Comunicar, n. 84, v. XXXIV, 2026 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com

The Role of Deepfake, Deception, and Disinformation in Conflict Zones Based on DL for NLP: A Critical AI-Era Perspective

El papel del deepfake, el engaño y la desinformación en zonas de conflicto basado en aprendizaje automático para PNL: una perspectiva crítica de la era de la IA

Pascal Muam Mah, AGH University of Krakow, 30-059 Krakow (Poland) (mah@agh.edu.pl) (https://orcid.org/0000-0001-6851-1518)

ABSTRACT

The fast-paced growth of artificial intelligence (AI) has transformed contemporary information warfare, particularly in unstable, conflict-ridden regions. Among the most significant threats is the "3D-Sec" triad Deepfake, Deception, and Disinformation which exploits synthetic media, psychological manipulation, and fabricated narratives to weaken digital trust, disrupt institutions, distort learning, and influence cognitive and socio-psychological behavior. This research rigorously examines the role of 3D-Sec in conflict zones and introduces an AI-driven framework designed to identify and mitigate its repercussions on security, governance, education, cognition, and public perception. As NLP and AI-generated content becomes increasingly lifelike, the boundary between reality and fabrication blurs. Existing detection techniques often lack contextual understanding and fail to address the complex, multi-dimensional characteristics of 3D-Sec threats, including their cognitive and educational impacts. Uncontrolled 3D-Sec campaigns jeopardize privacy, obstruct peace efforts, misinform learners, and contribute to socio-psychological stress and geopolitical instability. Understanding these mechanisms is critical to protecting vulnerable populations, safeguarding education, and ensuring the integrity of digital communication during conflict. We propose a Deep Learning–Natural Language Processing (DL-NLP) approach grounded in the 'Multidimensional Knowledge Framework for Data Analysis (6-W)'. This framework incorporates six analytical dimensions, expressed as ($Wt = f(W_y, W_r, W_n, W_o, W_h)$) capturing the contextual, temporal, cognitive, and socio-psychological dependencies essential for detecting 3D-Sec activities. The proposed model enhances detection capabilities by integrating semantic, situational, and cognitive indicators, enabling accurate recognition of AI-induced deceit, misinformation, and educational content manipulation. A context-sensitive, interdisciplinary defense system such as this is vital for combating AI-enhanced information threats in war-torn regions, protecting cognition, education, and societal resilience against manipulation and misinformation.

RESUMEN

El rápido crecimiento de la inteligencia artificial (IA) ha transformado la guerra informativa contemporánea, especialmente en regiones inestables y afectadas por conflictos. Entre las amenazas más significativas se encuentra la tríada "3D-Sec": Deepfake, Decepción y Desinformación, que explota los medios sintéticos, la manipulación psicológica y las narrativas fabricadas para debilitar la confianza digital, perturbar las instituciones, distorsionar el aprendizaje e influir en el comportamiento cognitivo y socio-psicológico. Esta investigación examina rigurosamente el papel del 3D-Sec en zonas de conflicto e introduce un marco impulsado por IA diseñado para identificar y mitigar sus repercusiones en la seguridad, la gobernanza, la educación, la cognición y la percepción pública. A medida que el contenido generado por IA y los sistemas de PLN se vuelven cada vez más realistas, la frontera entre la realidad y la fabricación se difumina. Las técnicas de detección existentes suelen carecer de comprensión contextual y no abordan las complejas y multidimensionales características de las amenazas 3D-Sec, incluidas sus repercusiones cognitivas y educativas. Las campañas 3D-Sec incontroladas ponen en riesgo la privacidad, obstaculizan los esfuerzos de paz, desinforman a los aprendices y contribuyen al estrés socio-psicológico y a la inestabilidad geopolítica. Comprender estos mecanismos es fundamental para proteger a las poblaciones vulnerables, salvaguardar la educación y garantizar la integridad de

Received: 2025-07-25 | Reviewed: 2025-08-18 | Accepted: 2025-08-21 | Online First: 2025-11-25 | Published: 2026-01-04
DOI: https://doi.org/10.5281/zenodo.18115680 | Pages: 228-246

la comunicación digital en contextos de conflicto. Se propone un enfoque de Deep Learning–Procesamiento del Lenguaje Natural (DL–PLN) basado en el "Marco de Conocimiento Multidimensional para el Análisis de Datos (6-W)". Este marco incorpora seis dimensiones analíticas, expresadas como ($Wt = f(W_y, W_r, W_n, W_o, W_\beta)$) capturando las dependencias contextuales, temporales, cognitivas y socio-psicológicas esenciales para la detección de actividades 3D-Sec. El modelo propuesto mejora las capacidades de detección al integrar indicadores semánticos, situacionales y cognitivos, lo que permite un reconocimiento preciso del engaño, la desinformación y la manipulación de contenidos educativos inducidos por IA. Un sistema de defensa interdisciplinario y sensible al contexto como este es vital para combatir las amenazas informativas potenciadas por IA en regiones devastadas por la guerra, protegiendo la cognición, la educación y la resiliencia social frente a la manipulación y la desinformación.

KEYWORDS | PALABRAS CLAVE

3D-Sec, Deepfake Detection, Disinformation Warfare, Artificial Intelligence in Conflict, Context-Aware NLP, Multidimensional Knowledge Framework, Deep Learning.

3D-Sec, Detección de Deepfake, Guerra de Desinformación, Inteligencia Artificial en Conflictos, PLN Sensible al Contexto, Marco de Conocimiento Multidimensional, Deep Learning.

1. Introduction

Artificial intelligence has become a dual-edged sword in modern conflict, offering both strategic advantages and formidable threats. Among the most insidious are synthetic content tools capable of manipulating public perception, destabilizing governance, and fueling chaos. These threats manifest most acutely through *3D-Sec* a triad comprising Deepfakes, Deception, and Disinformation. In war-tone regions, where infrastructures are fragile and populations vulnerable, the exploitation of 3D-Sec undermines peace processes and fuels hostilities. Citron and Chesney (2019) examine deepfakes as a growing threat to privacy, democracy, and national security, emphasizing legal, technological, and societal implications of synthetic media manipulation. More so, Westerlund (2019) reviews the rise of deepfake technology, exploring its development, potential applications, and associated ethical, societal, and security risks, highlighting the urgent need for regulatory and technical countermeasures. Also, Jacobsen (2025) explores advances in algorithmic detection of deepfakes, evaluating efficacy of emerging tools, technical challenges, and implications for culture and trust in media authenticity amid evolving synthetic content. Additionally, Maronkova (2021) discusses NATO's strategic communication in countering hybrid warfare, highlighting its critical role in combating disinformation and propaganda through coordinated messaging, transparency, and resilience-building initiatives.

Interdisciplinary research integrating AI, NLP, social sciences, education theory, and communication studies is crucial (Alam, Mrida, & Rahman, 2025). Algorithms can be manipulated invisibly, while humans face cognitive and ethical limits. Combining human and machine intelligence enables deeper semantic detection, protecting knowledge integrity and trust amid evolving 3D-Sec disinformation in conflict zones. In this context, interdisciplinary research integrating AI, NLP, social sciences, education theory, and communication studies is essential (Alam et al., 2025). Algorithms can be manipulated beyond traceability, while humans face cognitive and ethical limits.

Combining human insight with machine intelligence fosters deeper semantic detection, safeguarding knowledge integrity, media literacy, and trust amid the evolving 3D-Sec disinformation landscape in conflict zones.

2. Literature Review

2.1. AI Techniques in the Field of Cybersecurity

The rise of AI-generated content has heightened worries regarding Deepfake, Deception, and Disinformation (3D-Sec), especially in areas of conflict. Research conducted by Rana et al. (2022) identifies deepfakes as an escalating danger to public trust, with the potential to modify political narratives and provoke violence. The DL-NLP framework is useful in war or conflict zones. Artificial intelligence is increasingly pivotal in addressing digital challenges across various domains, disciplines, and subjects like explainable AI bolsters the detection of misinformation and hate speech on social media, thereby enhancing transparency and trust.

In the field of cybersecurity, AI techniques provide protection against sophisticated attacks (Al Siam, Hassan, & Bhuiyan, 2025), while proactive malware detection ensures resilience (Agrawal, Pandey, & Lakshmi, 2025). Furthermore, deep learning applications extend beyond security, influencing innovative marketing strategies and informing future research directions. Exposing Propaganda and Misinformation: Conflict zones are often saturated with deepfake videos, altered audio, and misleading narratives that can escalate tensions, mislead civilians, or distort international views. Identifying these materials is vital for preventing misinformation from influencing opinions or inciting violence. 2) Protecting Civilians and Communities: By recognizing false narratives, authorities and humanitarian organizations can provide civilians with verified information, thus reducing panic, rumor-driven behavior, or manipulative recruitment by adversaries. 3) Aiding Policy and Decision-Making: The accurate identification of disinformation allows governments, NGOs, and international organizations to develop targeted interventions, communication strategies, and strategic responses. 4) Ensuring Educational Integrity: In conflict zones, educational institutions and media may be exploited to disseminate propaganda. Detecting manipulated content helps educators maintain reliable learning environments, even in times of crisis.

Shu et al. (2020) categorize disinformation detection into three methods: knowledge-based, style- based, and propagation-based, highlighting the necessity for multimodal detection frameworks. Con- currently, Zhou et al. (2021) investigate the forensic examination of altered videos through GAN-detection algorithms. Research on misinformation, such as that by Lazer et al. (2018), points to source credibility, linguistic markers, and social network patterns as critical indicators. In regions prone to conflict, false information frequently

disseminates more rapidly than verified content due to increased emotional engagement (Vosoughi, Roy, & Aral, 2018). Recent developments combine NLP and DL techniques to identify narrative bias, fake identities, and coordinated online behavior. Together, these studies emphasize the pressing need for comprehensive, multidimensional strategies to address AI-enabled 3D-Sec threats.

2.2. Deepfakes & Synthetic Media in Conflicts

During the 2025 clash between India and Pakistan, deepfake videos produced by AI erroneously illustrated surrender statements and fabricated military operations were commonly seen reusing unrelated footage to escalate tensions and generate misconceptions DISA. Also, the Russo-Ukrainian conflict has witnessed the emergence of deepfake videos featuring President Zelenskyy and other prominent individuals, as well as AI-generated propaganda aimed at children; these artificial clips have eroded public trust and disrupted the dissemination of reliable information Wikipedia, DFRLab. Albader (2025) explores the potential impact of synthetic media in provoking genocide, contending that deepfakes and altered content can exacerbate hate speech, misrepresent reality, and dehumanize specific groups. The article advocates for the establishment of legal frameworks to combat the misuse of synthetic media in inciting mass violence and identity-driven atrocities. Also, an investigation on the conflict between free expression and the regulation of deepfakes as it pertains to the First Amendment (Bourgault, 2025). The article investigates the legal obstacles in managing harmful synthetic media while safeguarding constitutional rights, suggesting equitable strategies to mitigate misinformation without violating essential speech liberties.

2.3. Deception & Disinformation Tactics

Conflict participants often reuse archival footage (from video games or historical conflicts), create fabricated images, and construct misleading narratives to portray fictitious occurrences such as air strikes, coups, or mass defections. Such tactics have been noted in Sudan, Ukraine, Sudan, Cameroon, and various other conflict zones Wikipedia-Ukraine, Wikipedia-Sudan, Wikipedia-Cameroon. Disinformation entities such as Russia's" Doppelg¨anger" initiative replicate reputable media sources in various languages to disseminate pro-Russian narratives worldwide Wikipedia. A study that provides an in-depth examination of societal resilience against disinformation stemming from deepfakes (Carpenter, 2024). They elucidate definitions, evaluate the dual-use capabilities of synthetic media, survey cutting-edge detection technologies, scrutinize European legal structures, and propose educational and policy measures aimed at equipping the public to recognize altered content. Also, Samoilenko (2017) analyzes strategic deception in the context of 'truthiness,' a phenomenon where emotional appeal surpasses factual integrity. This chapter scrutinizes the motivations that drive deceptive communication, the techniques for recognizing such deception, and the manipulation of behavior through persuasive yet misleading messaging strategies in today's information environments.

2.4. Automated Disinformation & Computational Propaganda

Operations like "Operation Overload" (commonly referred to as Matryoshka) have amplified disinformation through the utilization of consumer-grade AI technologies for the cloning of images, videos, and voices; they dispatched tens of thousands of emails to fact-checkers to enhance visibility, even when the content was fabricated WIRED. Also, the application of bots, intensified messaging, and coordinated campaigns is essential to today's computational propaganda deepfakes are a vital resource for scalable influence operations Wikipedia. In their 2022 work, Chadwick and Stanyer (2021) articulate deception as a central theme that interlinks disinformation, misinformation, and misperceptions. Their proposed framework incorporates the intentions of various actors, the flow of information, and cognitive biases, thereby delivering a detailed typology for assessing the effects of deceptive strategies on digital communication and societal power relations.

Additionally, O'Hara (2022) investigates the ways in which bots and computational propaganda undermine truth and knowledge in the online environment, stressing the importance of information literacy education. The research underscores the duties of academic librarians in preparing students to critically assess digital information in a time influenced by automation and misinformation. In the study by Olanipekun (2025), the focus is on how AI technologies contribute to computational propaganda and misinformation, functioning as influential mechanisms for media manipulation. This analysis delves into the effects of AI-driven content on shaping public perceptions and highlights the importance of ethical oversight alongside digital media literacy.

The work of Plikynas, Rizgelienė and Korvel (2025) presents a systematic review based on PRISMA principles, leveraging machine and deep learning to scrutinize fake news, propaganda, and disinformation within online social networks. Their study investigates the various propagators, including both authors and bots, the attributes of the content, and its societal repercussions, while also charting the methods utilized and proposing potential future research paths.

2.5. Psychological & Strategic Impact

A study on media regarding wartime deepfakes revealed that designating authentic footage as fabricated can diminish trust in genuine documentation—potentially jeopardizing the credibility of journalism LERO. Deepfakes have surfaced as a means to erode military morale, trigger civilian anxiety (such as through deceptive ceasefire communications), and strategically mislead opponents or the public raising profound ethical and legal dilemmas under International Humanitarian Law (IHL). Nawaz (2025) explores the concept of psychological warfare in the context of the digital era, outlining tactics such as targeted messaging, emotional manipulation, and disinformation campaigns. The research evaluates the effects on personal cognition and social unity, while promoting countermeasures that include digital resilience training, education in critical thinking, and regulation through policy.

Hancock and Bailenson (2021) investigate the social implications of deepfakes, emphasizing dangers such as diminished trust, altered perceptions, and possible damage to reputations. They analyze the ways in which synthetic media challenges social norms and promote the need for media literacy, ethical standards, and technological measures to reduce misuse and safeguard public trust.

A great importance is given to achieve a balance between the detection of extremist material using AI and the safeguarding of freedom of expression and privacy, as evidenced by Germany's NetzDG legislation and global case studies. For instance, Khan et al. (2023) investigate the role of AI in counter-terrorism through the use of sophisticated predictive analytics aimed at identifying, preventing, and addressing terrorist actions on a global scale. Their study investigates the strategies that terrorist entities adopt in leveraging deepfakes to magnify their recruitment narratives and fortify extremist ideologies. This study underscores the importance of algorithmic pattern recognition, risk assessment frameworks, and strategies for data integration, while also highlighting the dual aspects of improved security advantages and the ethical implications associated with the use of AI. In another study, Matar (2025) explores the phenomenon of AI-assisted terrorism, highlighting the new challenges associated with automated weaponization, radicalization, and operational planning. The dissertation suggests counterterrorism strategies that incorporate regulatory frameworks, international collaboration, predictive surveillance, and the integration of human oversight with AI systems to reduce risks while maintaining ethical standards.

2.6. Rumor Politics & State Propaganda

Since the onset of Boko Haram's incursions in the northern region in 2014, public sentiment has been significantly shaped by rumors and conspiracy theories. Often state-sponsored, these narratives have portrayed foreign powers, especially France, as conspiring for regime change, which has bolstered President Biya's authoritarian control by casting him as a victim of foreign intervention (Nounkeu, 2020). Moreover, the exercise of suspicion in narratives promotes political mobilization through the fabrication of imagined foes; thus, rumors function as instruments of public political action, affecting elite discussions and international relations

The government of Cameroon has labeled separatist combatants as "terrorists" and "criminals," primarily portraying the internal strife as a counter-terrorism operation. This approach has masked genuine political concerns and intentionally merged the separatist uprising with the threats posed by Boko Haram, leading to confusion among both national and global observers inkl. Svetoka (2016), author of the NATO Strategic Communications Centre of Excellence report social media as a Tool of Hybrid Warfare, contends that social media has evolved into a significant weapon in hybrid conflicts. The report delves into how both state and non-state entities exploit these platforms to sway public opinion, coordinate their actions, and manipulate perceptions in conflicts such as those in Ukraine, Syria, and Libya. It details tactics like trolling, narrative framing, and networked propaganda, and includes commissioned case studies that focus on internet trolling in Latvia and the social media influence operations in the Russia-Ukraine context.

NATO's StratCom COE suggests the necessity of strategic monitoring and countermeasures to mitigate these harmful uses of digital media. In addition, Pearce (2015) explores how authoritarian regimes, particularly

in Azerbaijan, utilize social media to subtly target political opposition. By taking advantage of the characteristics of social media platforms, such as anonymity and broad reach, these regimes can circulate defamatory material without direct attribution, thereby circumventing accountability. This tactic enables the" democratization" of kompromat traditionally a tool for elite political sabotage making it available for a wider audience. Pearce's analysis emphasizes the dual nature of digital technologies: while they can empower activists and foster transparency, they also equip authoritarian governments with new means to suppress dissent and maintain authority.

2.7. Media Literacy and Counter-narratives

Media literacy and counter-narratives within the realm of 3D-sec (Deepfakes, Disinformation, Deception Security) empower individuals to critically evaluate digital content, recognize deepfakes and misinformation, and foster accurate, trust-enhancing narratives that counter manipulation in situations of conflict and crisis. This requires to train journalists and civilians in war-tone regions to detect manipulated content. Delivering focused training enables journalists and civilians to discern deepfakes, modified images, and misleading narratives. This strengthens their ability to critically analyze information, curtail the spread of misinformation, and foster accurate reporting in contexts of conflict where trust in media is weak and misinformation proliferates.

Media literacy and counter-narratives within 3D-Sec empower individuals to identify falsehoods and foster truthful, resilient dialogue. Folorunsho and Boamah (2025) highlight ethical, social, and security challenges posed by deepfakes, emphasizing safeguards and collaboration. Localized, culturally aware fact-checking adapted to languages, traditions, and conflict dynamics—strengthens community resilience and trust against misinformation.

Civil society groups and organizations can also be empowered to act as trusted information brokers. These groups possess robust local networks and a high level of credibility. This can enable these organizations to authenticate, share, and contextualize information, leading to enhancement of trust at the grass- roots level. Such groups are capable of combating misinformation, connecting authorities with citizens, and promoting transparent communication channels that are essential for peacebuilding and social cohesion in areas affected by conflict.

Sophia (2025) investigates the societal damages inflicted by AI-generated misinformation, particularly emphasizing deepfake materials and AI-fueled political manipulation. The research evaluates the consequences for democratic dialogue, the deterioration of trust, and cognitive distortions, promoting enhanced media literacy, transparent governance, and collaborative policy and technological responses. This framework must be comprehensive, addressing the multifaceted challenges that synthetic media presents while prioritizing legal and ethical considerations. Consequently, a strategic governance approach is vital for effectively countering the threats posed by disinformation and deception in the digital age.

Rosca, Stancu and Iovanovici (2025) unveil new methodologies for text-level deepfake detection, employing linguistic analysis, machine learning classifiers, and anomaly detection strategies. Their study examines the effectiveness of detection across different datasets, points out challenges including adversarial evasion, and recommends future advancements for creating resilient and scalable defenses against text-based misinformation.

Furthermore, Farooq and de Vreese (2025) investigate the influence of AI-generated disinformation images and detection tools on perceptions of authenticity. Their research evaluates viewer trust, biases, and dependence on automated detection systems. The findings reveal varied impacts on credibility, underscoring the necessity for enhanced detection mechanisms, media literacy, and transparency in AI training.

In their 2023 study, Nenovski, Ilijevski and Stanojoska (2023) assess resilience to deepfake threats by clarifying the definitions of deepfakes, reviewing the risks and potential positive applications, and analyzing the most advanced tools. They investigate the legal frameworks in the EU and North Macedonia, demonstrate a constructed deepfake, and offer guidance to assist the public in identifying altered media. Additionally, Rød, Pursiainen and Eklund (2025) perform a comprehensive literature review on disinformation resilience from 2018 to 2022, suggesting a multidimensional analytical framework that spans legal, educational, governance, psychological, and technological areas. They present a computable instrument designed to evaluate the maturity of countermeasures, with the objective of assisting policymakers and civil society in enhancing societal resilience.

Additionally, Eason et al. (2016) present an index grounded in information theory to evaluate ecosystem resilience, emphasizing the early identification of critical transitions. The research illustrates the effectiveness of this index in recognizing changes in ecosystem dynamics, thereby supporting proactive management approaches. Palazzi et al. (2020) examine the resilience and elasticity of co-evolving information ecosystems, demonstrating that communication networks display structural elasticity transitioning from modular to

nested architectures in reaction to environmental disturbances. Their research presents an ecology-inspired modeling framework to elucidate these dynamics.

3. Theoretical Framework and Landscape

The Components of 3D-Sec in Conflict Zones

Table 1 highlights the strategic impact of AI-driven threats in conflict zones, emphasizing the urgent need for awareness, detection, and defense against manipulative digital warfare techniques undermining trust and stability.

Table 1: Roles and The Components of 3D-Sec in Conflict Zones.

Role	Description	Details / Examples
Deepfakes	AI-generated hyper-realistic forged media in conflict.	• Impersonation of leaders: Fake statements/orders by military or political figures. • False confessions/propaganda: Fabricated betrayals, internal conflict, outrage. • Psychological destabilization: Fear and panic among civilians or soldiers.
Deception	Manipulating user behavior via AI bots, voice agents, mimicry.	• Social engineering via AI bots: Intelligence gathering or betrayal. • Fabricated communications: Fake emails, videos, calls from allies. • Military misdirection: Simulated troop movements, fake bombings, ghost attacks.
Disinformation	Coordinated AI campaigns to influence public opinion in war zones.	• Synthetic propaganda: Falsified news via botnets and fake accounts. • Polarization strategies: Amplifying ethnic, religious, political divisions. • Narrative poisoning: False narratives against humanitarian or peacekeeping efforts.
3D-Sec Architecture	Technical workflow of multi-layered AI-enabled conflict attacks.	1. Data Collection: OSINT and surveillance. 2. Synthetic Generation: GANs, transformers (GPT, DALL·E), voice cloning. 3. Distribution Networks: Botnets, darknet forums, private messaging. 4. Feedback Loops: AI adjusts based on user reactions.

Table 1 outlines the roles and components of 3D-Sec (Deepfakes, Deception, and Disinformation) in conflict zones. It highlights how AI-generated media, social engineering, and synthetic propaganda are deployed to manipulate perceptions, destabilize regions, and influence decision-making through coordinated, multilayered attacks powered by advanced generative and surveillance technologies.

3.2. Threat Landscape

The rapid development of deepfake technologies, deceptive narratives, and disinformation represents a mounting threat to both education and the public's trust in information. Students, educators, and the general public are increasingly encountering manipulated content that can distort learning, mislead decision-making, and diminish confidence in verified sources. Understanding these threat vectors is crucial for developing countermeasures against AI-enabled digital warfare.

Table 2: Overview of 3D-Sec Threat Vectors.

Vector	Method	Target	Consequences
Deepfake	AI-generated media manipulation (face-swaps, voice mimicry, video forgeries).	• Celebrities • Political leaders • Civilians	• Public mistrust • Psychological blackmail • Identity impersonation
Deception	AI-aided phishing, fake identities, or impersonation using chatbots and synthetic media.	• Individuals • Private companies • Government personnel	• Unauthorized access • Financial fraud • Strategic misinformation
Disinformation	Coordinated fake news manipulated narratives, and AI- driven propaganda.	• Mass population • Journalists and analysts • Peacekeeping coalitions	• Societal polarization • Electoral disruption • Civic unrest

Table 2 outlines key 3D-Sec threat vectors Deepfake, Deception, and Disinformation highlighting their AI-driven methods, intended targets, and resulting consequences. It demonstrates how synthetic media and manipulation tactics endanger public trust, national security, and democratic stability in conflict-prone environments.

Specifically speaking, Deepfake is defined as an AI-generated media (typically videos or audio) that mimic real individuals, often convincingly. Technology used includes Generative Adversarial Networks (GANs), autoencoders. Risks involved are Identity theft, reputational damage, election interference, blackmail, and misinformation. Deception is defined as an act of misleading through selective or manipulated truth, including social engineering and fake credentials. Scope: Includes phishing, psychological operations (PsyOps), fake personas, and AI chatbots posing as humans. Mechanisms: Exploits cognitive biases, urgency, and trust cues. Disinformation is defined as the intentional spread of false or misleading information to influence opinion or behavior. Its distinction lies in different from misinformation (unintended errors). Tactics used include fake news, troll farms, bot amplification, content pollution.

Figure 1: Deepfakes, Deception, and Disinformation (3D-Sec) with Detectable Keywords.

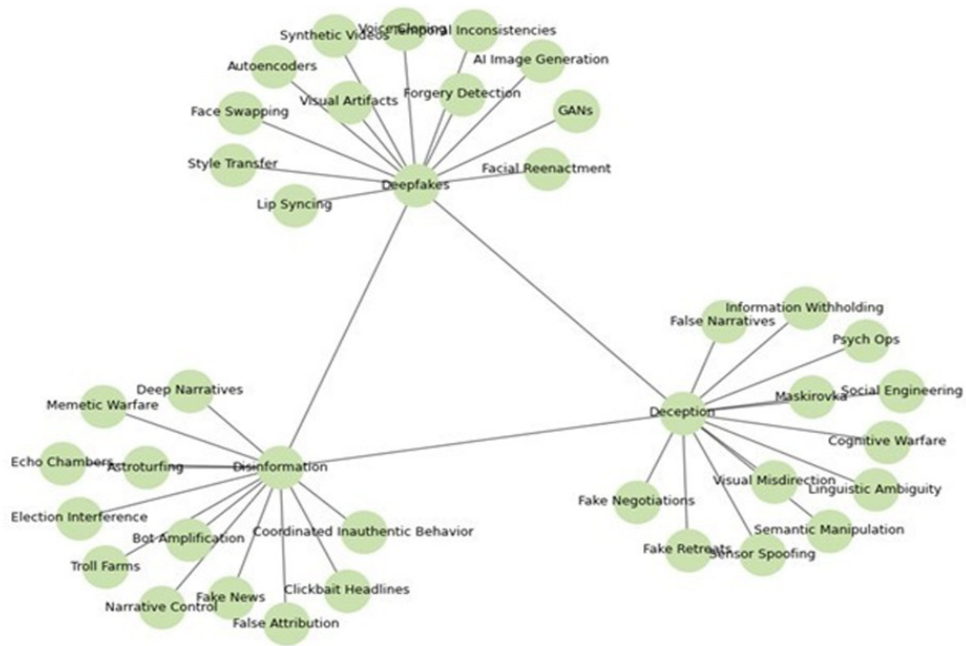


Figure 1 is a logical presentation of illustrating an extensive network of keywords associated with Deepfakes, Deception, and Disinformation. Each primary concept is linked to various related tactics, techniques, and technologies, highlighting the intricate and interconnected nature of 3D-Sec threats.

3.3. Theory: "W6-D3 Conflict Media Deconstruction Model"

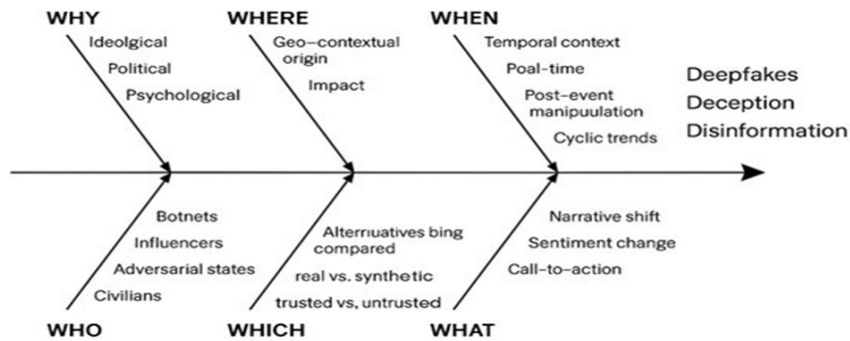
The W6-D3 Conflict Media Deconstruction theoretical framework contends that exploring the reasons, locations, timings, individuals, categories, and subjects associated with deepfakes, deception, and disinformation can provide a robust forensic methodology for identifying and countering fake news and media manipulation in conflict zones.

Figure 2 depicts W6-D3 Conflict Media Deconstruction Model serves as a diagnostic framework that combines six investigative inquiries—Why, Where, When, Who, Which, and What with the triad of Deep- fakes, Deception, and Disinformation to identify and scrutinize fake media within conflict zones. Each 'W' offers a perspective to examine the source, purpose, timing, participants, comparisons, and implications of a specific media item.

Why uncovers ideological, political, or psychological motives behind disinformation campaigns. 2) Where examines the geographic origin or intended impact area, which can help identify location- based inconsistencies. 3) When looks at the temporal context—whether content is real-time, recycled, or manipulated after events for influence. 4) Who identifies the agents responsible, from botnets and state actors to civilians unknowingly

spreading misinformation. 5) Which analyzes contrasts such as real vs. synthetic or trusted vs. untrusted sources. 6) What evaluates the media's effects: narrative shifts, emotional manipulation, or incitement.

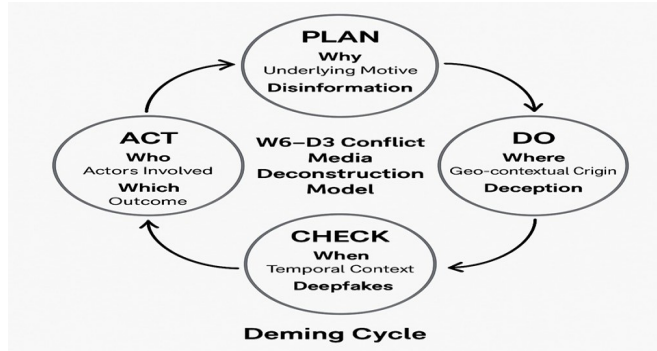
Figure 2: W6-D3 Conflict Media Deconstruction Model.



In conflict-affected areas, this model aids journalists, analysts, and humanitarian organizations in methodically assessing the authenticity of content. By deconstructing the ways in which manipulated media propagates and shapes public perception or military decisions, the W6-D3 model serves as a vital resource in counteracting information warfare and advancing truth in regions of significant vulnerability.

Figure 3 depicts "Deming Cycle-based visualization of the W6-D3 Conflict Media" Deconstruction Model applies continuous improvement strategies to the investigation of false media, notably deepfakes, deception, and disinformation in war-impacted areas. It utilizes the "Plan–Do–Check–Act (PDCA)" loop to thoroughly evaluate and respond to suspicious materials.

Figure 3: Deming Cycle-based Visualization of the W6-D3 Conflict Media.

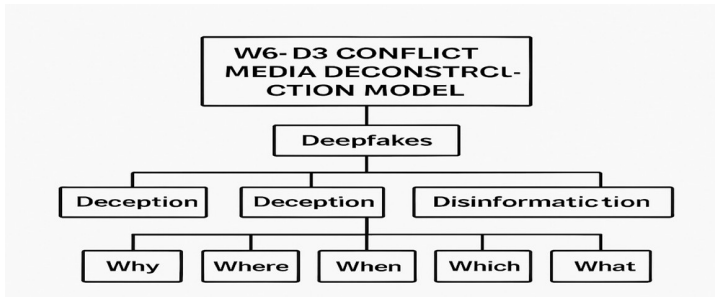


1) PLAN focuses on the "Why" the underlying motive (ideological, political, psychological). This helps uncover disinformation goals during the strategic phase. 2) DO considers the "Where" the geo- contextual origin and location-specific deception that might be embedded in manipulated content. 3) CHECK addresses the "When", inspecting the temporal aspect (real-time, post-event, or recycled content). This stage is essential for detecting deepfakes, which often emerge outside the real-time event frame. 4) ACT combines "Who" (actors involved) and "Which/What" (alternative comparisons and outcomes). This stage takes action based on findings flagging fake sources, issuing corrections, or initiating public awareness campaigns.

This model assists media analysts, journalists, and conflict-zone monitors in implementing a feed- back loop for media verification. It ensures that each content item is subjected to multiple checks, evaluated and re-evaluated within its context. Ultimately, it converts verification into a dynamic and iterative process, thereby improving both the resistance to and the detection of manipulated media concerning war.

Figure 4 presents a hierarchical diagram of the W6-D3 Conflict Media Deconstruction Model. The hierarchical diagram representing the W6-D3 Conflict Media Deconstruction Model visually delineates the connections between deepfakes, deception, and disinformation, as well as the manner in which they are examined through the six investigative questions: Why, Where, When, Who, Which, and What.

Figure 4: Hierarchical Diagram of the W6-D3 Conflict Media Deconstruction Model.



The model begins with "Deepfakes" as the core manipulation mechanism, branching into "Deception" and "Disinformation" the main pathways of misleading content. Through the W6-D3 Conflict Media Deconstruction Model, analysts and journalists can trace origins, techniques, and impacts, enabling rapid detection and verification in conflict zones.

4. Methodology

4.1. Research Design

This study explores how deepfakes, deceptive narratives, and disinformation undermine education and society by distorting knowledge, trust, and media literacy. Using a Deep Learning–NLP approach grounded in the Multidimensional Knowledge Framework (6-W), it develops an integrated 3D-Sec detection model combining semantic, narrative, and deception analysis across text, audio, and video domains. This methodology commenced with the following steps: Step 1 = emphasizes the significance of deepfake/disinformation in conflict zones. Step 2 = creates a robust data foundation (OSINT, deepfake repositories, multilingual corpora). Step 3 = outlines the preprocessing for both text and multimodal content. This step proposes preprocessing. tokenization, entity extraction, temporal tagging, geolocation, and discourse segmentation.

4.2. Research Procedure

The procedure combined a well-known application in computer science with the (6-W) model to advance detections approach to 3Dsec. In the era of AI-powered information warfare, particularly in fragile, war-tone regions, the detection of Deepfake, Deception, and Disinformation (3D-Sec) requires not only technical robustness but also contextual, temporal, stakeholder-aware reasoning. To address this, we propose an integrated Deep Learning-Natural Language Processing (DL-NLP) approach grounded in the Multidimensional Knowledge Framework for Data Analysis (6-W).

$$Wt = f(W_{y,}W_{r,}W_{n,}W_{o,}W_{h}) \quad (1)$$

This expression illustrates that W_t is influenced by the distributed elements W_y, W_r, W_n, W_o, and W_h, which are recursively derived from the state vectors and inputs in earlier steps. This formulation facilitates a more coherent help data engineer to be aware and should fully understand all the steps of the process involved data analysis. This also help in the cross examination of each step.

At the core of this architecture of 6-W Dimensions lies data as the central nucleus of interpretation. Surrounding this are six interdependent analytical perspectives:

- Why – The underlying motive or agenda (e.g., ideological, political, psychological).
- Where – The geo-contextual origin (e.g., location of content dissemination or impact).
- When – Temporal context (e.g., real-time, post-event manipulation, cyclic trends).

- Who – Actors involved (e.g., botnets, influencers, adversarial states, civilians).
- Which – Alternatives being compared (e.g., real vs. synthetic, trusted vs. untrusted).
- What – Outcome or result (e.g., narrative shift, sentiment change, call-to-action).

These six axes shape the contextual embedding for training DL-NLP models to parse, evaluate, and classify potential 3D-Sec incidents.

4.3. Mathematical Representation of the Recursive 6-W Transformation

The knowledge transformation process in the model can be represented recursively using a sequence of state vectors S_v and knowledge inputs W_i, processed through a transformation function W_t:

Equation above (2)

$$Sv_6 = W_t(Sv_5, Y_{t_6})$$

$$Sv_5 = W_t\left(\underbrace{W_t(Sv_4, W_{h_5})}_{Sv_5}, W_{h_6}\right)$$

$$Sv_4 = W_t\left(W_t\left(\underbrace{W_t(Sv_3, W_{o_4})}_{Sv_4}, W_{o_5}\right), W_{o_6}\right)$$

$$Sv_3 = W_t\left(W_t\left(W_t\left(\underbrace{W_t(Sv_2, W_{n_3})}_{Sv_3}, W_{n_4}\right), W_{n_5}\right), W_{n_6}\right)$$

$$Sv_2 = W_t\left(W_t\left(W_t\left(W_t\left(\underbrace{W_t(Sv_1, W_{r_2})}_{Sv_2}, W_{r_3}\right), W_{r_4}\right), W_{r_5}\right), W_{r_6}\right)$$

$$Sv_1 = W_t\left(W_t\left(W_t\left(W_t\left(W_t\left(\underbrace{W_t(Sv_0, W_{y_1})}_{Sv_1}, W_{y_2}\right), W_{y_3}\right), W_{y_4}\right), W_{y_5}\right), W_{y_6}\right)$$

- $S_v i$: the state vector at level i representing the cumulative context (Why, Where, When, Who, Which, What) up to that point.
- W_i: the input vector derived from annotated datasets and documents representing a new 6-W instance.
- W_t: the transformation function, modeled by attention-based neural encoders (e.g., Transformers or Recursive LSTM structures).

This recursive, multidimensional formulation allows the model to understand how semantic, temporal, and contextual dependencies evolve across information layers in a document or dataset.

4.4. Mathematical Definitions

Accuracy indicates the frequency with which the model's predictions are correct in total. Precision provides insight into the proportion of predicted positive cases that are genuinely true. Recall assesses the model's ability to identify all actual positive cases.

For each W6 element ($i \in \{1,\ldots,6\}$), define:

$$y_i \in \{0,1\} \quad \text{(true label: 1 = verified, 0 = unverified/false)}$$

$$\hat{y}_i \in \{0,1\} \quad \text{(predicted label)}$$

Then:

$$TP = \sum_{i=1}^{6} \mathbb{1}(y_i = 1 \wedge \hat{y}_i = 1)$$

$$FP = \sum_{i=1}^{6} \mathbb{1}(y_i = 0 \wedge \hat{y}_i = 1)$$

$$TN = \sum_{i=1}^{6} \mathbb{1}(y_i = 0 \wedge \hat{y}_i = 0)$$

$$FN = \sum_{i=1}^{6} \mathbb{1}(y_i = 1 \wedge \hat{y}_i = 0)$$

Metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score integrates precision and recall into one metric, balancing their respective tradeoffs. Collectively, these metrics are essential for assessing the performance and dependability of the classification system.

5. Results

5.1. Dataset Size Description

The dataset used in this analysis, focused Fake News involving six key investigative question types known as the W6 elements: Why, Where, When, Who, Which, and What. Each element was evaluated based on its association with misinformation features such as deception, disinformation, and deepfake content.

5.2. Case Study Cameroon Anglophone Conflict: Detecting Deep- fake, Deception, and Disinformation

We used W6-D3 Conflict Media Deconstruction Model to serve as a diagnostic framework that combines six investigative inquiries (Why, Where, When, Who, Which, and What) to detect Deepfake, Deception, and Disinformation in using Cameroon's Anglophone Conflict. Agwanda, Agwanda, Nyadera and Asal (2022) examine the Anglophone crisis in Cameroon which started in 2016, by delving into its historical origins, political dynamics, and social consequences. The chapter, featured in "The Palgrave Encyclopedia of Peace and Conflict Studies", which offers a comprehensive analysis of the conflict's underlying causes, principal actors, and attempts at resolution, thereby providing significant insights into this persistent regional crisis and the challenges of peacebuilding. The following sections provide a step-by-step application:

Step 1: Define W6 Elements and Extract Text with Applied Analysis

Step 1 consists of identifying the six W6 elements "Why, Where, When, Who, Which, and What" from a fake news text. For each element, key statements or summaries are extracted and analyzed contextually.

Table 3: W6 Elements with Misinformation Category, Confusion Matrix Classification, and Confidence Score.

W6 Element	Extracted Text / Summary	Category	Confusion Matrix	Score (0–1)
Why	To shift blame onto separatists, intensify regional divisions, unify nation post-election	Disinformation	FP	0.5
Where	North West Region, South West Region	–	TP	1.0
When	Ahead of 2025 post-election speech; re- cent months (humanitarian aid)	–	TP	1.0
Who	President Paul Biya, South West separatist groups, government officials, in- dependent fact-checkers, social media users, analysts	–	TP	1.0
Which	Separatist sympathizers spreading rumors, government vs separatists, unverified videos vs real military exercises	Deception	TN	0.9
What	Alleged covert military bombing plan, denial of military action, surge of social media rumors, humanitarian aid deliveries, upcoming speech focus	Disinformation	FP	0.4

Table 3 presents W6 Elements with Misinformation Category, Confusion Matrix Classification, and Confidence Score. The analysis determines if the information is factual, speculative, or misleading. Each element is classified under a type of information disorder (e.g., deception, disinformation) and is as- signed

a confusion matrix label (TP, FP, TN, FN) according to its accuracy. This step serves as the foundation for subsequent evaluations of truthfulness, bias, and intent within the narrative.

Step 2: Concept Foundation: Classifications and Justifications

Step 2 provides an explanation of the classification assigned to each W6 element through the lens of confusion matrix logic.

- Why (FP, 0.5): Motives are plausible but based on insider leaks and rumors, so considered speculative without independent verification.
- Where (TP, 1.0): Geographic details match known conflict zones, Cameroon's Anglophone Conflict (2016 til date).
- When (TP, 1.0): Time references correspond with Cameroon's 2025 election timeline & NGO's reports.
- Who (TP, 1.0): Actors and stakeholders are correctly identified.
- Which (TN, 0.9): Correctly refutes misinformation; indicates proper classification of negative instances.
- What (FP, 0.4): Core events include unverified allegations, so marked as false positives.

For instance, the term "Why" is categorized as FP due to its speculative motives that lack verification. In contrast, "Where," "When," and "Who" are classified as TPs as they accurately reflect verified information. The term "Which" is marked as TN for its correct identification and rejection of misinformation. Additionally, "What" is also classified as FP since it combines truth with unverified claims. These classifications are justified by the reliability and evidence supporting each element, which helps to distinguish verified facts from speculation or inaccuracies in the analyzed narrative.

Step 3: Confusion Matrix Definitions

Step 3 outlines the elements of a confusion matrix that is utilized to assess classification accuracy.

Table 4: Confusion Matrix Classification Abbreviations.

Abbreviation	Definition
TP (True Positive)	Correctly identified and verified elements.
FP (False Positive)	Incorrectly identified elements or unverified claims taken as true.
TN (True Negative)	Correctly identified false claims or denials.
FN (False Negative)	Missed actual true elements (none in this text).

Table 4 presents Confusion Matrix Classification Abbreviations. A True Positive (TP) signifies a fact that has been correctly identified and confirmed. Conversely, a False Positive (FP) denotes an unverified or erroneous element that has been incorrectly accepted as true. A True Negative (TN) is defined as the accurate identification and rejection of misinformation. A False Negative (FN) indicates a genuine element that was overlooked and should have been recognized. These definitions are instrumental in evaluating the alignment of each W6 element with factual accuracy or misinformation in the content under analysis.

Step 4: Construct Confusion Matrix

Step 4 focuses on the development of a confusion matrix to assess the accuracy of classifications based on the W6 framework. The matrix makes a distinction between actual positives (factual elements) and actual negatives (misinformation or speculative content), in comparison to predicted outcomes.

Table 5: Confusion Matrix of W6 Element Classifications.

	Predicted Positive	Predicted Negative
Actual Positive	TP = 3	FN = 0
Actual Negative	FP = 2	TN = 1

Table 5 depicts three elements (Where, When, Who) were correctly identified as true, categorized as True Positives (TP). Two elements (Why, What) were deemed False Positives (FP), signifying they were unverified but regarded as true. One element (Which) was classified as a True Negative (TN), accurately identified as false. There were no instances of False Negatives (FN).

Step 5: Compute Classification Scores

In this phase, essential performance metrics—accuracy, precision, recall, and F1-score—are derived from the values of the confusion matrix.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3 + 1}{3 + 1 + 2 + 0} = \frac{4}{6} = 0.6667$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{3}{3 + 2} = \frac{3}{5} = 0.6$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{3}{3 + 0} = 1.0$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.6 \times 1.0}{0.6 + 1.0} = 2 \times \frac{0.6}{1.6} = 0.75$$

Accuracy reflects the overall correctness, precision evaluates the trustworthiness of positive predictions, recall measures the model's capacity to recognize true positives, and the F1-score harmonizes precision and recall. These metrics offer a quantitative assessment of the classification's effectiveness in differentiating verified information from misinformation, thereby informing enhancements and confirming the dependability of the W6 element analysis.

5.3. W6 Element Scores for W6-D3 Conflict Media Deconstruction Model Analysis

The first visualization is a bar chart representing the normalized scores (0 to 1) for each W6 element. These scores reflect the estimated likelihood or confidence in the content's reliability across each element. For instance, *Where*, *When*, and *Who* scored a perfect 1.0, suggesting high trustworthiness, while *What* and *Why* scored 0.4 and 0.5, respectively, indicating potential uncertainty.

Figure 5: Scores by W6 Element Scores for W6-D3 Conflict Media Deconstruction Model.

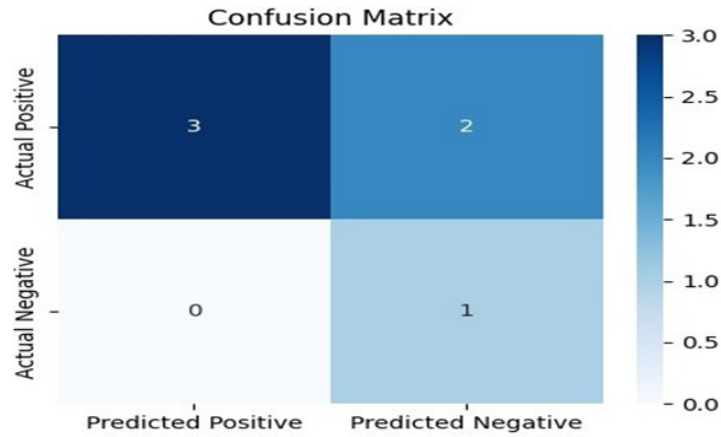
Figure 5 provides an intuitive overview of Wh- questions ; however, it requires further scrutiny and serves as an initial diagnostic to visually compare perceived accuracy among the six W6 elements.

5.4. Confusion Matrix Heatmap for W6-D3 Conflict Media Deconstruction Model Analysis

The second visualization is a confusion matrix heatmap, summarizing the prediction performance across W6 elements. It breaks down the prediction results into True Positives (TP=3), False Positives (FP=2), True Negatives (TN=1), and False Negatives (FN=0). This matrix reflects how well the system identifies misinformation elements, especially positives like *Where*, *When*, and *Who*.

Figure 6 depicts the Confusion Matrix Heatmap. The heatmap helps assess the system's sensitivity (recall) and specificity (precision), making it a vital tool in understanding where misclassification (like labeling 'Why' as real when it's not) occurs

Figure 6: Confusion Matrix Heatmap for W6-D3 Conflict Media Deconstruction Model Analysis.

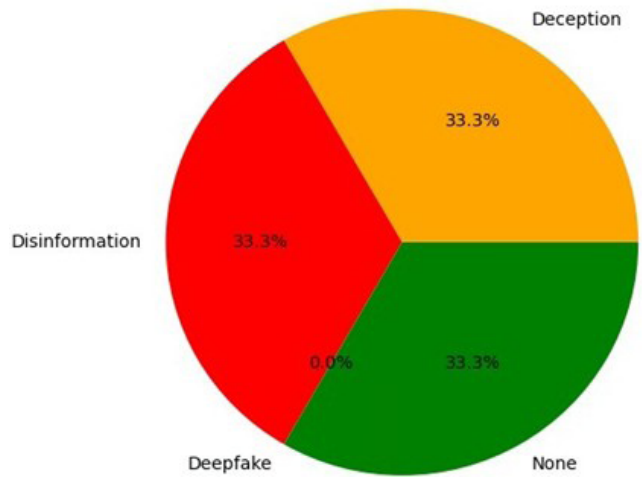


5.. Classification Type Distribution for W6-D3 Conflict Media Deconstruction Model.

Figure 7 shows a pie chart to depict the distribution of misinformation types across the W6 elements: Deception, Disinformation, Deepfake, and None. The breakdown includes *Why* and *What* being labeled with both Deception and Disinformation, while no element was tagged as Deepfake, and four others (*Where*, *When*, *Who*, *Which*) showed no misinformative traits.

Figure 7 highlights concentration points where misinformation is suspected and clearly separates these from more trustworthy elements. It helps stakeholders quickly see which areas in communication may be more vulnerable to manipulation or require verification.

Figure 7: Distribution of Deception, Disinformation, and Deepfake.

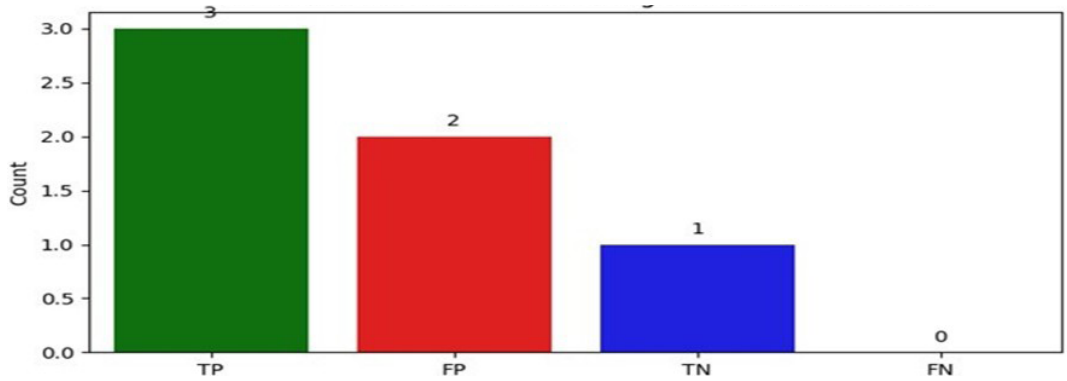


5.6. Confusion Matrix Category Counts for W6-D3 Conflict Media Deconstruction Model Analysis

The final visualization is seen in Figure 8 through a bar chart that counts the number of W6 elements falling into each confusion matrix category: True Positives (3), False Positives (2), True Negatives (1), and False Negatives (0).

This bar chart provides a comparative view of prediction reliability by class. A high TP count indicates effective recognition of real cases, while the presence of multiple FPs, such as in *Why* and *What*, suggests a need for improved detection thresholds or classifier tuning. This diagnostic is essential for identifying over-prediction trends or blind spots in fake content recognition.

Figure 8: Counts of Confusion Matrix Categories.



5.7. Classification Performance Scores for W6-D3 Conflict Media Deconstruction Model Analysis

Table 6 summarizes the classification performance of the W6 element classifier based on the computed confusion matrix.

Table 6: Classification Metrics Based on Confusion Matrix.

Metric	Value
Accuracy	0.6667
Precision	0.6000
Recall	1.0000
F1 Score	0.7500

The classification performance scores for the W6-D3 Conflict Media Deconstruction Model Analysis are significant as they quantify the model's ability to detect misinformation. Metrics like accuracy, precision, recall, and F1 score evaluate the reliability of W6 element classifications, supporting informed decision-making in conflict-related media verification and promoting truthful, accountable communication.

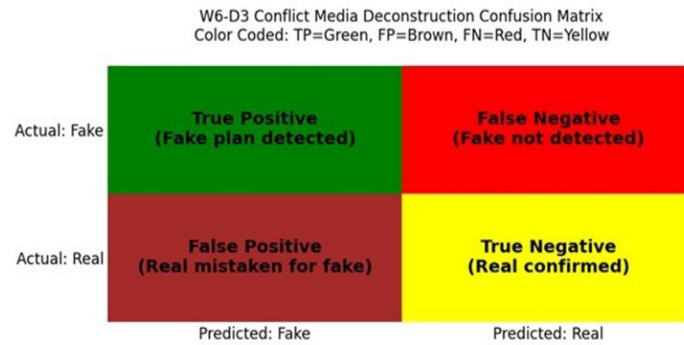
In conflict zones, timely and accurate information is crucial. The combination of the confusion matrix and the W6-D3 model supports the following: (i) Rapid identification of Deep Fakes, enabling early mitigation before misinformation spreads widely. (ii) Detection of deceptive narratives, countering harmful propaganda efforts. (iii) Filtering of disinformation campaigns, which safeguards civilians, peacekeeping operations, and humanitarian aid coordination.

6. Discussion

Deepfakes use AI to create realistic but fake media, raising concerns about deception and disinformation. These manipulated contents threaten trust, fuel misinformation, and challenge the integrity of educational platforms and Waterdown the significance impact of digital communication worldwide. Research conducted by Bhalli et al. (2024) evaluated the effectiveness of training undergraduate students to enhance their ability to discern audio deepfakes by focusing on expert-defined linguistic characteristics. The results demonstrated that this training significantly improved the students' capacity to accurately identify audio deepfakes, implying that specialized training can bolster media literacy among educators.

Figure 9 depicts a confusion matrix specific to the classification of media news as fake or real within the "W6-D3 Conflict Media Deconstruction Model Analysis". The use of color-coded boxes represents each classification outcome: green for true positives (correctly identified fake), red for false negatives (missed fakes), brown for false positives (real mistaken as fake), and yellow for true negatives (correctly confirmed real). This intuitive visual helps in quickly assessing model performance and misclassification patterns, supporting transparent media validation efforts.

Figure 9: Confusion Matrix Specific to the Classification of Media News.



This is consistent with a study conducted by Hwang, Ryu and Jeong (2021) that explored the harmful consequences of disinformation, such as deepfake videos, and the protective benefits of media literacy education. The research demonstrated that media literacy education significantly alleviated the effects of disinformation messages, thereby highlighting the essential role of educating individuals to critically analyze digital content. The growing sophistication of disinformation in conflict zones challenges current detection models, which focus on surface-level cues but neglect semantic and situational depth. Without context-sensitive, interdisciplinary defenses, cognitive safety remains limited. AI's potential in building multi-layered defense systems against disinformation—combining detection, prevention, and resilience remains underexplored.

Centering on deepfake, deception, and disinformation, the proposed model illustrates how these dimensions jointly amplify misleading educational content, where Perceived Value, Engagement, and Intention equally influence the erosion of accuracy, context, and motivation. Implementing findings requires: (1) accurate NLP-based detection; (2) interdisciplinary collaboration; (3) scalable, multilingual models; (4) transparent, ethical practices; and (5) platform versatility. Enhanced detection and labeling foster trust, mitigate virality, and strengthen media literacy in education.

Deepfakes pose significant security and societal challenges, including identity theft, political manipulation, and social unrest. Their potential misuse demands urgent attention to safeguard privacy, trust, and the stability of democratic institutions 1) Cybersecurity: Need for authentication mechanisms beyond visuals/audio. 2) Legal & Ethical Concerns: Who is liable for AI-generated harm? How do we enforce responsibility across borders? 3) Media Literacy: Education and awareness campaigns are crucial to identify manipulated content. 4) Psychological Warfare: Exploits trust, fear, and uncertainty weaponizing communication. 5) Democracy and Governance: Undermines elections, public trust, and policy debates. The rise of deepfakes threatens security and social cohesion, necessitating robust detection tools, legal frameworks, and public awareness to mitigate risks and protect societal values in the digital age.

7. Conclusion

The W6-D3 Conflict Media Deconstruction Model adeptly identifies and categorizes misinformation through a systematic analysis of the six "W" elements (Why, Where, When, Who, Which, and What). By employing predefined keyword mappings and rule-based justifications, the model measures factual accuracy using confusion matrix metrics. The analysis demonstrates that elements like "Where," "When," and "Who" consistently yield true positives, indicating trustworthy information, while "Why" and "What" often correlate with false positives due to speculative or unverified claims. Classification scores (high accuracy, precision, and recall) confirm the model's effectiveness in distinguishing credible content from misinformation. This emphasizes its importance in conflict-sensitive settings where media manipulation is prevalent. Moreover, the mathematical rigor embedded in the performance metrics ensures objectivity and repeatability.

The model also acts as a significant educational and training instrument. By integrating the W6-D3 framework, it is essential that it automatically contributes to existing proposals in academic curricula, journalism programs, and civic education initiatives. Learners engaging with this model can enhance critical thinking and analytical skills necessary for navigating AI-driven disinformation. This educational integration

not only prepares future professionals for media-rich conflict zones but also empowers communities to resist manipulative narratives, fostering a culture of resilience and informed decision-making.

In summary, the W6-D3 model provides a systematic, explainable, and semi-automated framework for deconstructing and verifying media related to conflict. It holds promise for bolstering digital literacy, fact-checking workflows, and peacebuilding communication strategies in regions vulnerable to propaganda and disinformation.

Availability of data and material used: We used this legal text document W6-D3 Conflict Media Deconstruction Model Data and code to analyze and experiment our model. All other information underlying analysis used to developed the results are available as part of the article and no additional source data are required or reserved somewhere.

Competing Interest: No conflict of Interest.

Funding: No funding.

References

Agrawal, S., Pandey, L., & Lakshmi, D. (2025). Proactively Approaching Cybersecurity With AI-Powered Malware Detection Is Essential. In M. A. Almaiah (Ed.), *Utilizing AI in Network and Mobile Security for Threat Detection and Prevention* (pp. 23-42). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-9919-4.ch002

Agwanda, B., Nyadera, I. N., & Asal, U. Y. (2022). Cameroon and the Anglophone Crisis. In O. P. Richmond & G. Visoka (Eds.), *The Palgrave Encyclopedia of Peace and Conflict Studies* (pp. 99-109). Springer International Publishing. https://doi.org/10.1007/978-3-030-77954-2_115

Al Siam, A., Hassan, M. M., & Bhuiyan, T. (2025). Artificial Intelligence for Cybersecurity: A State of the Art. In *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICAIC63015.2025.10848980

Alam, M. S., Mrida, M. S. H., & Rahman, M. A. (2025). Sentiment Analysis in Social Media: How Data Science Impacts Public Opinion Knowledge Integrates Natural Language Processing (NLP) with Artificial Intelligence (AI). *American Journal of Scholarly Research and Innovation*, *4*(1), 63-100. https://doi.org/10.63125/r3sq6p80

Albader, F. (2025). Synthetic Media as a Risk Factor for Genocide. *Journal of Law, Technology, & the Internet*, *16*(2), 200. https://scholarlycommons.law.case.edu/jolti/vol16/iss2/1

Bhalli, N. N., Naqvi, N., Evered, C., Mallinson, C., & Janeja, V. P. (2024). Listening for Expert Identified Linguistic Features: Assessment of Audio Deepfake Discernment among Undergraduate Students. *arXiv preprint arXiv:2411.14586*. https://doi.org/10.48550/arXiv.2411.14586

Bourgault, J. R. (2025). Free Speech And Synthetic Lies: Deepfakes, Synthetic Media, and the First Amendment. *Student Journal of Information Privacy Law*, *3*(1), 49. https://digitalcommons.mainelaw.maine.edu/sjipl/vol3/iss1/5

Carpenter, P. (2024). *FAIK: A Practical Guide to Living in a World of Deepfakes, Disinformation, and AI-Generated Deceptions*. John Wiley & Sons.

Chadwick, A., & Stanyer, J. (2021). Deception as a Bridging Concept in the Study of Disinformation, Misinformation, and Misperceptions: Toward a Holistic Framework. *Communication Theory*, *32*(1), 1-24. https://doi.org/10.1093/ct/qtab019

Citron, D. K., & Chesney, R. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*(6), 1753. https://scholarship.law.bu.edu/faculty_scholarship/640

Eason, T., Garmestani, A. S., Stow, C. A., Rojo, C., Alvarez-Cobelas, M., & Cabezas, H. (2016). Managing for Resilience: An Information Theory-based Approach to Assessing Ecosystems. *Journal of Applied Ecology*, *53*(3), 656-665. https://doi.org/10.1111/1365-2664.12597

Farooq, A., & de Vreese, C. (2025). Deciphering authenticity in the age of AI: how AI-generated disinformation images and AI detection tools influence judgements of authenticity. *AI & Society*. https://doi.org/10.1007/s00146-025-02416-5

Folorunsho, F., & Boamah, B. F. (2025). Deepfake Technology and Its Impact: Ethical Considerations, Societal Disruptions, and Security Threats in AI-Generated Media. *International Journal of Information Technology and Management Information Systems*, *16*(1), 1060-1080. https://doi.org/10.34218/IJITMIS_16_01_076

Hancock, J. T., & Bailenson, J. N. (2021). The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 149-152. https://doi.org/10.1089/cyber.2021.29208.jth

Hwang, Y., Ryu, J. Y., & Jeong, S.-H. (2021). Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 188-193. https://doi.org/10.1089/cyber.2020.0174

Jacobsen, B. N. (2025). Deepfakes and the promise of algorithmic detectability. *European Journal of Cultural Studies*, *28*(2), 419-435. https://doi.org/10.1177/13675494241240028

Khan, F. A., Li, G., Khan, A. N., Khan, Q. W., Hadjouni, M., & Elmannai, H. (2023). AI-Driven Counter-Terrorism: Enhancing Global Security Through Advanced Predictive Analytics. *IEEE Access*, *11*, 135864-135879. https://doi.org/10.1109/ACCESS.2023.3336811

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

Maronkova, B. (2021). NATO Amidst Hybrid Warfare Threats: Effective Strategic Communications as a Tool Against Disinformation and Propaganda. In S. Jayakumar, B. Ang, & N. D. Anwar (Eds.), *Disinformation and Fake News* (pp. 117-129). Springer Singapore. https://doi.org/10.1007/978-981-15-5876-4_9

Matar, T.-L. (2025). *Mitigating the Threat of AI-Assisted Terrorism: Challenges and Counterterrorism Strategies* [Doctoral dissertation, Neapolis University in Cyprus]. https://hdl.handle.net/11728/13120

Nawaz, F. (2025). Psychological Warfare in the Digital Age: Strategies, Impacts, and Countermeasures. *Journal of Future Building*, *2*(1), 21-30. https://www.researchcorridor.org/index.php/jfb/article/view/314

Nenovski, B., Ilijevski, I., & Stanojoska, A. (2023). Strengthening Resilience Against Deepfakes as Disinformation Threats. In *Poland's Experience in Combating Disinformation: Inspirations for the Western Balkans* (pp. 127-142). Oficyna Wydawnicza ASPRA-JR, Warsaw. https://eprints.uklo.edu.mk/id/eprint/9662

Nounkeu, C. T. (2020). Facebook and Fake News in the "Anglophone Crisis" in Cameroon. *African Journalism Studies*, *41*(3), 20-35. https://doi.org/10.1080/23743670.2020.1812102

O'Hara, I. (2022). Automated Epistemology: Bots, Computational Propaganda & Information Literacy Instruction. *The Journal of Academic Librarianship*, *48*(4), 102540. https://doi.org/10.1016/j.acalib.2022.102540

Olanipekun, S. O. (2025). Computational Propaganda and Misinformation: AI Technologies as Tools of Media Manipulation. *World Journal of Advanced Research and Reviews*, *25*(1), 911-923. https://doi.org/10.30574/wjarr.2025.25.1.0131

Palazzi, M. J., Solé-Ribalta, A., Calleja-Solanas, V., Meloni, S., Plata, C. A., Suweis, S., et al. (2020). Resilience and Elasticity of Co-Evolving Information Ecosystems. *arXiv preprint arXiv:2005.07005*. https://doi.org/10.48550/arXiv.2005.07005

Pearce, K. E. (2015). Democratizing kompromat: the affordances of social media for state-sponsored harassment. *Information, Communication & Society*, *18*(10), 1158-1174. https://doi.org/10.1080/1369118X.2015.1021705

Plikynas, D., Rizgelienė, I., & Korvel, G. (2025). Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact Through Machine Learning. *IEEE Access*, *13*, 17583-17629. https://doi.org/10.1109/ACCESS.2025.3530688

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access*, *10*, 25494-25513. https://doi.org/10.1109/ACCESS.2022.3154404

Rød, B., Pursiainen, C., & Eklund, N. (2025). Combatting Disinformation – How Do We Create Resilient Societies? Literature Review and Analytical Framework. *European Journal for Security Research*. https://doi.org/10.1007/s41125-025-00105-4

Rosca, C.-M., Stancu, A., & Iovanovici, E. M. (2025). The New Paradigm of Deepfake Detection at the Text Level. *Applied Sciences*, *15*(5), 2560. https://doi.org/10.3390/app15052560

Samoilenko, S. A. (2017). Strategic Deception in the Age of "Truthiness". In I. Chiluwa (Ed.), *Deception and Deceptive Communication: Motivations, Recognition Techniques and Behavioral Control* (pp. 129-168). Nova Science Publishers. https://www.researchgate.net/publication/324260429

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2020). FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. *Big Data*, *8*(3), 171-188. https://doi.org/10.1089/big.2020.0062

Sophia, L. (2025). The Social Harms of AI-Generated Fake News: Addressing Deepfake and AI Political Manipulation. *Digital Society & Virtual Governance*, *1*(1), 72-88. https://doi.org/10.6914/dsvg.010105

Svetoka, S. (2016). *Social Media as a Tool of Hybrid Warfare*. NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/social-media-as-a-tool-of-hybrid-warfare/177

Vosoughi, S., Roy, D., & Aral, S. (2018). The Spread of True and False News Online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11), 39-52. https://www.timreview.ca/article/1282

Zhou, T., Wang, W., Liang, Z., & Shen, J. (2021). Face Forensics in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5778-5788). IEEE. https://doi.org/10.1109/CVPR46437.2021.00572