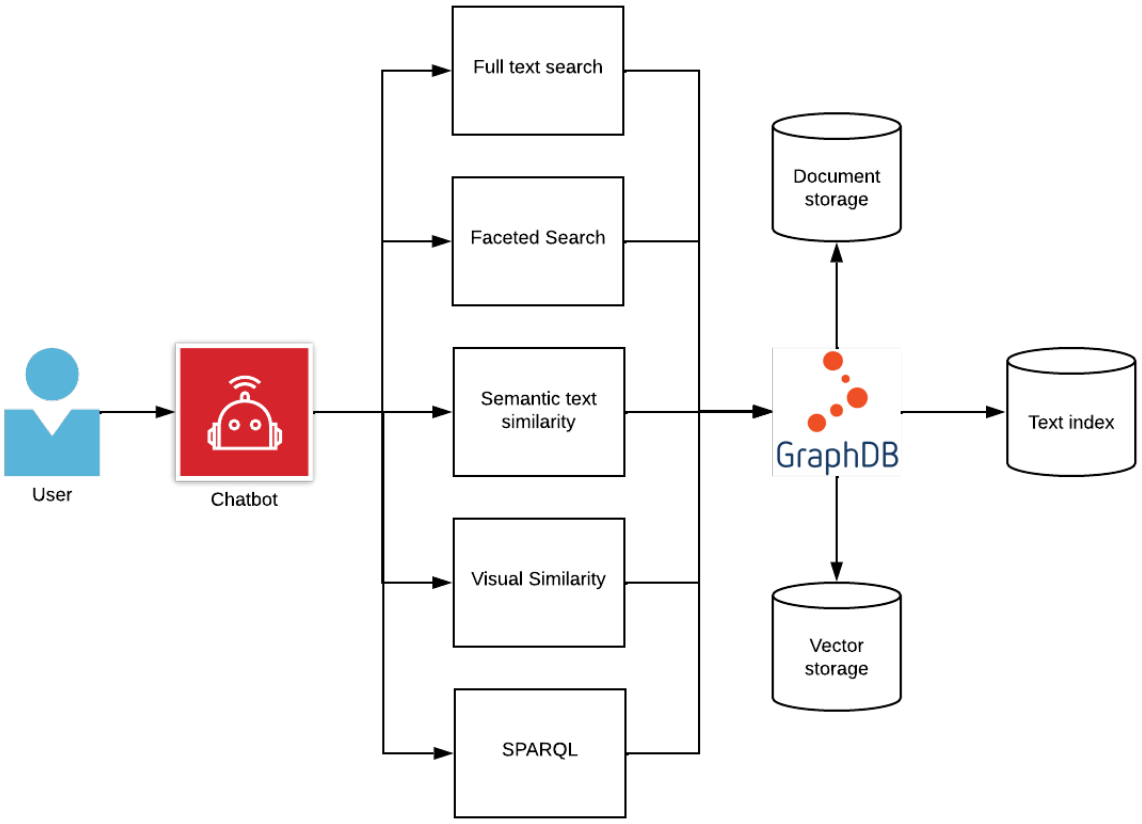


Figure 17 Search tools, used by the DBKF chatbot.

Second, an agentic chatbot is available to assist users. This chatbot has access to all the search functionalities available through the DBKF user interface, as well as direct database access. It is designed to help users perform complex, multi-tool interrogations of the data, providing a more interactive and guided experience.

Together, these accessibility features ensure that users can effectively explore and analyse the clustering results, whether through self-directed search or assisted interactions with the chatbot.

ONTO's clustering and protonarrative discovery approach addresses the challenge of identifying and analysing repeated disinformation claims across different languages, geographical regions, and time periods. By integrating metadata, such as language, dates, and sources, it provides the ability to track the evolution and spread of disinformation. Accessible through both a detailed search interface and an interactive chatbot, the output of the clustering system can be a powerful tool for journalists and investigators to uncover and understand the spatio-temporal dynamics of disinformation campaigns. While currently limited in handling streaming data, the system's robust clustering capabilities offer significant insights into the persistent and cross-border nature of disinformation.

4.3 Implementation

Temporal Analysis Visualization Tool

USFD has developed a web service to visualize the document relationship graphs for the temporal analysis of disinformation narratives. This visualization tool enables users to analyze their own data and interact with the results, with the goal of providing interpretable and actionable insights. The user interface (UI) requires two input files: (1) a CSV file containing the user's data, with one column for texts and another for dates in a consistent format and (2) the output from the topic modeling tool. To reduce response time, users are expected to generate the topic modeling results themselves by following the provided instructions. Figure 18 shows the initial page of the UI.

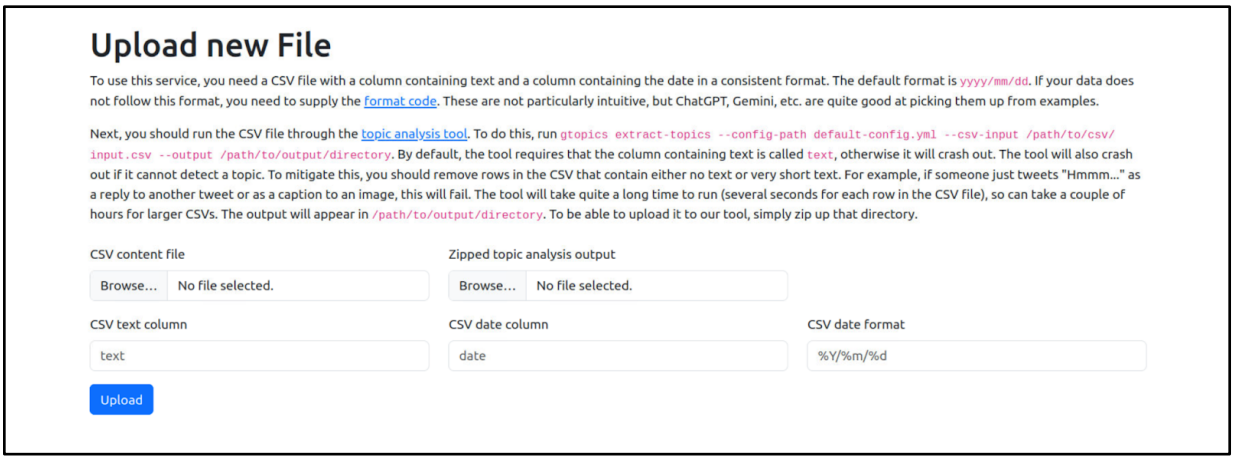


Figure 18 Screenshot of the initial page of the visualization tool. It allows users to upload a CSV file with the user data, and the output from the topic modelling tool.

We use the pre-trained document pair relationship classifier to predict relationships between input documents and construct graphs of related document groups. Additionally, the topic analysis tool assigns hierarchical topic labels to documents where a topic can be identified. After processing, the UI displays the stream graphs for each top-level topic over time (texts in blue). It shows the number of related documents under each topic and the median date when the cursor hovers over a specific topic. The topics are ranked from top to bottom along the y-axis based on the median time of their associated documents. Each first-level topic is expandable by clicking on the "+" symbols to show all second-level topics under it (texts in orange). Similarly, each second-level topic is expandable if there is any third-level topic under it (texts in green). Figure 19 shows the topic-level stream graphs.

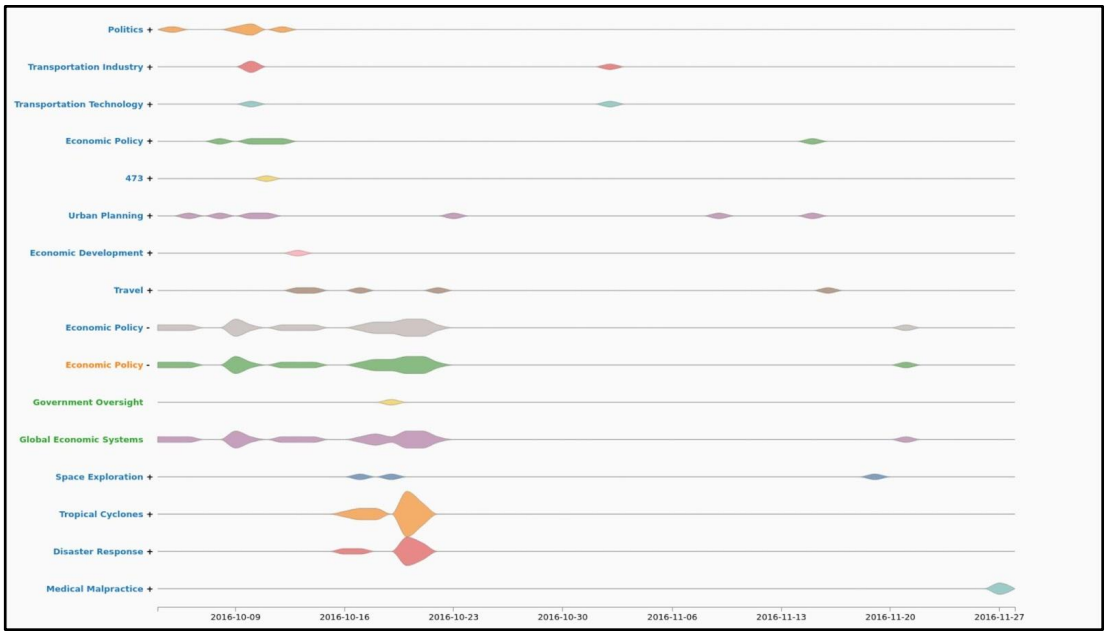


Figure 19 Screenshot of stream graphs of top-level topics after processing users' input data. It displays all expandable ("+") first-level topics (in blue) and the second- (in orange) and third-level topics (in green) under a certain parent topic. The stream graphs indicate the number of documents over time.

When users double-click on a fine-grained topic, they are directed to the document relationship page, where related documents (represented as dots) are shown in the same color. Related documents are connected by arrows, pointing from earlier to later documents. Each document is positioned along the x-axis according to its timestamp, while its position on the y-axis is randomly distributed for visualization purposes. Figure 20 displays the document-level relationship graph.

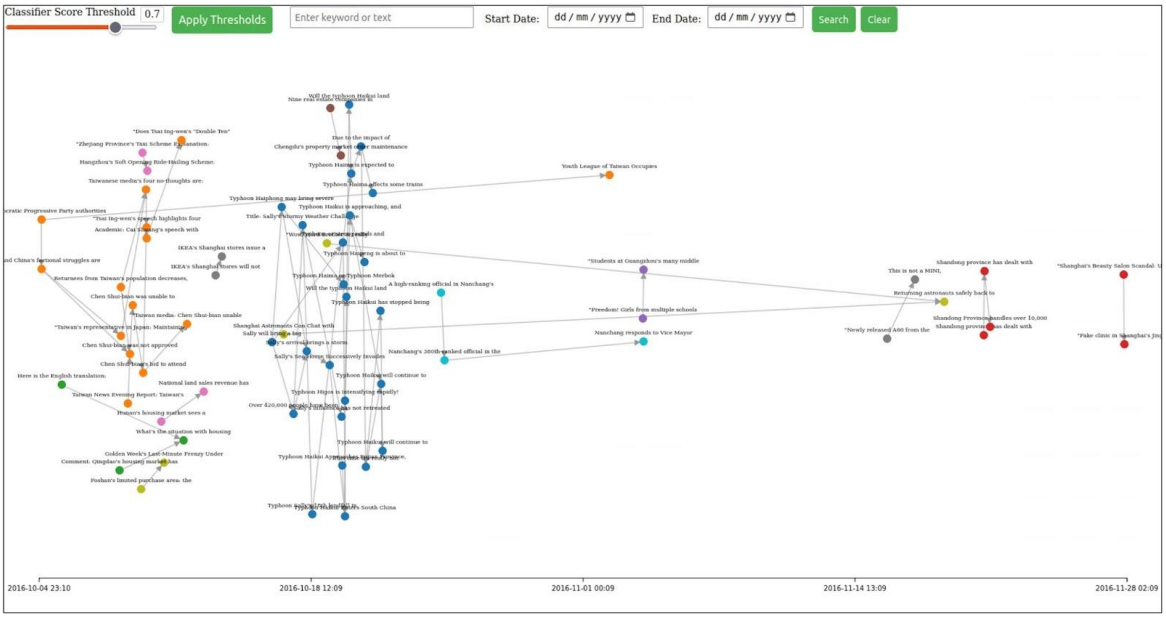


Figure 20 Page of document relationship graphs after users click on a certain topic.

For the document relationship graph, we have the following functionalities:

- Single Click: Clicking on a node displays the document's details—keywords, creation time, and content—on the right side of the screen.
- Double Click: Double-clicking a node reveals the group to which that node belongs. This allows users to explore specific groups in more detail.
- Text Search: Enter a keyword, phrase, or text snippet to find relevant documents. The tool returns the complete group containing the matching document(s), with the matches highlighted (see the following screenshot for text searching results).
- Date Search: Select a start date, an end date, or both to filter documents within a specific time range.
- Threshold Setting: Document relationships are determined based on a predictive score. The initial view uses a default threshold of 0.5. Users can adjust this threshold using the slider in the top left corner—a higher threshold results in fewer groups and fewer documents within each group.

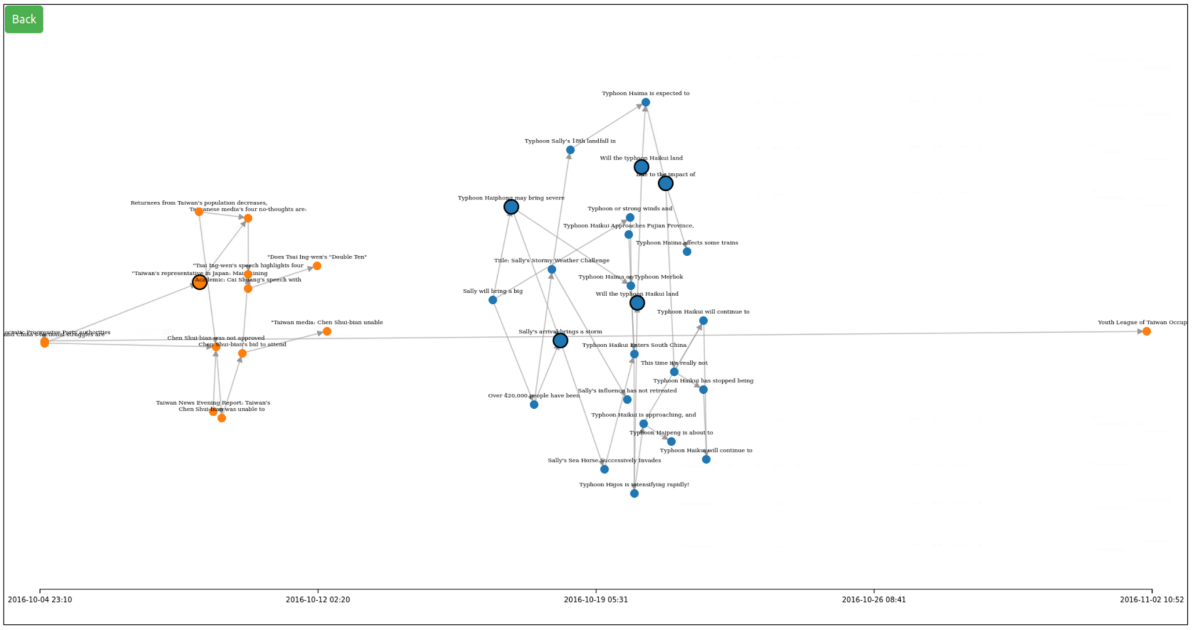


Figure 21 Screenshot of topic searching results with matching documents highlighted.

4.4 Evaluation

The KPIs for the narrative detection with spatio-temporal analysis are user-centric and include a qualitative assessment of (a) user satisfaction with the feedback collection process and the overall number of users involved in both (b) controlled and (c) "in-the-wild" evaluation activities. We evaluate the proposed method using a benchmark test set and by user evaluation feedback for the user interface. Since the topical and temporal analysis has not yet been integrated into the user-facing plugin tool, we are currently reporting only the quantitative performance of the topic modeling and document pair relationship prediction. To calculate the KPIs for the spatio-temporal narrative detection approach, we

will use data collected from the user questionnaire administered during the final evaluation cycle. The detailed evaluation for each component is the following.

Topic modelling

Compared to traditional, probabilistic topic models, LLM-based topic modelling directly generates human-interpretable topic names rather than a list of topic words. Therefore, traditional probabilistic topic modelling evaluation metrics such as topic coherence (Chang et al., 2009; Newman et al., 2010) are not valid for evaluating LLM-based topic modelling. Therefore, we propose a set of automatic evaluation metrics to assess the generated topics in the dimensions of naming adherence, human expectation alignment, and risk of hallucination.

- Metric 1 (M1): Number of Unique Topics is an indicator of the topic naming adherence rate. A better model should produce a smaller number of unique topics because similar yet duplicated topics are merged.
- Metric 2 (M2): Similarity Among Top N Topics is used to evaluate the informativeness of the generated topics. We argue that the generated topics should maximise the semantic differences among each other to convey the most topical information in the documents.
- Metric 3 (M3): Mutual Information (MI) measures the alignment between model-generated topics and human expectations. Specifically, we use MI for quantifying the similarity between topics across two lists: the topics generated by the model and the list of human labelled topics from the datasets.

We compare our proposed method with the off-the-shelf topic model in three standard benchmark datasets with human labelled topics, and demonstrate our proposed method (Topic Mistral) achieved the best results compared with baseline methods (see Table 9, for more results and detail please refer to (Mu, Bai, et al., 2024).

The benchmark dataset includes:

- 20 News Groups[24] (20NG) dataset is an open domain dataset that contains documents spanning 20 categories, which is widely recognised as a standard benchmark for various NLP downstream tasks.
- Wiki dataset (Merity et al., 2017) includes 14k Wikipedia articles that have been annotated into one of 15 general categories.
- Bills dataset comprises summaries of 33k bills from the 110th to the 114th U.S. Congresses, annotated with 21 general topics by humans. We use the processed dataset produced by Pham et al. (2024).

The four baseline methods are the following:

24 https://www.google.com/url?q=http://qwone.com/~jason/20Newsgroups/&sa=D&source=docs&ust=1752223238613182&usg=AOvVaw0tjX97RUy_Uzp-hYaK1dWb

- GPT-3.5 (GPT) is one of the most popular API-based LLM developed by OpenAI, renowned for its strong ability to generate human-like text.
- TopicGPT: Pham et al. (2024) use a postprocessing approach by adding additional prompts on top of the LLM's outputs to merge near-duplicate topics. We adapted their prompts to fit the Mistral-7B model for a fair comparison.
- LDA and BERTopic: For reference, we also compared against two widely-used topic modelling approaches: LDA (Blei et al., 2002) and BERTopic (Grootendorst, 2022). Since the outputs of LDA and BERTopic are unnamed, we employed an LLM (ChatGPT 3.5) to assign names to topics based on a list of representative words for each topic.

Table 9 Topic Evaluation results on 3 benchmark datasets. Three metrics are indicated as M1 (Number of Unique Topics, lower is better), M2 (Similarity Among Top N Topics, lower is better), and M3 (Mutual Information, higher is better). Our model performs the best across 3 benchmark datasets using all metrics except Wikipedia using M3.

	20NG			**Bills**			**Wikipedia**		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
LDA	-	0.160	0.202	-	0.156	0.254	-	0.125	0.284
BERTopic	-	0.171	0.286	-	0.171	0.296	-	0.135	0.288
GPT3.5	3,058	0.178	0.210	2,426	0.210	0.292	2,367	0.134	0.313
TopicGPT	-	-	0.286	-	-	0.315	-	-	0.388
Topic Mistral (Ours)	**373**	**0.112**	**0.449**	**192**	**0.104**	**0.488**	**219**	**0.101**	0.534

Document Pair Relationship Prediction

The evaluation metrics we used for document pair relationship prediction include Homogeneity, Completeness, V-measure score, and Normalized Mutual Information (NMI).

- Homogeneity is an evaluation metric based on conditional entropy that assesses the purity of clusters with respect to ground-truth class labels. Specifically, it measures the extent to which each cluster contains only members of a single class. A high homogeneity score indicates that most or all documents within a cluster share the same label. This aligns with the intuitive expectation that a well-formed cluster should group together similar instances from the same category.
- Completeness measures the extent to which all members of a given class are assigned to the same cluster. That is, completeness is high if all documents with the same true label end up in the same predicted cluster.
- V-measure score is the harmonic mean between homogeneity and completeness.

- NMI measures the amount of statistical information shared by the ground truth cluster labels and the cluster assignment results. The NMI is 1 if the cluster results perfectly match the truth labels, and it is close to 0 when the cluster labels are randomly generated.

Since there is only one publicly available dataset from Liu et al. (2020), in which each document is labeled with a story ID (documents sharing the same story ID indicate they pertain to the same topic), we use this dataset for both training and testing. We evaluate the pre-trained document pair relationship predictor on a subset of 200 documents, as demonstrated in our visualization tool. The results are presented in Table 10 using different predictive score threshold values.

Table 10 Evaluation results with different predictive score threshold values.

Predictive Score Threshold	Homogeneity	Completeness	V-measure	NMI
0.1	0.695	0.860	0.769	0.769
0.2	0.756	0.863	0.806	0.806
0.3	0.779	0.867	0.821	0.821
0.4	0.808	0.868	0.837	0.837
0.5	0.831	0.869	0.849	0.849
0.6	0.841	0.868	0.854	0.854
0.7	0.865	0.869	0.867	0.867
0.8	0.880	0.868	0.874	0.874
0.9	0.925	0.869	0.896	0.896

5 Conclusions

This deliverable documents the comprehensive development and successful implementation of coordinated sharing behavior detection and disinformation campaign discovery and analysis methods within the vera.ai project's Work Package 4. The research has achieved significant methodological breakthroughs while addressing critical challenges in the rapidly evolving landscape of information manipulation and platform governance. In the next subsections we will discuss the main achievements in light of the Tasks 4.2 and 4.3 objectives and KPIs.

5.1 Key Achievements, Validation Protocols, and Methodological Contributions

This deliverable marks a major step forward in coordinated behavior detection research, shifting from reactive, platform-dependent approaches to scalable, platform-independent methodologies suitable for global analysis. Central to this advancement is the development of **CooRTweet**, a universal coordination detection engine that overcomes the constraints of earlier tools tied to specific platform APIs. Performance evaluations against ground truth datasets show 92% accuracy in identifying coordinated networks, establishing new benchmarks and offering greater analytical flexibility across varied social media environments.

Complementing this, the **CIB Detection Tree framework** developed by EU DisinfoLab introduces a structured methodology for assessing Coordinated Inauthentic Behavior (CIB). It evaluates campaigns through four key dimensions: coordination, authenticity, source, and impact. Applied to major investigations—including Operation Overload, Russian TikTok influence operations, and QAnon campaigns—the framework produced quantified CIB likelihood scores ranging from 32% to 64%. These results demonstrate the framework's versatility and its potential to standardize cross-campaign comparisons.

Significant progress was also made in **narrative discovery**, particularly in tracking narratives over time and across linguistic or geographic boundaries. A novel LLM-based topic modeling approach using hierarchical clustering outperformed traditional probabilistic models, achieving a V-measure of 0.896 in document relationship prediction tasks. Simultaneously, the multilingual claim clustering system developed by ONTO processed nearly 200,000 entries from the Database of Known Fakes, uncovering persistent patterns of cross-border disinformation while preserving high precision in claim grouping. Integration of interactive visualization tools and chatbot interfaces ensures these advanced capabilities remain accessible to non-technical users.

The expansion into **multimodal detection** further extends the research frontier. Techniques such as visual similarity analysis, audio provenance detection, and cross-modal validation enable the identification of campaigns leveraging images, audio, and videos to bypass conventional text-based detection systems. This multimodal focus fills a critical gap in existing methodologies and responds directly to the growing sophistication of disinformation tactics.

The methodologies were successfully deployed across multiple operational scenarios. From October 2023 to August 2024, the **VeraAI Alert System** scaled significantly, expanding from an initial 1,225 monitored

accounts to identify over 10,000 coordinated links and 2,126 new accounts across 207 coordinated communities—ultimately encompassing 14,832 Facebook accounts, a 443% increase in scope. The **TikTok Coordinated Detection Project** analyzed over 1.26 million posts in 136 days, detecting 2,521 coordinated posts and mapping 2,248 distinct coordination networks. These deployments confirm the system's adaptability to video-first platforms and varying social media architectures.

Operational findings were equally impactful. Investigations revealed the exploitation of large authentic communities, the emergence of casino engagement bait networks, complex political propaganda infrastructures, and AI-generated disinformation campaigns. One notable case—the detection of a false health narrative involving Pope Francis that generated over 400,000 posts in a single week—illustrates the system's capacity to track large-scale, AI-driven threats in real time.

AI is deeply embedded across the detection pipeline. The integration of **GPT-4o** for automated network labeling, AI-driven content analysis, and machine learning-enhanced pattern recognition significantly reduced analysis time without sacrificing interpretability. Detection of exponential posting surges within gambling promotion networks following the adoption of generative AI provides concrete evidence of AI's growing role in coordination efforts. Additionally, the development of cross-modal validation techniques—capable of detecting inconsistencies between audio and text—introduces a novel strategy for identifying synthetic manipulation.

Finally, user-centric design guided the entire development process. The **CooRTweet** web interface underwent extensive iterative testing with data journalists and OSINT researchers, leading to streamlined workflows and improved usability. Closed beta testing and public usability surveys ensured that even complex detection functionalities remain accessible to practitioners without advanced technical training. New visual assessment methods and interactive network maps help translate analytical findings into intuitive insights, empowering fact-checkers, journalists, and researchers to identify and act on coordinated behavior with confidence.

5.2 Limitations and Future Directions

Despite the considerable progress made, several limitations remain that warrant attention in future research. While platform-independent methodologies have significantly expanded detection capabilities, the effectiveness of these systems is still hindered by evolving platform policies and restrictive API environments. The abrupt deprecation of research-supportive tools underscores the persistent vulnerability of academic research to commercial platform decisions and shifting access protocols.

The project has yielded vital insights into the growing challenges surrounding **platform transparency and research access**. The phase-out of CrowdTangle in August 2024, coupled with a comprehensive evaluation of Meta's replacement tools, revealed deep-seated limitations in current transparency initiatives. Proprietary content ID systems disrupt continuity between research data and live platforms, while the absence of automated monitoring features critically undermines real-time detection capabilities. Furthermore, constraints tied to digital clean room environments prove incompatible with the operational needs of alert systems, and limited search and export functionalities inhibit longitudinal analysis—particularly crucial for tracking evolving disinformation campaigns over time.

These findings underscore the urgent need for robust regulatory intervention. Frameworks like the **Digital Services Act** are essential for safeguarding research access and ensuring meaningful transparency. At the same time, they validate the foresight of investing in platform-independent methodologies, which now serve as a buffer against the erosion of research access rights. The shift from real-time automated monitoring to manual, retrospective analysis reflects a significant loss in the research community's capacity to proactively identify and respond to coordinated information campaigns as they unfold.

Although current implementations have demonstrated scalability and robustness, they may require **architectural refinements** to keep pace with the exponential growth of coordination networks—particularly those amplified by generative AI technologies. While the project has successfully addressed multilingual detection challenges, full coverage of all European languages and culturally specific contexts remains a complex, ongoing endeavor. The rapid advancement of generative AI also demands constant methodological evolution, particularly to detect increasingly subtle and evasive techniques such as the spread of low-quality but high-volume content known as “AI slops.”

Drawing from these insights and operational experience, several strategic recommendations emerge. First, the research affirms the critical need for **comprehensive platform transparency mandates** that balance sustainable research access with user privacy protections. While the Digital Services Act offers a foundational legal framework, effective implementation must directly respond to the technical constraints surfaced by this research—such as the need for automated monitoring, traceable content IDs, and interoperable data environments.

Second, the scale and complexity of contemporary coordination detection efforts call for **collaborative research infrastructures** that integrate academic institutions, civil society organizations, and independent verification bodies. Pooling expertise, access capabilities, and methodological innovations will be essential for maintaining responsiveness in an increasingly complex disinformation ecosystem.

Finally, the continued development of **platform-agnostic detection methodologies** is central to the sustainability and resilience of this field. The research demonstrates the value of hybrid models that integrate automated systems with human expertise—offering a scalable, interpretable, and adaptable framework. These approaches should guide future collaborations between AI-powered detection technologies and professional verification practices, ensuring both precision and accountability in the fight against coordinated disinformation.

5.3 Broader Implications for Information Integrity and Final Reflections

The work documented in this deliverable lays a critical foundation for preserving information integrity in an increasingly complex digital landscape. By demonstrating that coordinated influence operations can be effectively detected, analyzed, and tracked at a global scale, the research offers both a technological breakthrough and a source of optimism for those committed to defending democratic discourse. At the same time, it highlights the growing sophistication and organizational complexity of contemporary threats.

The identification of 207 distinct coordinated communities—each exhibiting systematic operational specialization—reveals the industrial scale and strategic adaptability of modern influence operations.

These are no longer isolated or amateur efforts; they represent highly structured campaigns capable of pursuing diverse objectives across multiple platforms and geopolitical contexts. Addressing this scale and sophistication requires equally coordinated responses from the research community, technology platforms, and regulatory bodies.

The **platform-independent architecture**, **multimodal detection capabilities**, and **robust theoretical frameworks** developed through this research position the *vera.ai* project—and the broader community of disinformation researchers—to meet these evolving challenges. As influence operations increasingly rely on AI-generated content, cross-platform orchestration, and evasive tactics, the methodologies presented here offer foundational capabilities for sustaining the integrity of the digital information ecosystem.

This deliverable goes beyond technical achievement. It introduces essential tools for protecting public trust and safeguarding democratic institutions. The successful integration of automated detection systems with human expertise provides a scalable and interpretable model for identifying coordinated inauthentic behavior, while preserving the nuanced analysis required to understand complex disinformation dynamics.

The methodologies have been thoroughly validated through real-world deployments, peer-reviewed evaluation, and extensive user testing with professional practitioners. Their ability to adapt to emerging threats—while remaining accessible to non-technical users—ensures practical utility across the broader verification ecosystem, from fact-checkers and journalists to OSINT researchers and civil society actors.

As techniques of information manipulation become more advanced—leveraging generative AI and operating across diverse digital infrastructures—the **platform-agnostic, multimodal detection systems** developed here provide a resilient foundation for ongoing defense efforts. This research not only advances methodological innovation but also establishes actionable frameworks for collaborative protection against disinformation.

Crucially, the *vera.ai* project's commitment to **open science**, **transparent methodology**, and **user-centered design** ensures that these innovations contribute meaningfully to the collective mission of preserving democratic resilience in digital environments. By equipping researchers, journalists, and watchdog organizations with sophisticated yet accessible tools, this work helps build the infrastructure needed to uphold information integrity in the 21st century.

Ultimately, the future of information verification does not lie in replacing human expertise with machines, but in **thoughtfully integrating technological capabilities with professional judgment and democratic values**. The methodologies developed through this research provide a blueprint for that integration—ensuring that efforts to defend the public sphere remain both technically robust and fundamentally human in purpose, application, and oversight.

References

Project's outputs cited in the deliverable are reported in **bold**.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet Allocation. In Advances in Neural Information Processing Systems 14 (pp. 601–608). The MIT Press. https://doi.org/10.7551/mitpress/1120.003.0082

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining (pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Cinus, F., Minici, M., Luceri, L., & Ferrara, E. (2025). Exposing cross-platform coordinated inauthentic activity in the run-up to the 2024 U.s. election. Proceedings of the ACM on Web Conference 2025, 541–559. https://doi.org/10.1145/3696410.3714698

DiResta, R., & Goldstein, J. A. (2024). How spammers and scammers leverage AI-generated images on Facebook for audience growth. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-151

Gerhardt, M., Cuccovillo, L., & Aichroth, P. (2023). Advancing audio phylogeny: A neural network approach for transformation detection. 2023 IEEE International Workshop on Information Forensics and Security (WIFS), 1–6. https://doi.org/10.1109/wifs58808.2023.10375058

Gerhardt, M., Cuccovillo, L., & Aichroth, P. (2024). Audio Provenance Analysis in Heterogeneous Media Sets. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4387–4396. https://doi.org/10.1109/CVPRW63382.2024.00442

Giglietto, F. (2024). tiktok_csbn: TikTok coordinated account live map. Github. https://github.com/fabiogiglietto/tiktok_csbn

Giglietto, F., Farci, M., Marino, G., Mottola, S., Radicioni, T., & Terenzi, M. (2022). Mapping Nefarious Social Media Actors to Speed-up Covid-19 Fact-checking. https://doi.org/10.31235/osf.io/6umqs

Giglietto, F., Marino, G., Mincigrucci, R., & Stanziano, A. (2023). A Workflow to Detect, Monitor and Update Lists of Coordinated Social Media Accounts Across Time: The Case of 2022 Italian Election. Social Media + Society.

Giglietto, F., Olaniran, S., Mincigrucci, R., Marino, G., Mottola, S., & Terenzi, M. (2022). Blowing on the Fire: An Analysis of Low Quality and Hyper Partisan News Sources Circulated by Coordinated Link Sharing Networks in Nigeria. https://doi.org/10.2139/ssrn.4162030

Giglietto, F., Righetti, N., & Rossi, L. (2020a). CooRnet. Detect coordinated link sharing behavior on social media. https://ora.uniurb.it/handle/11576/2675493

Giglietto, F., Righetti, N., Rossi, L., & Marino, G. (2020b). It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections. Information, Communication and Society, 23(6), 867–891. https://doi.org/10.1080/1369118X.2020.1739732

Giglietto, F., Terenzi, M., Chakraborty, A., & Marino, G. (forthcoming). Synthetic Seduction: Evolving Visual Persuasion in Coordinated Online Gambling Promotion with Generative AI. In S. Papadopoulos,

K. Bontcheva, V. Mezaris, & R. Rogers (Eds.), Countering Disinformation in the Era of Generative AI. Springer.

Gleicher, N. (2018). Coordinated inauthentic behavior explained. Retrieved August, 19, 2019.

Graham, T., Bruns, A., Zhu, G., & Campbell, R. (2020). Like a virus: The coordinated spread of Coronavirus disinformation. https://eprints.qut.edu.au/202960/

Graham, T., & QUT Digital Observatory. (2020). Coordination Network Toolkit. Queensland University of Technology. https://doi.org/10.25912/RDF_1632782596538

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2203.05794

Gruzd, A., Mai, P., & Soares, F. B. (2022). How coordinated link sharing behavior and partisans' narrative framing fan the spread of COVID-19 misinformation and conspiracy theories. Social Network Analysis and Mining, 12(1), 118. https://doi.org/10.1007/s13278-022-00948-y

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (July-Aug 1998). Support vector machines. IEEE Intelligent Systems and Their Applications, 13(4), 18–28. https://doi.org/10.1109/5254.708428

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2019). Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. Political Communication, 1–25. https://doi.org/10.1080/10584609.2019.1661888

Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2019, October). ViSiL: Fine-grained spatio-temporal video similarity learning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South). https://doi.org/10.1109/iccv.2019.00645

Kordopatis-Zilos, G., Tolias, G., Tzelepis, C., Kompatsiaris, I., Patras, I., & Papadopoulos, S. (2023). Self-supervised video similarity learning. In arXiv [cs.CV]. arXiv. http://arxiv.org/abs/2304.03378

Kordopatis-Zilos, G., Tzelepis, C., Papadopoulos, S., Kompatsiaris, I., & Patras, I. (2022). DnS: Distill-and-select for efficient and accurate video indexing and retrieval. International Journal of Computer Vision, 130(10), 2385–2407. https://doi.org/10.1007/s11263-022-01651-3

Kulichkina, A., Righetti, N., & Waldherr, A. (2024). Protest and repression on social media: Pro-Navalny and pro-government mobilization dynamics and coordination patterns on Russian Twitter. New Media & Society. https://doi.org/10.1177/14614448241254126

Liu, B., Han, F. X., Niu, D., Kong, L., Lai, K., & Xu, Y. (2020). Story Forest. ACM Transactions on Knowledge Discovery from Data, 14(3), 1–28. https://doi.org/10.1145/3377939

Luceri, L., Salkar, T. V., Balasubramanian, A., Pinto, G., Sun, C., & Ferrara, E. (2025). Coordinated inauthentic behavior on TikTok: Challenges and opportunities for detection in a video-first ecosystem. In arXiv [cs.SI]. arXiv. https://doi.org/10.48550/arXiv.2505.10867

Magelinski, T., Ng, L., & Carley, K. (2022). Synchronized action framework for detection of coordination on social media. Journal of Online Trust & Safety, 1(2). https://doi.org/10.54501/jots.v1i2.30

Maksimović, M., Aichroth, P., & Cuccovillo, L. (2021). Detection and localization of partial audio matches

in various application scenarios. Multimedia Tools and Applications, 80(15), 22619-22641.

Marino, G., Almeida Paroni, B., & Giglietto, F. (2025). The Brazilian digital battlefield: Investigating the dynamics of political information campaigns in post-Bolsonaro era. AoIR Selected Papers of Internet Research. https://doi.org/10.5210/spir.v2024i0.13999

Merity, S., Keskar, N. S., & Socher, R. (2017). Regularizing and optimizing LSTM language models. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1708.02182

Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., Nayak, C., Persily, N., State, Bogdan, & Wilkins, A. (2020). Facebook Privacy-Protected Full URLs Data Set [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/TDOAPG

Mu, Y., Bai, P., Bontcheva, K., & Song, X. (2024). Addressing topic granularity and hallucination in large language models for topic modelling. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2405.00611

Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2403.16248

Nimmo, B., Zhang, A., Richard, M., & Hartley, N. (2025). Disrupting malicious uses of AI. OpenAI. https://openai.com/global-affairs/disrupting-malicious-uses-of-ai/

Panizio, E. (2024). Disinformation narratives during the 2023 elections in Europe. EDMO. https://edmo.eu/publications/secondedition-march-2024-disinformation-narratives-during-the-2023-elections-in-europe/

Pearson, G. D. H., Silver, N. A., Robinson, J. Y., Azadi, M., Schillo, B. A., & Kreslake, J. M. (2025). Beyond the margin of error: a systematic and replicable audit of the TikTok research API. Information, Communication and Society, 28(3), 452–470. https://doi.org/10.1080/1369118x.2024.2420032

Pham, C., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2024). TopicGPT: A prompt-based topic modeling framework. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico. https://doi.org/10.18653/v1/2024.naacl-long.164

Righetti, N. (2025). The multiple nuances of online firestorms: The case of a pro-Vietnam attack on the Facebook digital embassy of China in Italy amidst the pandemic. Italian Sociological Review, 15(1), 1–1. https://doi.org/10.13136/ISR.V15I1.850

Righetti, N., & Balluff, P. (2025). CooRTweet: A generalized R software for coordinated network detection. Computational Communication Research, 7(1), 1. https://doi.org/10.5117/ccr2025.1.7.righ

Righetti, N., Giglietto, F., Kakavand, A. E., Kulichkina, A., Marino, G., & Terenzi, M. (2022). Political advertisement and coordinated behavior on social media in the lead-up to the 2021 German federal elections. medienanstalt-nrw.de. https://www.medienanstalt-nrw.de/fileadmin/user_upload/NeueWebsite_0120/Zum_Nachlesen/BTW22_Political_Advertisement.PDF

Righetti, N., Kulichkina, A., Almeida Paroni, B., Cseri, Z. F., Aguirre, S. I., & Maikovska, K. (2025).

Mainstreaming and transnationalization of anti-gender ideas through social media: the case of CitizenGO. Information, Communication and Society, 1–24. https://doi.org/10.1080/1369118x.2025.2470229

Romero-Vicente, A. (2025). Visual assessment of CIB in disinformation campaigns. EU Disinfo Lab. http://disinfo.eu/visual-assessment-of-cib-in-disinformation-campaigns/

Roselle, L., Miskimmon, A., & O'Loughlin, B. (2014). Strategic narrative: A new means to understand soft power. Media War & Conflict, 7(1), 70–84. https://doi.org/10.1177/1750635213516696

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. Proc. ACM Hum.-Comput. Interact., 3(CSCW), 1–26. https://doi.org/10.1145/3359229

Terenzi, M. (2025). Cryptocurrencies from Mainstream to Fringe Platforms. Media Manipulation and Deceptive Schemes on Facebook and Telegram. Comunicazione Politica. https://doi.org/10.3270/116611

Annex I: Complete VeraAI Alert System Network Analysis

Coordinated Facebook Communities Identified (207 Networks, 14,832 Accounts)

This appendix provides a comprehensive listing of all 207 coordinated communities identified through the VeraAI Alert System Facebook network analysis, organized by size and categorized by operational type.

Summary Statistics

- Total Communities Detected: 207
- Total Facebook Accounts: 14,832
- Average Community Size: 72 accounts
- Largest Community: 1,842 accounts (Pro-AMLO political network)
- Smallest Communities: 2 accounts (multiple micro-clusters)
- Network Expansion: 443% beyond core vera.ai monitoring scope

Very Large-Scale Networks (500+ accounts)

Table AI- 1 Major influence operations and state-scale coordination.

Community ID	Size	Network Label
1	1,842	Pro-AMLO, Morena, and Fourth Transformation (4T) supporter groups
45	1,665	African entertainment, news, community groups, music, religion, and marketplace
2	1,133	Brazilian political groups: pro-Bolsonaro, pro-Lula, left vs right
6	889	Pro-Gustavo Petro, Colombia Humana, Pacto Histórico, anti-Uribe activism
8	804	Argentine Kirchnerist and Peronist political support and discussion groups
100	795	Online casino freeplay groups featuring Juwa, Orion Stars, Fire Kirin
21	728	Latin American interest groups, religious devotion, celebrity fans, local classifieds
40	615	African Christian ministries, gospel celebrities, fashion, and entertainment fan groups
22	609	Celebrity and TV Show Fan Groups, Housing, Pets, and Faith

Large Networks (100-499 accounts)

Table AI- 2 Major influence operations and thematic amplification groups.

Community ID	Size	Network Label
19	474	Peruvian left-wing politics, anti-Fujimorismo, regional news, employment groups
13	437	Ukrainian news, patriotism, diaspora, and job opportunities in Europe
15	384	Rwanda and Burundi music, media, sports, and fan communities
4	345	Southern Africa and Kenya social groups, jobs, music, jokes, love
3	338	Buy-sell groups, celebrity fan pages, local classifieds, and communities
14	270	Indonesian home design, online shopping, recipes, Islamic inspiration, social aid
11	231	Liberal and anti-conservative political groups in US and Poland
41	220	Indonesian celebrity fans, religious groups, giveaways, business and kuliner
5	213	Czech and Slovak nationalist, anti-EU, pro-Russia, anti-establishment groups
60	208	Local news, classifieds, fan groups, and marketplace communities Latin America
42	197	Sabah and Malaysian local community, marketplace, and news groups
168	176	Hausa language entertainment, religious, and celebrity fan groups and news
7	136	Pro-Trump, Conservative, Anti-Biden, Republican Support and Meme Groups
120	107	Brazilian online casino and betting platforms with signup bonuses

Medium Networks (50-99 accounts)

Table AI- 3 Tactical coordination units and specialized operations

Community ID	Size	Network Label
47	99	Used motorcycle sales groups in Bogor, Bekasi, Cikarang areas
30	82	Japanese business networking, side jobs, and nationalist political groups
12	70	Diverse interest groups_ religion, business, fandoms, sales, and local news
32	70	Pro-AfD, German nationalist, anti-government, and anti-Green political groups

Community ID	Size	Network Label
147	66	Buy and sell groups Palu and Central Sulawesi region
69	65	Yamaha two-stroke motorcycle sales and communities in Indonesia
116	65	Myanmar online shopping, jobs, literature, and Thailand migrant support
46	61	Samarinda and Kutai community groups and local marketplace networks
63	57	Gorontalo regional news, marketplace, and community groups network

Small Networks (20-49 accounts)

Table AI- 4 Specialized coordination clusters.

Community ID	Size	Network Label
31	49	Armenian news, politics, diaspora, and community information network
161	48	Buy and sell groups for Indonesian cities and regions
114	43	Ehsaas Program, Imran Khan Fans, Pakistani Regional Groups
35	41	Bitung news and online marketplace groups
87	40	Real estate and local trading groups in Myanmar
172	37	Crypto airdrops, mining, and Indonesian cryptocurrency communities
82	36	French Yellow Vests and Insoumise Political Activism Network
56	35	Timor-Leste online marketplaces, job listings, and media pages
107	34	Anti-BJP, AAP and Opposition Supporters, Political Activists, Journalists Fans
140	34	Burmese news and media community groups and pages
193	34	Indian National Congress and Rahul Gandhi supporter groups
17	33	Used motorcycle and auto parts trading groups in Central Java
39	33	Norwegian nationalist and anti-globalist political activism groups
115	32	Spanish right-wing political discussion and anti-government activism
55	31	Buy and sell groups for Solo Raya and Central Java
34	30	Cikampek area buy, sell, and food marketplace groups

Community ID	Size	Network Label
16	28	Local community groups and marketplaces in East Java, Indonesia
79	26	Lost and found pets groups in Argentina
178	24	SportyBet betting tips and groups, Ghana-focused
57	23	Pro-NUG Myanmar news and resistance support groups
84	21	Pro-Putin and BRICS Support Network

Micro Networks (10-19 accounts)

Table AI- 5 High-synchronization clusters and specialized units

Community ID	Size	Network Label
207	18	West Kalimantan local information and buy-sell groups
10	17	Swedish nationalist and anti-Islam political activism network
9	16	Epoch Times and NTD media network accounts
25	15	Caldas Novas local news, classifieds, and community promotion accounts
75	15	Used motorcycle sales groups in Kediri, Indonesia
171	15	English football clubs and players fan pages and updates
59	14	Polish political activism and Kukiz'15 supporters network
92	14	Santa Cruz Bolivia classifieds, jobs, real estate, and autos network
174	14	Italian political leaders and local Padova community groups
18	13	Pro-Bukele and Salvadoran Nationalist Political Groups
110	13	Italian conspiracy and activism discussion groups and public figures
190	13	Local Buy_Sell, Lost Pets, and Community Groups Mexico
77	12	Shan State news, groups, and organizations (RCSS, SSPP, SSA)
144	12	Philippines Buy and Sell, Job Hiring, and Classifieds Groups
26	11	Neighborhood classifieds and commerce groups in South São Paulo
97	11	Yellow Vests and French Protest Movement Accounts

Community ID	Size	Network Label
121	10	Pro-Gbagbo and Pan-African Political Advocacy and News Accounts
167	10	Tanzanian and Kenyan entertainment, memes, and university communities
191	10	Bali Online Marketplaces and Buy-Sell Groups

Nano-Clusters (3-9 accounts)

Table AI- 6 High-synchronization clusters and specialized units.

Community ID	Size	Network Label
62	9	Online Slot Gambling Information and Sharing
94	9	Central Java vehicle and marketplace buy-sell groups
58	8	Pro-Trump and Conservative News Pages by Western Journal
128	8	Devotees of Habib Umar and Islamic Scholars
158	8	Pakistani celebrity and fan groups network
98	7	Italian political and media discussion groups
141	7	Islamic faith, education, and online commerce in Indonesia
143	7	Nigeria News and Breaking News Pages
148	7	Indonesian Islamic Preachers and Religious Content
177	7	East Nusa Tenggara Regional Community and News Groups
101	6	Pro-Trump and Conservative Political Groups and Fan Pages
118	6	Online Slot Gambling Promotions and Freebet Sharing
169	6	Cat and Dog Enthusiast Pages and Videos
179	6	News, greetings, education, and regional updates in Russian and Ukrainian
198	6	Argentina rural lifestyle and local buy-sell-trade groups
43	5	Serbian agriculture and protest groups with Russian news content
66	5	Slovak Politics and Public Figures Discussion Network
96	5	Local Indonesian Marketplace and Business Information Accounts

Community ID	Size	Network Label
117	5	Motorcycle and electronics trading and repair Myanmar network
137	5	North Nias regional news and community information
150	5	Czech political discussion and satire groups
188	5	Love, heartbreak, and youth-themed emotional expression accounts
23	4	Bulukumba commerce and job opportunities network
36	4	German left-wing political discussion and anti-AfD pages
48	4	Local Republican Party and County Political Groups
65	4	Peruvian News, Politics, and Local Community Groups
67	4	Parung-Ciseeng Area Mobile Phone Buy and Sell Groups
68	4	US Conservative Politics and Kansas City Chiefs Fans
71	4	Yellow Vests Movement Groups in France
72	4	Buleleng Online Marketplaces and Local Commerce Accounts
78	4	Pro-Milei Argentine Libertarian Political Network
91	4	Brazilian local news and blogs network
104	4	Hairstyles, Tattoos, and Daily Quotes Content Network
108	4	Ukrainian Community, Health Tips, and Marketplace Accounts
125	4	Myeik Region News and Betel Nut Trading Accounts
132	4	Pro-Colombia Peace and Political News Support Network
163	4	Nigerian Celebrity and Entertainer Fan Clubs
170	4	Wattpad stories, memes, and badminton interest network
49	3	Pro-Sandinista and leftist Nicaraguan political advocacy accounts
54	3	Ukrainian Greetings and Well-Wishes Groups
61	3	Jujuy Local Marketplace and Used Cars Network
64	3	Brazilian military and pro-Bolsonaro political advocacy network

Community ID	Size	Network Label
70	3	Supporters of Alexis Tsipras and SYRIZA Party
73	3	Marisa Pohuwato local news and marketplace portals
81	3	Christian Religious Inspiration and Prayer Pages
95	3	Love Poems and Gaming News Accounts
119	3	Brazilian Christian and religious-themed accounts
122	3	Yellow Vests movement supporters and activism in Lyon
129	3	Polish Agriculture and Farming Accounts
138	3	Gbagyi Community and Political News Network
164	3	South China Sea and Philippines News Network
181	3	Elon Musk and SpaceX Spanish Fan Pages
199	3	Pro-Trump and Anti-Biden Political Supporters
206	3	Good morning and blessing quotes and images

Dyads and Triads (2-3 accounts)

Table AI- 7 Tight coordination dyads and triads

Community ID	**Size**	**Network Label**
20	2	Brazilian sports and history enthusiasts
24	2	Uruguayan politics and economy discussion pages
27	2	Colombian Marketplace and Police Criticism Accounts
28	2	French COVID-19 Protest and Freedom Discussion Accounts
29	2	Pro-Ted Cruz and Conservative Political Content
33	2	German-language knowledge and information sharing accounts
37	2	Caldas Novas and Goiás Local News and Community
38	2	Cat Rescue and Community Cat Organizations
44	2	Slovak Public Affairs News and Discussion

Community ID	Size	Network Label
50	2	Polish Patriotism and National Identity Accounts
51	2	Chalco community news, politics, sports, and disability topics
52	2	Serbian Opposition and Anti-Government Activism
53	2	Ukrainian Communities in Italian Cities
74	2	Brigantine New Jersey local community groups
76	2	Pro-Bolsonaro Brazilian Political Advocacy Pages
80	2	Carapungo-Calderón Local Sales and Community Network
83	2	Norwegian pro-car and climate skepticism network
85	2	Polish anti-geoengineering and conspiracy theory network
86	2	Ukrainian-themed accounts referencing Zelensky and Odnoklassniki
88	2	Swedish News and Media Outlets
89	2	Atibaia community and local interest pages
90	2	Papua New Guinea-China news and friendship network
93	2	Buy and Sell Groups Indonesia
99	2	Parigi City Local News and Community Updates
102	2	Traditional Catholicism and Latin Mass Advocacy
103	2	Norwegian political discussion and activism accounts
105	2	Accounts referencing Acayucan and Juan Diaz Covarrubias locations
106	2	Mendoza Local Classifieds and Community Groups
109	2	Spanish News and Public Opinion Pages
111	2	French leftist activist and workers' debate groups
112	2	Danish Anti-EU and World Economic Forum Transparency Network
113	2	Brazilian Army and Military Veterans Community
123	2	Anti-wind power and climate skepticism Sweden network

Community ID	Size	Network Label
124	2	Odessa and Bilhorod-Dnistrovskyi local news and community updates
126	2	Norwegian Economy and Anti-Trade Agreement Activism
127	2	Croatian Family and Homeland Interest Groups
130	2	Mexican Regional News and Entertainment Accounts
131	2	Ukrainian diaspora communities in Europe
133	2	Neighborhood and Community Groups in Spanish-Speaking Areas
134	2	Argentinian Libertarian and Javier Milei Supporters
135	2	Massachusetts Local News and Political Groups
136	2	Spanish-Language News and Personal Profile Accounts
139	2	Delaware Republican political news and advocacy accounts
142	2	Norwegian media critics and alternative news supporters
145	2	Russian-speaking community in Korea, job postings and assistance
146	2	Catholic Church Traditionalism and Ministry Network
149	2	Nezahualcóyotl local community and lifestyle pages
151	2	Angola Photography and Culture Pages
152	2	Pro-Trump Supporter Groups in US States
153	2	Polish Media Access and Nationwide Announcements
154	2	Swedish Nationalist and Anti-Immigration Advocacy Accounts
155	2	Hindi News and Bahujan Unity Discussion Accounts
156	2	Local news and community groups in Apaseo El Grande
157	2	Local Mexican Community and News Pages
159	2	Lincoln community and local heroes network
160	2	Indonesian Islamic Religious Community and Scholar Followers
162	2	Polish pro-democracy and media supporter groups

Community ID	Size	Network Label
165	2	Brazilian regional news and community organizations
166	2	Islamic News and Religious Teachings
173	2	Supporters of Mahabodhi Monastery Buddhist Monk
175	2	Christian Prayer and Saint Devotion Groups
176	2	Pro-Trump and Anti-Biden Political Groups
180	2	Anti-Trump Political Activism Accounts
182	2	Peruvian political news and Phillips Butters support network
183	2	German Regional Political Activism Accounts
184	2	Pro-Putin and Russia Supporter Network
185	2	Cikampek Local News and Community Updates
186	2	Polish Unity and Civic Movement Network
187	2	Brazilian Northeastern Regional and Leftist Political Pages
189	2	Local News and Community Pages in Northern Mexico
192	2	Puebla Local News and Community Updates
194	2	US Conservative and Christian Activism Pages
195	2	Guerrero News and Local Updates
196	2	Anti-Trump and Pro-Democrat Political Advocacy
197	2	Leslar fan and entertainment community
200	2	Cat-themed community and fan pages
201	2	California Political Advocacy and Election Support Accounts
202	2	Local Property Rentals and Services in Argentina
203	2	Autism Awareness and Acceptance Advocacy Network
204	2	Pro-Trump 2024 American Political Supporters
205	2	Cat-themed religious humor and parody accounts

Annex II: Expert Opinion on the Adaptation of the Coordinated Account Detection Workflow to Current Meta Data Access Tools

This annex provides my[25] expert assessment regarding the adaptation of the methodology described in the paper "A Workflow to Detect, Monitor, and Update Lists of Coordinated Social Media Accounts Across Time: The Case of the 2022 Italian Election" to the current data access environment provided by Meta through the Meta Content Library (MCL) & API.

Executive Summary

The 9-step workflow we developed detects, monitors, and updates lists of coordinated social media accounts spreading problematic content. Our methodology leverages CrowdTangle's API features—specifically its real-time post search, overperforming score metrics, and public API with key authentication—to enable scheduled monitoring and automated alerts. With CrowdTangle's discontinuation, significant adaptations would be required to implement this methodology using the Meta Content Library.

Based on my assessment, while some aspects of the workflow could potentially be adapted to the current environment, several critical steps face substantial limitations or cannot be replicated using the Meta Content Library interfaces. These limitations primarily stem from:

- Proprietary ID System Disconnect: MCL uses an internal ID system that cannot be used on Meta platforms, creating a critical disconnect between research data and live content that prevents tracking and cross-referencing of coordinated networks.
- Absence of Automated Monitoring: The workflow critically depends on scheduled, automated monitoring processes that are not supported in either Meta's Research Platform or the SOMAR VDE.
- Digital Clean Room Isolation: The clean room approach prevents creating a real-time alert system, which was essential to the original workflow for timely notification to fact-checkers.

Meta Content Library and API Overview

Before assessing each step of our workflow, it is important to understand the current tools provided by Meta that would replace CrowdTangle in our methodology:

Meta Content Library: a web-based tool that allows researchers to explore and understand data across Facebook, Instagram and Threads by offering a comprehensive visual and searchable collection of publicly accessible content. It provides a user interface to browse and search public content.

Meta Content Library API: Enables querying and analyzing Meta's full public content archive. This API is available within Meta's Researcher Platform and approved third-party cleanrooms.

[25] This expert opinion was authored by Fabio Giglietto.

Both Content Library and API are controlled-access environments. The functionality to download the public data subset is only available in the Content Library user interface (not in the API), and only if the Inter-university Consortium for Political and Social Research (ICPSR) has approved the application and appropriate contract terms have been agreed upon.

Programmatic analysis of public content can be performed using the Content Library API in the Researcher Platform or in an approved third-party cleanroom environment. At the time of writing, the SOMAR Virtual Data Enclave (VDE) at the University of Michigan is the only approved third-party cleanroom environment available. Each researcher is limited to 60 searches per minute and 500,000 total search results per 7-day rolling window, with an additional limitation of 1,000 multimedia query results in the same period.

Given these premises, the workflow could theoretically be implemented using one of two combinations:

- Meta Content Library + API accessed through Meta's Researcher Platform digital cleanroom
- Meta Content Library + API accessed through the SOMAR VDE digital cleanroom

Meta's Researcher Platform operates within a secured environment, running a modified version of Jupyter inside an Amazon WorkSpaces Secure Browser instance. The only destination enabled in the secure browser is the website hosting the Jupyter instance. Users are permitted to copy and paste text from their local computers into the digital cleanroom, but not in the reverse direction. Although researchers can install packages from official repositories (CRAN for R and PyPI for Python), packages dependent on internet access will not operate. The Research Platform automatically deletes all Meta Content Library API research output data and generated local files every 30 days during a 24-hour maintenance window on the first day of each month. While notebook files and input cells are preserved, all output cells, non-notebook files, and query results are removed, requiring researchers to rerun queries and regenerate outputs monthly.

Meta's Researcher Platform offers both CPU and GPU servers. When logging in, you can select either server type and switch between them as needed. GPU servers support specialized packages like TensorFlow. Researchers can download pre-trained models from Hugging Face for various analytical tasks. Available models include Facebook's NLLB (200-3.3B and 200-distilled-600M), Google's T5 (standard and small), Google BERT Base Model, and Facebook mBART for multilingual translation. The features and limitations of Meta's Researcher Platform are comprehensively documented in the public resources available at https://developers.facebook.com/docs/researcher-platform.

The SOMAR VDE was designed to allow researchers to access highly sensitive data remotely, eliminating the need to travel to a physical clean room. According to the official documentation, programmatic analysis of multimedia content (photos and videos) attached to posts is only available "when using the Content Library API in an approved third-party cleanroom that supports the specific functionality" such as the SOMAR VDE.

To access the Meta Content Library API in this environment, approved researchers must:

1. Create an account in the University of Michigan system
2. Enable two-factor authentication
3. Download and install VPN software
4. Connect to a Windows virtual machine (VM) via remote desktop through the VPN

5. Launch a Linux VM from within the Windows VM
6. Connect to the Linux VM via remote desktop from the Windows VM

The Linux VM operates in a completely isolated environment with no internet connectivity, and unlike the Research Platform which allows one-way copying (from local computer to cleanroom), the SOMAR VDE blocks all copy/paste functionality in both directions—users cannot transfer text either to or from the VM. Multiple researchers from the same team share this VM, which comes pre-configured with both Python and R development environments. While researchers can install packages from official repositories (CRAN for R and PyPI for Python), any packages requiring internet connectivity will not function.

The SOMAR VDE requires users to manually delete all content associated with invalid Meta Content Library IDs every 180 days and certify this deletion through a formal process, even if they had no invalid data to delete. Users receive a list of invalid IDs (content removed from Meta platforms for various reasons) and must complete the deletion within 30 days or have their API access suspended until they comply.

Features, limitations, and access protocols described here reflect the state of these systems as of April 2025 and may be subject to change.

The 9-Step Workflow for Detecting Coordinated Accounts

Figure AII- 1 A visualization of the circular process workflow

As detailed in our paper (Giglietto et al., 2023) and illustrated in Figure AII- 1 above, our methodology employs a circular 9-step workflow designed to detect, monitor, and update lists of coordinated social media actors. The coordinated network of Facebook groups analyzed in our paper was discovered through

this workflow. This iterative approach addresses the dynamic nature of coordinated operations on social media. Rather than providing merely a detailed but static picture of account behavior, the workflow tracks operations over time, enabling academic researchers, fact-checkers, and civil society organizations to observe how coordinated networks evolve.

The workflow begins with compiling initial lists of known problematic accounts (Step 1) identified through academic research, verified fact-checking databases, and investigative findings, followed by scheduled monitoring of their posts (Step 2). The workflow then evaluates post performance (Step 3), analyzes content to extract key features (Step 4), and searches for identical or near-duplicate content across platforms (Step 5). Steps 6-7 involve detecting coordination patterns and matching newly discovered accounts with existing ones. The final steps (8-9) automatically integrate frequently surfacing accounts into the monitoring pool and continuously update the list of problematic actors in near-real time.

A critical component of our workflow is the automated alert system that operates between steps 3 and 7. In our original implementation, the system identified and archived the top three over-performing posts and up to ten coordinated over-performing posts. These alerts were automatically sent to fact-checkers through a dedicated Slack bot and RSS feed, providing real-time notification of potentially problematic content that required immediate attention. This feature enabled rapid response to emerging coordinated campaigns during crucial periods such as elections.

This circular design creates a self-updating detection system that remains effective beyond peak activity periods, making it particularly valuable for election monitoring. The workflow was successfully implemented during the 2022 Italian snap election, revealing three distinct types of coordinated information operations: a politically motivated network of 90 M5S-supporting entities that spread hyperpartisan content and fabricated polls to influence voter behavior; a click-economy driven operation of 46 seemingly religious Pages that exposed nearly 800,000 followers to misleading political content through clickbait tactics; and a religiously motivated network of 1,390 public groups that used sophisticated Messenger bots to conduct covert proselytism activities for the Church of Almighty God.

Building on this experience, we expanded the monitoring process globally in the context of the Vera AI Horizon EU project, starting with a seed of 1,225 coordinated and problematic Facebook public accounts (Pages or public groups). These accounts were identified because they repeatedly coordinated to spread at least four of the 36,091 web pages flagged as false by Meta's third-party fact-checkers between 2017 and 2022 (Messing et al., 2020). Between October 2023 and August 2024, this implementation of the workflow (which we called "Vera AI alerts") identified 7,068 coordinated posts, 10,681 coordinated links, and 2,126 new coordinated accounts organized across 17 networks. The operation was discontinued on August 14, 2024, due to the deprecation of CrowdTangle.

We are currently investigating three distinct cases of coordinated information operations that were surfaced by this global deployment of our alert system: a pro-Putin propaganda network, a scheme promoting online gambling, and adult content distribution in poorly moderated groups.

Assessment of Each Workflow Step

Step 1: Compilation

Original approach: Compile initial lists of known problematic social media accounts to be monitored.

Adaptation to MCL + API in Meta Research Platform:

- This step can be completed outside the research platform
- Lists can be imported into the Library via the “Import from CSV” feature using the producer (e.g. account, group, Page) URL

Adaptation to MCL + API in SOMAR VDE:

- This step can be completed outside the VDE
- Lists can be imported into the Library via the “Import from CSV” feature using the producer (e.g. account, group, Page) URL

Limitations for both environments: None. It's worth noting that this import feature converts the externally usable URL (an ID that can be used to reference a specific post on Facebook, Instagram and Threads) to the internal MCL ID. This demonstrates that such conversion between external URLs and internal MCL IDs is technically possible.

Step 2: Monitoring

Original approach: Access the APIs of social media analytics platforms (CrowdTangle) to periodically monitor posts published by the initial list of accounts through a scheduled process (every 6 hours).

Adaptation to MCL + API in Meta Research Platform:

- Import the producer list MCL internal ids in the Research Platform environment using the Share/Create API list ID
- The API allows querying posts created by a producer (or set of producers) in the last six hours
- No native scheduling capabilities exist within the platform as of April 2025 (crontab is not installed)
- User sessions have time limits, preventing continuous monitoring
- Manual execution would be required for each monitoring interval

Adaptation to MCL + API in SOMAR VDE:

- Import the producer list MCL internal ids in the Research Platform environment using the Share/Create API Search ID
- Similar querying capabilities for posts from specific accounts
- As of April 2025, the platform lacks native task scheduling capabilities, with system administrators having restricted access to the crontab utility through permission controls
- SOMAR provides slightly more flexibility for session persistence

Limitations for both environments:

- Neither environment supports the automated, scheduled monitoring that was essential to our original workflow

- A manual real-time or near-real-time monitoring at 6-hour intervals would be impractical.

In CrowdTangle, we used to retrieve the 100 best-performing posts (sorted by over-performing score) created during the last six hours by each producer list. Given the absence of the over-performing score in MCL, it is possible to sort the posts by views instead. However, sorting by views will return posts from the most popular producers in the large majority of cases. The over-performing score enabled a more fair comparison of post performance across the monitored accounts, possibly surfacing well-performing trendy posts from smaller producers. While it is theoretically possible to calculate this score using MCL data, it would require multiple API calls, thus consuming a significant portion of the limited quota available

Query rate limits and volume restrictions apply in both environments.

Step 3: Evaluation

Original approach: Collect and evaluate the early performance of content (actual) based on the historical performance of posts published by monitored accounts (expected), using CrowdTangle's overperforming score. This evaluation powered our alert system, which automatically identified and sent the top three over-performing posts to fact-checkers through a dedicated Slack bot and RSS feed.

Adaptation to MCL + API in Meta Research Platform:

- Engagement metrics are available (reactions, comments, shares)
- View counts are available, providing a potentially more accurate measure of content reach than engagement metrics alone
- Custom calculation would require significant additional data processing
- No way to programmatically export the results of an analysis

Adaptation to MCL + API in SOMAR VDE:

- Engagement metrics are available (reactions, comments, shares)
- View counts are available, providing a potentially more accurate measure of content reach than engagement metrics alone
- Custom calculation would require significant additional data processing
- No way to programmatically export the results of an analysis

Limitations for both environments:

While view counts provide valuable performance data, they lack the comparative benchmarking against historical account performance that the overperforming score provided (see the limitations in step 2)

Need for manual development and validation of custom performance metrics

Historical baseline data collection is hindered by API query quotas and would be labor-intensive in both environments

Creating an alert system is impossible in the Meta Content Library because it is not currently possible (both in SOMAR VDE and Meta Research Platform) to programmatically export the results of an analysis due to the closed nature of the digital clean room approach. Additionally, even if this were possible, communicating to an external user the internal Meta Content Library ID of a post would be ineffective

because fact-checkers would need to see the post in its original context via the URL or the platform-specific ID of the post.

Step 4: Analysis

Original approach: Analyze the content of overperforming posts to extract characterizing features (text, image text, shared links). The goal of this step was to extract features that could be used for platform-wide search in the next step.

Adaptation to MCL + API in Meta Research Platform:

- Basic content features (text, shared links) are accessible
- No programmatic access to the multimedia content (images, reels)
- No native OCR for extracting text from images
- GPU and some pre-trained model available

Adaptation to MCL + API in SOMAR VDE:

- Basic content features (text, shared links) are accessible
- Programmatic access to the multimedia content (images, reels)
- No native OCR for extracting text from images
- May support limited image processing through available packages

Limitations for both environments:

Neither environment offers native image text extraction comparable to CrowdTangle

In the original workflow, relying on image text extraction for image similarity detection was a limited but well-functioning practical workaround that enabled coordination detection without requiring complex image processing

While the GPU servers and pre-trained models could theoretically enable custom OCR implementation, this would require significant development effort and expertise

Custom solutions would be restricted by API query quotas and computational resource limitations

Processing limitations for high volumes of content would severely impact analysis speed and scalability

The inability to extract text from images creates a critical blind spot for detecting coordinated behavior that relies on image-based messaging

The task of creating a custom image analysis pipeline introduces technical complexity and variability in results that would compromise the reliability of the coordination detection system

Step 5: Search

Original approach: Use extracted features to search for recent identical or near-duplicate posts currently circulating on the platforms. This is a critical data gathering step to identify platform-wide content that is identical or nearly identical to posts shared by the monitored accounts.

Adaptation to MCL + API in Meta Research Platform:

- Limited search capabilities for exact text matches
- URL-based searches are supported but constrained by requiring a keyword parameter ("q") and only returning exact URL matches

Adaptation to MCL + API in SOMAR VDE:

- Limited search capabilities for exact text matches
- URL-based searches are supported but constrained by requiring a keyword parameter ("q") and only returning exact URL matches

Limitations for both environments:

Search for sentences beyond five words is not supported, severely limiting the ability to find matching text-based posts

The API only supports searching for one URL at a time and requires a keyword parameter ("q")

Image-text search is limited to approximately the last 180 days and requires specific parameter settings

Search operations would likely hit quota limitations

Step 6: Detecting

Original approach: Run a coordinated detection algorithm on the posts returned by search to identify coordinated posts and signs of coordination among the monitored accounts and potentially additional accounts that joined the coordinated networks. This step also powered our alert system, which automatically identified and sent the top ten over-performing coordinated posts to fact-checkers through a dedicated Slack bot and RSS feed. In the original implementation this step was performed with CooRnet. In the adaptation, we would use the more flexible and platform-agnostic CooRTweet (Righetti & Balluff, 2025).

Adaptation to MCL + API in Meta Research Platform:

- CooRTweet can be installed

Adaptation to MCL + API in SOMAR VDE:

- CooRTweet can be installed

Limitations for both environments: None

Step 7: Matching

Original approach: In the original CrowdTangle-based workflow, this step involved matching accounts showing coordinated behavior with the initial monitoring list and storing information on newly appearing accounts. This was a critical bridge between detection and list updating.

Adaptation to MCL + API in Meta Research Platform:

- Basic matching operations are feasible
- Limited persistent storage capabilities for tracking results

- Results would need to be exported within platform restrictions

Adaptation to MCL + API in SOMAR VDE:

- Similar basic matching capabilities
- Limited persistent storage capabilities for tracking results

Limitations for both environments:

Both environments can technically support matching operations if previous steps can be completed.

The effectiveness is dependent on the success of steps 5 and 6.

Persistent storage of results for iterative analysis presents challenges.

Step 8: Merging

Original approach: Add accounts that surface multiple times to the list of monitored accounts.

Adaptation to MCL + API in Meta Research Platform:

- Manual merging of account lists is possible
- No automated detection of repeated appearances
- Limited data persistence between sessions complicates tracking
- Data retention policies may cause loss of historical detection information

Adaptation to MCL + API in SOMAR VDE:

- Similar manual merging capabilities
- Better data persistence between sessions
- Potentially more sophisticated list management through available tools
- Data retention policies may cause loss of historical detection information

Limitations for both environments:

Without automated, scheduled processes, maintaining and updating account lists is labor-intensive

The iterative nature of our original methodology is significantly compromised

Data retention policies undermine the ability to identify repeatedly coordinating accounts over time

Step 9: Updating

Original approach: Continuously update the list of problematic actors in near-real-time.

Adaptation to MCL + API in Meta Research Platform:

- Data deletion policies may cause loss of historical detection information
- Producer lists can’t be managed programmatically (from the API)

Adaptation to MCL + API in SOMAR VDE:

- Data deletion policies may cause loss of historical detection information

- Producer lists can't be managed programmatically (from the API)

Limitations for both environments:

An API that allows management of producer lists (currently the API is read-only) would enable researchers to programmatically update a producer list with newly discovered accounts. Additionally, this solution would make it possible to monitor the activity of these new accounts in real time through the Meta Content Library's graphical user interface.

Both environments impose strict data deletion requirements that further complicate the workflow. The Research Platform automatically deletes all output data every 30 days, while the SOMAR VDE requires users to manually delete content associated with invalid IDs every 180 days. These policies disrupt the continuity needed for tracking coordinated behaviors over time, as historical detection data must be constantly regenerated or is permanently lost. A potential solution would be to exempt social media platforms from the "right to be forgotten" directive of the GDPR when data is shared for legitimate research purposes. This would reduce the burden of constant data deletion and regeneration, allowing researchers to maintain more consistent datasets for tracking coordinated behavior patterns over time.

Overall Feasibility Assessment

The implementation of this workflow using current MCL tools is not feasible in a way that preserves the original methodology's effectiveness and purpose.

While individual parts of the workflow can be implemented within the current Meta Content Library & API environments and can benefit from improved features such as expanded coverage, the availability of view count metrics and programmatic access to comments (both environments) and multimedia content (only available in the SOMAR VDE), a full adaptation of the original methodology is not currently feasible due to three critical limitations:

Proprietary ID System Disconnect: MCL uses a proprietary ID system that cannot be used on Meta platforms, creating a fundamental disconnect between research data and live content. This prevents seamless tracking and cross-referencing of coordinated networks. Even if analysis results could be exported (which they currently cannot), the internal MCL IDs would be unusable for fact-checkers who need to see posts in their original context through platform-specific URLs.

Practical Example: Maria, a fact-checker monitoring the Polish presidential election, previously received real-time alerts via CrowdTangle, including direct platform links. With MCL, she would only receive internal IDs unusable for direct access to posts, forcing her into impractical manual searches without guaranteed success. This disrupts rapid verification and effective intervention.

Absence of Automated Monitoring Capabilities: The workflow's effectiveness depends critically on scheduled, automated monitoring processes running at regular intervals (every 6 hours in the original implementation). Neither Meta's Research Platform nor the SOMAR VDE provides native scheduling capabilities, and session time limits prevent continuous monitoring. Manual execution at each interval would be impractical and defeat the purpose of the real-time detection system.

Practical Example: Tomasz, monitoring Romanian elections, benefited from continuous automated detection. Now, he must manually initiate monitoring, facing frequent session timeouts and human

resource limitations, creating significant detection blind spots—particularly during critical periods such as nights and weekends.

Digital Clean Room Isolation: The clean room approach fundamentally prevents the creation of a real-time alert system, which was a cornerstone of the original workflow. The closed nature of these environments, operating in isolation from the internet with limited or disabled data export functionality, makes it impossible to share time-sensitive information with fact-checkers or other stakeholders outside the clean room environment.

Practical Example: Ana, working across Poland and Romania, previously received instant notifications through external channels like Slack. Under current MCL conditions, she cannot automate alerts or share findings promptly. Manual extraction from isolated environments significantly delays responses, hindering timely interventions and cross-border coordination.

Conclusion

The current suite of tools available in the Meta Content Library represents a significant step backward compared to CrowdTangle, specifically in terms of effectively monitoring coordinated inauthentic behavior during electoral periods using the established 9-step workflow. Although certain analytical enhancements, such as improved view-count metrics and multimedia capabilities within SOMAR VDE, are beneficial, these improvements do not offset the substantial deficiencies in core workflow functionality.

Looking ahead to upcoming EU elections, such as those in Poland and Romania in 2025, researchers will be unable to effectively deploy a real-time detection system for coordinated behaviors using the current tools and workflow. While a modified approach based on periodic batch analyses may be feasible, this significantly alters the methodology's original intent and effectiveness, particularly regarding early detection and rapid response to emerging coordinated networks.

Additionally, implementing even this limited version would require considerably more time and resources than previously needed with CrowdTangle. Researchers would face extensive manual intervention and the necessity for cumbersome workarounds, ultimately yielding results that are neither timely nor comprehensive. The critical advantage of the original workflow—delivering near-real-time alerts to fact-checkers about coordinated content—cannot currently be replicated.

To effectively monitor coordinated inauthentic behavior during sensitive electoral periods using the 9-step workflow, Meta must address these critical limitations. Essential improvements include enabling cross-system ID referencing, supporting automated and scheduled monitoring processes, and establishing secure yet practical mechanisms for sharing timely findings beyond the clean room environment.

It is necessary to emphasize that the Meta Content Library and its APIs were likely not initially designed with this specific monitoring workflow in mind. Their current limitations in supporting the 9-step workflow do not imply these tools lack utility for other research purposes. I should also recognize the significant investment of resources by both Meta and the SOMAR at the University of Michigan to develop and implement the advanced features that are currently available.

It is important to note that Meta and SOMAR regularly update their tools and services. The evaluation presented here reflects the state of these systems as of April 2025; future updates may resolve some of the issues identified.

Annex III: The Pope case Coordinated Sharing Detection Service (CSDS) Tutorial

In late February 2025, researchers at the University of Urbino uncovered a large-scale disinformation campaign on Facebook surrounding Pope Francis's health—just months before his passing. Over the course of a week, this operation generated more than 400,000 near-identical posts, each proclaiming "Breaking News: His Holiness Pope Francis has failed," often paired with an AI-augmented collage of images designed to evade detection. Thanks to the Vera AI alerts workflow, which continuously tracks and updates lists of coordinated accounts through behavioral and interaction metrics, our team flagged and extracted these posts into a single CSV for analysis. Since that initial discovery, the "Sick Pope" scenario has become our go-to example in multiple CooRTweet demonstrations and training sessions. In this tutorial, we'll guide the users through exactly how to upload, visualize, and explore that very dataset, so you can see firsthand how CooRTweet uncovers the hidden networks behind coordinated disinformation. In the next subsections, we will report the tutorial step-by-step.

Step 1 - Log In and Review the Guide

1. **Open the Guide**
   Go to https://coortweet.lab.atc.gr/.
2. **Check the CSV Template**
   The Guide highlights expected columns (account_id, post_text, timestamp, etc.) and links to a sample CSV. Make sure your cleaned file follows that format.

Step 2 - Upload & Process Your Data

1. **Go to "Graph"**
   Click **Graph** in the left sidebar.
2. **Select and Upload**
   Upload your cleaned dataset. We have chosen popesick.csv[26] for this specific tutorial.
3. **Wait for the Graph**
   Once processing completes, you'll see a network of circled clusters representing accounts that shared "Breaking News: His Holiness Pope Francis has failed."

[26] The dataset is not linked in the deliverable as we can't redistribute the dataset publicly per MCL sharing data agreement.

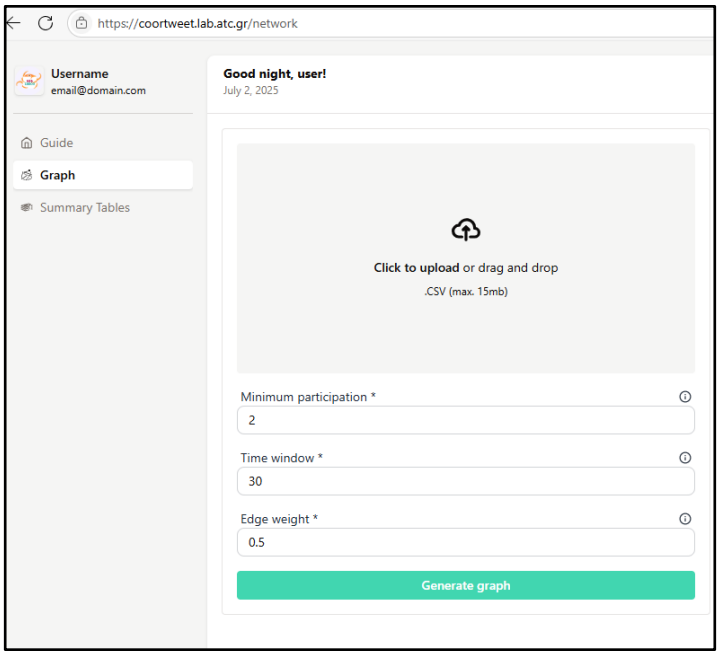


Figure AIII- 1 Snapshot of CSDS website

Step 3 - Explore the Network Visualization

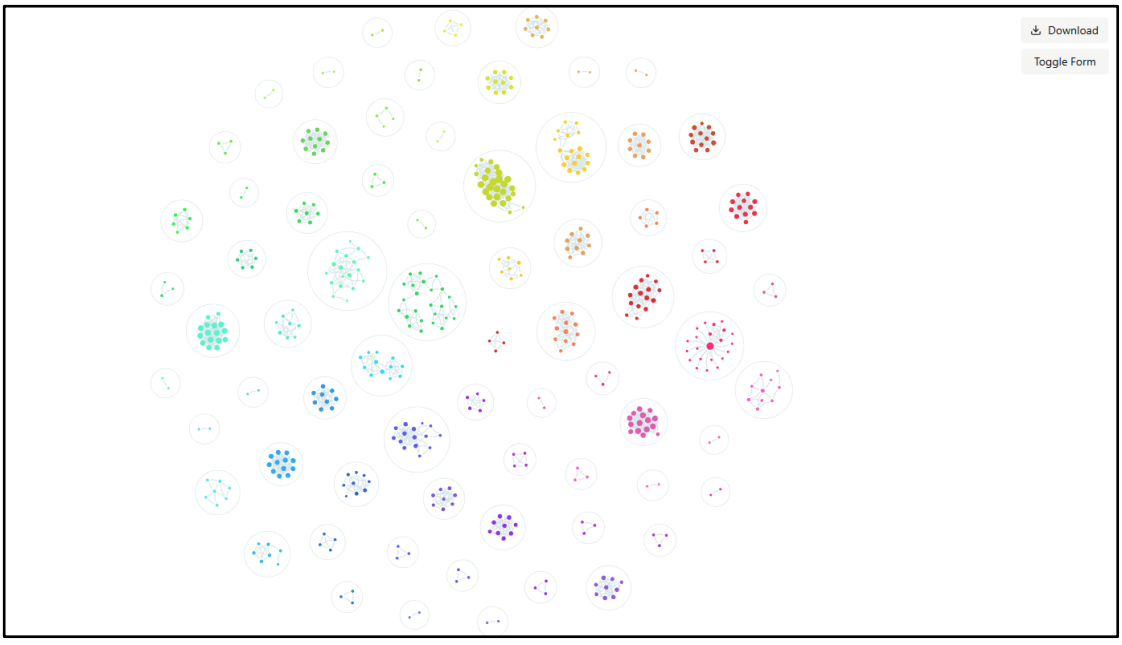


Figure AIII- 2 The Sick Pope coordinated network visualised by CSDS.

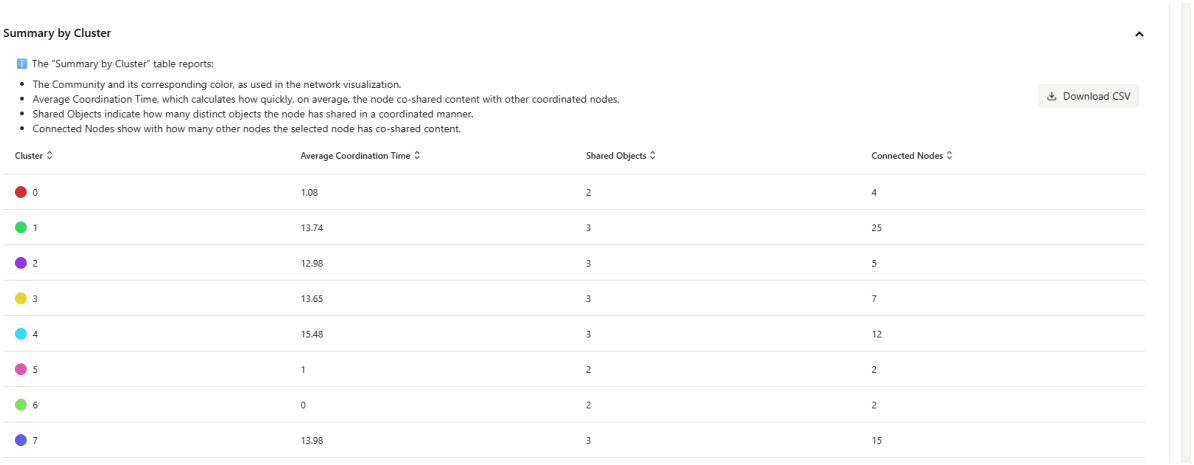
- **Zoom/Pan** with scroll and drag
- **Node Size** ∝ volume of coordinated posts
- **Colour** marks different clusters
- **Download** exports your cleaned data and a PNG snapshot

Step 4 - Navigate Summary Tables

Click **Summary Tables** in the sidebar or go to https://coortweet.lab.atc.gr/summary. You'll find three sections:

Step 4.1 - Summary by Cluster

Lists each detected cluster, its colour, average coordination time, number of shared objects, and connected nodes.

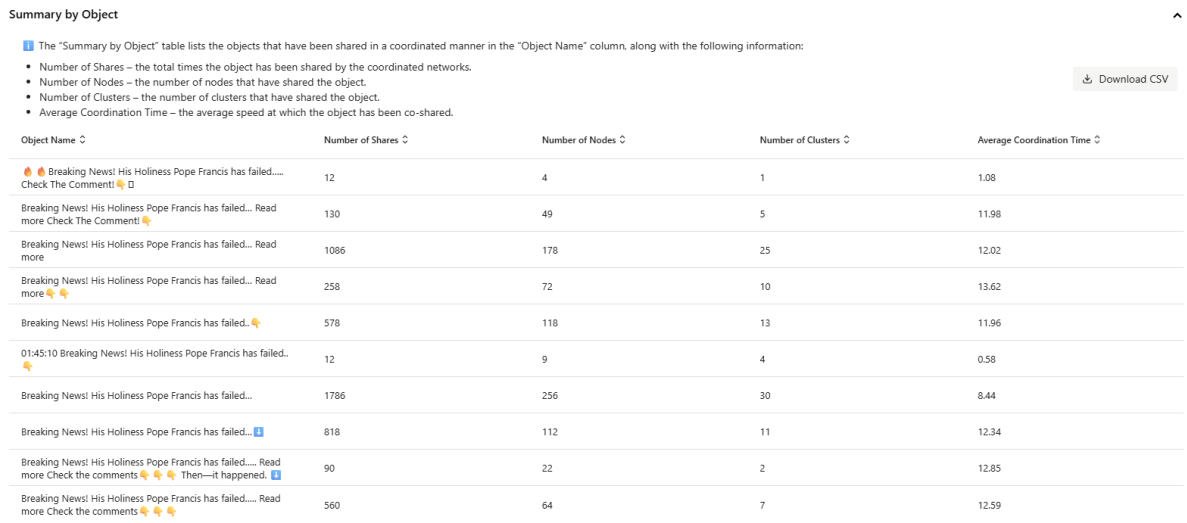


Cluster	Average Coordination Time	Shared Objects	Connected Nodes
0	1.08	2	4
1	13.74	3	25
2	12.98	3	5
3	13.65	3	7
4	15.48	3	12
5	1	2	2
6	0	2	2
7	13.98	3	15

Figure AIII- 3 Summary by cluster.

Step 4.2 - Summary by Object

Shows each unique post text ("object") that circulated in coordination, along with how many times it was shared, how many accounts and clusters participated, and average coordination speed.



Summary by Object

The "Summary by Object" table lists the objects that have been shared in a coordinated manner in the "Object Name" column, along with the following information:

- Number of Shares – the total times the object has been shared by the coordinated networks.
- Number of Nodes – the number of nodes that have shared the object.
- Number of Clusters – the number of clusters that have shared the object.
- Average Coordination Time – the average speed at which the object has been co-shared.

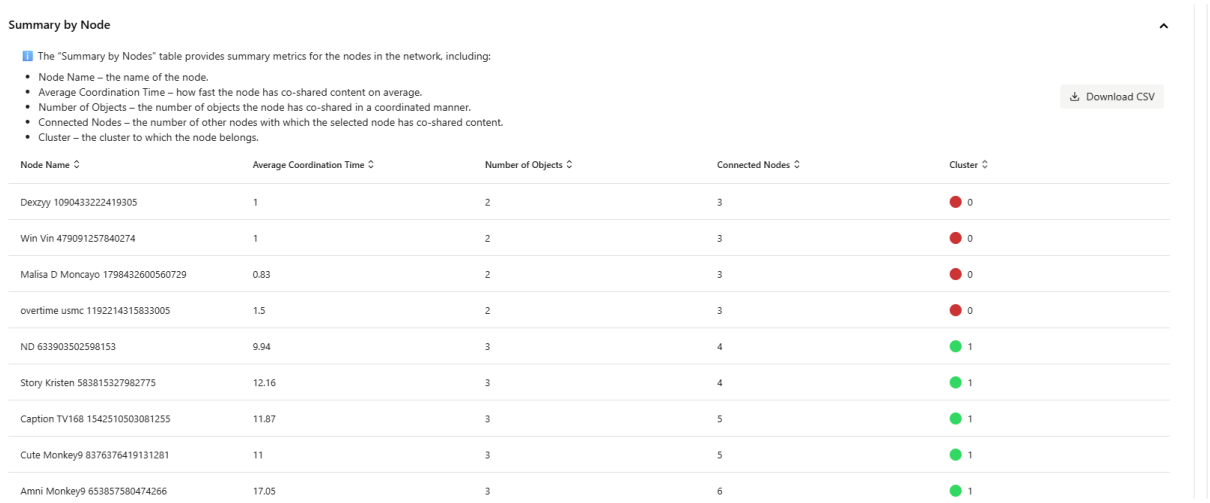
Download CSV

Object Name	Number of Shares	Number of Nodes	Number of Clusters	Average Coordination Time
🔥🔥Breaking News! His Holiness Pope Francis has failed..... Check The Comment!👇◘	12	4	1	1.08
Breaking News! His Holiness Pope Francis has failed... Read more Check The Comment!👇	130	49	5	11.98
Breaking News! His Holiness Pope Francis has failed... Read more	1086	178	25	12.02
Breaking News! His Holiness Pope Francis has failed... Read more👇👇	258	72	10	13.62
Breaking News! His Holiness Pope Francis has failed..👇	578	118	13	11.96
01:45:10 Breaking News! His Holiness Pope Francis has failed..👇	12	9	4	0.58
Breaking News! His Holiness Pope Francis has failed...	1786	256	30	8.44
Breaking News! His Holiness Pope Francis has failed...⬇	818	112	11	12.34
Breaking News! His Holiness Pope Francis has failed..... Read more Check the comments👇👇👇 Then—it happened. ⬇	90	22	2	12.85
Breaking News! His Holiness Pope Francis has failed..... Read more Check the comments👇👇👇	560	64	7	12.59

Figure AIII- 4 Summary by object.

Step 4.3 - Summary by Node

Ranks every account by average coordination time, number of distinct objects shared, connected nodes, and cluster membership.



Summary by Node

The "Summary by Nodes" table provides summary metrics for the nodes in the network, including:

- Node Name – the name of the node.
- Average Coordination Time – how fast the node has co-shared content on average.
- Number of Objects – the number of objects the node has co-shared in a coordinated manner.
- Connected Nodes – the number of other nodes with which the selected node has co-shared content.
- Cluster – the cluster to which the node belongs.

Download CSV

Node Name	Average Coordination Time	Number of Objects	Connected Nodes	Cluster
Dexzyy 1090433222419305	1	2	3	0
Win Vin 479091257840274	1	2	3	0
Malisa D Moncayo 1798432600560729	0.83	2	3	0
overtime usmc 1192214315833005	1.5	2	3	0
ND 633903502598153	9.94	3	4	1
Story Kristen 583815327982775	12.16	3	4	1
Caption TV168 1542510503081255	11.87	3	5	1
Cute Monkey9 8376376419131281	11	3	5	1
Amni Monkey9 653857580474266	17.05	3	6	1

Figure AIII- 5 Summary by node.

Step 5 - Focus on a Node (the Molly Katherine case)

1. **Hover** to see an account's tooltip (ID, share count, first/last seen).

2. **Click** to highlight its edges and circle. This also filters the Summary Tables view.

3. **Click again** to clear the filter.

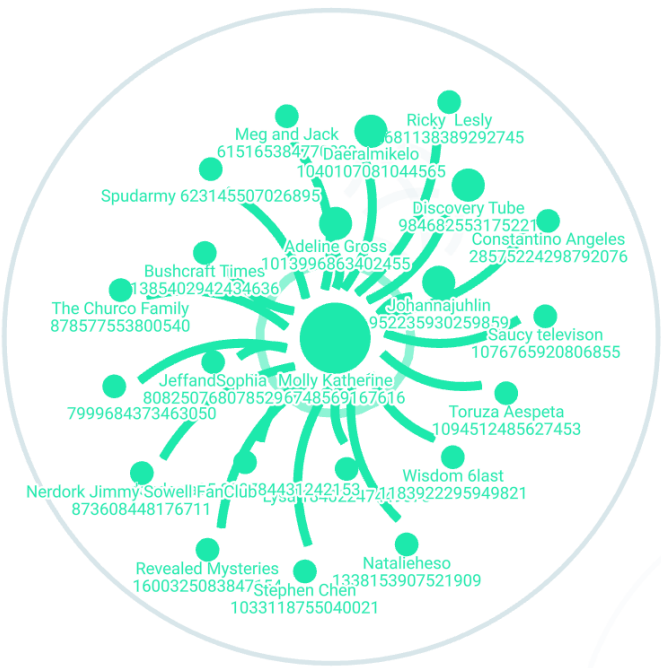


Figure AIII- 6 The Molly Katherine node.

Note: The significance of the Molly Katherine case

The Molly Katherine account is particularly illuminating when the user looks into the Facebook account:

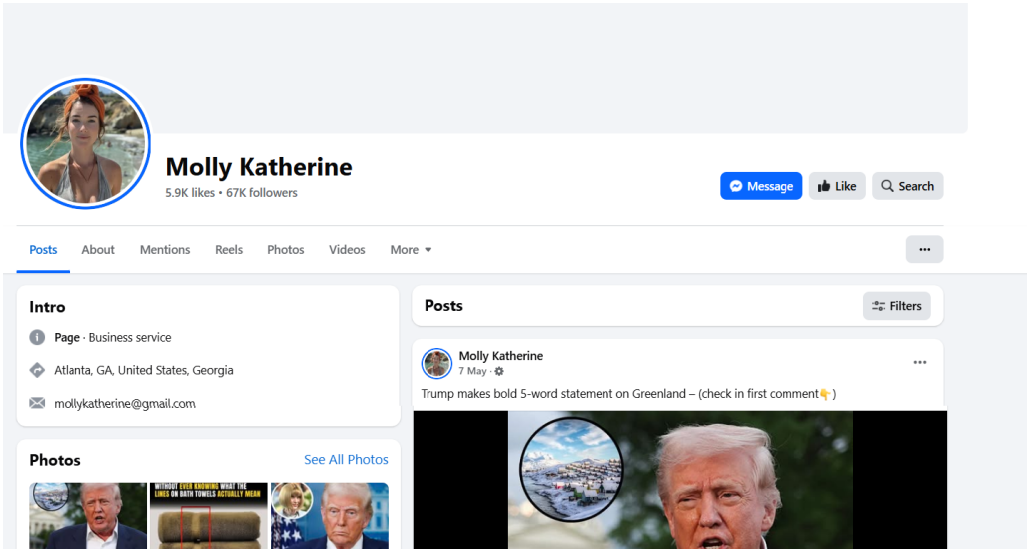


Figure AIII- 7 The Molly Katherine Facebook profile page.

The page is apparently run by a woman named Molly Katherine but the profile picture seems to possess the visual aesthetics of an AI-generated picture. Thus, to investigate further, a user can use the synthetic media detector[27] integrated in the Verification Plugin. Upon uploading the profile picture, the result shows that it is very likely that the image is AI-generated.

Figure AIII- 8 The Molly Katherine Facebook profile photo detected as AI-Generated.

This is also an emblematic example of how different tools developed and updated under the aegis of the vera.ai project can be used in tandem to uncover disinformation operations.

Step 6 - View Filtered Summaries for the Selected Node

When you click a node in the Graph, the Summary Tables page auto-filters to that account:

- **Cluster** section shows only the cluster that contains the selected node.
- **Object** section lists only the post texts that node shared in coordination.
- **Node** section shows metrics just for that single account.

[27] Developed by vera.ai partners as documented in D3.1 Explainable AI methods for analysis and verification of text, audio, image & video misinformation.

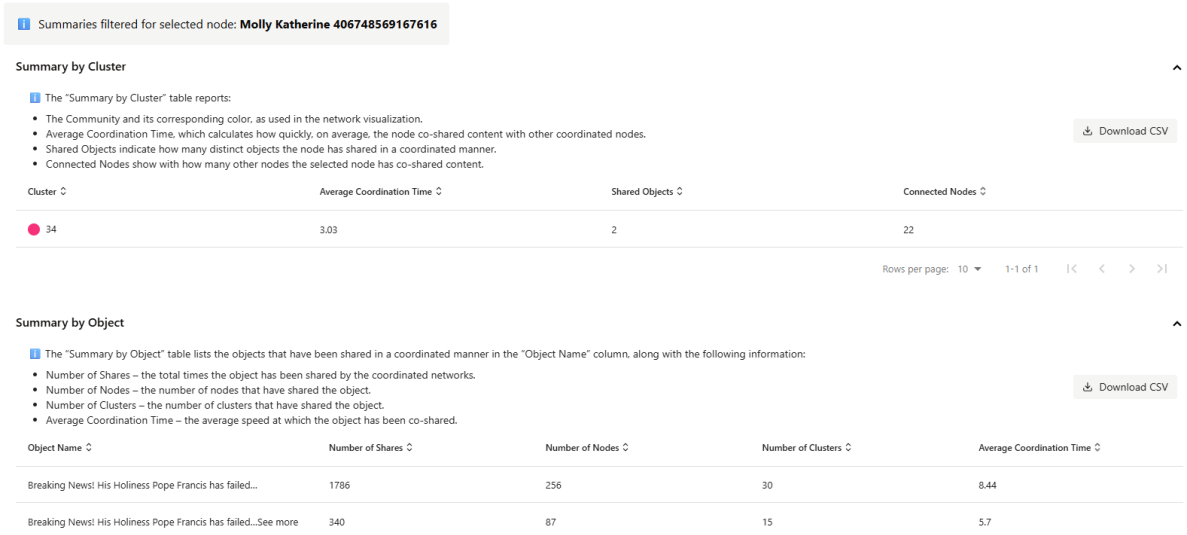


Figure AIII- 9 Summaries for the Molly Katherine node.

Step 7 - Sort, Filter & Export

- **Sort:** Click any column header to re-order.
- **Filter:** Use the small filter icon on each column to search for specific text.
- **Download CSV:** Click the **Download CSV** button at the top right of each summary panel to export the current view.

Step 8 - Next Steps & Best Practices

- **Combine Graph + Tables:** Use graph clicks to drive table filters, then export the exact subset you need for reporting.
- **Iterate Quickly:** Return to Graph, tweak your time/threshold parameters, and immediately see how clusters change.
- **Document Findings:** Capture both network snapshots and table exports to build a clear narrative for your investigation.

vera.ai

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.