Framing Deception: A Multi-Modal Analysis of Political Deepfake Narratives

Bich Ngoc Doan
EPFL
Switzerland
bich.doan@epfl.ch

Yasmine Kroknes-Gomez
EPFL
Switzerland
yasmine.kroknes@epfl.ch

Vittoria Meroni
EPFL
Switzerland
vittoria.meroni@epfl.ch

Khadija Tagemouati
EPFL
Switzerland
khadija.tagemouati@epfl.ch

Abstract

Political deepfakes, AI-generated manipulations of video and image content, are emerging as powerful tools for influencing public perception and political discourse. While most existing research has focused on detection or technical realism, less attention has been paid to the narratives embedded within these artifacts and their impact on engagement. This study analyzes a curated collection of politically oriented deepfakes involving prominent public figures, combining methods from narrative analysis, visual emotion recognition, and symbolic image interpretation. Using topic modeling on textual summaries, we identify common themes and archetypes, while emotion and symbol detection reveal how facial expressions and imagery are used to construct persuasive messages. Additionally, we examine dissemination patterns, highlighting the role of key actors and individual users in amplifying these narratives across social networks. Overall, the study demonstrates how political deepfakes function as carefully engineered narratives designed to exploit emotional cues and symbolic frames, influencing discourse within digitally mediated environments.

Keywords

deepfake, political framing, misinformation, symbolic analysis, narrative analysis, dissemination patterns

1 Introduction

In recent years, deepfakes, hyper-realistic audio and video manipulations generated by artificial intelligence, have emerged as a disruptive force in the political landscape. Once requiring expensive equipment and expert skills, the production of deepfakes has become widely accessible due to advances in generative AI. This technological shift has collapsed traditional barriers of cost and expertise, allowing virtually anyone to create and disseminate synthetic media [2]. From fabricated scandals to AI-generated commentary, political figures like Trump, Biden, and Macron are frequently the targets of manipulated media, which has led to real-world consequences [7]—such as market disruptions [18] and political unrest [4]. Most recently, the U.S. witnessed AI-generated robocalls mimicking President Joe Biden's voice [21], while several synthetic videos targeted European candidates during national elections [19]. These incidents have amplified global concerns regarding AI-driven election interference and public trust in democratic processes.

While prior research has focused heavily on deepfake detection [8, 13] and its persuasiveness [2, 16], these studies are often conducted in controlled settings and fail to capture how deepfakes circulate and resonate within both online and offline environments. Little is known about the kinds of political narratives deepfakes encode, how these narratives engage audiences, or the sociotechnical conditions that influence their spread. Drawing from the Political Deepfake Incident Database (PDID) [23], a publicly available resource of annotated political deepfakes, our work bridges this gap by examining how these media artifacts are embedded in and shaped by broader online and offline sociotechnical systems. We first investigate the online artifacts themselves, analyzing the narratives embedded in the shared deepfakes:

- **RQ1:** What are the dominant narrative elements—textually and visuall—used in political deepfakes?
- **RQ2:** Which attributes of a political deepfake predict higher levels of virality and engagement?

We further explore how political deepfakes travel through the information ecosystem and intersect with real-world events:

- **RQ3:** What are the dissemination patterns of political deepfakes, and how do they correlate with real-world events?

By exploring these questions, we aim to advance theoretical understanding of the societal, political, and psychological implications of political deepfakes. Our findings can inform future research on media manipulation, support platform governance efforts, and contribute to media literacy initiatives aimed at navigating the challenges of synthetic content.

Team Contributions: The project involved close collaboration across all stages. Bich Ngoc Doan led the textual narrative and engagement analyses. Yasmine Kroknes-Gomez carried out the

Report'25, June 2025, EPFL, Lausanne, Switzerland

ACM ISBN

facial expression analysis. Vittoria Meroni analyzed visual symbols. Khadija Tagemouati examined dissemination dynamics. All members jointly contributed to the writing process.

2 Data

This project uses the Political Deepfakes Incident Database (PDID), curated by researchers at Purdue University [23] and hosted on AirTable[1]. The database specifically includes politically significant deepfake content, focusing on deepfake incidents coded explicitly as such, excluding less sophisticated 'cheapfake' entries. As of March 7, 2025, our dataset comprises 976 deepfake instances featuring prominent US-based political figures (e.g., Donald Trump, Kamala Harris, Emmanuel Macron).

Data collection commenced in June 2023 and remains ongoing, encompassing incidents from 2017 onward. Data sources primarily include English-language social media platforms and popular news websites. Inclusion criteria mandate that an incident must have gained visibility in public news channels, subsequently disseminated through linked articles and social media.

The database's structure consists of 191 variables organized into thematic categories, which are presented in Table 1.

For analytical rigor and data protection, sensitive personal information from the Source and Sharer category, such as explicit identifiers and coder comments, were excluded from analysis. Once anonymized, aggregated attributes (e.g., occupation, country, platform type) were retained. Public figures targeted by deepfakes are exempt from anonymization due to their inherently public status.

Our analysis includes retrieving image media via direct links in the database, explicitly omitting audio and video files to maintain project scope. All visual content used in presentations and reports will be clearly labeled (e.g., AI-generated), avoiding misrepresentation. Explicit or sensitive material is categorically excluded from public dissemination.

3 Methods

To examine how political deepfakes function within broader sociotechnical systems, our methods integrate multiple analytical perspectives. To this end, we look into narrative construction, audience engagement, and dissemination patterns—each offering a distinct lens on how deepfakes are crafted, received, and circulated.

3.1 Narrative Analysis

To understand the kinds of narratives political deepfakes encode, we conduct both textual and visual analyses, reflecting the multimodal nature of these media. Deepfakes are not only shaped by what is said (textual content) but also by how political figures are visually portrayed through expressions, symbols, and staging. Analyzing both modes allows us to capture a more comprehensive picture of the persuasive strategies and embedded ideological framings.

3.1.1 Textual Extraction - Content Analysis. The dataset provides a `content summary` for each deepfake instance, offering detailed textual descriptions of what is literally depicted in the media (e.g., "Kamala Harris speaking out into a communist crowd"). These summaries serve as rich face-value annotations that capture not just who is shown, but also what actions or settings are emphasized—making them a valuable entry point for analyzing the narrative structures underlying deepfakes.

To analyze these descriptions, we first applied BERTopic, a modern topic modeling framework that outperforms traditional models like LDA on short texts due to its use of contextualized sentence embeddings [14]. Each summary was treated as a document and embedded using `all-MiniLM-L6-v2`, a lightweight but effective model. These embeddings were then clustered via UMAP and HDBSCAN to derive a set of semantically distinct topics.

The resulting topics were then qualitatively analyzed [3]. Two coders reviewed the topic representations and associated documents to identify broader thematic groupings. Through discussion and iterative merging, we distilled these topics into a higher-level narrative topology focused on two core dimensions: dominant actors (who is depicted) and archetypes (what role or behavior is being portrayed). This pipeline enables us to systematically analyze how deepfakes construct meaning through storytelling.

3.1.2 Visual Extraction - Facial Expression Analysis. Facial expressions act as visual affective cues in political deepfakes, subtly shaping how viewers interpret intent, authenticity, and threat. By quantifying these cues we can trace patterns in emotional representation across political figures and narratives.

Tool Choice: Py-Feat (ResMaskNet) We use the open-source *Py-Feat* toolbox—specifically its *ResMaskNet* (ResNet) backbone—as it offers a pragmatic balance between accuracy, transparency, and ease of deployment. ResMaskNet achieves an F1 score of approximately 0.55 on the AffectNet dataset, outperforming traditional methods such as SVM (0.39) and FACET (0.40), and remains one of the top-performing freely available toolkits, although more recent specialized deep networks surpass it on benchmarks like CK+ and JAFFE [6, 17]. Py-Feat is widely adopted in HCI and multimodal research pipelines, including applications in social robotics [5] and large language model-based audio-prosody fusion systems [10]. It produces calibrated probabilistic emotion outputs, which supports dominant/secondary label assignment and mixed-affect modeling. Furthermore, it offers a lightweight and fully open-source Python API that supports detection, alignment, and inference on commodity GPUs, facilitating reproducibility and bias auditing.

Pipeline The procedure consists of two main steps:

Step 1: Face Detection and Significance Ranking. To quantify the emotional cues embedded in political deepfakes, we first detected all visible faces in each image using Py-Feat's face detector. We then ranked these faces by narrative salience, computing a composite score that weighted both their size (bounding-box area, normalized by the largest face in the image) and centrality (inverse Euclidean distance from the image center), with weights of 0.6 and 0.4, respectively. This prioritization reflects the visual grammar of deepfakes, in which prominent figures are typically large and centrally positioned. We retained only the top-ranked face per image, unless a second face had an area within 95% of the first, and discarded any that fell below a minimum size threshold to avoid noisy side-profile detections.

Step 2: Emotion Extraction & Post-Processing. Next, we extracted emotional content by predicting the probability distribution over

[1]The dataset can be accessed at http://bit.ly/pdid

Category	Description
Descriptors	Basic metadata such as URLs and **media** formats.
Social Media	Engagement metrics, including views, **likes**, comments, and shares.
Source and Sharer	Information about the **content's originators and disseminators**, including **entity type**, occupation, and follower count.
Target(s)	**Details about individuals or entities addressed or attacked**, including response behaviors.
Verification	Indicators of content authenticity; excluded from analysis due to limited relevance.
Context & Content	**Textual content summary**, depicted harm (e.g., violence, discrimination), and communicative intent (e.g., satire, reputational harm), perceived synopsis (e.g plot, hero, villain)
Real-world Connections	**Associations with actual events or harms** (e.g financial losses, political interference).
Politics & Policy	Broader framing related to policy sectors, narrative strategies, and political implications.

Table 1: Overview of PDID variables. Highlighted text are the variables we used for the analysis.

seven basic emotions—anger, disgust, fear, happiness, sadness, surprise, and neutral—for each retained face. If two faces were analyzed in an image, we computed the elementwise average of their emotion vectors to yield a single composite representation. From this, we identified both the dominant and secondary emotions based on their probabilities (p_1, p_2), recording their labels and scores. Each image thus contributed four features: the dominant and secondary emotion labels, along with their respective probabilities. These served as both categorical and continuous predictors in our downstream regressions linking facial affect to engagement metrics (likes, shares, comments) and broader symbolic or narrative frames.

3.1.3 Visual Extraction - Symbol Detection. Symbols are powerful narrative tools in political deepfakes. Unlike explicit verbal messages, visual symbols work on a subconscious level, instantly triggering emotional and cultural associations in the viewer [12]. Their inclusion is rarely accidental—instead, they serve to anchor the deepfake's political, ideological, or nationalistic message.

Tool Choice: YOLOv8 and CLIP Utilizing this context, we developed a hybrid pipeline using YOLOv8 for bounding box proposal and CLIP (ViT-L/14). This pairing balances localization precision with computational efficiency, based on our internal benchmarking.

Step 1: Prompt-Based Symbol Matching. For each image, YOLOv8 identifies object regions, which are then cropped and encoded via CLIP's image encoder. We encode a predefined set of textual prompts (e.g., "army," "prisoner," "American flag," "police," "rainbow pride flag," "communist symbol") using CLIP's text encoder, and compute cosine similarity between each cropped region and each textual prompt. Regions with a confidence score above 0.90 are retained as symbolic detections.

Step 2: Reference-Based Communist Symbol Detection. Upon preliminary exploration, we found a high visual diversity and contextual ambiguity of communist iconography. For this reason, we added a curated set of reference images of communist symbols (e.g., the hammer and sickle) to encode into the CLIP embeddings. Each YOLO-detected region is then matched against this reference set via cosine similarity to capture nuanced symbol usage that may evade prompt-based classification.

This dual approach enables the detection of both explicit and subtle symbolic content, accommodating a broad spectrum of politically charged imagery often embedded in deepfake media.

3.2 Popularity Modeling

Understanding how online users engage with different types of narratives is crucial for gauging the persuasive power and reach of political deepfakes. Beyond narrative construction, virality speaks to which stories are amplified, by whom, and to what effect.

To explore this, we adapt the content characterization and engagement modeling framework proposed by Mejova et al. [11], originally developed to study the virality of Twitter memetic warfare. Drawing from their approach, we focus on content shared on the most common platform in our dataset—X/Twitter—and use the number of likes as our proxy for engagement.

We construct linear regression models to estimate how specific narrative elements influence popularity. The dependent variable is the log-transformed number of likes, while the independent variables are drawn from our annotated narrative features, including dominant actors, archetypes (textual), emotions, and symbols (visual). Categorical variables are encoded via one-hot encoding, using "none" as the reference category. This allows us to assess whether certain narrative framings—such as depictions of sadness, patriotic symbolism, or controversial figures—systematically correlate with greater engagement.

3.3 Dissemination Patterns

Understanding how political deepfakes disseminate is key to uncovering the pathways through which they gain visibility, the types of actors involved in their propagation, and the sociopolitical dynamics that shape their spread. Analyzing dissemination patterns not only reveals who shares this content and when it gains traction, but also provides insight into influence, intent, and potential coordination behind its circulation.

To investigate this, we used the dataset's records of content transmission—tracking interactions between `source` and `target` entities involved in the dissemination of deepfakes. Each interaction record includes metadata such as the source and target account types , and timestamps. We represented these interactions as a directed graph, where nodes correspond to users or entities, and edges represent instances of content sharing from a source to a target. Using NetworkX, we constructed and analyzed this dissemination graph, with preprocessing steps to anonymize users and clean malformed or noisy records.

4 Results

We present a comprehensive analysis of political deepfakes, structured around three key research questions, providing a multifaceted understanding of how political deepfakes function both as narrative tools and as social phenomena within digital media ecosystems.

4.1 Narrative Frames in Deepfake Politics (RQ1)

4.1.1 Textual Extraction - Content Analysis. Our BERTopic modeling revealed a total of 37 narrative topics from the summary texts. These topics were organized hierarchically via HDBSCAN, revealing two prominent dimensions of variation: the central actor (i.e., the individual featured in the image) and the associated archetype or storyline (i.e., the narrative framing of their role). Figure 1 illustrates a representative subset of the topic tree, where two major branches correspond to Kamala Harris and Donald Trump — the most frequently depicted figures in the dataset.

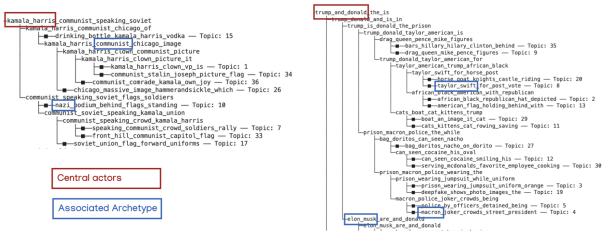

Figure 1: Subset of the BERTopic-generated topic hierarchy, highlighting two dominant subtrees centered on the key political figures and their associated narrative archetypes.

Kamala Harris is primarily associated with narratives involving ideological extremism (e.g., communism, Nazism), often exaggerated or distorted. In contrast, Donald Trump appears across a more varied set of storylines, frequently co-depicted with other high-profile figures (e.g., Elon Musk, Taylor Swift, Joe Biden), suggesting portrayals of social alliances, endorsement, or satire.

To systematically characterize these narratives for downstream tasks (e.g., engagement prediction), we performed a qualitative coding of the 37 topics to identify and validate dominant actors and overarching archetypes. While Kamala Harris and Donald Trump dominated the actor space, as expected, several others (e.g., Biden, Macron) appeared less frequently, being categorized as "Others". More substantively, the coding surfaced four major narrative archetypes (Figure 2):

- **Character Caricature:** Overdrawn or symbolic depictions aligning figures with fictional or mythical personas (e.g., Macron as the Joker).
- **Scandal Fabrication:** Imaginary depictions of personal controversy or misbehavior (e.g. Trump in fictitious romantic scandals).
- **Public Endorsement:** Situations where figures receive support or validation from celebrities (e.g., Taylor Swift) or the general public, often for political or ideological messaging.
- **Ideological Demonization:** Assigning extreme or oppositional political affiliations (e.g., Nazism) to elicit moral outrage or delegitimize.

Figure 2: Demonstration of each dominant arcs depicted in the deepfake media.

A small number of outlier topics (e.g., Kamala on a Doritos bag, Biden walking with Pikachu) were marked as "absurdist" and grouped into a fifth category ("Other") for completeness. These narrative archetypes constitute a structural typology of deepfake content and are leveraged in subsequent sections for analyzing engagement and dissemination.

4.1.2 Visual Extraction - Facial Expression Analysis. The aggregate distribution of facial affects (Fig. 3) shows a pronounced skew towards **anger** as the *dominant* expression, appearing in nearly twice as many frames as the next most common category (*happiness*). Secondary emotions, in contrast, cluster around *sadness*, *neutral*, and *surprise*, suggesting that deepfake creators often layer a subtler secondary affect beneath an overtly negative or activating primary cue.

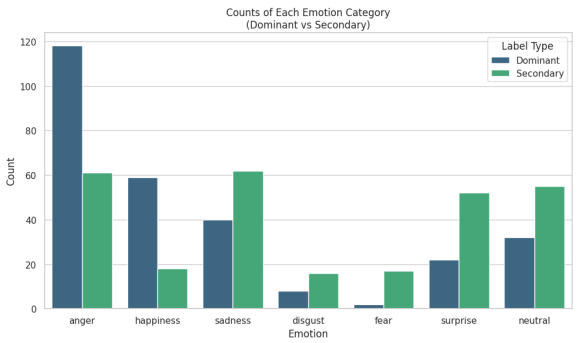

Figure 3: Frequency of dominant vs. secondary emotions across the full dataset.

When we examine mean confidence scores (Fig. 4), two trends emerge:

- **High-certainty positivity:** *Happiness* yields the highest average dominant score ($\bar{p} \approx 0.80$), indicating that when a smiling face is present, the model assigns it with near-maximal confidence.
- **Ambivalent negativity:** *Anger, disgust,* and *fear* manifest with lower dominant-probability plateaus ($\bar{p} \approx 0.55$–0.72) and

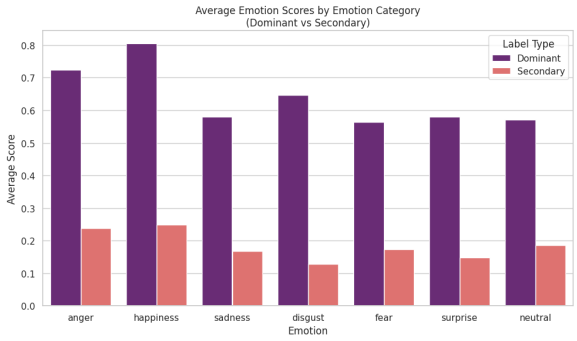

Figure 4: Average probability score for each emotion (dominant vs. secondary).

markedly lower secondary scores, suggesting that negative expressions are often accompanied by residual uncertainty or blended affect.

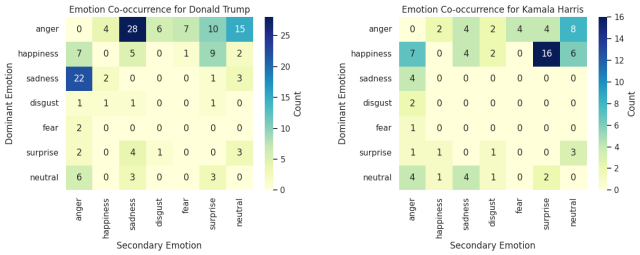

Figure 5: Emotion co-occurrence matrices for the two most frequent targets (*Donald Trump* left, *Kamala Harris* right). Rows = dominant, columns = secondary.

We also conduct target-specific heatmaps (Fig. 5), which focus on the two aforementioned major actors, to highlight their systematic framing differences:

- **Donald Trump.** The modal pairing is *anger* ⇒ *sadness* (28 instances), followed by *anger* ⇒ *neutral*. Such blends reinforce a portrayal of volatile grievance—an angry façade edged with either defeat or stoicism.
- **Kamala Harris.** The dominant pattern is *happiness* ⇒ *surprise* (16 instances), framing Harris as upbeat but perhaps naïve or caught off-guard; a secondary cluster *anger* ⇒ *neutral* suggests episodic portrayals of contained indignation.

Box-plots of score distributions (Fig. 6) reveal three additional insights:

(1) **Positive extremes.** *Happiness* shows both high medians ($\tilde{p} \approx 0.95$) and narrow spreads, implying deepfakes seldom deploy half-hearted smiles.
(2) **Mid-confidence menace.** *Disgust* and *surprise* are expressed with wider interquartile ranges, consistent with their perceptual ambiguity and the model's lower baseline certainty for these categories.

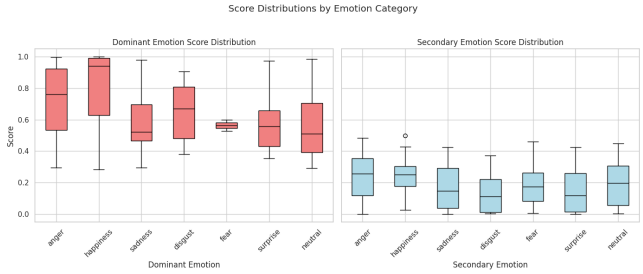

Figure 6: Score distributions for dominant and secondary emotions. Boxes = IQR, whiskers = 1.5 IQR.

(3) **Sparse but decisive fear.** Although *fear* is the least frequent dominant category, its scores cluster tightly around ∼ 0.57 with low variance—whenever fear is detected, it is rendered in a consistent, recognisable form.

Synthesis. Collectively, these emotion patterns suggest that political deepfakes favour **high-arousal affect**. Designers either amplify anger (often tinged with sadness or neutrality) to provoke outrage, or deploy unequivocal happiness to signal victory or charisma. The co-occurrence asymmetries between targets indicate deliberate tailoring: Trump deepfakes accentuate grievance narratives, whereas Harris deepfakes oscillate between charm and surprise, perhaps aiming to portray inexperience. Such fine-grained emotional engineering underscores the strategic sophistication of contemporary visual misinformation.

4.1.3 Visual Extraction - Symbol Detection. The analysis of symbolic content in political deepfakes reveals several critical insights regarding narrative construction, emotional framing, and targeted messaging.

First, as seen in Figure 7, the most frequently detected symbols include the *communist symbol*, *army*, and *prisoner*, each appearing in over 30 incidents. These symbols serve as narrative anchors, signaling ideological opposition, authority, and victimization, respectively.

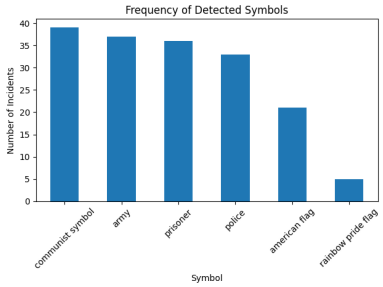

Figure 7: Frequency of detected political symbols across all analyzed deepfakes.

Symbol-villain associations show a clear alignment with political figures:

- The **prisoner** symbol is strongly tied to *Donald Trump*, reinforcing themes of persecution or martyrdom.

- The **communist symbol** is frequently linked to *Kamala Harris*, suggesting a portrayal of ideological extremism.
- An unexpected but notable pairing emerges between **police** symbolism and *Emmanuel Macron*, possibly reflecting criticism of state authority.

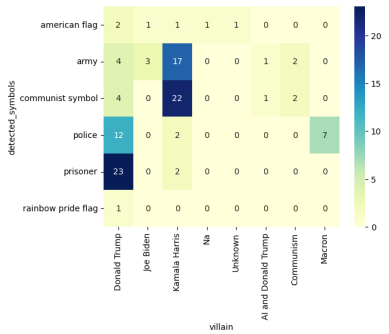

Figure 8: Detected symbols in relation to the villain role assigned in deepfakes.

Conversely, hero-symbol mappings highlight how symbols like the **American flag** and **army** are used to frame individuals—especially Donald Trump—in a patriotic or protective role. Meanwhile, the **Justice System** emerges as the hero in many prisoner-related deepfakes, pointing to legal or moral narratives.

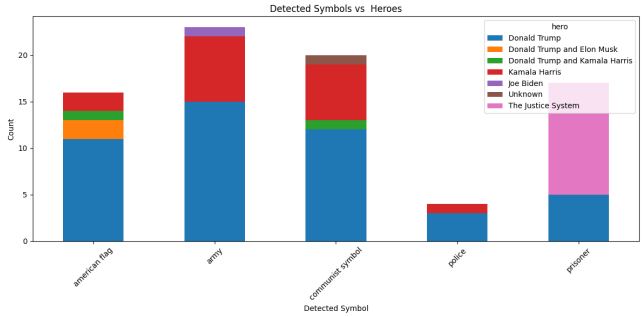

Figure 9: Detected symbols in relation to hero roles across the dataset.

The harm annotation analysis demonstrates that:

- The **communist** and **army** symbols are often tied to *political interference* or ambiguous, overlapping harms.
- The **prisoner** symbol is more strongly associated with *violence and discrimination.*
- Multiple harm categories frequently co-occur within the same symbolic instance, indicating complex and layered narrative strategies.

Taken together, these results suggest that political deepfakes do not rely solely on textual deception, but actively deploy symbolic imagery to shape perception, reinforce bias, and amplify ideological divides.

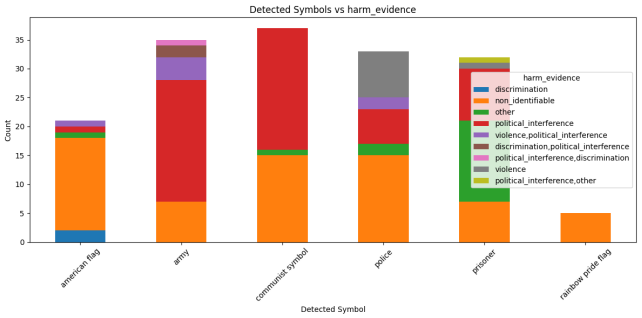

Figure 10: Detected symbols mapped to harm categories such as discrimination, violence, or political interference.

4.2 Engagement Patterns Across Narratives (RQ2)

To investigate how different narrative elements influence public engagement, we modeled the log-transformed number of likes (base 2) as the dependent variable, using linear regression. This transformation enables intuitive interpretation: a coefficient of +1 indicates that a given feature doubles the number of likes, while a coefficient of -1 implies it halves.

Each deepfake instance was characterized using one dominant narrative feature, derived from above analysis, per modality:

- **Textual:** dominant actor (2-class) and narrative archetype (4-class),
- **Visual (Emotion):** predicted facial emotion (7-class),
- **Visual (Symbol):** most confidently detected symbol (5-class).

Figure 11 presents the model coefficients with 95% confidence intervals. Several features were significantly associated with changes in engagement levels ($p < 0.05$). Most notably:

- **Rainbow Pride Flag** symbols were associated with a **66.6**× increase in likes ($\beta = 6.057$, $p < 0.01$).
- **Police** imagery saw a **32.86**× increase ($\beta = 5.038$, $p < 0.01$).
- **Prisoner** and **Army** depictions also significantly boosted engagement ($\beta = 3.545$ and 3.292, respectively).
- The presence of **Donald Trump** as the central actor was associated with a **7.44**× increase ($\beta = 2.895$).
- In contrast, the **Scandal Fabrication** archetype corresponded with a significant **decrease** in engagement, with likes reduced to approximately **0.12**× ($\beta = -3.003$).

These results suggest that engagement with political deepfakes is highly sensitive to symbolic imagery and the portrayal of specific actors, with ideological symbols and polarizing figures driving higher visibility online.

4.3 Tracing Dissemination Pathways (RQ3)

4.3.1 Central Actor in the Dissemination Network. To examine how deepfakes spread, we constructed a directed source–target network in which edges represent reposting behavior—i.e., a source account disseminating a deepfake originally targeting a specific figure. Within this network, Donald Trump emerged as the most central node based on both the frequency of his mentions (34) and his centrality (PageRank: 0.026; Degree Centrality: 0.059). We

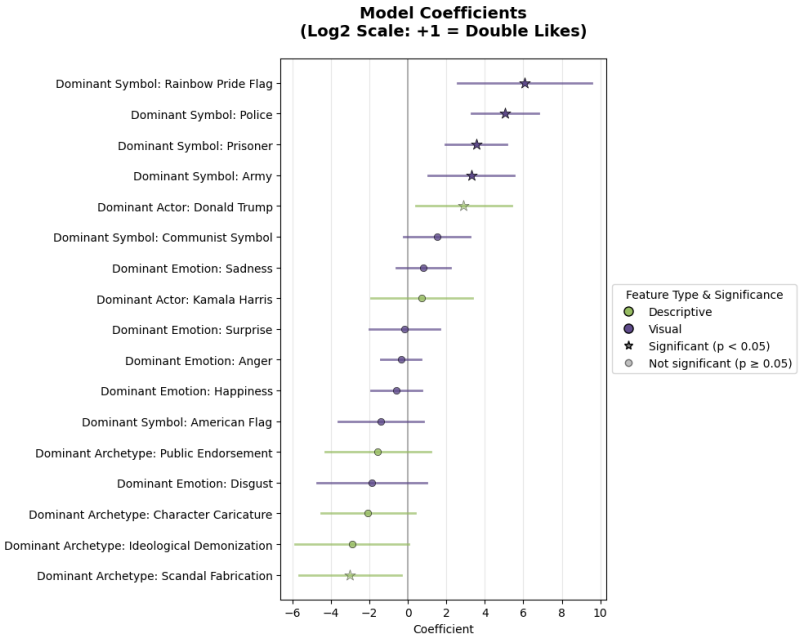

Figure 11: Coefficients and 95% confidence intervals for the regression model of political deepfake popularity. The target variable is the log2 of the number of retweets: a coefficient of +1 indicates that a feature is associated with double the amount of likes.).

further aggregated usernames to infer whether deepfakes were circulated by supporters or critics. Notably, a large portion of deepfakes targeting Trump appeared to be shared by users with "trump" in their usernames (compared with other presented public figures - Table 2)—suggesting they were either supporters or aligned with his persona. In several cases, the content was also disseminated by Trump's own account. This pattern indicates that deepfakes portraying Trump were not solely oppositional attacks but were also co-opted and circulated within sympathetic communities, potentially to mock, defuse, or reframe the narrative.

Table 2: Top 5 Deepfake Targets by Sharers Containing 'Trump'

Target	Count
Donald Trump	34
Taylor Swift	6
Kamala Harris	5
Joe Biden	4
Elon Musk	3

4.3.2 Temporal Dynamics of Deepfakes. Temporal analysis revealed sharp increases in deepfake activity during critical political moments (Fig 12), notably in March 2023 when Donald Trump was indicted by a New York grand jury over a hush money payment made to adult film star Stormy Daniels during the 2016 election [20] and between July and September 2024 when Joe Biden decided to withdraw from re-election [24]. These temporal spikes suggest that deepfakes are not randomly distributed but strategically timed to coincide with events where public opinion is particularly malleable. This pattern indicates a calculated effort to sway voters or derail political conversations during moments of national focus.

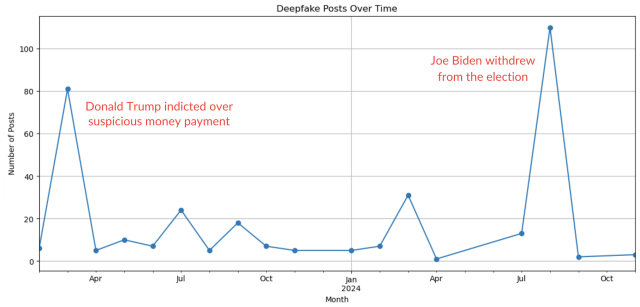

Figure 12: Distribution of deepfakes shared over time

4.3.3 Human-Driven Dissemination. Contrary to widespread fears about automated bots [9, 15], the top spreaders of deepfakes were mostly individual users. This finding implies a high degree of intentionality in the sharing of deceptive content. These users may be driven by ideological motives, desire for engagement, or efforts to provoke reactions. The human-driven nature of deepfake dissemination complicates efforts to combat it, as it reflects genuine community engagement rather than algorithmic amplification. This suggests that addressing deepfakes will require not only technical solutions but also public education and media literacy initiatives to reduce susceptibility to manipulation.

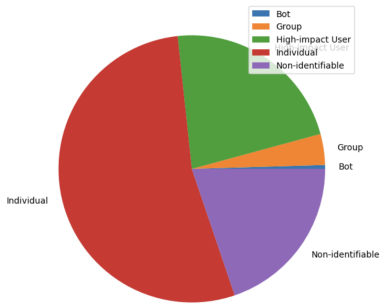

Figure 13: Types of deepfake posters

5 Discussion

Summary of Key Findings. Our analysis shows that political deepfakes online are deeply entangled with real-world contexts. They do not exist in isolation but emerge as responses to significant public events, mirroring how figures like Trump and Harris are perceived and contested. These synthetic artifacts reflect how public narratives are shaped—not only by their content but also by who shares them and when. We observe repeated patterns: Kamala Harris is often associated with extreme ideologies or ironic scandals, while Donald Trump is depicted both as an empowered icon and a victim of systemic repression.

These portrayals are not neutral. They consistently rely on emotionally charged expressions—especially anger and sadness—and symbolic elements that evoke themes of ideology, national identity, and authority. The use of emotionally and symbolically rich content suggests that deepfakes function not just as deception, but as tools of narrative construction. Dissemination is largely driven by individuals, many of whom are supporters of the figures depicted, further enhancing the perceived authenticity of the content. Together, these patterns point to a sophisticated mechanism of influence—one that weaponizes cultural symbolism and emotional appeal to align synthetic media with pre-existing political fault lines.

Limitations and Methodological Considerations. While our findings are robust within the scope of our analysis, several limitations must be acknowledged. First, our facial expression analysis relies on the seven-class Ekman taxonomy provided by Py-Feat, which excludes more nuanced political emotions like contempt or pride. Second, our heuristic for identifying the main subject—prioritizing the largest and most central face—is effective for simple compositions but may misinterpret complex scenes with multiple actors or unconventional layouts.

The primary methodological limitation is the absence of human validation for the model's outputs. Without human Facial Action Coding System (FACS) coders to verify a sample of Py-Feat's classifications, our confidence in the tool's accuracy does not guarantee its perceptual validity in the context of ambiguous or satirical deepfakes. Furthermore, while the PDID provides rich metadata, the social media metrics it contains (likes, shares, comments) measure only surface-level engagement, not genuine persuasion or attitude change in the audience. Finally, our study analyses plausible forgeries alongside overtly fantastical caricatures; this mixing of genres complicates inferences about how convincing deepfakes function in the wild.

Implications of the Work. This project demonstrates that a lightweight, fully local, and open-source pipeline can effectively extract meaningful behavioral signals from a large, unstructured database of political misinformation like the PDID. Our analysis pipeline addresses this gap by quantifying the broader thematic framings along with visual affects—dimensions not captured by existing metadata—to enables new lines of inquiry into how sensual cues align with patterns of dissemination and narrative construction. More broadly, our workflow offers a replicable and transparent methodology for misinformation forensics that avoids reliance on proprietary cloud APIs, making it accessible to academic and investigative communities alike.

Beyond technical utility, our findings point to important societal implications. As synthetic media becomes increasingly common on social platforms, users encounter emotionally charged and symbolically loaded content at scale. While some deepfakes appear humorous or absurd, others tap into victimhood, anger, or ideological tension, shaping how political figures are perceived and engaged with. These patterns are shown to be not accidental, rather timed and tailored for maximum impact. In this context, increasing public awareness of such mechanisms is critical. Media literacy initiatives—widely implemented across the US [1] and the EU [22]—has been promoting citizens' critical evaluation skills and awareness of synthetic media strategies. Beyond individual literacy, however, platforms, civil society groups, and policymakers must collaborate on verification infrastructures and moderation strategies grounded in empirical insights. Our work contributes actionable data that can help bridge communication gaps between technical experts, decision-makers, and the broader public.

Future Directions. Building on our current analysis centered primarily on the U.S. political context, future work should extend to other political landscapes and cultural settings to better understand how deepfake narratives manifest globally. Political misinformation and synthetic media reflect diverse social, cultural, and ideological dynamics that vary across regions; comparative studies could reveal unique patterns and cross-cultural similarities. Additionally, scaling the analysis to incorporate larger datasets and leveraging more advanced, up-to-date models will be crucial for capturing the evolving nature of deepfake narratives. Continuous monitoring with scalable methods can enhance the generalizability of findings and provide timely insights into how misinformation strategies adapt over time across different sociopolitical environments.

6 Conclusion

This work sheds light on the complex narratives embedded in political deepfakes and their strategic dissemination by influential actors and grassroots users alike. By combining multimodal analysis of emotion, symbolism, and narrative themes, we reveal how deepfakes manipulate public perception and engagement around key political figures. Our findings underscore the urgent need for comprehensive approaches to misinformation that address both the content and its pathways of spread, informing future research, policy, and media literacy efforts.

References

[1] 2022. S.4490 - Digital Citizenship and Media Literacy Act. https://www.congress.gov/bill/117th-congress/senate-bill/4490/text

[2] Soubhik Barari, Christopher Lucas, and Kevin Munger. 2025. Political deepfakes are as credible as other fake media and (sometimes) real media. *The Journal of Politics* 87, 2 (2025), 510–526.

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[4] Sarah Cahlan. 2020. How misinformation helped spark an attempted coup in Gabon. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/

[5] Pedro Cardoso, Joao Rodrigues, and Rui Novais. 2023. *Multimodal Emotion Classification Supported in the Aggregation of Pre-trained Classification Models*. 433–447. https://doi.org/10.1007/978-3-031-36030-5_35

[6] Jin Hyun Cheong, Eshin Jolly, Tiankang Xie, Sophie Byrne, Matthew Kenney, and Luke J. Chang. 2023. Py-Feat: Python Facial Expression Analysis Toolbox. arXiv:2104.03509 [cs.CV] https://arxiv.org/abs/2104.03509

[7] Erin Cook. 2019. Deep fakes could have real consequences for Southeast Asia. *Lowy Institute* 23 (2019).

[8] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2110013119.

[9] Karen Hao. 2021. A deepfake bot is being used to "undress" underage girls. https://www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/

[10] C. Kumar, N. Gowtham, Mohammed Zakariah, and Absulaziz Almazyad. 2024. Multimodal Emotion Recognition Using Feature Fusion: An LLM-Based Approach. *IEEE Access* PP (01 2024), 1–1. https://doi.org/10.1109/ACCESS.2024.3425953

[11] Yelena Mejova, Arthur Capozzi, Corrado Monti, and Gianmarco De Francisci Morales. 2025. Narratives of war: Ukrainian memetic warfare on twitter. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–28.

[12] Anat Rafaeli and Iris Vilnai-Yavetz. 2004. Instrumentality, aesthetics and symbolism of physical artifacts as triggers of emotion. *Theoretical Issues in Ergonomics Science* 5, 1 (2004), 91–112.

[13] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. 2022. Deepfake detection: A systematic literature review. *IEEE access* 10 (2022), 25494–25513.

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[15] Adam Satariano and Paul Mozur. 2023. The People Onscreen Are Fake. The Disinformation Is Real. https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html

[16] Kaylyn Jackson Schiff, Daniel S Schiff, and Natalia S Bueno. 2025. The liar's dividend: can politicians claim misinformation to evade accountability? *American Political Science Review* 119, 1 (2025), 71–90.

[17] Harisu Abdullahi Shehu, Will N. Browne, and Hedwig Eisenbarth. 2025. Emotion categorization from facial expressions: A review of datasets, methods, and research directions. *Neurocomputing* 624 (2025), 129367. https://doi.org/10.1016/j.neucom.2025.129367

[18] Andrew Ross Sorkin, Bernhard Warner, Sarah Kessler, Michael J. de la Merced, Lauren Hirsch, and Ephrat Livni. [n. d.]. https://www.nytimes.com/2023/05/23/business/ai-picture-stock-market.html

[19] Sam Stockwell. 2024. *AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections*. Technical Report. The Alan Turing Institute. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections

[20] The Guardian. 2023. Donald Trump indicted latest news charges protests court updates.

[21] The Guardian. 2024. Company that sent fake Biden robocalls in New Hampshire agrees to $1m fine. https://www.theguardian.com/technology/article/2024/aug/22/fake-biden-robocalls-fine-lingo-telecom Accessed: 2025-06-10.

[22] Riina Vuorikari, Natalia Jerzak, Zbigniew Karpinski, Artur Pokropek, Jadwiga Tudek, et al. 2022. Measuring digital skills across the EU: digital skills indicator 2.0. *Joint Research Centre Publications Office of the European Union* (2022), 4–26.

[23] Christina P Walker, Daniel S Schiff, and Kaylyn Jackson Schiff. 2024. Merging AI incidents research with political misinformation research: Introducing the political deepfakes incidents database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23053–23058.

[24] Wikipedia. 2024. Donald Trump indicted latest news charges protests court updates.

Our full data and code are publicly available at this link

In the appendix, we included example results regarding each of our data analysis procedure, demonstrating with real examples found in the set.

A Textual Feature Analysis

Topic	Count	Name	Representation	Representative_Docs	Dominant_Actor	Dominant_Archetype
15	13	15_drinking_bottle_kamala_harris	['drinking', 'bottle', 'kamala', 'harris', 'vodka', 'glass', 'outside', 'house', 'white', 'next']	['someone who looks like kamala harris is sitting in an alley drinking liquor next to a cardboard sign that says wanna be president ', 'kamala harris is depicted drinking a bottle of alcohol at a rally', 'kamala is drinking ', 'kamala harris drinking vodka from a bottle ', 'kamala harris drinking a bottle of vodka']	Kamala Harris	Scandal Fabrication
34	7	34_communist_stalin_joseph_picture	['communist', 'stalin', 'joseph', 'picture', 'flag', 'she', 'it', 'shaking', 'harris', 'her']	['it is an image of kamala harris and joseph stalin standing together and laughing', 'it is a picture of kamala harris in a communist uniform and holding a communist flag', 'it is a picture of kamala harris in a communist uniform and she is standing with a communist flag in her hand', 'it appears to be a picture of kamala harris shaking hands with joseph stalin she has a communist flag wrapped around her', 'it is a picture of kamala harris with a communist flag wrap around her and she is shaking hands with joseph stalin']	Kamala Harris	Ideological Demonization
9	16	9_drag_queen_mike_pence	['drag', 'queen', 'mike', 'pence', 'figures', 'various', 'pride', 'an', 'walking', 'dressed']	['mitch mcconnell as a drag queen', 'ted cruz is depicted as a drag queen ', 'mike pence is depicted as a drag queen wearing a bright pink outfit ', 'the video shows donald trump mike pence and lindsey graham dressed in drag donald trump is asking his republican supports to stop drag bans because he loves dressing in drag ', 'mike pence is dressed in drag']	Others	Character Caricature
8	16	8_taylor_swift_for_post	['taylor', 'swift', 'for', 'post', 'vote', 'swifties', 'shirts', 'sam', 'uncle', 'endorsing']	['a picture of trumps post with ai generated crowd pictures of taylor swift fans wearing swifties for trump shirts', 'taylor swift is dressed in an uncle sam outfit inciting people to vote for trump ', 'it is a screenshot of donald trumps post of an ai generated photo of taylor swift endorsing trump', 'it is a picture of trumps post that has ai generated taylor swift fans for trump', 'taylor swift is depicted in an uncle sam outfit to incite people to vote for him as taylor is ']	Donald Trump	Public Endorsement
21	11	21_deepfake_could_they_online	['deepfake', 'could', 'they', 'online', 'videos', 'deepfakes', 'because', 'obama', 'warning', 'barack']	['a deepfake of barack obama is sharing a message to his audience that they should be wary about videos they see on the internet because they could be a deepfake even if they look real ', 'an extremely realistic deepfake of barack obama is warning people to be wary of videos they see online because they could be deepfakes', 'a deepfake of barack obama is warning viewers to be careful when watching videos online because they could be deepfakes ', 'a deepfake of barack obama is warning the audience to take care when watching a video online because it could also be a deepfake ']	Barack Obama	Others

Figure 14: An excerpt of annotated BERTopic result (topic, count, representation) with corresponding dominant actors and archetypes

B Visual Feature Analysis

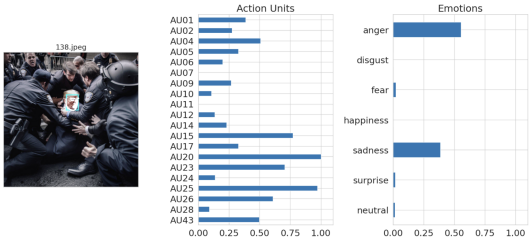

Figure 15: An example of facial emotion detection

Figure 16: An example of symbol detected - American Flag

Figure 17: Examples of images used to detect the communist symbol

C Dissemination Patterns

Figure 18: Sharer Network: Top 20 Communities with Top Spreaders, anonymized. Later analysis aggregating dominant public figures in sharers' names revealed the central nodes being largely related to Trump.