Supplementary Material

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

Lucas Stampe[1] and Christian Grimme[2]

[1]Department of Information Systems, University of Münster, Germany; Email: lucas.stampe@uni-muenster.de; ORCID: https://orcid.org/0009-0005-3134-6121

[2]Department of Information Systems, University of Münster, Germany; Email: christian.grimme@uni-muenster.de; ORCID: https://orcid.org/0000-0002-8608-8773

I. Method demonstrator

As described in the paper, the interviews were conducted using a methodological demonstrator in the form of an analysis dashboard, which assists in identifying inauthentic behavior in social media data streams. The four categories identified in our study based on the literature (see Sec. 3) are reflected in the demonstrator, along with a representative method for detecting disinformation. To reflect the expert competencies and approaches documented by Unger et al. (2025), we ensured that each of the six disinformation feature groups from Table 1 is represented by several of the implemented methods. As shown in Table 3, the demonstrator comprises several analysis components that serve as indicators of coordination and automation. These include, among others, visualizations of content clusters over time, a selection of graph-based network analysis approaches, a detection mechanism for artificially generated text in social media posts, and the ability to generate an AI summary of filtered content via a prompt interface.

Table 1. Disinformation features derived from expert interviews (*here:* color-coded to correspond to tables 3 and 4)

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

Table 3. Descriptive components and methods used in the demonstrator to detect disinformation, and their corresponding features of disinformation and disinformation campaigns.

Method-Demonstrator Components		Disinformation Features			Disinformation Campaign Features		
Component group	**Representation in Method-Demonstator**	**Content**	**Language**	**Source**	**Temporal**	**Organizational**	**Dissemination**
Dataset description	Dataset description - source, topic, data collection	-	-	-	-	-	-
	KPIs for filter selection - number of posts, unique users, time frame	-	-	-	-	-	-
	Interactive data table – timestamp, user, text, #,@,www, clusterNo, AI%	-	-	-	-	-	-
Data aggregation	Bar chart - top hashtags, mentions, and hyperlinks	X	X	X		(X)	
Data over time	Line chart - number of posts over time				X	(X)	
	Line chart - top user posting patterns				X	X	
Unsupervised learning	Line chart - stream cluster weights over time				X	X	X
	Table column – investigate cluster, filter data by stream cluster number per post	X	X	X			
Supervised learning	Line chart - GPT-detector probability over time				X	X	X
	Table column – filter by GPT-detector probability per post, investigate	X	X	X			
Network analysis	Network graphs - shared links, hashtags, mentions, separate or in combination	X		X		X	X
	Network graphs - users with shared links, hashtags, mentions, separate or in combination	X		X		X	X
Large language models	Text field and user interface - AI summary based on user-selected columns and text field to prompt	X	X	X	X	X	X

Notes: The bold black lines separate dataset description from the commonly used views of OSN datasets and representative selection of methods from the literature.

Prior to these, several simpler components were added to describe the dataset (for exploration purposes). These are listed under the categories "Dataset description," "Data aggregation," and "Data over time," which include simple statistics, aggregations, and time-series charts. These were included due to their simplicity and frequency of their use in describing OSN datasets, including in related research areas. This allowed respondents to quickly gain an overview of the scenario at hand. The overall layout of the demonstrator is shown in Figure 5.

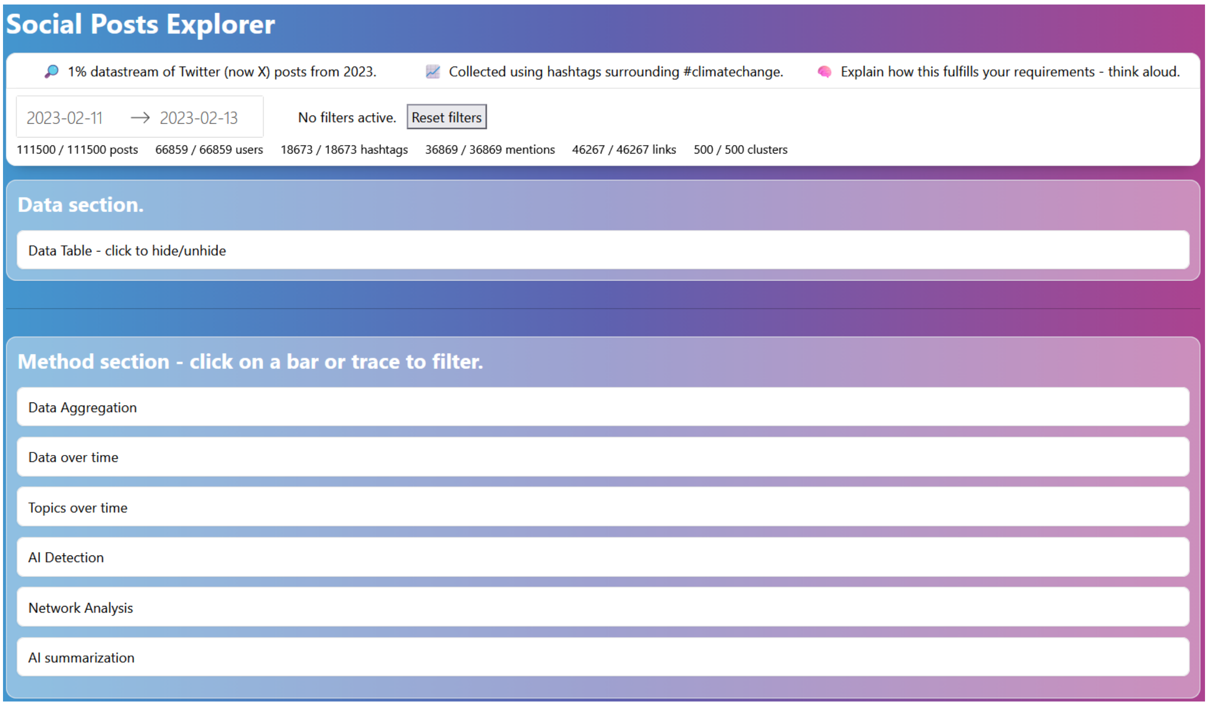

Figure 5. Overview of the method demonstrator. General statistics on the dataset are presented at the top, followed by a data table and a method section for analysis and filtering.

Each analysis component and detection method is visualized in accordance with standard practice in the relevant literature. An example is shown in Figure 6. We used the method demonstrator during our semi-structured expert interviews to gain insights into the experts' problem-solving processes. In a practical test environment, the experts were asked to indicate whether and how the respective approach is used, or could be used, in their work. Based on real data, our prepared scenario naturally contains noise, missing values, and inconsistencies. To ensure relevance, we selected a major social network (X, formerly Twitter) and a widely discussed topic - climate change - for our dataset. The method demonstrator allows the interviewee to freely use each component (including in combination) to filter the data.

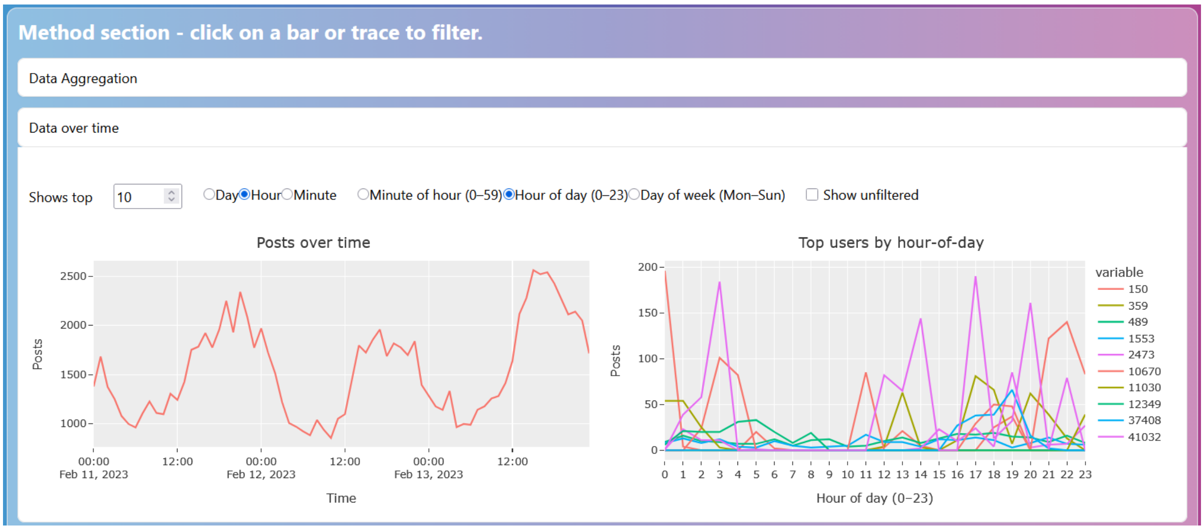

Figure 6. An example component from the method demonstrator. Note the options for aggregation levels via radio buttons. Left: Number of posts over time. Right: Number of posts by top users aggregated by hour.

We explicitly note that some of the analytical components were developed as part of an earlier research project, which accordingly influenced the choice of methods. For each method used here, there are other, context-specific, and potentially more specialized tools, such as Gephi for network analysis (Bastian et al., 2009). However, despite the limited time frame of the interviews, the experts were sufficiently informed about the representative nature of the methods, which was also reflected in their responses. Regarding the influence of interface design on the study, the following argument can be made: According to Nielsen & Landauer (1993), a small sample size of participants (3–7) is sufficient to identify most usability problems in interfaces. Although not the focus of the study, the most pressing usability issues regarding the method demonstrator were likely identified during the interviews and excluded from the coding. To further ensure neutrality, only the data table and aggregated statistics were visible at the start; other visualizations had to be activated via buttons to better reveal the experts' thought processes.

Please refer to our GitHub repository (https://github.com/ItsLukeHimself/OSINT_dashboard) for the code of the method demonstrator. This includes a dummy data generator for the initial setup, as we are unable to share the dataset.

References for I. Method Demonstrator

Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international AAAI conference on web and social media* (Vol. 3, No. 1, pp. 361-362).

Nielsen, J., & Landauer, T. K. (1993, May). A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206-213).

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

II. Interview Guideline

(Translated to English – all interviews were conducted in German)

Opening

- *- introduction of interviewer -*

Introduction to the Interview

- *"The purpose of this interview is to understand the thought processes of experts specifically yours – when dealing with inauthentic behavior in open online media and any tools you may use.*
- *I will start with questions concerning your current work processes as they are right now.*
- *To dive deeper, I will later present you with a concrete case to learn how you would approach it to determine the conceptual fit of representative detection methods from the literature.*
- *Feel free to ask any questions and offer any criticism you might have - we want to learn about the spectrum of tasks, and that includes finding the limits of applicability.*
- *With your permission, I will record this interview. The recording will be saved locally and encrypted. In case we wish to quote you by name, we will approach you separately."*

General Questions

- *"Where do you work and what is your job title?"*
  - *"Can you describe your domain and role?"*
  - *"How much experience do you have?"*
- *"Where in your work do you encounter disinformation and what is your goal in dealing with it?"*
  - *"What features of disinformation do you look for?"*
  - *"What is your current workflow for this?"*
  - *"What is your method of intervention?"*
- *"What OSNs, websites or tools do you use?"*
  - *"What parts of the process are left up to you?"*
  - *"Is this based on expertise, experience, or something similar?"*

Task Introduction and Sub-Tasks

- Scenario: *"You will be presented with a snapshot dataset from an online social network. A selected set of detection methods for identifying disinformation in online social networks has been used to create indicators for disinformation features. These are presented in tables and/or visualizations typically used for these methods."*
- Task: *"The dataset includes inauthentic behavior that can be argued to be an attempt at spreading disinformation. Instead of having to rely on a regular user account and conducting research outside of the online social network, explore the data and method output with the help of the provided demonstrator to arrive at an initial assessment. I will lead you through the method blocks."*
- Process: "*Now for the process: As you perform the task, please think aloud. Where would the methods help? What would need to be different? Verbalize your reasoning, the assumptions you make, and any difficulties you encounter. There are no right or wrong answers since I am specifically interested in* ***your*** *thought process*."

General Prompts

- *"Tell me what you're thinking right now."*
- *"Why did you choose that option?"*
- *"What are you expecting to happen next?"*
- *"What are you noticing about this screen/module?"*

Open-Ended Questions:

- *"Overall, how easy or difficult was it to complete the task using these methods?"*
- *"What were the most intuitive disinformation features?"*
- *"What disinformation features were confusing or frustrating?"*

Targeted Questions (based on observations, referring to specific moments during the TA):

- *"I noticed you hesitated when using method Y. Can you tell me more about what you were thinking at that point?"*
- *“It seemed you found method Z particularly helpful. What made it so effective?”*
    - *“What disinformation feature(s) were you looking for specifically?”*
- *“If you couldn’t use method A, how would you proceed? “*
    - *“If data quality drops, which method would fail first?”*

Round-up

- *“Given what you have told me at the beginning of the interview regarding your own work processes, how could what you have seen today impact your work? Where would you use this? What would need to happen for you to be able to use this?”*

Closing

- *“Would you be available for follow-up questions or another short test with improved features?”*
- *“Is there anyone else you think I should talk to?”*

III. Literature Overview

Table 4 presents a summary of the principal findings from selected literature reviews on methods with applications in disinformation detection. This summary serves both to enhance transparency regarding our research strategy as described in Sec. 4 and to inform readers of the specific challenges and opportunities identified in the existing literature as referenced in Sec. 5.5. The former specifically supports our arguments in Sec. 3.

The studies are assigned to one of the six feature categories as defined by Unger et al., (2025): c*ontent*, *language*, and *source* features (*disinformation* features), and *temporal*, *coordination*, and *dissemination* features (*disinformation campaign* features). For the full table including lower-level features, please consult the paper. Note that the assignment of the studies to features is not necessarily exclusive: we aimed for the best match considering the overall topic while also accounting for the respective studies' findings. Regarding the former, the literature review on "content-based fake news detection […]" by Capuano et al. (2023) was assigned to *content* features, and the review on "misinformation dissemination on social media […]" by Comito et al. (2023) was assigned to *dissemination* features. Regarding the latter, the review by Tufchi et al. (2023) was assigned to *temporal* features due to its usage of a "disinformation lifecycle" to frame the review (p.7), as well as its implications for research which include, among others, temporal dynamics, time-consuming manual fact-checking processes, auxiliary information (including source and time), and early detection. Likewise, Küçük and Can (2020) list the visual presentation of stances' temporal evolution to support decision making, application to data streams, and modelling of spatial and temporal locality for improved context-sensitivity; therefore, their review was also assigned to *temporal* features.

This overview serves as a reference point to relate the findings from our interview study to issues discussed in the literature. However, as we point out in our paper, finding exact conceptual matches between disinformation (campaign) features and representative methodological developments for their detection in the literature would require separate research.

Table 4. Selected literature references integrated into the framework of disinformation and disinformation campaign features (based on Unger et al., 2025).

	Content features	Language features	Source features
Disinformation	*Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review* (Capuano et al., 2023): • data issues (no standard process; access, verification, and storage problems), • homogenous datasets (lack of multi-topic and multi-language datasets, bias towards English), • continuous learning (address evolution of malicious techniques), • decontextualized analysis (associate news with context), • attacks on NLP systems (distortion of facts, confusion of causes and actor/target) *A Comprehensive Comparative Analysis of Deepfake Detection Techniques in Visual, Audio, and Audio-Visual Domains* (Bekheet et al., 2024): • expanding the scope of detection (incorporate emerging facets like text overlays and motion patterns, context videos are presented in, go beyond cross-modal inconsistencies), • advanced DL techniques (investigate newer architectures, specifically transformers) • data availability (foster creation of diverse datasets and public availability, including various manipulation techniques, real-world variations, diverse speakers)	*Emotion detection for misinformation: A review.* (Liu et al., 2024): • dataset collection (address platform affordances, data availability, language diversity), • annotation (reliability of automated annotation), • multimodality (address visual and language understanding), • benchmarks (lack of benchmarks for multiple languages, platforms, modes of data), • interpretability (make methods transparent, explainable, understand errors), • LLMs (combined approaches, LLMs as guides, exploit external knowledge) *Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions* (Pangtey et al., 2025): • multimodal and cross-modal (alignment of cues in different modes to address contextual humor), • zero-shot and few-shot generalization (unseen topics, low-resource languages, address with meta-learning, prompt-tuning, synthetic data), • cross-cultural and multilingual (account for varying opinions, address with RAG or language-specific adapters), • bias and fairness (address biases in data with adversarial training and debiasing), • integrate with fact-checking and misinformation detection (into existing pipelines, use knowledge-graphs, real-time retrieval, include reasoning and verifiability)	*Fake profile detection on social networking websites: a comprehensive review* (Roy and Chahar, 2021): • security and privacy (users unknowingly share data), • rumor detection (backtracking to source), • spam detection (high user engagement), • data volume (efficient processing), • code-mixed data (formats and languages), • abusive content detection (abusive users, cross-language), • cyberbullying detection (trace posts, early detection), • fake accounts (include more platforms) *Graph-based fake account detection: A survey* (Dehkordi and Zehmakan, 2025): • dataset development & data-centric research (few benchmark datasets, bias towards Twitter, account for platform affordances and multimodality, add additional labels like spammers), • focus on heterophilic edges (address feature mixing between real and fake accounts by incorporating causality), • few-shot & zero-shot (for node-classification, integrate into existing detection frameworks), • cold start (early detection hindered by new nodes with limited connections), • explainability (increase transparency of detection models to support decisions and build trust) • adversarial attacks (design sophisticated attack strategies resembling fake accounts in real-world networks)

	Temporal features	Organizational features	Dissemination features
Disinformation Campaign	*Detecting deceptive engagement in social media by temporal pattern analysis of user behaviors: a survey* (Xiao and Bertino, 2017): • data integration (integrate temporal data with contextual data), • application focus (evaluation in further scenarios), • data limitations (selection of appropriate methods), • lack of datasets (limitations in labelling), • problem formulation (potentially more important than method selection) *Stance Detection: A Survey* (Küçük and Can, 2020): • cross-lingual and multilingual (account for differences between languages, limited range of languages covered by datasets), • other media content and bots (explore stances in visual and speech content, detect bots with human-like conversational capabilities), • decision making (benefits for retrieval systems, visualize frequency and temporal evolution), • data streams (close-to real-time processing), • deep NLP (cost vs. performance, pretraining and fine-tuning), • datasets and evaluation (annotation, reliability, availability, account for different languages and text genres), • context-sensitivity (model spatial and temporal locality, incorporate conversational interactions) *A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities* (Tufchi et al., 2023): • multimodality (potentially conflicting data, multi-technique frameworks), • explainability (decrease mistrust in systems, enable understanding of results), • temporal dynamics (account for changing dynamics over time), • manual fact-checking process (labor-intensive, time-consuming, error-prone), • lack of multi-lingual datasets (bias towards English), • auxiliary information (include source and time), • early detection (timely detection to mitigate consequences, integrated and multi-method approaches)	*Detection and characterization of coordinated online behavior: a survey* (Mannocci et al., 2024): • multi-platform (platform affordances, data collection issues), • temporal variability (dynamics, stability, group compositions, adaptation to countermeasures), • heterogeneity (different characteristics of coordinated behavior), • multimodality (develop features and methodologies), • GenAI (AI detection, simulate coordinated behavior), • scalability (approximation algorithms, sampling techniques, graph summarization), • validation and formalization (simulations, formal and testable frameworks), • subjectivity (largely subjective indicators, prioritize transparency), • attribution (collaboration with platforms, multidisciplinary techniques), • multidisciplinarity (collaboration between diverse stakeholders), • ethical dilemmas (concerns about surveillance and consent, potential for misuse) *Fake News Propagation: A Review of Epidemic Models, Datasets, and Insights* (Raponi et al., 2022): • model complexity (balance performance vs. diminishing utility, complexity reduces reusability, limited scalability), • generalization (achievability of modelling many states and transitions vs. performance-driven modelling), • time factor (time parameter often ignored, human behavior changes and so do features and classes), • limits of epidemiological models (propagation in real-life networks characterized by factors not captured by graphs) • limited datasets (too narrow in scope relative to user base) • cross-platform modelling (weak transferability across platforms)	*Multimodal fake news detection on social media: a survey of deep learning techniques* (Comito et al., 2023): • datasets (include more modalities, languages, styles, and topics), • finer classification (go beyond binary fake news detection, allow reasoning), • scalability (scale with modalities, consider storage and processing power), • enhance multimodal classifiers (multimodal embeddings, focus on relevant parts within data, generate synthetic datasets), • source verification and author credibility (potential for automation), • cross-domain (domain shift, adaptive learning), • explainability (transparent decisions and suggestions, justifications), • integration of additional features (creators of news, network characteristics, • propagation-based features), exploitation of emotions (embed emotions provoked by fake news), • exploitation of statistical features (synthetic representation, distribution patterns) *Misinformation dissemination on social media: key research themes and evolutionary paths between 2013 and 2023* (Xu et al., 2025): • impacts of social media misinformation (economic and social disruptions) • misinformation governance (process steps of detection, blocking, verification, correction) • detection as a major research focus (early detection, study propagation & dissemination characteristics) • blocking strategies (reduce nodes activated by misinformation, target influential nodes or links, hinder diffusion) • correction and debunking (important governance tools, measures transparent but costly, multi-party collaboration, help users understand) • governance challenges on social media (autonomy, anonymity, openness, diverse dissemination patterns, gap between posting and escalation) • future directions (stronger real-time monitoring, improved automated detection, analysis of misinformation characteristic, stronger engagement on social media, collaboration with opinion leaders, proactive dissemination of corrective information, guiding users to reduce misinformation impact)

References for III. Literature Overview

Bekheet, A. A., Ghoneim, A., & Khoriba, G. (2024, July). A Comprehensive Comparative Analysis of Deepfake Detection Techniques in Visual, Audio, and Audio-Visual Domains. In *2024 Intelligent Methods, Systems, and Applications (IMSA)* (pp. 122-129). IEEE.

Capuano, N., Fenza, G., Loia, V., & Nota, F. D. (2023). Content-based fake news detection with machine and deep learning: a systematic review. *Neurocomputing*, *530*, 91-103.

Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining*, *13*(1), 101.

Dehkordi, A. S., & Zehmakan, A. N. (2025). Graph-based Fake Account Detection: A Survey. *arXiv preprint arXiv:2507.06541*.

Küçük, D., & Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys (CSUR)*, *53*(1), 1-37.

Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2024). Emotion detection for misinformation: A review. *Information Fusion*, *107*, 102300.

Mannocci, L., Mazza, M., Monreale, A., Tesconi, M., & Cresci, S. *Detection and characterization of coordinated online behavior: a survey (2024)*.

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

Raponi, S., Khalifa, Z., Oligeri, G., & Di Pietro, R. (2022). Fake news propagation: a review of epidemic models, datasets, and insights. *ACM Transactions on the Web (TWEB)*, *16*(3), 1-34.

Roy, P. K., & Chahar, S. (2021). Fake profile detection on social networking websites: a comprehensive review. *IEEE Transactions on Artificial Intelligence*, *1*(3), 271-285.

Tufchi, S., Yadav, A., & Ahmed, T. (2023). A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, *12*(2), 28.

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

Xiao, Q., & Bertino, E. (2017). Detecting deceptive engagement in social media by temporal pattern analysis of user behaviors: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(5), e1210.

Xu, G., Qian, M., & Meng, L. (2025). Misinformation dissemination on social media: key research themes and evolutionary paths between 2013 and 2023. *Humanities and Social Sciences Communications*, *12*(1), 1775.