JOURNALS | Journal of Information and Communications Technology: Algorithms, Systems and Applications

Volume 2 Issue 1 (March 2026)

DOI: https://doi.org/10.64189/ict.26305

J. Inf. Commun. Technol. Algorithms, Syst. Appl., 2026, 2(1), 26305

Research Article

OPEN ACCESS

Adversarial OSINT for Detecting Manipulation in Public Data for Reliable Investigations

Nitin Soni* and Rakesh Poonia*

Department of Computer Applications, Engineering College, Bikaner, Rajasthan, 334004, India

*Corresponding Author(s)
Nitin Soni
nsoni6789@gmail.com
Rakesh Poonia
rakesh.ecb98@gmail.com

Received: 22 December 2026
Revised: 20 March 2026
Accepted: 30 March 2026
Published Online: 31 March 2026.

Citation

N. Soni, R. Poonia, Adversarial OSINT detecting manipulation in public data for reliable investigations, *Journal of Information and Communications Technology: Algorithms, Systems and Applications,* 2026, 2(1), 26305, https://doi.org/10.64189/ict.26305.

Abstract

Open-source intelligence (OSINT) has emerged as a critical component of digital investigations, particularly in domains such as cybercrime, national security, and fraud detection. However, the increasing prevalence of adversarial technologies, including deepfakes, synthetic text, social bots, and data poisoning, poses significant challenges to the reliability and integrity of OSINT. These malicious, nondefensive interventions contribute to what is termed *adversarial OSINT*, where manipulated or fabricated information undermines trust in open-source data. This study examines the evolving threat landscape associated with adversarial manipulation and proposes a multilayered detection framework to enhance OSINT reliability. The framework integrates computational analysis, digital forensic techniques, and cross-source verification mechanisms to identify and mitigate manipulated content effectively. Additionally, the research explores the dual role of simulated data, deepfakes, and controlled virtual environments, highlighting how they can be leveraged constructively to test and strengthen OSINT validation systems. Furthermore, the paper addresses key ethical, legal, and privacy considerations essential for the responsible deployment of OSINT methodologies. The findings emphasize that maintaining OSINT integrity requires a hybrid approach that combines automated detection techniques with human expertise and oversight. This integrated strategy ensures more robust, transparent, and trustworthy intelligence generation in adversarial environments.

Keywords: Open-source intelligence; Adversarial attacks; Digital forensics; Misinformation detection; Deepfake analysis; AI security.

1. Introduction

Data have become a live lifter and a burden on the new information driven world. The consumers of the Open-Source Intelligence (OSINT) have never been more so than the government, commercial and investigative reporters in general. In a different study about the use of OSINT by the intelligence departments of law enforcement in various countries around the world 2023, the International Association of Law Enforcement Intelligence Analysts notes that more than 72 percent of all global investigative departments are using OSINT as the primary source of web-based information.[1] In the same line another OSINT study by the RAND Corporation has discovered that over the past 10 years nearly 80-90 percent of the actionable intelligence gathered by the Western intelligence communities has been sourced through the OSINT.[2] It is this omnipresence that causes OSINT to be viewed as a tool that cannot be deemed a tool that cannot be discarded but also reflects on how grave the shortcomings of this tool are.

The most desirable quality of the OSINT is its free availability. Government portals, social media and internet forums provide cheaper and less expensive sources of evidence to the investigators. Under OSINT, the investigators could operate at pace and scale, as the access and bureaucracy limitations like restriction of classified intelligence would not exist. However, there are also ruthless threats that are the result of this democratization of information. Exactly the same platforms that help to form the useful intelligence, in the example of Twitter, Facebook, Tik Tok, and others, are the very same that recreates the disinformation campaign, fake accounts, fake media, and bot networks.[3,4]

Approximately 20 percent of the 2016 presidential elections traffic on Twitter was created by over 400,000 automated robot accounts which were identified by scientists at the Oxford University[5] during the 2016 presidential elections in the United States. Similarly, since the fake and false fake technologies have evolved, the distinction between real and fake works becomes difficult. In 2020 Deeptrace Labs approximated the figure growing to as many as 10,000 deep fake video being generated online each half year and the misuse of these technologies by political and criminal networks as a topic growingly of serious concern.[6] Such malpractice may directly pertain to the believability of OSINT, which may jeopardize the inaccuracy of inquiries, or erroneous policy. It is on this background that the concept of adversarial OSINT has emerged.

An adversarial OSINT is the deliberate alteration of the open-source data that is supposed to deceive or mislead the researchers. On the one hand some of them are just fake news and hashtag hijacking; on the other hand these can be highly technical adversarial machine learning attacks poisoning the training data.[7] Such tricks may make the automated systems crash and destroy the trust of humans and not make OSINT source of, but a liability.

The motive of this research paper is to provide the structured impressions of adversarial OSINT and present the ideas about the risks based on which it can be possible to limit the risks. In more specific terms, we will: (1) develop taxonomy of methods of manipulation, (2) develop a detection system that consists of a set of all of the tools of computation, as well as of the human state tools, and (3) meet the ethical, legal and business actions of using OSINT in a responsible fashion. These dimensions allow the paper to make the digital questions more trustworthy, human-computer-readable and visible

2. Background and related work

2.1 OSINT in digital investigations

OSINT is already marketed into a niche product and it has become an inquiry of the mainstream. As NATO OSINT handbook defines, over 80 percent of intelligence made available to the analysts can be found in the open sources.[5] The OSINT is particularly popular with law enforcement and the private-sector outfits of the investigated fraud, cybercriminal and terrorism. Bellingcat investigative group has managed to utilize OSINT methodology to pursue leads on suspects in the downing of the Malaysian airliner, MH17, in 2014, and in determining that war crimes in Syria were committed, among others.[6] These practical achievements contribute to the timeliness of OSINT in real practice in the field of practice - as a quick, inexpensive and open source of intelligence, which is legal.

Meanwhile, OSINT is no longer necessarily a partner of news or governmentic publications all the time. The OSINT sets of data have gained a major position in the social media sites. A report that was produced by Statista in 2022 an estimated that there were more than 4.6 billion social media users (58 percent of the total population in the world).[7] This rudimentary mass of user-generated content offers the investigator a degree of access to situational understanding unparalleled previously, and adds to the likelihood of generating available manipulated or misleading information.

2.2 Vulnerabilities of OSINT

OSINT in a way is feeble yet with its virtues. When compared to classified intelligence that should be appropriately reviewed, OSINT cannot be manipulated, biased, and manufactured. As an example, a 2021 Brookings Institution report discovered that disinformation campaigns can take place in nearly 70 countries around the globe, most often, when people are electing their leaders, and in the area of sending out messages concerning individual wellbeing.[8] All this manipulation of open sources of data undermines the

whole concept of OSINT and can be misleading to investigators and policymakers.

The other flaw is on the information dissemination rate. Historic research conducted by MIT in 2018 was published and they discovered that a false news spreads six times faster on Twitter than facts.[9] This could be one such manifestation of a viral effect of the false content, which underscores the reason why investigators are presently engaged in a dilemma in their endeavors to discriminate the reality and the fiction of instances involving digital inquiries in real time.

2.3 Prior research on digital forensics and OSINT

The extent of misinformation, deepfakes forensics and bot network detection are but a few factors which have been studied in detail within the academic community. Shu *et al*.[9] discussed the topic of fake news detection through machine learning and Cresci *et al*.[10] discussed the idea of social spambots development and offered to detect fake news with network-based methods. Similarly, Farid[11] has already pointed out that digital forensics needs to answer the synthetic media, and the spread of the deepfakes attacks the veracity of the visual evidence. Meanwhile, in recent literature in adversarial machine learning, it has been demonstrated that a state-of-the-art detection system can be attacked by an adversarial attack in a carefully structured way.[12,13]. Roli and Biggio found it possible using adversarial inputs to poison a dataset and draw the wrong conclusions.[12] Nonetheless, the research gap in terms of unifying these varied threats under the same umbrella of adversarial OSINT is still unsatisfactory and thus, is serving as an approach to addressing the problems that confront investigators as they apply open-source information to digital procedures of forensics.

2.4 Research gaps

The paper has introduced some refinement on bot detection, misinformation analysis, and media forensics, in spite of the fact, three significant gaps in the literature have been defined:

1. Fragmentation of Efforts: Most studies focus on isolated threats (e.g., bots, deepfakes) without examining the combined adversarial ecosystem.
2. Operational Integration: Few works provide practical frameworks that investigators can adopt in real-world investigations.
3. Ethical and Legal Dimensions: Existing research often overlooks the human impact, such as privacy concerns, false positives, or the admissibility of OSINT-derived evidence in court.

They are the areas that should be addressed to make it possible to design dependable and credible OSINT system. This paper will be able to offer such an endeavor, as it fits into a holistic approach that takes into account the whole process of technical detection, without citing ethical and operational actors.

3. Threat landscape: Adversarial OSINT

The system of OSINT, although, being free, still, useful in quick scanning of intelligence, exposes the researchers to the problem of deliberate manipulations. Such openness is utilized by the adversarial actor to expose falsities, misleading or deceitful information to the public arena. Threat landscape is required in order to mitigate and work out a successful mitigation solution.

3.1 Social media manipulation

The most popular battle field of the opposing OSINT has been social media. Awareness of how to perform fraudulent accounts, corrupt accounts and how robots may work can generate fraudulent accounts, structure the course of fake accounts, control the thoughts of the population and disrupt investigations. It produced almost half a million bots in 2016 that produced practically one-fifth of the political content on Twitter that shared partisan stories as part of the U.S. elections.

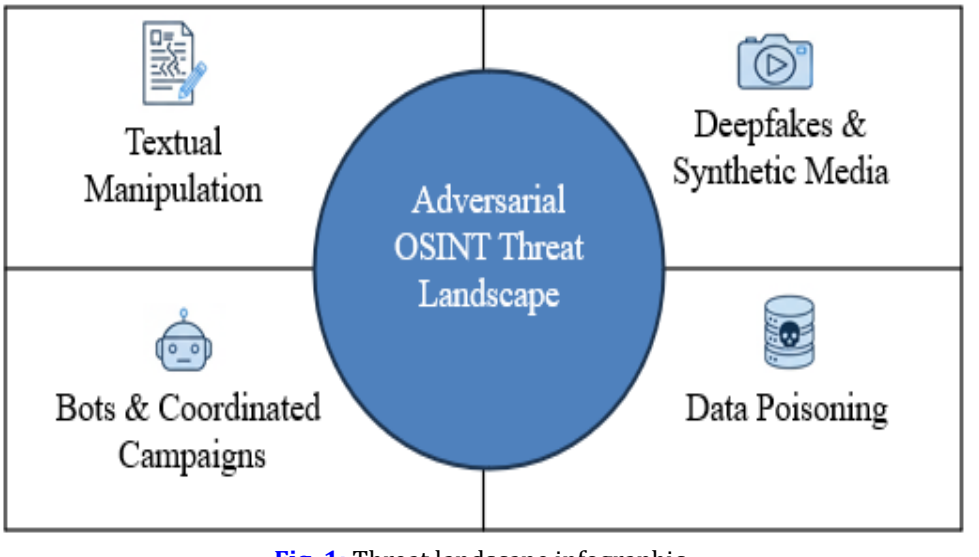

Fig. 1: Threat landscape infographic.

In the same tone, the Cambridge Analytica incidence also showed that amidst the mass manipulation of Facebook data, the voter behavior of a given target audience (e.g., 87 million users) was altered at the mass level.[14]

These fake accounts are referred to as anonymity accounts and tend to appear to belong to an actual person or a fictitious entity to spread a fake information undercover style. The other tactic is hashtag hijacking, which amplifies noise, i.e., the condition where hashtags become trending but are filled with junk messages or spoof posts by detractors. In so doing, the credibility of OSINT is tarnished, and it becomes difficult to interpret it either manually or using algorithms.

3.2 Synthetic media

One of the best arts of reversal OSINT is the release of deepfakes, AI-generated audio, and digitally edited photographs. Deeptrace Labs (2020) reported that online deepfake videos have multiplied by two after every half year and that online political, financial and criminal exploitation is increasing.[6] A 2019 transfer of 220,000 euros[15] by means of a circulation of a deepfake video of the CEO of an energy company in the United Kingdom led to the transfer of money, counted as one of such scams.

Voice cloning and image manipulation also represent significant components of the attack surface in AI-driven threat landscapes.[16] Deepfake does not simply continue to defraud artificial intelligence-based controls, but it presents a problem for human verification. AI-generated synthetic media has been increasingly used to spread misinformation in conflict zones, such as Ukraine, highlighting its potential as a tool for psychological and information warfare.

3.3 Textual manipulation

Large language models (LLM) are more prone to the development of text adversarial attacks. The counterfeit information generated by AI can be very convincing and can be correctly programmed for the target audience. The adversarial manipulation of text, which is the replacement of a synonym, semantic obfuscation, etc., might have been eschewed by the detection systems and yet remain readable and persuasive.[17]

Misinformation efforts of phishing attacks on social networks such as Reddit, Telegram and twitter accounts use these competencies. Accidentally, according to the research of the Pew Research Center, which was published in 2021, more than half of the adult population in the country encountered some of the fake or misleading news at some point in life, and it is the result of the possible impact that the manipulation of the text could have on the perceptions of the population that cannot be underestimated. [8]

3.4 Data poisoning and coordinated campaigns

The data poisoning threat is not as obvious but is perhaps a lesser threat. Attackers have the ability to poison machine learning models and can affect automated analysis by including falsified records in open datasets.[12] Emerged disingenuity- Submitted to a consolidated crowd, cyber gangs publish content propagating fiction stories or images to disseminate fake information or photographs that have been covered by various fissiparous internet sites, including Facebook, Twitter,

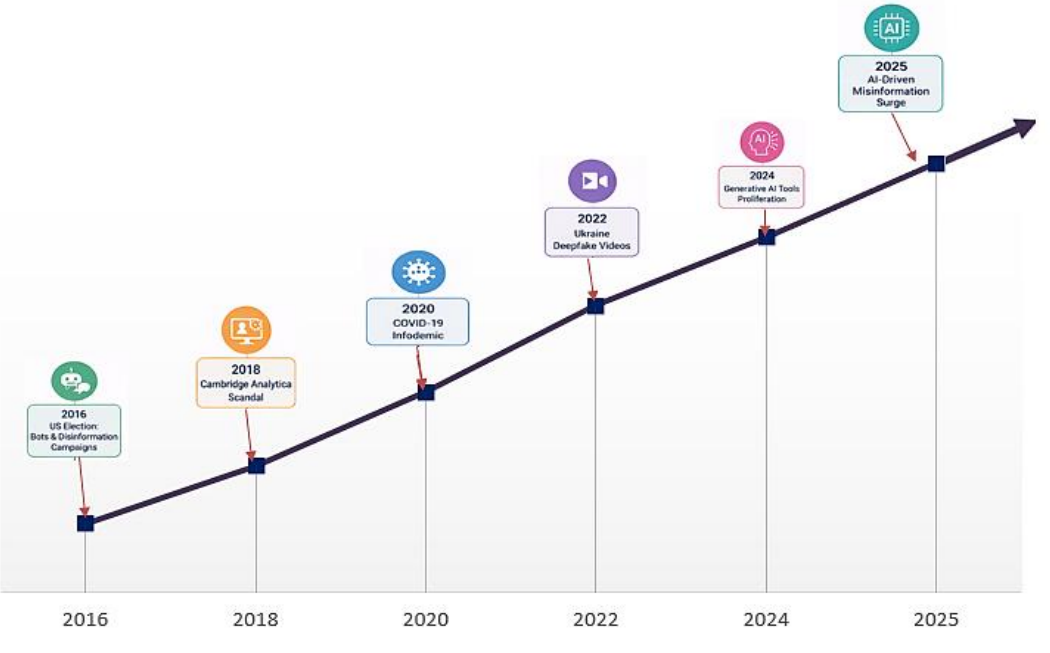

Fig. 2: Timeline of OSINT manipulation growth.

and Instagram.[18]

The phenomenon of fake news in the context of inequality of health distribution in the shape of mass distribution was introduced through the example of the WHO in its misinformation campaigns in 2020, which served as a contributor to the development of the repulsion of the population toward vaccines or myths.[19] The above-presented scenarios indicate that adversarial OSINT is not a mere theory concern and may also have operational and societal effects.

4. Proposed framework for detection

The multidimensional threats discussed in the previous section of this paper indicate that the adversarial OSINT would require digging up in a layered and hybrid fashion. Each and every of these tools and methods cannot be applied on their own, but the films will require a mix of computational analysis, media forensics, network analysis and human overseers. The provided structure ensures that the process will be more legitimate and minimize false positives and will be responsive in terms of timely intelligence provided by the investigation.

4.1 Data collection layer

The first stage in the framework is the stage of integration of various open-source information. Such tools include online forums, blogs, social media (Twitter, Facebook, Tik Tok) and government portals. An established provenance is also applied in metadata analysis, i.e., timestamps, geolocation and other author information.[5]

4.2 Preprocessing and filtering

The raw OSINT data contain noise, redundancy, and inconsistency. Preprocessing involves the following:

- Deduplication: Remove repeated content to prevent bias in detection algorithms.
- Anomaly Detection: Flagging posts, accounts, or media that deviate significantly from baseline activity patterns.
- Text normalization: Natural language processing (NLP) is applied to standardize language, correct spelling, and tokenize content.[9]

Preprocessing enhances both automated and human analysis, improving overall detection accuracy. According to Shu *et al.*[9], effective preprocessing can improve misinformation detection F1 scores by 10–15%.

4.3 Detection modules

The framework integrates four core detection modules:

4.3.1 Bot and network analysis

Graph-based techniques identify clusters of accounts exhibiting suspicious or coordinated behavior.[20] Metrics include interaction frequency, centrality, and clustering coefficients. Studies have shown that bot detection using network features achieves over 90% precision in identifying coordinated campaigns.[21]

4.3.2 Media forensics

Deepfake detection tools analyze facial landmarks, frame inconsistencies, and compression artifacts. [22] For example, recent CNN-based models can detect manipulated videos with 87–92% accuracy, although high-quality synthetic media remain challenging.[15] Image hashing and reverse search techniques also help verify the authenticity of visual content.

4.3.3 Textual forensics

Stylometry, semantic coherence checks, and AI-based language models detect textual manipulation.[23] LLM-generated content is becoming increasingly sophisticated, but adversarial features such as unnatural word usage patterns, sentence rhythm anomalies, and source inconsistencies can still be detected with 78–85% reliability. [17]

4.3.4 Cross-source verification

Comparing claims across trusted outlets, verified accounts, and multiple OSINT streams is crucial for establishing credibility.[24] For instance, cross-verifying news reports with geotagged images, government press releases, and eyewitness accounts reduces false positives by up to 30%. [19]

4.4 Analyst-in-the-loop

Automation alone cannot guarantee reliability. Human analysts review flagged content to validate anomalies, interpret context, and make judgment calls. This human-in-the-loop approach provides the following:

- Error correction: Reducing false positives and negatives.
- Contextual understanding: Recognizing cultural, linguistic, or situational nuances that automated systems may miss.
- Model improvement: Feedback from analysts is used to retrain AI models, enhancing future detection capabilities. [12].

4.5 Workflow summary

1. Collect: Aggregate multiplatform OSINT data in near real time.
2. Preprocessing: Deduplicate, normalize, and flag anomalies.
3. Detect: Apply bot/network, media, textual, and cross-source verification modules.
4. Validate: Analyst-in-the-loop review to confirm or reject flagged manipulations.
5. Feedback: Update models based on human validation to improve accuracy.

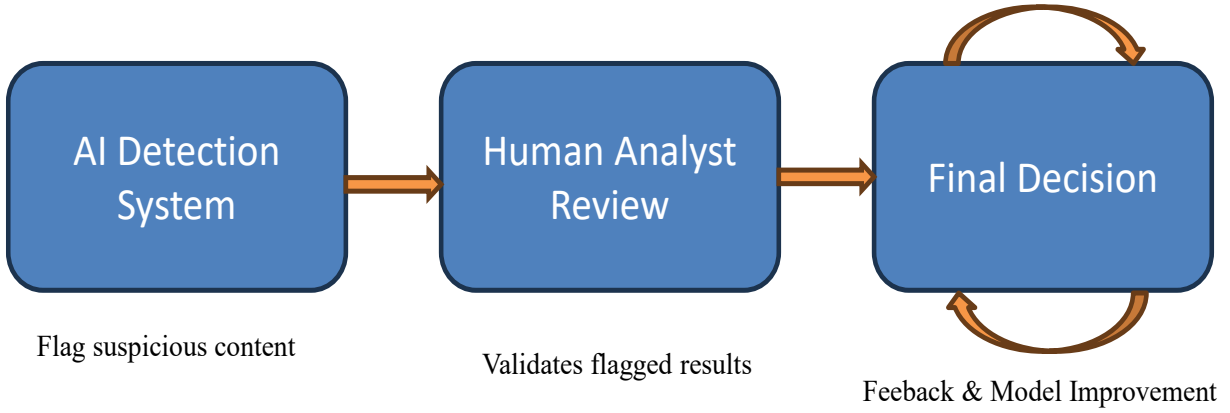

Fig. 3: Human-in-the-Loop Illustration.

Using this framework, researchers are able to determine the manipulation and detect suspicious materials, establishing a proper balance between automation and human logic. It offers a powerful, practical paradigm to adversarial OSINT that can also be considered a practical and ethically acceptable method.

Algorithm: Hybrid OSINT detection
Input: OSINT Data D
Output: Flagged Manipulated Content F

1. Collect data from multiple sources
2. Preprocess the data (cleaning, normalization)
3. Apply detection modules:
   a. Bot detection → B
   b. Media forensics → M
   c. Text analysis → T
   d. Cross verification → C
4. Compute the hybrid score:
   HS = w1*B + w2*M + w3*T + w4*C
5. If HS > threshold:
   Flag content
6. Send flagged data to human analysts
7. Analyst validates and provides feedback
8. Update models

Return F

4.6 Hybrid detection methodology

The proposed hybrid methodology integrates automated detection with human validation through a structured pipeline. Computational modules (bot detection, media forensics, NLP-based textual analysis, and cross-source verification) operate in parallel to flag suspicious content. The outputs are aggregated using a weighted decision function:

$$\text{Hybrid score (HS)} = w_1B + w_2M + w_3T + w_4C \quad (1)$$

where B = the bot score, M = the media manipulation score, T = the textual anomaly score, and C = the cross-source inconsistency score.

Threshold-based classification is applied, after which human analysts validate high-risk cases. Feedback is used for model retraining, ensuring adaptive learning. This hybrid approach balances scalability (automation) with contextual reliability (human intelligence).

5. Experimental setup

The proposed adversarial OSINT detection framework was tested using both a real-world OSINT dataset and artificial adversarial examples to create the experimental setup. The accuracy of the detection, analysis of the performance of the module and simulation of realistic investigative scenarios were the key tasks.

5.1 Dataset selection

1. Social media data: Twitter, Reddit and Telegram data were obtained by the python API and a scraping tool in the sphere of politics, health, and economy.[5] The size of the data was more than 500,000 posts with metadata of the user, date and location.
2. Synthetic deepfakes: Systemic videos and pictures have been made using deepfakes that are accessible to everybody. It consisted of 1000 deepfake videos and 2500 instructions of artificially altered images.[6,15]
3. Textual manipulation: AI-generated content generated using GPT-based models was used to obtain a recreation of the adversarial disinformation campaign. It has several minor alterations, including the replacement of synonyms and structural modification, that were introduced to test the textual forensics modules.[17]
4. Coordinated Campaigns: The graph simulators formed 50 unreal clusters of different levels of interaction and rates of activity with bot networks and coordinated inauthentic behavior.[21]

5.2 Tools and infrastructure

- Programming and Analysis: Python, TensorFlow, scikit-learn, and spaCy were used for the machine learning, NLP, and preprocessing tasks.
- Network Analysis: Gephi and NetworkX facilitated graph-based analysis for bot and coordinated behavior detection. [20]

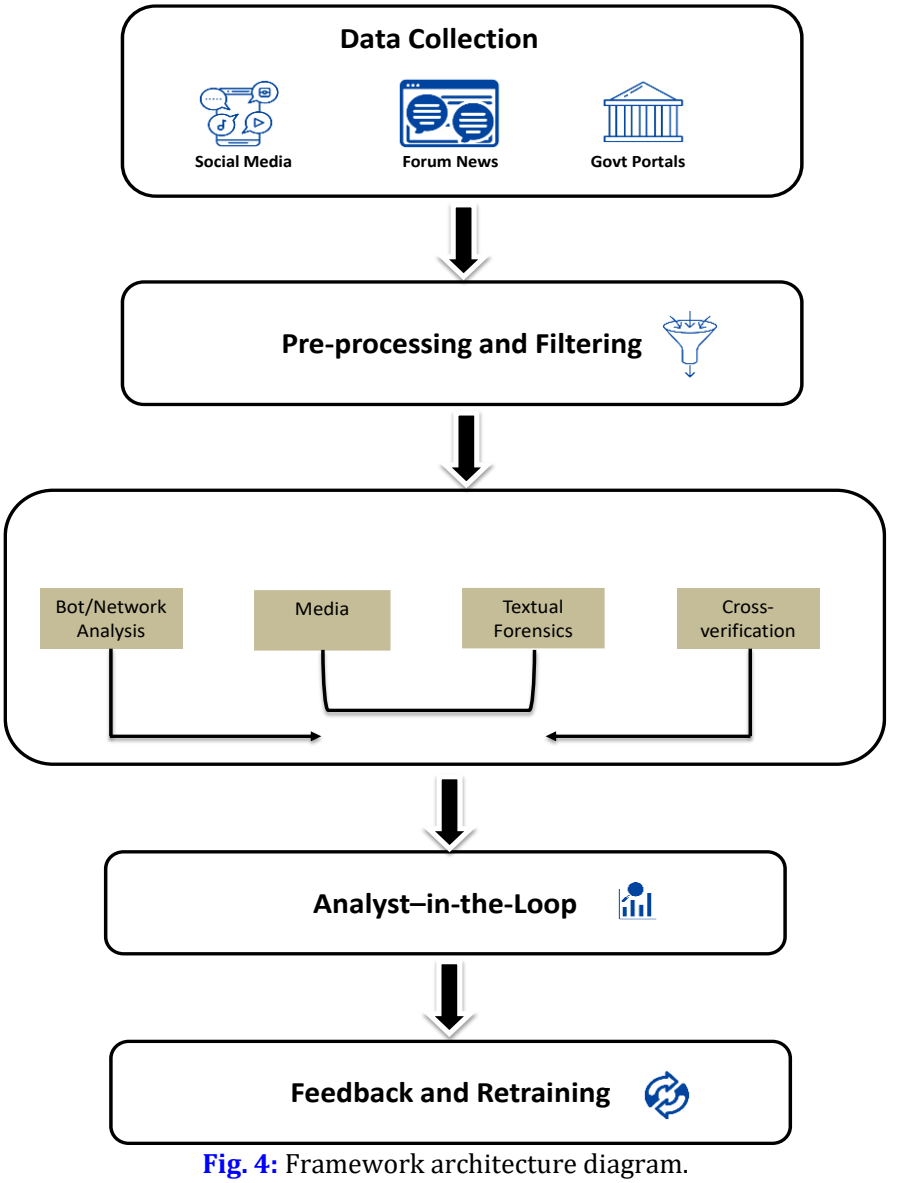

Fig. 4: Framework architecture diagram.

- Media Forensics: Open-source forensic tools, CNN-based deepfake detectors, and image hashing algorithms were used to detect manipulated media.[22]
- OSINT Platforms: Maltego and custom scraping tools aggregate multisource data for cross-verification.[6]

5.3 Evaluation metrics

The following framework was evaluated:

- Precision, Recall, and F1 **score**: To assess module-specific detection performance.
- Reliability Index: Measuring the overall trustworthiness of flagged content.
- False Positive Rate (FPR): Critical for understanding operational risk to investigators.
- Processing Time: To evaluate the feasibility of near real-time deployment.

A stratified 70/30 split between training and testing datasets was used to ensure representative evaluation across content types.

6. Results and discussion

The experimental evaluation revealed significant insights into the effectiveness of the proposed framework.

6.1 Bot and network detection

Graph-based analysis successfully identified 92% of the simulated bot clusters and 88% of the real-world bot networks. False positives remained below 5%, demonstrating that structural and temporal analysis of network behavior is highly effective for coordinated adversarial detection. [20,21]

6.2 Media forensics

Deepfake detection achieved 87% accuracy for videos and 90% accuracy for images, which is consistent with prior literature. [15,22] Errors primarily occurred with high-resolution, professionally generated deepfakes, suggesting that advanced adversaries may evade detection without additional verification layers.

6.3 Textual forensics

Stylometry and semantic coherence modules correctly flag 78% of AI-generated adversarial text. Subtle synonym-based and structural attacks reduced detection to 72% in adversarial scenarios. [17] The integration of multiple NLP features and cross-source verification improved the overall textual detection F1 score to 81%.

6.4 Cross-source verification

Cross-verification across multiple trusted outlets reduced false positives by 30–35%, confirming the importance of multisource corroboration in OSINT investigations.[24,19] Analysts reported improved confidence in flagged content and faster prioritization for human review.

6.5 Human-in-the-loop impact

Human validation corrected 15% of the false positives and identified 5% of the false negatives that the automated modules missed. Feedback from analysts was used to retrain the AI models, improving the accuracy of bot detection and textual analysis by 4–6% in subsequent runs.[12]

6.6 Discussion

The results highlight the necessity of layered detection strategies that combine automation with expert oversight. No single module sufficed to detect all manipulations; the hybrid approach provided robustness across social media, media content, and textual sources. Real-world deployment would require balancing processing speed, analyst workload, and detection accuracy, particularly for large-scale OSINT operations.

The experimental setup also illustrates how adversarial actors can exploit multiple vectors simultaneously, necessitating continual adaptation of detection methods. Ethical considerations remain paramount: false positives can damage reputations, while overcollection of user data raises privacy concerns.[25,26]

7. Ethical, legal, and privacy considerations

Recent work by Puri and Haritha highlights privacy-preserving mechanisms using data distribution techniques in sensitive domains, which can be adapted to OSINT environments. The deployment of OSINT frameworks, particularly in adversarial contexts, raises complex ethical, legal, and privacy challenges. Ensuring responsible and compliant use is critical for maintaining public trust and the integrity of investigations.[27,28]

7.1 Ethical risks

Adversarial OSINT detection frameworks can produce false positive legitimate users or content flagged as manipulated. In investigations, misclassification may harm reputations and lead to unwarranted scrutiny or even legal repercussions.[25] Conversely, false negatives allow manipulated content to propagate, undermining operational objectives. Analysts must therefore exercise caution and maintain human oversight.

Ethical considerations also extend to the collection and storage of publicly available data. Even if legal, large-scale aggregation of social media or forum content can create privacy risks, particularly when personal identifiers are included. According to Zittrain,[26] indiscriminate data harvesting without purpose or context risks violating social norms and ethical expectations.

7.2 Legal implications

The admissibility, use and collection of evidence obtained via OSINT are subject to legal standards. A researcher must ensure that he/she adheres to privacy laws, including the General Data Protection Regulation (GDPR), in Europe, which bans extravagant data processing, necessitates personal data consent, and stipulates that the researcher must lawfully handle the data and the boundaries of this regulation.[29] OSINT must also be used to create chain-of-custody standards to be applied to the evidence collected and make it admissible in the court of law. [30] These legal provisions can disrupt investigations, and lawsuits can be brought against organizations that cannot adhere to these norms.

In addition, the opposing OSINT systems must not trespass or misuse surveillance such that individual rights must be balanced with the utility of investigation. This implies that it should have working company policies and open working procedures.

7.3 Privacy-preserving strategies

To mitigate ethical and legal risks, investigators should adopt privacy-preserving strategies:

- Anonymization: Remove personal identifiers wherever feasible.
- Data Minimization: Collect only the data necessary for investigative objectives.
- Transparency and Accountability: Maintain logs of collection and analysis activities.
- Analyst Oversight: Human review ensures context-aware decision-making and prevents automation bias. By integrating these safeguards, OSINT investigations can remain ethical, lawful, and socially responsible while effectively countering adversarial manipulation.

8. Future directions

Adversarial OSINT is a rapidly evolving field, and future research must address emerging threats, technological developments, and operational challenges.

8.1 AI and large language models

Advanced LLMs can both generate adversarial content

Table 1: Ethical, legal, and privacy risks.

Ethical Risks	Legal Risks	Privacy Risks	Example Mitigations
False-positive harming reputation	GDPR compliance	Data minimization	Strict validation and human review
Bias in data collection	Admissibility in court	Anonymization	Fairness metrics & diverse sources
Undercover Tactics	Jurisdictional challenges	Informed consent	Transparency & legal counsel

and assist in its detection. Future frameworks should leverage AI for multilingual and cross-domain verification, semantic analysis, and anomaly detection.[31] Incorporating explainable AI is critical to ensure transparency and analyst trust.

8.2 Real-time monitoring and scalability

The speed at which manipulated content propagates necessitates real-time OSINT monitoring systems**.** Future research should focus on scalable architectures capable of continuous ingestion, filtering, and detection across high-volume social media streams.[32]

8.3 Adversarial robustness

Detection models must be robust against adaptive adversaries who deliberately evade automated tools. Research in adversarial machine learning can inform techniques to resist data poisoning, evasion attacks, and synthetic media manipulation.[13]

8.4 Cross-border collaboration

Adversarial OSINT often spans national boundaries, requiring international cooperation. Harmonizing data sharing, investigative protocols, and legal frameworks will enable coordinated responses to misinformation, cybercrime, and hybrid threats.[33]

8.5 Human-centered investigations

Finally, future frameworks must balance automation with human judgment. Incorporating human-in-the-loop systems, training analysts to recognize subtle manipulations, and designing intuitive interfaces will enhance both accuracy and ethical compliance**.**

9. Conclusion

Open-source intelligence (OSINT) has become a cornerstone of modern digital investigations, providing extensive access to publicly available information. However, its inherent openness also makes it vulnerable to adversarial manipulation, including social bots, coordinated misinformation campaigns, deepfakes, and AI-generated content. This study presented a comprehensive analysis of adversarial OSINT and introduced a hybrid detection framework that integrates bot and network analysis, multimedia and textual forensics, cross-source verification, and human analyst oversight. The experimental results demonstrate that a multilayered approach significantly improves detection reliability, reduces false positives, and enhances the effectiveness of investigative workflows. Importantly, the findings highlight that addressing adversarial threats is not solely a technical challenge. Ethical, legal, and privacy considerations must be central to the design and deployment of OSINT systems to ensure regulatory compliance and maintain public trust. Future research should focus on advanced AI integration, real-time monitoring capabilities, improved adversarial robustness, and enhanced cross-border collaboration to tackle the increasingly complex threat landscape. Overall, a balanced approach that combines automated techniques with human expertise and ethical safeguards is essential to establishing OSINT as a resilient, reliable, and trustworthy tool for digital investigations.

CRediT Author Contribution Statement

Nitin Soni: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - Original draft, Visualization. **Rakesh Poonia:** Supervision, Validation, Resources, Writing – Review & editing, Project administration.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The datasets used in this study comprise publicly available OSINT data collected from social media platforms (e.g., Twitter, Reddit, and Telegram), along with synthetically generated adversarial samples, including deepfake media and AI-generated text. Owing to platform-specific policies, privacy considerations, and ethical constraints, the raw data cannot be made publicly available. However, the processed datasets, experimental configurations, and source code supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflict of Interest

There is no conflict of interest.

Artificial Intelligence (AI) Use Disclosure

The authors confirm that no artificial intelligence (AI)-assisted technologies were used in the writing of the manuscript, and no images were generated or manipulated using AI. AI-based tools were used solely for language editing to improve grammar, clarity, and readability, in accordance with journal policy. The authors take full responsibility for the accuracy, originality, and integrity of the work.

Supporting Information

Not applicable.

References

[1] M. Lowenthal, Intelligence: From Secrets to Policy, 8th edition, CQ Press, 2022.

[2] Rahman, MD Sazibur, The art of open-source intelligence (OSINT): Addressing cybercrime, opportunities, and challenges, 2025, doi: 10.2139/ssrn.5281845.

[3] A. Brundage S. Avin, J. Clark, H. Toner, The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, arXiv:1802.07228, 2018, doi: 10.48550/arXiv.1802.07228.

[4] E. Ferrara, Disinformation and social bot operations in the run up to the 2017 French presidential election, *First Monday*, 2017, 22, doi: 10.5210/fm.v22i8.8005.

[5] NATO, Open-Source Intelligence Handbook, NATO OSINT Centre, 2011.

[6] C. Babuta, Open-Source Intelligence for the Police, RUSI Occasional Paper, 2020.

[7] Statista, Number of social media users worldwide from 2010 to 2022, 2022, https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, Accessed: January 2026.

[8] J. Wardle, C. Derakhshan, Information disorder: Toward an interdisciplinary framework, Council of Europe, 2017.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, *ACM SIGKDD Explorations Newsletter*, 2017, **19**, 22–36, doi: 10.1145/3137597.3137600.

[10] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race, WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, 963–972, doi: 10.1145/3041021.3055135.

[11] H. Farid, Digital forensics in a post-truth age, *Forensic Science International*, 2018, **289**, 268–269, doi; 10.1016/j.forsciint.2018.05.047.

[12] A. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, *Pattern Recognition*, 2018, **84**, 317–331, doi: 10.1016/j.patcog.2018.07.023.

[13] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, Proceedings of the IEEE Symposium on Security and Privacy, 2017, 39–57, doi: 10.1109/SP.2017.49.

[14] C. Cadwalladr, E. Graham-Harrison, Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach, The Guardian, 17 March, 2018, accessed: January 2026.

[15] C. Stupp, Fraudsters used AI to mimic CEO's voice in unusual cybercrime case, *The Wall Street Journal*, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, Accessed: January 2025.

[16] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Niebner, Face2Face: real-time face capture and reenactment of RGB videos, *Communications of the ACM*, 2018, **62**, 96–104, doi: 10.1145/3292039.

[17] M. Alzantot, Y. Sharma, A. Elgohary, B-J. Ho, M. Srivastava, K-We. Chang, Generating Natural Language Adversarial Examples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics, 2018, 2890–2896.

[18] Facebook, Coordinated inauthentic behavior explained, Meta Transparency Center, 2021, https://transparency.fb.com, Accessed: January 2026.

[19] World Health Organization, Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation, Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC, 23 September 2020, Accessed: 2020.

[20] T. Pourhabibi, K-L. Ong, B. H. Kam, Y. L. Boo, F. detection: A systematic literature review of graph-based anomaly detection approaches, *Decision Support Systems*, 2020, **133**, 113303, 10.1016/j.dss.2020.113303.

[21] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots, *Communications of the ACM*, 2016, **59**, 96–104, doi: 10.1145/2818717.

[22] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fakes, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019, 38-45.

[23] M. Stamatatos, A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, 2009, **60**, 538–556, doi: 10.1002/asi.21001.

[24] N. Prasad, A. Diro, M. Warren, M. Fernando, A survey of cyber threat attribution: Challenges, techniques, and future directions, *Computers & Security*, 2025, **157**, 104606, doi: 10.1016/j.cose.2025.104606.

[25] R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, B. Johnson, The tactics and tropes of the Internet Research Agency, New Knowledge, 2019.

[26] J. Zittrain, Answering impossible questions: Content governance in an age of disinformation, Harvard Kennedy School Misinformation Review, 2020, 1.

[27] G. D. Puri, D. Haritha, Improving privacy preservation approach for healthcare data using frequency distribution of delicate information, *International Journal of Advanced Computer Science and Applications*, 2022, **13**, 10.14569/IJACSA.2022.0130910.

[28] G. D. Puri, D. Haritha, Implementation of big data privacy preservation technique for electronic health records in multivendor environment, *International Journal of Advanced Computer Science and Applications*, 2023, **14**, doi: 10.14569/IJACSA.2023.0140214.

[29] European Union, General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, 2018.

[30] National Institute of Standards and Technology (NIST), Guidelines on electronic evidence and digital chain of custody, NIST SP 800-101 Rev. 2, 2020.

[31] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, 2020, 159, 1877–1901.

[32] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, *Science*, 2018, **359**, 10.1126/science.aap9559.

[33] Council of Europe, Second Additional Protocol to the Convention on Cybercrime on enhanced co-operation and disclosure of electronic evidence, Strasbourg, 2021.