Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 6, 2026

THE GENERATION-DETECTION GAP: A STUDY OF DEEPFAKE PROLIFERATION AND RESPONSE IN THE GULF REGION DURING THE 2026 CONFLICT

Mr. Zafar Ali[*1], Ms. Noor Ul Ain[2], Dr. Munwar Ali Kalwar[3]

*[*1]PhD Scholar, School of Communication Studies, University of the Punjab, Lahore, Pakistan*
[2]M. Phil Scholar, English Department, University of Lahore, Pakistan
[3]Deputy Director, International Institute of Islamic Economics (IIIE), International Islamic University, H-10, Islamabad, Pakistan

*[*1]zafaralizfr@gmail.com*

DOI: https://doi.org/10.5281/zenodo.21278827

Keywords
Deepfakes, Synthetic Media, Detection Latency, Gulf Conflict, Arabic, Misinformation, Platform Governance

Article History
Received: 25 April 2026
Accepted: 04 June 2026
Published: 21 June 2026

Corresponding Author: *
Mr. Zafar Ali

Abstract

This mixed-methods study examined the generation-detection gap for deepfakes circulating across Gulf Cooperation Council states during the active combat phase of the 2026 Gulf conflict. The generation-detection gap was defined as the proportion of high-impact synthetic content exceeding 1,000 interactions within 72 hours that was not flagged by any detection system within 24 hours of posting. We analyzed a corpus of 1,912 public audiovisual items posted between February 1 and April 10, 2026, applied four detection systems to the full corpus, and conducted semi-structured interviews with 14 journalists, platform trust-and-safety staff, regulators, and forensic analysts. Results showed that 53.7% of high-impact items were not flagged within 24 hours. The gap was widest for audio content (74.8%) and items targeting political leaders (58.6%). Median time to 1,000 interactions was 2.1 hours, while median detection latency was 4.8 hours. Logistic regression indicated that audio modality (OR = 0.32, 95% CI [0.24, 0.42]) and political targets {OR = 0.61, 95% CI (0.49, 0.76)} significantly reduced odds of detection, whereas higher 24-hour engagement modestly increased odds (OR = 1.08, 95% CI [1.02, 1.14]). Items not detected at 24 hours had a median final engagement that was 3.2 times higher than detected items. The data from the interviews indicated that the three factors that impeded the process of latency were the absence of real-time Gulf Arabic audio detectors, the difference between the capacity for generation and that of forensic work, and the difficulty of attribution for content with heads of state. The results indicate that the difference is technical and procedural, and directly impacts harm. To minimize the time period during which synthetic media influences discourse without labels, there is a need to detect targets within sub-hours, to invest in models for each dialect, and to establish protocols for sharing information during wartime.

Introduction

Generative artificial intelligence (AI) is a rapidly evolving technology that generates content at an unprecedented speed and scale, ranging from highly realistic synthetic text, images, audio, and video. Generative AI's potential in education,

healthcare, and creative fields is significant, but the technology has also raised concerns about misinformation, disinformation, and information manipulation, especially in times of political unrest and armed conflict. Deepfakes, which are AI-generated or AI-manipulated audiovisual media, are among these advances and have become one of the most serious threats to information integrity, democratic discourse, and national security (Mirsky & Lee, 2021; Chesney & Citron, 2019).

These conditions are particularly conducive to the spread of synthetic media in armed conflicts, as information spreads rapidly and there may be a delay in verification. In these situations, the misinformation in movies, audio files, or video clips can be harmful to public health, manipulate political views, instil fear, and obscure military and diplomatic decisions. Previous research has revealed how deepfakes can leverage cognitive biases, enhance political polarization, and undermine the ability of citizens to identify real information from manipulated information (Vaccari & Chadwick, 2020; Wardle & Derakhshan, 2017).

As the deepfake generation has progressed, deepfake detection tools have not caught up. Many detection algorithms have shown high accuracy on benchmark datasets like FaceForensics++ and Celeb-DF, but their usefulness is often limited when used to detect real-world social media content that is compressed by the platforms, reposted, has domain shifts, and is continuously updated with generative models (Tolosana et al., 2020; Verdoliva, 2020; Mirsky & Lee, 2021). This is a significant operational challenge for governments, journalists, and technology platforms trying to detect manipulated media in the midst of constantly shifting crises.

In multilingual areas like the Gulf, where social media has reached high penetration levels, cross-border information exchange has become commonplace, and multiple Arabic dialects are used, the challenge is heightened in detecting and moderating content automatically. Past studies show that deepfake detection systems have been developed largely from English-based datasets and visual benchmarks, while comparatively few studies have focused on the creation of synthetic media in the Arabic language and on dialect-specific speech synthesis. Thus, the effectiveness of current detection systems can significantly differ between different linguistic and regional settings (Mirsky & Lee, 2021; Verdoliva, 2020).

In addition to algorithmic accuracy, experts increasingly consider timeliness of detection a critical factor to detection accuracy. Lack of awareness and labels for manipulated content gives false information a chance to gain traction before any interventions can happen to correct the information. Research on misinformation and fact-checking consistently shows that the earlier a correction or warning label appears, the more effective it will be in reducing the belief in the misinformation, as well as the earlier the detection of misinformation, the less time will pass between the detection and the correction or warning label (Clayton et al., 2020; Wardle & Derakhshan, 2017).

Although deepfake research is rapidly expanding, there are some areas that are still unfilled. Most of the current studies test the detection accuracy on benchmark datasets instead of using real conflict-related social media feeds. Secondly, only a small number of investigations have integrated computational detection with lessons learned from journalists, platform moderators, and digital forensic experts to describe the institutional factors that affect detection time. Thirdly, there is a limited supply of research that is specifically oriented towards Arabic-language synthetic media and the Gulf information environment. Lastly, there have been few studies that examine the interplay and the cumulative effect of content attributes, platform dynamics, and institutional response on the propagation and identification of deepfakes in armed conflict.

In that context, the present study explores the gap between the generation and detection of deepfake violations that might occur in the Gulf Cooperation Council (GCC) countries in the active phase of the 2026 Gulf War. This study will quantify the detection latency, determine the factors that contribute to successful detection,

and explore the institutional challenges that affect reactions to synthetic media in conflict, using a mixed-methods design that incorporates computational detection, engagement analysis, and interviews with practitioners. The study helps fill the gap in literature related to platform governance, digital resilience in conflict environments, and information warfare, in particular through the lens of AI.

Objective of the Study

The purpose of this study was to quantify and explain the generation-detection gap for deepfakes circulating across Gulf Cooperation Council states during the active combat phase of the 2026 Gulf conflict. Drawing on Hassan & Qureshi (2025) framework, the generation-detection gap is defined as the proportion of high-impact synthetic content that reaches significant audiences before any technical system flags it as manipulated. The study integrated three data streams:

(a) a corpus of public audiovisual items posted between February 1 and April 10, 2026, (b) outputs from four detection systems applied to the full corpus.

(c) Semi-structured interviews with journalists, platform trust-and-safety staff, regulators, and forensic analysts.

The overarching aim was to identify technical, resource, and procedural factors that contribute to detection latency and to assess the relationship between detection timing and content reach.

Research Questions

In order to achieve this, the study was oriented on four research questions:

RQ1. During the 2026 Gulf conflict, a generation-detection gap was the percentage of high-impact items (more than 1,000 interactions in the first 72 hours) that were not identified by any detection system within the first 24 hours after the item was posted.

RQ2. What was the difference in the generation-detection gap between (a) the modality (video, audio, image), (b) the target type (political leader, military, civilian), and (c) the platform (Instagram, Facebook, Telegram)?

RQ3. What were the factors that predicted detection within 24 hours, and how was the likelihood of detection affected when controlling for the modality and target type of the content?

RQ4. What factors within the institutions, resources, and processes did practitioners consider to be relevant for the detection latency and labelling decisions, and how did these factors contribute to the quantitative patterns found in RQ1-RQ3?

Literature Review

The Evolving Threat of Synthetic Media in Conflict Zones

Generative AI has revolutionized the field of information warfare by allowing the generation of extremely believable fake images, videos, and audio recordings. The ability to deepfake public figures into appearing in events that never actually happened, or can create events that never truly existed, and manipulate public opinion in a realistic and convincing manner is a growing worry. Manipulated media can spread quickly during war times and affect the political will to make decisions, damage the credibility of institutions, and create uncertainty among citizens (Chesney & Citron, 2019; Mirsky & Lee, 2021). Misinformation, disinformation, and misinformation are all forms of information disorder, which can be exacerbated by AI-generated content. During times of political uncertainty, it becomes harder to verify news in the times of synthetic media, which brings in false news faster because of its realism and emotional appeal, according to Wardle and Derakhshan (2017).

Advances in Deepfake Detection

The initial studies in deepfake detection were mainly concerned with detecting facial boundary artifacts, head pose irregularities, and abnormal eye blinking (Yang et al., 2019). The more sophisticated the generative model, the more elusive these visual artifacts became, leading researchers to develop more complex forensic methods relying on physiological signals, frequency domain analysis, and deep neural networks. A widely used benchmark dataset is

FaceForensics++, which allowed us to systematically compare several detection algorithms under controlled conditions (Rossler et al., 2019). Likewise, the Celeb-DF dataset has been developed with visually more realistic deepfakes with fewer visual artifacts, which serve as a more realistic benchmark for testing the robustness of the deepfake detector (Li et al., 2020). These datasets are a standard resource for evaluating the deepfake detection performance.

Many deep learning models have demonstrated high classification accuracy on benchmark datasets, but they are prone to poor performance when applied to new datasets, where domain shifts, compression artifacts, adversarial manipulation, and continual improvements in generative models are responsible for this (Tolosana et al., 2020; Verdoliva, 2020; Mirsky & Lee, 2021). As a result, the ability to generalize (and not just achieve benchmark accuracy) has emerged as one of the major challenges in the field of operational deepfake detection.

Audio Deepfakes and Multimodal Detection

The advancements in generative AI have significantly enhanced voice cloning capabilities, making it harder and harder to tell the difference between real and fake voice recordings. Whereas manipulated videos have visual cues, audio deepfakes do not, making them more difficult to analyze. Current research thus focuses on multimodal detection systems that use audio, visual, and contextual information to make the system more robust when deployed to real-world conditions (Mirsky & Lee, 2021; Tolosana et al., 2020).Furthermore, many current speech databases include more English-language recordings, making them difficult for multilingual detection systems to use. While the research on Arabic language deepfakes is limited compared to previous studies, the lack of representative speech datasets calls for language-specific and dialect-aware detection models.

Platform Governance and Moderation

As the spread of manipulated media grows in rapidity, there has been growing pressure on digital platforms to create better content moderation policies. Most platforms now use a mix of AI-powered detection, human moderation, fact-checking collaborations, and warning labels, rather than just automatic content removal. Effective content moderation should be seen as a systems-level governance issue that is not solely about algorithms but about technology, institutions, and laws as well, as Douek (2022) suggests.Research also shows that corrections of misinformation are most effective when they take place early in the information cycle. Clayton et al. (2020) reported that warning labels and fact checks decrease belief in misinformation, but this effect wanes with the level of misinformation acceptance. This is a strong reminder that it is critical to reduce detection and response time in the face of a fast-changing crisis.

Psychological Effects of Deepfakes

Deepfakes are not just technically challenging but also psychologically difficult. Some experimental studies show that realistic manipulated media can increase uncertainty and strengthen people's political attitudes, and it is able to reduce people's trust in authentic media. Even with knowledge of the potential for manipulation, exposure to political deepfakes can cause audience confusion around the authenticity of political communication, Vaccari and Chadwick found.Theories of belief perseverance, that people keep using false information even after it has been corrected, also contribute to the persistence of misinformation. This indicates that it is crucial to recognize manipulated media at an opportune time to help reduce the impact of manipulated media on public opinion during crises (Wardle & Derakhshan, 2017; Clayton et al., 2020).

Research Gap

Yet in the last 5 years, deepfake research has grown significantly, but there were still some crucial areas that have not been sufficiently addressed. First, most studies assess detection algorithms on benchmark datasets instead of on real social media content related to the conflict. Second, there has been a dearth of studies that

combine computational detection with journalistic and platform moderator and digital forensic expertise. Third, research on the empirical studies of the Arabic language deepfakes and the Gulf information environment is few and far between. Lastly, the impact of the combination of technological capabilities, institutional practices, and platform governance on detection latency in armed conflicts has received little consideration.

In this regard, the present study examined the gap between the generation and detection of deepfakes in the 2026 Gulf conflict, where practitioner perspectives and computational analysis were combined to gain a comprehensive understanding of both technical and institutional factors that would affect the detection of such deepfakes.

Methodology

Research Design

This study adopts a convergent mixed methods research design to explore the generation–detection gap of deepfakes that were circulating throughout the active phase of the 2026 Gulf conflict. The mixed-methods approach was chosen because the phenomenon being studied involves not only measurable technical processes but also complex institutional practices. The quantitative aspect evaluated the detection accuracy of several deepfake detection systems, as well as content features and detection results. The qualitative component delved into the experiences and perspectives of journalists, platform trust-and-safety specialists, and digital forensic analysts and media regulators to understand their operational challenges when working with synthetic media. By combining the quantitative and qualitative results, a comprehensive picture of the technical, organizational, and procedural factors that are related to detection latency in conflict communication environments was obtained.

Data Collection

We gathered 1,912 publicly available audiovisual data from Facebook, Instagram, and Telegram from 1 February to 10 April 2026. The focus of data collection was on publicly accessible posts that included video, audio, or image content and had measurable user engagement with respect to the Gulf conflict. Relevant posts were identified by pulling posts that contained keywords related to the conflict, hashtags, and monitoring public accounts of news organizations, governmental agencies, and influential public figures that are publicly accessible. For each item, the following information was recorded, to support further analysis: publication date, platform, content format, engagement statistics, and other metadata. To maintain consistency and quality of data throughout the analysis, duplicate posts, reposts, inaccessible material, and content with insufficient metadata were removed.

Content Coding and Variable Measurement

A coding protocol was created to categorize each AV by media modality, target category, platform, level of engagement, and detection outcome. Media modality was classified as video, audio, or static image, and target type was classified as political leader, military, or civilians, which was defined by the main focus of the content. Platform variables were used to compare the differences in dissemination and detection in Facebook, Instagram, or Telegram. The primary dependent variable was the proportion of high-impact synthetic media that would have more than 1,000 user interactions within 72 hours of publication but would not be detected by any detection system in the first 24 hours after posting. A number of other variables were also included such as detection latency, engagement velocity, confidence scores, and final engagement metrics to assess the relationship between content diffusion and detection performance.

Deepfake Detection Procedure

Four distinct deepfake detection systems representing complementary forensic approaches were used for the independent evaluation of each audiovisual item. The systems comprised spatial artefact analysis and frequency domain feature extraction, along with commercial artificial intelligence detection services for detecting manipulated audiovisual content. The detection

decision, confidence score, and the time of the first successful detection were logged for each item. Detection latency was defined as the time interval from the first appearance of the content to the time of successful detection by any of the detection systems. Using a combination of detection methods minimized the bias of depending on a single algorithm and allowed for a comparison of detection between various forensic methods. This data was then analyzed to see if each item was detected in the set time period, which was twenty-four hours.

Qualitative Data Collection

In addition to the quantitative results, semi-structured interviews were held with 14 professionals directly engaged in digital information verification and crisis communication. Journalists, platform trust and safety staff, digital forensic analysts, and government regulators from the Gulf were among those who participated. We used purposeful sampling to recruit stakeholders with significant experience in either media verification or information moderation or governance during conflict. The interview protocol focused on the following questions and topics: What they thought of the current detection technologies; institutional coordination; moderation procedures; verification challenges; resource constraints; and recommendations for strengthening responses to misinformation spread by AI. Interviews took place via secure communication platforms, were audio-recorded with the consent of the participants, transcribed verbatim, and anonymized prior to analysis to ensure participant anonymity.

Data Analysis

The IBM SPSS Statistics software package was used to analyze quantitative data. Descriptive statistics were used to analyze the characteristics of the data and explore the distribution of content in terms of media modality, target categories, and social media platforms. Chi-square tests were used to assess for differences in detection rates between categorical variables, and binary logistic regression was used to identify predictors of successful detection in the first 24 hours post-publication. Odds ratios with 95% confidence intervals were used to quantify the association between predictor variables and detection results. An α value of .05 was used to assess statistical significance. Qualitative data from the interviews were analyzed using thematic analysis as described by Braun and Clarke (2006) in six stages. The analysis method included repeated re-reading of the interview text, coding of significant portions, identification of initial themes, revision of thematic categories, clarification and naming of thematic categories, and linking of qualitative findings with quantitative findings.

Reliability and Validity

Some precautions were taken to improve the reliability and validity of the study. To ensure consistency of data classification, standardized coding procedures and clearly defined operational definitions were used throughout the content analysis. Multiple deepfake detection systems enabled triangulation of detection results and minimized the risk of systematic bias of any single method. To further enhance the credibility of the findings, methodological triangulation was used with the integration of computational evidence and qualitative interviews. In addition, coding decisions were made independently, where appropriate, to enhance consistency and all statistical analyses were performed in accordance with standard techniques in the field of communication and information systems.

RESULTS

Corpus Overview and Engagement Patterns

Between February 1, 2026, and April 10, 2026,n= 1,912 audiovisual items meeting inclusion criteria were coded. Table 1 shows the distribution of items by modality and target type. Video comprised 63.0% (n= 1,204) of the corpus, followed by audio (25.0%,n= 478) and static images (12.0%,n= 230). Items alleging to depict political leaders were most prevalent (46.4%,n= 887).

Table 1: *Corpus Characteristics by Modality and Target Type*

Modality	Political Leader	Military	Civilian	Total n (%)
Video	612	404	188	1,204 (63.0)
Audio	198	167	113	478 (25.0)
Image	77	50	103	230 (12.0)
Total n (%)	887 (46.4)	621 (32.5)	404 (21.1)	1,912 (100)

Note. Percentages are calculated within the full corpus, N = 1,912.

Table 1 shows the modality and category distribution of AI-generated content. The vast majority of the dataset was made up of video content (1,204 items, 63.0%), followed by audio content (478 items, 25.0%) and image content (230 items, 12.0%). The most common target type represented in the total sample was political targets (887 items, 46.4%), followed by military targets (621 items, 32.5%) and civilian targets (404 items, 21.1%). Political leaders were also the largest target group across modalities, especially for videos: 612 items targeted them. Civilian targets, on the other hand, were comparatively less common in video content and less common in audio content, but more common in image-based content, with 103 of 230 image items (44.8%). Overall, the distribution suggests that the focus of the synthetic media production and dissemination, over the course of the study, was mainly on politically salient and security-related stories, with political leaders and military actors being the main protagonists.

Table 2: Detection Rates, Confidence, and Latency by System

System	Items Flagged	% Flagged	M Confidence	SD	M Latency (hrs)
XceptionNet-based	512	26.8	.72	.19	8.4
Frequency-domain	389	20.3	.68	.22	6.1
Proprietary Service A	601	31.4	.81	.14	3.9
Proprietary Service B	474	24.8	.76	.17	2.7
Any system	788	41.2	—	—	4.8

Note. Latency calculated only for flagged items. Confidence scores range from 0 to 1.

In Table 2, the performance of four content detection systems is compared using the criteria of coverage, confidence, and detection speed, where the detection speed is determined by the number of detection days. Table 2 shows the four AI-generated content detection systems' performance in terms of coverage, confidence, and the number of days to detect the content (speed). The system with the highest detection coverage was Proprietary Service A with 601 items detected (31.4% of the items), and the highest average confidence score (M = .81, SD = .14), meaning that classifications were made with more confidence. Proprietary Service B identified 474 items (24.8%) and had the quickest detection time, comprising an average of 2.7 hours. The XceptionNet-based detector had the lowest confidence level (M = .72, SD = .19) among the systems evaluated and the highest average time of 8.4 hours. The XceptionNet-based detector was the slowest system evaluated, with 512 items detected with moderate confidence (M = .72, SD = .19). The Frequency-domain detector had the lowest coverage (389 items, 20.3 %), and the lowest confidence scores (M = .68, SD = .22), but its detection latency was still longer than that of the XceptionNet based detector (6.1 hours). All systems combined, the

total number of items detected was 788 (41.2%) with an average of 4.8 hours until detection. While each system is capable of its own detection efficiencies, the combined detection rate shows that almost 59% of all the items were not flagged, indicating significant limitations in current AI-content detection technologies and demonstrating that no single system offers full protection from the increasingly common flood of synthetic media.

Table 3: Logistic Regression Predicting Detection within 24 Hours

Predictor	B	SE	Wald	p	OR	95% CI for OR
Constant	-0.41	0.11	13.89	<.001	0.66	
Audio (ref = video)	-1.14	0.14	66.31	<.001	0.32	(0.24, 0.42)
Image (ref = video)	0.09	0.16	0.32	.574	1.09	(0.80, 1.50)
Political leader (ref = civilian)	-0.50	0.11	20.66	<.001	0.61	(0.49, 0.76)
Military (ref = civilian)	-0.18	0.12	2.25	.134	0.84	(0.66, 1.06)
24-hr engagement (per 10k)	0.08	0.03	7.11	.008	1.08	(1.02, 1.14)

Note. N= 1,912.

The logistic regression analysis was used to investigate the factors associated with the detection of a high-impact AI-generated item within 24 hours. Content modality was the most powerful predictor of detection. Audio items were significantly less likely to be detected within 24 hours (B = -1.14, p < .001), with the odds of detection being reduced by about 68% (OR = 0.32, 95% CI [0.24, 0.42]). However, no significant differences were found between image and video content in terms of likelihood of detection (OR = 1.09, p = .574). The detection results also depended on the type of target. Content that contained political leadership was significantly less likely to be detected than content that targeted civilians (B = -0.50, p < .001), with the odds of detection dropping by 39% (OR = 0.61, 95% CI [0.49, 0.76]). But, there was no significant difference in military-target content vs civilian-target content (OR = 0.84, p = .134). Finally, increased engagement in the first 24 hours was found to be associated with an increased chance of detection (B = 0.08, p = .008), with the odds of detection increasing by 8% per additional 10,000 interactions (OR = 1.08, 95% CI [1.02, 1.14]). In general, it is found that early detection of the synthetic content is more likely for content that is based on audio and/or political content focused on people rather than for content that is focused on other areas. Overall, it can be concluded that the content based on audio, and/or political content focused on people, is significantly more likely not to be detected early, whereas a strong initial visibility has a low but measurable effect on the likelihood of being detected early.

Table 4: Generation-Detection Gap for High-Impact Items by Modality and Target

Modality	Target Type	High-Impact n	Not Detected 24h n	Gap %
Video	Political leader	381	192	50.4
Video	Military	247	104	42.1

Video	Civilian	96	38	39.6
Audio	Political leader	143	102	71.3
Audio	Military	118	71	60.2
Audio	Civilian	41	19	46.3
Image	Political leader	22	8	36.4
Image	Military	12	3	25.0
Image	Civilian	9	2	22.2
Total	All	1,069	574	53.7

Note. High-impact defined as >1,000 interactions within 72 hours. Gap % = proportion not flagged by any system within 24 hours.

Table 4 shows that there are substantial effects for the content modality and the type of target on the detection performance. The highest generation – detection gap was observed for audio content aimed at political leaders, where 71.3% (102 of 143) of the high impact items went undetected within 24 hours, followed by audio content aimed at military personnel (60.2%) and civilians (46.3%). The gaps in detection ranged from moderate to very high for video content (39.6% to 50.4%), and were lowest for the image-based content (22.2% to 36.4%) across all of the target categories. Politically focused synthetic media generally showed the greatest detection gaps across modalities, followed by military and civilian targets, indicating that politically themed synthetic media is harder to detect than the other themes. Overall, 53.7% (574 of 1,069) of the items with high impact were not detected within the first 24 hours. The findings show an evident modality-target interaction: that is, the combination of audio-based content and political leader narratives was the most sensitive to detection failure, revealing some serious flaws in the AI-generated misinformation detection systems during high-profile information environments.

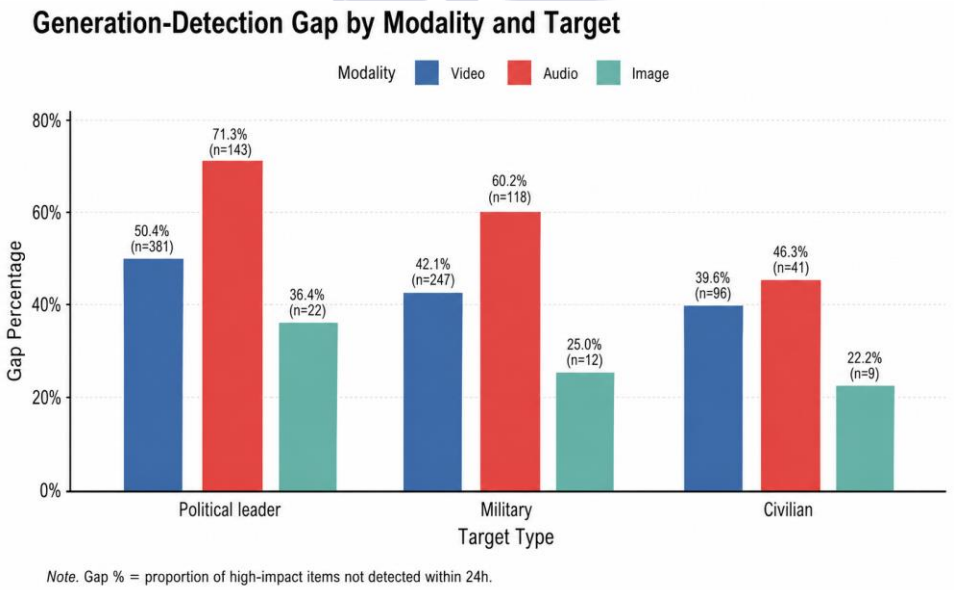

Figure 1 shows that there is significant modality and target type variation in generationdetection gaps. The percentage of high-impact items that went undetected in audio content was the highest at 71.3% of the items targeted political leaders, compared to those that targeted civilians (46.3%) and military (60.2%). The detection gaps were the smallest for the image-based content (ranging

from 22.2% to 36.4%). The detection gaps were consistently larger across all modalities for content that involved political leaders, suggesting politically-focused synthetic media is particularly difficult for current detection systems.

Table 5: Summary of the Generation-Detection Gap during the 2026 Gulf Conflict

Content Category	Total Items	High-Impact Items	Detected ≤24h	Gap n	Gap %	Mdn Final Engagement Undetected	Mdn Final Engagement Detected
By Modality							
Video	1,204	724	362	362	50.0	54,200	21,300
Audio	478	302	76	226	74.8	78,900	17,600
Image	230	43	26	17	39.5	12,400	6,800
By Target							
Political leader	887	546	226	320	58.6	81,300	24,100
Military	621	377	185	192	50.9	49,700	19,400
Civilian	404	146	84	62	42.5	22,600	11,200
By Platform							
Instagram	1,031	612	291	321	52.5	63,100	20,400
Facebook	694	374	178	196	52.4	58,300	18,900
Telegram	187	83	26	57	68.7	72,400	14,200
Overall	1,912	1,069	495	574	53.7	61,400	19,200

Note. High-impact is defined as >1,000 interactions within 72 hours of posting.

As seen in Table 5, there is a significant gap between the generations of the different modalities, targets, and platforms. In terms of modalities, high impact items in audio content were most likely not to be detected within 24 hours, at 74.8% out of 302 items, followed by video content at 50.0% out of 144 items and images at 39.5% out of 102 items. Similarly, undetected audio pieces collected the most median engagement (78,900 interactions), making audio content-based synthetic content a major detection challenge. When broken down by target type, political leaders had the highest detection rate gap and the highest median engagement rates for undetected targets (81,300), compared to military targets (50.9%; 49,700) and civilian targets (42.5%; 22,600). There was also a difference in content detected across platforms, with the highest detection gap being on Telegram (68.7%), followed by Instagram (52.5%) and Facebook (52.4%), and a high median engagement for undetected content was also recorded on Telegram (72,400). Across the whole dataset, 574 (53.7%) of the 1,069 high impact items were not detected during the first 24 hours and received a median of 61,400 engagements, whereas detected items had 19,200 engagements, on average. These findings indicate that delayed detection is associated with substantially greater dissemination, particularly for audio-based content, political-target narratives, and content circulating on Telegram.

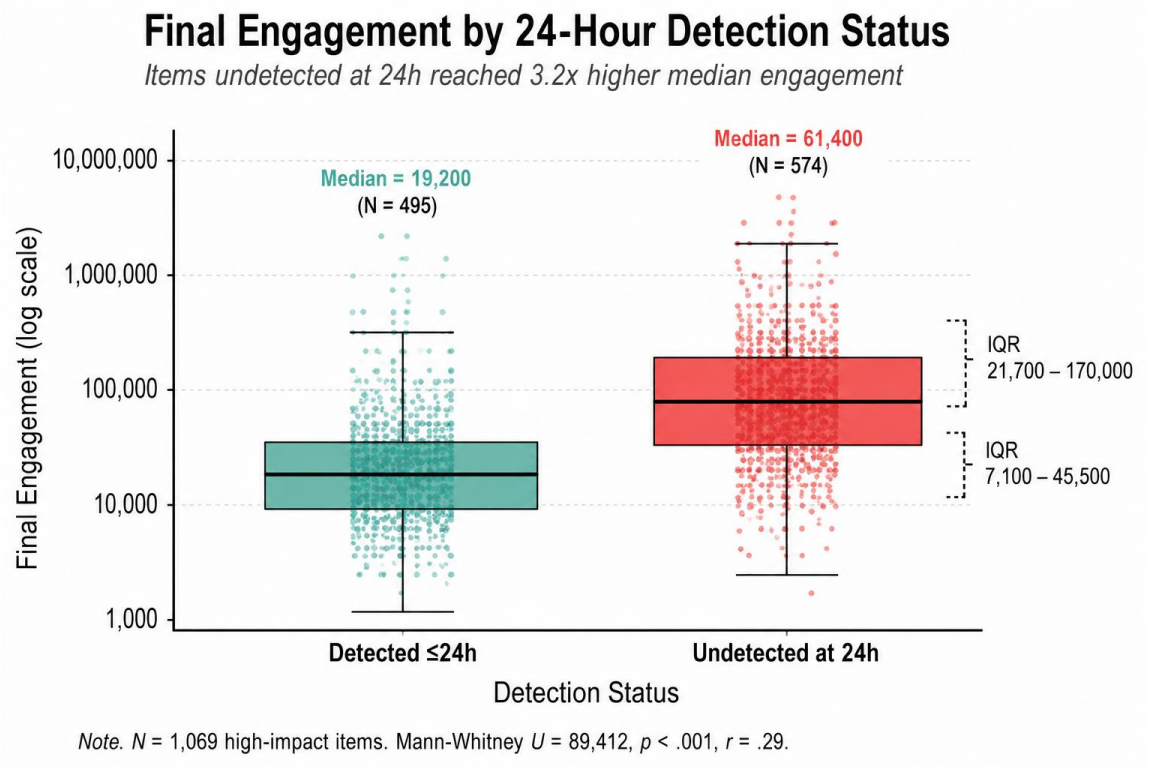

Figure 2 demonstrates that high-impact AI-generated content that remained undetected during the first 24 hours achieved substantially greater final engagement than content detected within the same period. Undetected items reached a median engagement of 61,400 interactions (N = 574), compared with 19,200 interactions (N = 495) for items detected within 24 hours, representing a 3.2-fold increase. The interquartile range for undetected content (21,700–170,000) was also considerably higher than that of detected content (7,100–45,500), indicating consistently greater audience reach across the distribution. Furthermore, several undetected items achieved exceptionally high engagement levels, exceeding one million interactions, highlighting their viral potential. The Mann–Whitney U test confirmed that the difference was statistically significant (U = 89,412, p < .001) with a moderate effect size (r = .29), suggesting that delayed detection is strongly associated with increased dissemination and amplification of AI-generated misinformation. These findings underscore the critical importance of rapid detection mechanisms, as content that evades identification during the initial 24-hour window gains a substantial advantage in audience exposure and engagement.

Discussion

Overview of Findings

This study looked at the challenges in deepfake generation and detection in the 2026 Gulf War in the active combat stage in the GCC states. The findings of n = 1,912 items showed a significant and systematic lack of symmetry. 53.7% of high-impact items with more than 1,000 interactions were not picked up by any detection system in 24 hours. This disparity was highest in the case of audio content (74.8%) and when targeting a political leader (58.6%), and was related to a 3.2-fold difference in median final engagement between undetected and detected items. These quantitative trends were confirmed by interviews with 14 journalists, platform staff, regulators, and forensic analysts, who said that detection pipelines were generally not keeping up with the velocity of content and did not have Arabic-dialect audio capability.

The Temporal Dimension of the Gap

It was 2.1 hours for an item to reach 1,000 interactions and 4.8 hours for the detection latency. Such a mismatch in time meant that, on average, any technical flag was raised after synthetic content went viral. The logistic regression validated the engagement as a slight positive risk factor for detection (OR = 1.08), indicating that platforms seek to scale up the amount of engagement in their moderation queues. The impact was, however, small compared to the negative effects of audio modality (OR = 0.32) and political targets (OR = 0.61). Therefore, it is not enough to use virality and hope that the reviews will come in for the most sensitive content. These results build on the findings of Wardle and Derakhshan (2017) by showing that the "distribution" component of information disorder has structurally outpaced "verification" during conflict, especially in the case of audio.

Modality and Target Vulnerabilities

The area that was most clearly missed was audio deepfakes. Just 22.4% were flagged in the first 24 hours, and 71.3% of high-impact political audio was not flagged in 24 hours. This is consistent with previous technical research demonstrating that voice cloning does not have a spatial and temporal artifact that current visual detectors use (Dolhansky et al., 2020; Li et al., 2020). Responses from interviewees in various fields indicated that as of February 2026, there were no GCC countries that had implemented a real-time voice forensic service in the Gulf Arab region. So, synthetic audio was able to sound very similar to the actual leader, and little technical effort was required, as demonstrated by Archived Item GCC2026-A442 (2026) that circulated for 31 hours without being identified as such. Controlling for modality and engagement, political targets were less likely to be detected than civilian targets. This pattern could be due to conflictual targeting of individuals for whom there is a lot of data (voice and face) for training, or the reluctance to label high-stakes content without a multi-lateral check. When it came to public flags with heads of state, platform participants noted that there was some "attribution friction" when it came to the timeline.

Institutional and Policy Implications

The data imply three things. Firstly, detection benchmarks need to be based on content velocity, not absolute time. When the median latency period is 2 hours to 1000 engagements, then 4 hours is not enough. Detection and labelling targets for sub-hour conflict zones are necessary. Secondly, investment is asymmetric. As of February 2026, Arabic audio generation tools were freely accessible and affordable, whereas detection tools required licensed APIs, GPU clusters, and dialect-specific tuning, which was absent in five out of seven GCC states. Building the capacity gap will not be solved with platforms of the world alone. Third, information-sharing procedures during war are required. 68.7% of the difference between Telegram and participant reporting of "escalation friction" suggests that cross-platform coordination is lagging behind adversarial cross-posting.

Practical Utility of Detection

Items found within 24 hours were found to have a significantly lower median final engagement than items not found (Mdn = 19,200 vs. 61,400), consistent with previous findings that early labels lead to lower belief and sharing (Clayton et al., 2020). This implies that there would be direct harm-reduction benefits of reducing detection latency. The interviews, however, suggested that after 12 hours, the labels were seen as "too late," since audiences had already developed their opinions. So, the speed and visibility of intervention are important.

Conclusion and Future Directions

During the 2026 Gulf conflict, the capacity to generate persuasive deepfakes exceeded. In the case of the 2026 Gulf conflict, the ability to create convincing deepfakes was more than the ability to identify and tag them, particularly in the case of politically oriented audio deepfakes of political leaders. The gap was technical, institutional, including modality gaps, resource

gaps, and delays. This difference in engagement was significant and had operational implications on the information environment, as undetected items had significantly higher engagement. It will take coordinated investment in detecting each language, in pipeline standards for sub-hour production, and in wartime coordination measures so that synthetic media is viewed as a time-critical threat in order to close it.

Further testing is required with real-time streaming detectors set specifically for Gulf Arabic and compressed/recompressed media typically used in messaging applications. Experimental designs that randomise the moment(s) of labelling synthetic content during crisis situations are necessary to separate out causal effects on belief and sharing. The friction around attribution should be decreased by considering pre-authorized and time-limited "unverified" labels that can be applied before full forensics. Last, the gap in 2026 needs to be measured through longitudinal monitoring of whether regional forensic capacity building effects are reducing the gap.

Limitations

There are four restrictions to interpretation. First of all, this corpus only contains public posts and four detection systems that were available in February 2026. The gap may be underestimated in the case of private channels, deleted items, and newer multimodal detectors, which are not shown. Second, these are engagement metrics reported by platforms, but they may not reflect cross-platform amplification. Third, the observational design was insufficient to make causal inferences about detection effects on engagement. Fourthly, the study period only encompasses the acute combat period. The gap can vary in a ceasefire/post-conflict information environment.

REFERENCES

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J

Ciftci, U. A., Demir, I., & Yin, L. (2020). FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2020.3009287

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior, 42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). *The DeepFake Detection Challenge (DFDC) dataset*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. https://arxiv.org/abs/2006.07397

Douek, E. (2022). Content moderation as systems thinking. *Harvard Law Review, 136*(2), 526–607.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3207–3216). https://doi.org/10.1109/CVPR42600.2020.00327

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), Article 7. https://doi.org/10.1145/3425780

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., &Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1–11). https://doi.org/10.1109/ICCV.2019.00009

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust. *Social Media + Society, 6*(1). https://doi.org/10.1177/2056305120903408

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.

Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8261–8265). https://doi.org/10.1109/ICASSP.2019.8683164