International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 08 | Aug 2026 www.irjet.net p-ISSN: 2395-0072

Deepfake Technology in Information Warfare: A Multimodal Detection and Automated Counter-Propaganda Framework

Capt. Premdeep Dagar

[1]*Student Officer, Military College of Telecommunication Engineering (MCTE), Mhow, Madhya Pradesh, India*

Abstract – *Advances in generative deep learning have made photorealistic image, video, and voice synthesis widely accessible, turning deepfakes into a practical instrument of information warfare and computational propaganda. Existing tooling largely treats deepfake detection and disinformation-response as separate problems. This paper presents a working prototype that unifies both: a multimodal deepfake detector (image, video, and audio) is combined with a sentiment- and polarization-analysis module over simulated social media reaction, and the two are fused into an automated Information Warfare Threat Score together with drafted counter-propaganda messaging (community notes, headline rebuttals, and verified talking points). Every machine-learning-dependent stage is paired with a deterministic, dependency-light fallback, so the system is fully reproducible without a Graphics Processing Unit (GPU) or model downloads, while transparently reporting which backend (pretrained model vs. heuristic) produced each score. The pipeline is exposed through three interchangeable delivery surfaces a command-line script, a Representational State Transfer Application Programming Interface (REST API) built with FastAPI, and a native desktop Graphical User Interface (GUI) built with Tkinter that all share the same underlying detection and analysis code. We describe the system architecture, present illustrative results from an end-to-end run in both heuristic and real-model configurations, and discuss the limitations and ethical considerations of automating counter-disinformation response.*

Key Words: Deepfake detection, Information warfare, Computational propaganda, Multimodal forensics, Sentiment analysis, Disinformation, Counter-propaganda, Synthetic media

I. INTRODUCTION

Generative adversarial networks (GANs) [1] and, more recently, diffusion- and transformer-based generators have made it possible to synthesize faces, voices, and full video sequences that are difficult for both humans and automated systems to distinguish from authentic recordings. This capability is popularly termed "deepfake" technology. It has moved from a research curiosity to a documented instrument of political manipulation, fraud, and information warfare [2], [4].

Two properties of the modern information environment make synthetic media especially dangerous as a propaganda tool. First, false and emotionally charged content is empirically known to spread farther, faster, and more broadly than accurate content on social media platforms [6]. Second, the manipulation is frequently not fully automated it is computational propaganda, a hybrid of algorithmic amplification (bots, recommendation systems, coordinated accounts) and human curation designed to shape public opinion at scale [7]. A convincing deepfake dropped into a polarized, high-engagement social conversation can therefore cause damage long before any forensic analysis is completed, and long before a correction even a well-sourced one can catch up to the false claim's reach.

Most existing academic and open-source tooling addresses only one half of this problem: detecting whether a given piece of media is manipulated [2], [3]. Comparatively little tooling connects a detection verdict to the actual social dynamics of the content it concerns, i.e., how far it has spread, how angry or polarized the reaction is, and what an evidence-based counter-message should say. This paper describes a prototype system, developed for the project "Deepfake Technology in Information Warfare – Detection and Counter-Propaganda Strategies," that closes this loop end-to-end.

The contributions of this work are:

(1) A multimodal deepfake detector covering image, video, and audio, backed by pretrained HuggingFace classification models with a fully deterministic, dependency-light heuristic fallback for each modality.

(2) An information-warfare sentiment analyzer that goes beyond positive/negative polarity to estimate anger, verification-seeking behavior, amplification behavior, audience polarization, and susceptibility to the content.

(3) A threat-scoring and counter-propaganda generation module that fuses detection confidence with audience-reaction metrics into a single 0–100 Information Warfare Threat Score and auto-drafts a community note, a headline rebuttal, verified talking points, and an audience-engagement strategy.

(4) A zero-setup reproducibility design: every ML dependency is wrapped so the full pipeline runs immediately on a machine with no GPU and no model downloads, while a "backend" field on every result discloses whether a real model or a heuristic produced it.

(5) Three parallel delivery surfaces (CLI, REST API, desktop GUI) built on one shared codebase, so the same detection logic can be scripted, integrated into other systems, or operated interactively.

II. RELATED WORK

A. Generation and Detection of Synthetic Visual Media

The generative-adversarial-network framework of Goodfellow et al. [1] underlies most modern face-swap and face-reenactment techniques. Rössler et al. [2] introduced FaceForensics++, a benchmark of over 1.8 million manipulated facial images spanning four manipulation methods, and showed that domain-specific CNN classifiers substantially outperform human observers at detecting them. Tolosana et al. [3] provide a comprehensive survey of face-manipulation categories (entire-face synthesis, identity swap, attribute manipulation, expression swap) and the corresponding detection literature, noting that detector generalization across manipulation methods and compression levels remains an open problem a limitation our system addresses in practice by disclosing backend and confidence rather than presenting a single opaque verdict.

B. Synthetic and Spoofed Audio

The ASVspoof challenge series [5] established standardized datasets and evaluation protocols for detecting synthetic, voice-converted, and replayed speech attacks against speaker verification systems, and remains the reference benchmark for audio anti-spoofing research. Self-supervised speech representation models such as wav2vec 2.0 [12] have since become common backbones for downstream audio-classification tasks, including synthetic-speech detection, by providing strong pretrained acoustic features without requiring large labeled corpora.

C. Deepfakes as an Information-Warfare Instrument

Chesney and Citron [4] provide the foundational legal and policy analysis of deepfakes as a threat to privacy, democracy, and national security, surveying technological, legal, and market-based responses. Vosoughi et al. [6] empirically demonstrated, across roughly 126,000 verified Twitter stories, that false news reaches significantly more people and spreads significantly faster than true news — a finding directly motivating our decision to score reach and audience polarization alongside raw detection confidence rather than treating detection as sufficient on its own. Woolley and Howard [7] frame this dynamic within computational propaganda: the combination of automation, algorithmic amplification, and human curation used to manipulate public opinion at scale across both democratic and authoritarian contexts.

D. Sentiment and Social-Reaction Analysis

VADER [8] remains a widely used lexicon- and rule-based sentiment analyzer tuned specifically for short, informal social-media text, and is used in this system as an intermediate fallback layer. Transformer-based sentiment classifiers fine-tuned on tweets, such as the RoBERTa-based [9] models trained under the TweetEval benchmark [10], provide substantially higher accuracy on social-media-specific sentiment tasks and are used as the primary sentiment backend when available.

E. Positioning of This Work

Prior work largely treats (i) deepfake forensics, (ii) misinformation-diffusion measurement, and (iii) sentiment analysis of online reaction as separate research threads. The system described here is, to our knowledge, a rare example of an openly documented, end-to-end prototype that mechanically fuses all three into a single actionable score and drafts a corresponding counter-messaging response, while remaining fully runnable without specialized hardware.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

The overall pipeline is presented in Fig-1. Media input (image, video, or audio) is routed to the appropriate detector, simulated social-media posts are routed to the sentiment analyzer, and both outputs are fused by the Threat Report Generator into a numeric threat score, a categorical threat level, and drafted counter-propaganda content.

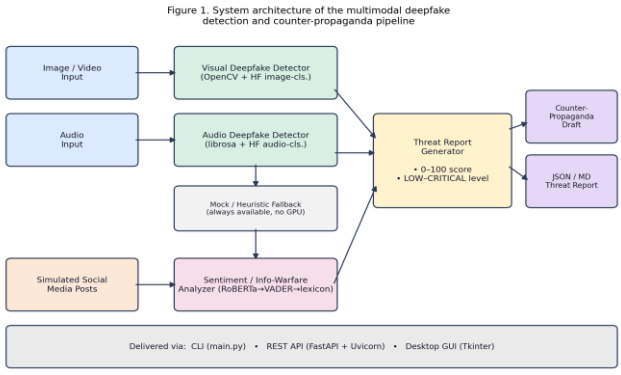

Fig -1: System architecture of the multimodal deepfake detection and counter-propaganda pipeline

A. Multimodal Deepfake Detection

For image and video input, OpenCV samples up to eight evenly-spaced frames (video) or loads the image directly. Each frame is classified by a pretrained HuggingFace image-classificationpipeline (dima806/deepfake_ vs_real_ image_detection by default [13]); per-frame fake-

probabilities are averaged into a single confidence score, thresholded at 0.5 to produce a REAL/FAKE label. For audio input, the waveform is loaded and classified by a pretrained HuggingFace audio-classification pipeline (MelodyMachine/Deepfake-audio-detection-V2 by default [14]), analogous in role to wav2vec2-based synthetic-speech classifiers built on the representation-learning approach of [12], producing an AUTHENTIC/SYNTHETIC label.

B. Deterministic Fallback for Reproducibility

Every call into torch, transformers, opencv, or librosa is wrapped in exception handling. If a package is missing, a model fails to download, or no GPU is available, the system falls back to a fast, deterministic heuristic: image frames are scored via Laplacian-variance edge statistics (a proxy for compression/generation artifacts), and audio is scored via spectral-flatness and zero-crossing-rate variance (a proxy for the unnaturally regular prosody of some synthetic voices). These heuristics are explicitly not validated forensic methods; they exist solely so the full pipeline is reproducible on any machine, including one with no internet access. Every result carries a backend field (e.g.,huggingface:dima806/deepfake_vs_real_image_detection or mock) so downstream consumers always know which method produced a given score.

C. Information-Warfare Sentiment Analysis

Simulated social-media posts (platform, author, text, engagement count) are scored along five axes. Sentiment polarity uses a three-tier cascade: a RoBERTa sentiment model fine-tuned on tweets [9], [10] is attempted first; if unavailable, VADER [8] is used; if that is also unavailable, a small built-in lexicon provides a final fallback. In parallel, three custom lexicon-based signals are computed directly from the text: an anger/outrage signal, a verification-seeking signal (words such as “verify,” “artifact,” “debunk,” “forensic”), and an amplification signal (words such as “share,” “spread,” “everyone”). These are aggregated into two composite metrics: a polarization score, which is high when both the amplifying camp and the skeptical camp are simultaneously large (genuine audience disagreement, not simply one-sided reaction), and a susceptibility score, which estimates how much the audience is primed to share content without first verifying it.

D. Threat Scoring and Counter-Propaganda Generation

The Threat Report Generator combines three weighted components into a 0–100 Information Warfare Threat Score: deepfake detection confidence (weight 0.45), a log-scaled engagement/reach proxy derived from total post engagement (weight 0.20), and the audience polarization score (weight 0.35). The resulting score is mapped to one of four bands LOW (0–25), MODERATE (25–50), HIGH (50–75), or CRITICAL (75–100) which in turn sets the recommended response urgency. Conditioned on the detector verdict and the susceptibility score, the system drafts a short community-note-style rebuttal, a one-line headline rebuttal, a set of verified-rebuttal talking points (e.g., requesting original source files, cross-referencing official channels), and an audience-engagement strategy tailored to whether the audience is share-first, mixed, or already verification-minded.

IV. IMPLEMENTATION

A. Technology Stack

The prototype is implemented in Python 3.10 and above, using PyTorch and HuggingFace Transformers for model inference, OpenCV for video frame extraction, librosa for audio feature analysis, pandas for data handling, FastAPI/Uvicorn for the REST API layer, and Tkinter for the desktop GUI. All components are pure-Python and run on CPU; GPU acceleration is used automatically when available but is not required.

B. Delivery Surfaces

Three interchangeable entry points share the same detector, analyzer, and report generator modules. The command-line script generates synthetic sample media if none is supplied and runs the full pipeline in one pass, printing a threat report. The REST API loads each detector/analyzer once at startup as a singleton and exposes granular endpoints (/detect/image, /detect/video, /detect/audio, /analyze/sentiment, /report/generate) as well as a single convenience endpoint (/pipeline/full) that accepts an uploaded media file and an optional JSON array of social posts, returning a complete threat report with interactive documentation at /docs. The desktop GUI is a pure HTTP client built on Tkinter it performs no detection logic itself, instead calling the REST API in a background thread so the interface remains responsive while pretrained models load or run inference, and presents results across Threat Summary, Counter-Propaganda, Markdown, and Raw JSON tabs.

C. Auditability

Every generated report is persisted to disk as both JSON and Markdown and assigned a report identifier, retrievable later via the API. Uploaded media files are retained under a dedicated uploads directory rather than deleted after inference, supporting downstream human forensic review.

V. ILLUSTRATIVE CASE STUDY

To validate the end-to-end pipeline, we ran the system against a procedurally generated test image and a

synthesized test audio clip (used only to exercise the pipeline, not as forensic ground truth) alongside a hand-authored 15-post simulated social-reaction dataset engineered to contain a mix of angry amplification, calm skepticism, and forensic commentary, with total engagement of 53,270 across posts. Table-1 reports two independent end-to-end runs: one using the deterministic heuristic backends only, and one using the pretrained HuggingFace models.

Table -1: End-to-end case study heuristic vs. pretrained-model backends

Metric	Heuristic (mock) mode	Pretrained-model mode
Visual detector backend	mock (Laplacian-variance heuristic)	huggingface: dima806/deepfake_vs_real_image_detection
Visual verdict (confidence)	REAL (47.9%)	FAKE (55.0%)
Audio detector backend	mock (spectral-flatness heuristic)	huggingface: MelodyMachine/Deepfake-audio-detection-V2
Audio verdict (confidence)	SYNTHETIC (56.6%)	AUTHENTIC (8.7%)
Sentiment backend	VADER	VADER*
Polarization score	0.48	0.48
Susceptibility score	0.37	0.37
Information Warfare Threat Score	47.5 / 100	70.8 / 100
Threat level	MODERATE	HIGH

*In this configuration DEEPFAKE_USE_REAL_MODELS gated only the deepfake detectors; the lightweight VADER sentiment backend, which requires no model download, was active in both runs.

Both runs completed without error and produced a complete, structured report (JSON and Markdown) in under two seconds for the heuristic run; the pretrained-model run additionally required a one-time download of the classification models. The visual and audio verdicts differ between runs because the procedurally generated test image and synthesized test tone carry no genuine deepfake signal in either direction this is expected and illustrates exactly the behavior the backend-disclosure design is intended to surface: a consumer of the report can see which method produced a verdict and weight its reliability accordingly, rather than treating every score as equally authoritative. In the pretrained-model run, the automatically drafted counter-propaganda output included a community-note-style rebuttal flagging the image as showing "manipulation-consistent artifacts (confidence 55%)" and recommending the audience "verify with a primary source," together with four verified-rebuttal talking points and a HIGH-urgency audience-engagement strategy demonstrating that the threat-scoring and messaging-generation stages correctly propagate detector uncertainty into their recommendations rather than presenting either verdict as certain.

VI. DISCUSSION

A. Strengths

The dual real-model/heuristic design gives the system two properties that are often in tension in academic prototypes: it can use state-of-the-art pretrained classifiers when available, while remaining fully installable and runnable in constrained or offline environments for teaching, evaluation, or rapid triage. Fusing detection confidence with audience-reaction metrics, rather than reporting detection in isolation, more directly reflects the actual risk surface described in the computational-propaganda and misinformation-diffusion literature [6], [7]: a low-confidence fake with a small, skeptical audience poses a materially different threat than the same confidence score attached to a highly polarized, rapidly amplifying conversation.

B. Limitations

The heuristic fallbacks (Laplacian variance, spectral flatness, lexicon counting) are intentionally simple and are not validated forensic techniques; they should not be used as evidence in any real moderation, legal, or security decision. The pretrained models used are general-purpose community models rather than models trained and evaluated on a held-out benchmark such as FaceForensics++ [2] or ASVspoof [5], so the accuracy figures in Table-1 describe a functional demonstration rather than a calibrated performance evaluation. The threat-scoring weights (0.45 / 0.20 / 0.35) are heuristically chosen for illustrative balance and have not been empirically calibrated against real-world outcome data. The lexicon-based anger, verification-seeking, and amplification signals are English-only and vocabulary-limited, and the social-reaction dataset used in the case study is hand-authored rather than sampled from a live platform.

C. Ethical Considerations

Automating counter-propaganda drafting carries a distinct risk: an overconfident or poorly worded automated rebuttal can itself become a source of misinformation, particularly if a detector's mock-heuristic backend is mistaken for its real-model backend. For this reason, the system always surfaces which backend produced a given verdict and frames every drafted rebuttal as provisional ("automated analysis indicates,"

"recommend manual forensic review") rather than as a final determination. We regard human review as a required step before any drafted counter-messaging is published, and recommend the system be positioned as a triage and drafting aid rather than an autonomous fact-checking authority.

D. Future Work

Priority extensions include benchmarking the visual and audio detectors against FaceForensics++ [2] and ASVspoof [5] to obtain calibrated accuracy figures, adding explainability output (e.g., Grad-CAM frame overlays) so a human reviewer can see which regions drove a verdict, extending sentiment analysis to multiple languages, connecting the sentiment analyzer to live social media APIs rather than simulated data, and conducting a controlled user study to measure whether the automatically drafted counter-messaging measurably changes audience belief or sharing behavior compared to a generic correction.

VII. CONCLUSION

This paper presented a working, reproducible prototype that unifies multimodal deepfake detection, social-media sentiment and polarization analysis, and automated counter-propaganda drafting into a single Information Warfare Threat Report, delivered through a command-line interface, REST API, and desktop GUI built on shared code. By pairing every ML-dependent stage with a transparent, deterministic fallback, the system remains runnable without specialized hardware while disclosing exactly which method produced each score. An end-to-end case study shows the pipeline producing coherent, internally consistent threat assessments and draft counter-messaging in both configurations. The largest open problems are benchmark-calibrated accuracy evaluation and empirical validation of the counter-propaganda messaging's real-world effectiveness, both identified as priorities for future work.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Premdeep Dagar: Conceptualization, Methodology, Investigation, Software, Validation, and writing – original draft, review and editing.

DECLARATION OF COMPETING INTEREST

The author declares that he has no known financial or personal conflicts of interest that could have influenced the research presented in this article.

DECLARATION OF GENERATIVE AI USE

During manuscript preparation, ChatGPT and Perplexity were used for paraphrasing assistance. After using these services, the author reviewed, revised, and self-edited the content as required, taking complete accountability for the content of the published article.

REFERENCES

1) I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. 27th Int. Conf. Neural Information Processing Systems (NIPS), vol. 2, 2014, pp. 2672–2680.

2) A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 1–11.

3) R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131–148, 2020.

4) R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California Law Review, vol. 107, pp. 1753–1819, 2019.

5) M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in Proc. Interspeech, Graz, Austria, 2019.

6) S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

7) S. C. Woolley and P. N. Howard, Eds., Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media. Oxford, U.K.: Oxford Univ. Press, 2018.

8) C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in Proc. Int. AAAI Conf. Web and Social Media (ICWSM), vol. 8, no. 1, 2014, pp. 216–225.

9) Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.

10) F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in Findings of the Assoc. Comput. Linguistics: EMNLP 2020, 2020, pp. 1644–1650.

11) M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc.

36th Int. Conf. Machine Learning (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.

12) A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Proc. 34th Conf. Neural Information Processing Systems (NeurIPS), 2020.

13) dima806, "deepfake_vs_real_image_detection" [Online image-classification model]. Hugging Face Model Hub. Available:https://huggingface.co/dima806/deepfake_vs_real_image_detection

14) MelodyMachine, "Deepfake-audio-detection-V2" [Online audio-classification model]. Hugging Face ModelHub.Available:https://huggingface.co/MelodyMachine/Deepfake-audio-detection-V2

15) CardiffNLP, "twitter-roberta-base-sentiment-latest" [Online sentiment-classification model]. Hugging Face ModelHub.Available:https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest