International Journal of Management, IT & Engineering
Vol. 16 Issue 07, July 2026,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Detection of Fake Identities in Social Media: A Comprehensive Literature Review

Sunita Gaur
School of Computer Science & Information Technology Devi Ahilya Vishwavidyalaya
Indore, India sunita9300@gmail.com

Yasmin Shaikh
International Institute of Professional Studies Devi Ahilya Vishwavidyalaya
Indore, India yasminshaikh01@gmail.com

***Abstract*—Fake identities, including fully automated bots, human-operated sockpuppets, and hybrid cyborgs, manipulate public discourse and amplify misinformation on social media. This review analyzes the technical evolution of fake identity detection from 2015 to 2026. As adversarial tactics advanced, defense mechanisms transitioned from feature-based machine learning to sequential deep learning and Graph Neural Networks (GNNs). Current state-of-the-art detectors face severe degrada-tion due to Generative AI, which produces synthetic personas capable of evading traditional NLP and structural filters. Along-side algorithmic evolution, this paper evaluates the statistical validity of modern benchmarks. It highlights how class imbal-ance and temporal drift artificially inflate reported accuracy metrics. Mitigating LLM-driven deception requires transitioning toward multimodal detection pipelines and robust statistical evaluation frameworks like MCC and AUC-PR.**

***Index Terms*—Fake Identity Detection, Social Bots, Graph Neural Networks, Generative AI, Cybersecurity**

I. Introduction

Social media platforms act as the primary infrastructure for global political discourse and news consumption. As a result, these networks are frequent targets for coordinated manipulation. The actors executing this manipulation rely on a variety of inauthentic accounts. These encompass fully automated social bots, human-operated sockpuppets, hybrid human-algorithmic cyborgs, and identity clones [1], [2].

The societal harms generated by these accounts are quan-tifiable and severe. False news spreads more rapidly and broadly than truth across all categories of information. The novelty hypothesis suggests that human psychology naturally prioritizes the distribution of shocking or unexpected informa-tion, a vulnerability that automated networks explicitly target [3]. Automated accounts disproportionately amplify this low-credibility content. They exploit algorithm mechanics during vulnerable periods like elections and public health crises [4]. Beyond discrete fake news dissemination, bots amplify hate speech and polarizing content to increase the baseline hostility of civil society [5]. This environmental degradation forces moderate voices out of online public squares. On a geopolit-ical scale, these identities facilitate computational propaganda and astroturfing campaigns executed by state and non-state actors across dozens of countries [6].

The ongoing arms race between platform manipulation and detection mechanisms directly dictates the integrity of public information [7]. This environment is no longer restricted to isolated scripts pinging platform APIs. It has evolved into complex participatory disinformation architectures where arti-ficial accounts feed narratives to witting or unwitting human influencers who then provide organic amplification [4].

This review synthesizes the technical literature on fake identity detection to address three core research questions:

1) How have detection algorithms evolved in response to adversarial evasion tactics, particularly the integration of Generative AI and Large Language Models?
2) What mathematical and structural flaws exist in current benchmark datasets and evaluation metrics that lead to real-world performance degradation?
3) Which technical methodologies offer the most viable defense against cross-platform, multimodal disinforma-tion networks operating in the current threat landscape?

Following the methodology outlined in Section II, the review defines the conceptual taxonomy of fake identities. Subsequent sections evaluate the technical progression of detection methodologies. We trace the path from traditional machine learning to modern Graph Neural Networks and detail the statistical validity of the metrics used to evaluate them. The review systematically assesses real-world platform deployments and defines the immediate technical requirements for detecting LLM-generated personas.

II. Review Methodology

To ensure a rigorous and reproducible synthesis of the existing literature, this paper adopts a Systematic Literature Review (SLR) methodology. This process was inspired by Kitchenham and Charters' structural guidelines for software engineering and computer science reviews. The methodology was executed in three primary phases: literature search, screening, and final selection.

A. *Search Strategy*

An exhaustive search was conducted across major academic databases. We queried IEEE Xplore, ACM Digital Library, ScienceDirect, and Springer. We also included arXiv to capture cutting-edge preprints on Generative AI capabilities, recognizing the rapid publication cycles required by modern machine learning developments.

The Boolean search queries were formulated us-ing combinations of targeted keywords. The primary string utilized was: (`"social bot detection"` OR

"fake identity detection" OR "sockpuppet" OR "cyborg account") AND ("Graph Neural Networks" OR "LLM" OR "Generative AI" OR "Coordinated Inauthentic Behavior" OR "feature engineering").

B. Inclusion and Exclusion Criteria

To maintain the quality and relevance of the review, strict inclusion criteria were applied. First, studies had to be pub-lished between 2015 and 2026. This window intentionally captures both the foundational rise of traditional social bots during the mid-2010s political events and the modern era of LLM-powered personas. Second, we restricted inclusion to peer-reviewed journal articles, high-impact conference pro-ceedings (e.g., NeurIPS, ACL, EMNLP, KDD, WWW), and highly cited preprints with demonstrated community consen-sus. Third, all papers had to be written in English.

Studies focusing solely on general network security, such as DDoS prevention or packet-level routing anomalies, were excluded. We required a specific social media or public digital forum context. Generic email spam filtering or basic CAPTCHA solving techniques were similarly discarded.

C. Study Selection and Data Extraction

The initial execution of the search strings yielded approxi-mately 320 records across all databases. Automated duplicate removal eliminated 85 records. The remaining 235 unique papers underwent a preliminary screening of their titles and abstracts by two independent reviewers. Discrepancies were resolved through a third-party consensus meeting. This phase eliminated 140 papers deemed tangentially related or outside the scope of the specific taxonomies under investigation.

The remaining 95 articles were subjected to a comprehen-sive full-text review. During this phase, studies were evaluated based on methodological rigor and dataset quality. We heavily weighted the introduction or utilization of major benchmarks like TwiBot-22 or MGTAB. Relevance to open challenges in the field, specifically adversarial evasion, was a critical factor. Ultimately, 51 primary studies were selected to form the core analytical foundation of this literature review.

III. Background & Conceptual Framework

A. Taxonomy of Fake Identities

Bot classification relies on two primary axes: the degree of human interference and the level of coordination. Secondary features such as temporal pattern replication further subdivide these categories into distinct operational models. It is critical to recognize that identities exist on a spectrum rather than in isolated silos.

Traditional spambots operate with rigid programming. They scrape content or repeat hardcoded phrases to drive traffic to external domains. Their lack of behavioral variance makes them highly susceptible to static threshold bans. Social spam-bots represent a step up in complexity. They curate personas, steal profile pictures, and maintain realistic posting schedules to evade basic scrutiny.

Cyborg accounts blend automation with manual interven-tion. A script might handle routine content scraping and network building, but a human operator takes control to engage in nuanced arguments or reply to high-profile figures [8]. This hand-off mechanism creates severe erratic spikes in stylometric analysis, confounding traditional NLP models.

Crowdturfers are distinct from algorithmic threats. These are real humans operating out of coordinated facilities or gig-economy platforms. They are paid micro-transactions to follow targets, like posts, or write synthetic reviews. Because the accounts have genuine device histories and localized IPs, they are entirely invisible to pure ML classifiers. Detection requires large-scale graph analysis to find the unnatural finan-cial centralization of their actions.

B. Motivations Behind Fake Identities

Fake identities serve specific functional purposes tied di-rectly to financial or political capital.

Political influence operations deploy bots to artificially inflate candidate support and suppress opposition. Roughly 20 percent of posts during the 2016 US presidential election originated from automated partisan accounts [9]. Distinct structural differences exist within these networks. Some iso-late efforts on amplifying specific narratives through heavy retweeting. Others flood networks with noise, a tactic known as "hashjacking," designed to dilute legitimate conversation and demobilize opposition [10].

TABLE I, Taxonomy of Inauthentic Identity Types

Identity Type	Controller	Primary Objective	Key Detection Vector	Resilience
Traditional Spambot	Script	Disseminate URLs / Malware	Content Repetition, URL Blacklists	Low
Social Spambot	Advanced Script	Influence Operations / Hype	Metadata, Graph Homophily	Medium
Sybil Cluster	Script / Human	Reputation Manipulation	Graph Connectivity, Lockstep Behavior	Medium-High
Sockpuppet	Human	Deception / Ban Evasion	Stylometry, Device Fingerprinting	High
Crowdturfer	Paid Human	Astroturfing / Fake Reviews	Network Anomaly, Payment Traces	High
LLM-Enhanced Bot	Generative AI	Conversational Influence	Adversarial Noise, Semantic Inconsis-tency	Extreme

Automated accounts act as early amplifiers for false stories. They generate momentum before human verification occurs [4]. This amplification effect identically applies to hateful and inflammatory content. By rewarding controversial statements with immediate algorithmic engagement, bots train human influencers to adopt more extreme positions to maintain their audience [5].

Financially motivated accounts operate in entirely different topologies. Commercial spam bots promote fake products, sell fraudulent engagement, and execute ad-click fraud. Cryp-tocurrency markets are particularly vulnerable. Massive bot networks coordinate pump-and-dump schemes by simultane-ously flooding specific tags with positive sentiment, driving algorithmic trending before liquidating their holdings at the expense of retail investors. Highly targeted deception, such as spear phishing, relies on impersonating journalists or institu-tions to extract data or funds directly.

C. *Historical Inflection Points in Bot Evolution*

Theoretical models of bot detection were forced to evolve primarily in response to massive real-world deployment events. Three distinct crises serve as the primary catalysts for architectural shifts in detection literature.

a) *Case Study 1: The 2016 US Election and the Failure of Feature ML:* The 2016 US Presidential Election represented the first widely documented instance of computational propa-ganda scaling to global geopolitical influence. Analysis of 20 million election-related tweets revealed that roughly 400,000 social bots generated 3.8 million posts. This accounted for nearly one-fifth of the entire political conversation [9].

This event exposed the terminal flaws in early machine learning detection. Prior to 2016, filtering relied heavily on catching high-volume spam accounts with sparse profiles. The bots deployed during the election were specifically engineered to bypass these exact models. They maintained balanced follower-to-following ratios. They interweaved organic stolen text with hyper-partisan links. They utilized programmed sleep cycles to mimic human circadian rhythms based on their stated geographic locations. Because they successfully spoofed the metadata that tools like Botometer relied upon, these accounts achieved survival rates exceeding 95 percent [11]. The failure to detect these networks in real-time forced the research community to abandon static profile features.

b) *Case Study 2: The COVID-19 Infodemic and Struc-tural Amplification:* The COVID-19 pandemic revealed how automated networks exploit structural topology during a fast-moving crisis. As public health organizations struggled to disseminate prevention measures, existing political botnets aggressively pivoted to amplify unverified health claims and alternative treatments.

During the pandemic, the detection challenge fundamen-tally shifted. Bots were no longer simply broadcasting spam. They acted as topological bridges. Research demonstrated that bots frequently embedded themselves at the core of misinformation diffusion networks. They served as primary sources that human users reposted 25 percent of the time. This effectively laundered the artificial origin of the claims [4]. This dynamic neutralized deep learning text classifiers.Once a human user quoted or reposted the misinformation, the text was indistinguishable from organic speech. To combat this "infodemic," detection architectures had to adopt Graph Neural Networks capable of identifying the coordinated bipartite structures of bots feeding content to human super-spreaders.

c) *Case Study 3: The Generative AI Explosion (2022-2024):* The widespread public release of robust Large Language Models like GPT-3.5 and LLaMA democratized high-tier deception. Previously, maintaining a persuasive sockpuppet required massive human labor or highly complex, brittle scripting. Generative AI allowed threat actors to spawn thou-sands of unique, contextually aware personas for fractions of a cent per prompt.

These models eliminated syntax errors, repetitive phrasing, and rigid posting schedules. Early detection models relying on stylometry collapsed. An LLM-driven bot could instantly adapt its tone from aggressive political discourse to casual sports commentary. It could hold sustained, multi-turn conver-sations with legitimate users, building deep trust over months before pivoting to payload delivery. This era definitively proved that analyzing individual account content in isolation was mathematically insufficient.

D. *Platforms & Ecosystems Under Study*

API access historically dictated detection research priori-ties. This skewed literature heavily toward open platforms and left massive blind spots in closed ecosystems.

E. *Twitter/X*

Twitter dominates the literature due to its historically per-missive streaming API. Foundational studies estimated active bot populations at 9 to 15 percent [12]. Research on this plat-form isolated pervasive electoral amplification, low-credibility content propagation, and coordinated partisan astroturfing [13]. The platform's structure, favoring public broadcast over private mutual connections, makes it the ideal testing ground for mass narrative injection.

F. *Meta (Facebook, Instagram, Threads)*

Restricted graph access on Meta platforms shifts detection toward internal structural network anomalies. Meta formalized this as Coordinated Inauthentic Behavior (CIB) [14]. CIB mapping relies on bipartite graphs tracing co-URL sharing and coordinated group administration rather than individual user profiling [15]. Because researchers cannot scrape com-plete social graphs, academic study relies on datasets volun-tarily released by the platform after takedown events.

G. *Decentralized and Visual Ecosystems (Telegram, TikTok)*

Platforms utilizing fundamentally different architectures pose severe challenges. Telegram's encrypted channels and API opacity prevent large-scale academic scraping. Threat actors use it as a command-and-control center, organizing human crowdturfers and coordinating cross-platform attacks. Visually driven platforms face unique threats. Algorithm engagement spam relies on view-botting and manufactured watch-time metrics rather than text generation. These visual

International Journal of Management, IT & Engineering
Vol. 16 Issue 07, July 2026,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

algorithms bypass traditional NLP-centric detection models entirely. Threat actors utilize generative AI to produce deep-fake audio and synthesized talking heads. These entirely short-circuit models designed to analyze syntax and metadata [16].

IV. Detection Methodologies

A. Feature-Based (Traditional ML) Approaches

Early detection relied on hand-engineered features processed by classifiers like Random Forests and Support Vector Machines [2], [17]. Systems categorized these features into distinct classes. Profile features measured account age, string entropy in usernames, and avatar presence. Behavioral features calculated the velocity of actions and the ratio of original posts to retweets.

The Botometer framework extracted over a thousand such variables across network topology, temporal rhythms, and sen-timent. It achieved high accuracy against early spam behavior [12], [18]. The models fundamentally assumed that malicious automation was sloppy. They looked for the cheapest, fastest execution paths.

This manual feature engineering failed outright against so-cial spambots. Attackers simply reverse-engineered the feature weights. If a model flagged accounts with high retweet-to-tweet ratios, the botnet controller added a script to scrape and post organic Wikipedia snippets to balance the ratio. They utilized profile picture generators to bypass default-avatar flags [11]. Consequently, traditional ML research pivoted toward scalable training data selection to force models into generalizing core traits rather than memorizing easily spoofed metadata [19].

B. Deep Learning Approaches

Deep learning eliminated the need for explicit metadata by analyzing raw behavioral sequences. Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) differentiate bots based purely on the chronological sequence of timeline events [20].

For a given timeline sequence $X = (x_1, x_2, ..., x_t)$, the LSTM memory cell updates its hidden state h_t via the forget gate $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ and input gate i_t. This architecture inherently models the temporal rhythm of posts without relying on static profile aggregates. A bot might execute actions with machine precision that an LSTM easily captures in its hidden states, even if the overall ratios look human. Integrating BiLSTMs with word embeddings provided early robust baselines against sophisticated social spambots [21].

Transformer architectures revolutionized content evaluation. BERT and RoBERTa provided deep contextual embeddings capable of identifying semantic contradictions across a user's timeline [22], [23]. Current pipeline architectures heavily favor multimodal integration. They combine dense text em-beddings from Transformers with temporal sequencing and image analysis using Convolutional Neural Networks (CNNs) in hybrid neural networks [24], [25].Conceptually, this acts as a mathematical neighborhood watch. The adjacency matrix A aggregates the behavioral traits of a user's direct contacts. The degree matrix D normalizes the data so highly connected hub accounts don't heavily skew the results. If an account is heavily embedded in a cluster of automated profiles, this propagation rule forces the central account to inherit a bot-like signature. To support these massive topological maps, benchmarks have expanded aggressively. They jumped from TwiBot-20, evaluating 229,000 users [27], to TwiBot-22 mapping 92 million users and over a billion edges [28].

C. Graph & Network Analysis

Because follower networks and interaction graphs resist localized profile manipulation, this sub-field has become the primary focus of modern detection architectures. An attacker can easily fake a bio or generate a human-like tweet, but forcing thousands of legitimate, high-reputation accounts to follow their bot requires immense capital.

Graph Convolutional Networks (GCNs) isolate bots by analyzing the structural asymmetry of their connections [26]. The core algorithm updates a user's profile representation layer-by-layer:

$$H^{l+1} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{l}W^{l}) \quad (1)$$

State-of-the-art models like the Relational Graph Trans-former (RGT) improve upon basic GCNs by recognizing that not all social ties are equal [29]. RGT calculates an attention coefficient α_{ij} based on the specific type of edge interaction r_{ij}. The model learns relational context. It easily differentiates between a synchronized retweet and an organic text reply, assigning higher threat weights to coordinated behavior. Graph theory is identically applied to Coordinated Inauthentic Be-havior detection via bipartite networks. These track accounts that share identical obscure URLs or amplify niche hashtags with microsecond synchronization [15], [30].

D. NLP & Linguistic Analysis

Early textual detection isolated bots through stylometric measurements. Models evaluated vocabulary richness, part-of-speech ratios, and grammatical error rates. Generative AI functionally bypasses these filters. LLMs generate text with perfect syntax and high perplexity variance.LLM-generated fake professional personas successfully evade both standard stylometry and specialized BERT-based classifiers [31]. Because modern synthetic text lacks the struc-tural anomalies previous NLP models were trained to flag, detecting AI-generated personas represents a hard failure point in current literature. Researchers are exploring completely new linguistic metrics, evaluating structural repetition at the token level or utilizing zero-shot detection via the source LLM's own probability distributions [16], [32].

E. Temporal & Behavioral Analysis

Biological sequence alignment algorithms can encode chronological user actions into fixed strings. A sequence of tweet, reply, retweet is encoded as T-R-RT. Algorithms compare these strings across thousands of users to identify automated uniformity across seemingly unconnected accounts

[33]. Time-series tools utilize warped correlation to detect synchronized activity bursts indicative of centrally controlled botnets reacting to an API command [34].

Longitudinal analysis isolates predictable operational life-cycles. It exposes networks that maintain total dormancy for months before executing highly aggressive activation spikes during targeted political events [35]. This dormancy is de-signed specifically to age the accounts, bypassing filters that aggressively scrutinize new registrations.

F. Comparative Performance of Key Detection Models

To synthesize the methodological evolution, it is necessary to benchmark how different architectural paradigms perform against modern datasets. Table 2 compares three highly influ-ential models representing the primary eras of detection.

Feature-based models achieved early success but fall prey to adversarial metadata manipulation. Deep learning handles se-quential text well but lacks awareness of the broader network topology. Graph Neural Networks currently achieve state-of-the-art results by mapping heterogeneous relations. However, they require immense computational overhead to process the entire graph and struggle with concept drift over time.

V. Datasets & Benchmarks

Manual annotation of bot datasets suffers from high com-putational cost and notably low inter-annotator agreement. Humans are terrible at identifying sophisticated bots. To achieve scale, researchers frequently default to convenience sampling. They aggregate accounts recently suspended by platforms or isolate users sharing identical anomalous hash-tags.

This methodology introduces severe statistical artifacts. Detectors trained on convenience samples learn the artifact rather than the underlying bot behavior. If all the positive ex-amples in a training set were scraped via a specific right-wing political hashtag, the model learns to associate conservative text with bots, not the automation itself. This causes immedi-ate accuracy collapse in out-of-distribution deployments [36]. Flawed ground truth generates systemic false-positive rates. It routinely results in the misclassification of hyper-active human users, particularly neurodivergent individuals or heavy gamers, as automated scripts [37].

These datasets also undergo rapid temporal degradation. Detectors trained on 2015 data perform near random chance against bots captured in 2020. The platform meta evolves, UI features change, and bot developers iterate. Persistent class imbalance and concept drift enforce strict expiration dates on all training data [38], [39].

A. The Formal Mathematical Flaws in Standard Evaluation Metrics

A critical vulnerability in fake identity detection literature is the reliance on mathematical frameworks that fail to account for deployment realities. Historically, researchers have leaned heavily on Accuracy and the Receiver Operating Character-istic Area Under the Curve (ROC-AUC) to claim algorithmic superiority.

In real-world network traffic, inauthentic accounts represent a severe minority class. Even during highly active computa-tional propaganda campaigns, bots rarely account for more than 5 to 10 percent of total platform users. When evaluated on highly imbalanced data, traditional metrics collapse.

Accuracy is wildly misleading. If a dataset contains 95 percent legitimate human users and 5 percent bots, a broken classifier that statically predicts "Human" for every account achieves 95 percent Accuracy. It identifies zero malicious identities, yet appears robust on paper.

While the F1-Score is widely adopted to address this, it completely ignores True Negatives. Successfully identifying legitimate users without accidentally suspending them is crit-ical for platform health. F1 provides an incomplete picture of moderation safety.

a) *The Case for MCC and AUC-PR:* To guarantee a detection model is learning genuine behavioral differences, literature must shift toward robust statistical indicators. The Matthews Correlation Coefficient (MCC) is increasingly rec-ognized as the most reliable single-value metric in imbalanced environments. MCC evaluates all four quadrants of the confu-sion matrix.

Because the equation normalizes the coefficient by the marginal sums, MCC generates a high score only if the classifier correctly identifies a high percentage of both the bot minority class and the human majority class. If a model guesses the majority class, the numerator zeroes out, exposing the model's ineffectiveness.

Similarly, ROC-AUC calculates the False Positive Rate against the True Negative count. Because the number of legitimate human accounts is astronomically high, the False Positive Rate remains artificially low even if thousands of humans are incorrectly classified as bots. Transitioning to the Area Under the Precision-Recall Curve (AUC-PR) resolves this by plotting Precision directly against Recall. This com-pletely isolates the model's performance on the minority bot

TABLE II, Comparison of representative bot detection models.

Paradigm	Base Architecture	Benchmark	Reported Metric	Primary Vulnerability
Feature ML	Random Forest (Botome-ter)	Varol-2017	High accuracy on early spam.	Evaded by metadata mimicking and tem-poral camouflage.
Deep Learning	Hybrid DNN + LSTM	Cresci-2017	AUC up to 0.97 on legacy data.	Lacks structural awareness; vulnerable to LLM text.
Graph Network	Relational Graph Trans-former	TwiBot-22	86% Acc, 97% F1.	Extreme class imbalance creates spurious edge aggregation.

class and strips away the artificially inflated True Negative count.

VI. Challenges & Open Problems

Current detection methods consistently lag behind adver-sarial capabilities. Closing this gap requires addressing deep technical limitations across generalizability, dataset construc-tion, and operational ethics.

Adversaries actively program around publicized detection thresholds. If a paper demonstrates that an account tweeting more than 50 times a day is likely a bot, the bot herder simply updates the script to cap out at 49 tweets [40]. LLMs eliminate the operational cost of generating persuasive personas entirely. Detectors inherently struggle with generalizability. Models memorize dataset quirks and fail catastrophically on cross-platform data. A model trained on Twitter data is useless on Reddit. Scalable training data selection mitigates minor degradation, but platform API changes and internet culture shifts strictly limit model deployment lifespans [39].

Automated moderation utilizing flawed detection systems routinely silences legitimate political speech. Platform-native CIB enforcement occurs opaquely, granting proprietary algo-rithms unchecked authority over content visibility without public oversight or appeal processes [6].

VII. Detection Tools & Real-World Deployments

Botometer scaled academic feature-based machine learning via a public API that processed millions of daily requests. Sustained evasion tactics forced complete architectural over-hauls across its operational lifespan. The dependency on this API within the social sciences necessitated published guardrails specifically dictating its limitations. Researchers built supplementary tools like Bot-Hunter to provide multi-tier detection characterization [41].

Twitter's API restrictions disabled real-time public detec-tion pipelines. Large-scale detection infrastructure is now almost exclusively proprietary and maintained by platform-native security teams. Meta's internal systems target the broader structural orchestration of Coordinated Inauthentic Behavior over individual node classification.

State-sponsored troll farms operate with complex strategic objectives. They attempt targeted demographic polarization rather than simple spam. This requires highly advanced tech-niques like Inverse Reinforcement Learning to map effectively [42]. All deployed systems face severe sociotechnical limita-tions where statistical false positives directly manifest as the censorship of real human users [40].

VIII. Future Research Directions

1) Generative AI Resilience: Existing NLP detectors fail systematically against GPT-era synthetic content. Research must prioritize specialized counter-LLM clas-sifiers and the continuous generation of adversarial synthetic training datasets. We need models capable of identifying LLM watermarks or token probability distributions natively.Multimodal Architectures: Disinformation utilizes di-verse media across multiple networks simultaneously. Pipelines must jointly evaluate linguistic intent, video artifacts, and temporal graphs concurrently rather than in isolated sub-routines [25].
2) Federated Detection: Platform API restrictions mandate a shift toward decentralized, privacy-preserving mod-els. Federated learning allows cross-platform training on local device data without exposing raw user records to centralized academic servers [43].
3) Standardization and Explainability: Methodological fragmentation requires standardized temporal drift evaluation frameworks. Opaque moderation demands algorithmic explainability. When an account is banned, the system must produce an auditable trace of the features that triggered the enforcement [44].
4) Collaborative Pipelines: Fully automated moderation is brittle. Implementations should shift toward Human-AI collaborative pipelines. They should pair machine anomaly flagging at scale with specialized human con-textual enforcement for borderline cases.

IX. Conclusion

The 2016 baseline of automated script detection relied entirely on the assumption that malicious actors operated cheaply and predictably [7]. Early researchers treated bots like static network intrusions, assuming they would endlessly repeat the same behavioral patterns. This paradigm has com-pletely collapsed. It has been superseded by the requirements of tracking generative AI personas and hyper-coordinated adversarial networks. The threat model is no longer a single script flooding a trending hashtag. It is an algorithmic orchestration involving deepfake media, synthetic language generation, and intricate network topology designed specif-ically to hijack human psychology. Modern disinformation campaigns operate like living organisms that actively adapt to whatever detection thresholds academics publish.

Graph-based topological models have rapidly advanced classification capabilities by targeting the one resource attack-ers cannot easily simulate. That resource is organic human connections [28]. It requires immense capital and time to trick thousands of real users into following a synthetic account. However, the heavy reliance on these massive computational networks exposes systemic vulnerabilities. Temporal degra-dation rapidly decays model accuracy as platform interfaces and cultural trends shift [44]. Furthermore, systemic dataset artifact bias heavily restricts real-world reliability [36]. The rigid statistical reliance on Accuracy and F1 scores mathe-matically hides severe false positive rates. When a model prioritizes overall accuracy in a highly imbalanced social environment, it inevitably generates collateral damage against legitimate human speech. We frequently end up algorithmi-cally suspending the very users the platforms are supposed to protect.

Purely computational improvements are insufficient to solve what

is fundamentally a sociotechnical vulnerability. The continuous cycle of adversarial machine learning mathemat-

ically favors the attacker in an open social system. They only need to find a single flaw or contextual blind spot in the detection model. Defenders must attempt to build models immune to infinite variance. It is becoming increasingly clear that you cannot solve a human deception problem entirely with a mathematical classification function.

Consequently, tracking global computational propaganda necessitates structural platform transparency alongside strict regulatory frameworks [6]. The black-box nature of current platform moderation APIs fundamentally stifles academic peer review. Right now, independent researchers are effectively locked out of the platforms they are attempting to study. Without transparent data access, the scientific community is relegated to studying the history of botnets rather than evalu-ating their current live deployments. We spend thousands of hours analyzing what bots did two years ago simply because that is the only data the platforms are willing to release to the public.

Future detection pipelines depend heavily on extreme inter-disciplinary integration. The isolated engineering silos need to break down. Computer scientists must build the multimodal transformer pipelines required to parse LLM text and graph topology. Sociologists must map the shifting cultural context of online polarization so the algorithms actually understand which societal narratives are currently vulnerable to injec-tion. Legal scholars must formally define the thresholds of acceptable automated moderation. Only by combining robust statistical paradigms like MCC with decentralized federated tracking can platforms protect public discourse without com-promising digital civil liberties [40]. The era of the easily detectable spammer is permanently over, and the current era of synthetic societal manipulation requires an equally sophis-ticated and collaborative defense.

References

[1] A. Alharbi, H. Dong, X. Yi, Z. Tari, and I. Khalil, "Social Media Identity Deception Detection: A Survey," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–35, 2021, doi: 10.1145/3446372.

[2] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Systematic Literature Review of Social Media Bots Detection Systems," *Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 6, p. 101571, 2023, doi: 10.1016/j.jksuci.2023.04.004.

[3] S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018, doi: 10.1126/science.aap9559.

[4] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The Spread of Low-Credibility Content by Social Bots," *Nature Communications*, vol. 9, no. 1, p. 4787, 2018, doi: 10.1038/ s41467-018-06930-7.

[5] M. Stella, E. Ferrara, and M. De Domenico, "Bots Increase Exposure to Negative and Inflammatory Content in Online Social Systems," *Proceedings of the National Academy of Sciences*, vol. 115, no. 49, pp. 12435–12440, 2018, doi: 10.1073/pnas.1803470115.

[6] S. C. Woolley and P. N. Howard, "Computational Propaganda: Polit-ical Parties, Politicians, and Political Manipulation on Social Media," *Computational Propaganda*. Oxford University Press, 2019.

[7] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The Rise of Social Bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016, doi: 10.1145/2818717.

[8] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting Automa-tion of Twitter Accounts: Are You a Human, Bot, or Cyborg?," in *IEEE Transactions on Dependable and Secure Computing*, IEEE, 2012, pp. 811–824. doi: 10.1109/TDSC.2012.75.

[9] A. Bessi and E. Ferrara, "Social Bots Distort the 2016 U.S. Presidential Election Online Discussion," *First Monday*, vol. 21, no. 11, 2016, doi: 10.5210/fm.v21i11.7090.

[10] L. Luceri, A. Deb, A. Badawy, and E. Ferrara, "Red Bots Do It Better: Comparative Analysis of Social Bot Partisan Behavior," pp. 1016–1021, 2019, doi: 10.1145/3308560.3316735.

[11] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 963–972. doi: 10.1145/3041021.3055135.

[12] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM)*, 2017, pp. 280–289. doi: 10.1609/ icwsm.v11i1.14871.

[13] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political Astro-turfing on Twitter: How to Detect Fake Grassroots Campaigns," *Political Communication*, vol. 37, no. 1, pp. 135–159, 2020, doi: 10.1080/10584609.2019.1661888.

[14] N. Gleicher, "Coordinated Inauthentic Behavior Explained," *Meta Newsroom (Facebook)*, 2018, [Online]. Available: https://about.fb.com/ news/2018/12/inside-feed-coordinated-inauthentic-behavior/

[15] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering Coordinated Networks on Social Media: Methods and Case Studies," vol. 15, pp. 455–466, 2021, doi: 10.1609/ icwsm.v15i1.18075.

[16] E. Ferrara, "Social Bots, Deepfakes, and Disinformation: AI for Political Influence Operations," *IEEE Internet Computing*, vol. 27, no. 5, pp. 11–17, 2023, doi: 10.1109/MIC.2023.3260622.

[17] M. Latah, "Detection of Malicious Social Bots: A Survey and a Refined Taxonomy," *Expert Systems with Applications*, vol. 151, p. 113383, 2020, doi: 10.1016/j.eswa.2020.113383.

[18] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "BotOrNot: A System to Evaluate the Credibility of Twitter Accounts," in *Proceedings of the 25th International World Wide Web Conference Companion*, 2016, pp. 273–274. doi: 10.1145/2872518.2889302.

[19] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and Gener-alizable Social Bot Detection through Data Selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 1096–1103. doi: 10.1609/aaai.v34i01.5460.

[20] S. Kudugunta and E. Ferrara, "Deep Neural Networks for Bot Detec-tion," 2018, pp. 312–322. doi: 10.1016/j.ins.2018.08.019.

[21] S. Yang, L. Gu, J. Jiang, and M. Luo, "Twitter Bot Detection Using Bidirectional Long Short-Term Memory Neural Networks and Word Embeddings," in *Proceedings of the 2020 First International Conference on Societal Automation (SA)*, 2020. doi: 10.1109/ SA51202.2020.9364014.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Train-ing of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT 2019*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[23] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019, doi: 10.48550/ arXiv.1907.11692.

[24] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb, and S. S. Mathew, "DeeProBot: A Deep Hybrid Profile-Based Model for Social Bot Detection," *Social Network Analysis and Mining*, vol. 12, no. 1, p. 43, 2022, doi: 10.1007/s13278-022-00872-z.

[25] F. Alam, H. Mubarak, W. Gao, and P. Nakov, "A Survey on Multi-modal Disinformation Detection," pp. 6625–6643, 2022, doi: 10.48550/ arXiv.2103.12541.

[26] F. Wei and U. T. Nguyen, "Twitter Bot Detection Using Graph Convo-lutional Network," 2019, doi: 10.1109/GC46384.2019.00020.

[27] S. Feng, H. Wan, N. Wang, J. Li, and M. Luo, "TwiBot-20: A Compre-hensive Twitter Bot Detection Benchmark," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Manage-ment (CIKM)*, 2021, pp. 4485–4494. doi: 10.1145/3459637.3482019.

[28] S. Feng *et al.*, "TwiBot-22: Towards Graph-Based Twitter Bot Detec-tion," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.48550/arXiv.2206.04564.

[29] S. Feng, Z. Tan, R. Li, and M. Luo, "Heterogeneity-Aware Twitter Bot Detection with Relational Graph Transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 3977–3985. doi: 10.1609/aaai.v36i4.20314.

[30] M. Cinelli, E. Brugnoli, A. L. Schmidt, F. Zollo, W. Quattrociocchi, and A. Scala, "Coordinated Inauthentic Behavior and Information Spreading on Twitter," *Decision Support Systems*, vol. 160, p. 113819, 2022, doi: 10.1016/j.dss.2022.113819.

[31] N. Ayoobi, S. Shahriar, and A. Mukherjee, "The Looming Threat of Fake and LLM-Generated LinkedIn Profiles: Challenges and Opportu-nities for Detection and Prevention," in *Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HT'23)*, 2023, pp. 1–13. doi: 10.1145/3603163.3609064.

[32] C. Chen and K. Shu, "Can LLM-Generated Misinformation Be De-tected?," 2024, doi: 10.48550/arXiv.2309.13788.

[33] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection," *IEEE Intelligent Systems*, vol. 31, no. 5, pp. 58–64, 2016, doi: 10.1109/MIS.2016.29.

[34] N. Chavoshi, H. Hamooni, and A. Mueen, "DeBot: Twitter Bot Detec-tion via Warped Correlation," in *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 817–822. doi: 10.1109/ICDM.2016.0096.

[35] L. Luceri, S. Giordano, and E. Ferrara, "Down the Bot Hole: Actionable Insights from a One-Year Analysis of Bot Activity on Twitter," *First Monday*, vol. 26, no. 1, 2021, doi: 10.5210/fm.v26i1.11174.

[36] C. Hays, A. Schofield, K. Yang, F. Menczer, and C. Callison-Burch, "Simplistic Collection and Labeling Practices Limit the Utility of Benchmark Datasets for Twitter Bot Detection," in *Proceedings of the ACM Web Conference 2023 (WWW)*, 2023, pp. 3660–3669. doi: 10.1145/3543507.3583214.

[37] A. Rauchfleisch and J. Kaiser, "The False Positive Problem of Auto-matic Bot Detection in Social Science Research," *PLOS ONE*, vol. 15, no. 10, p. e241045, 2020, doi: 10.1371/journal.pone.0241045.

[38] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "Bot Datasets on Twitter: Analysis and Chal-lenges," *Applied Sciences*, vol. 11, no. 9, p. 4105, 2021, doi: 10.3390/ app11094105.

[39] K.-C. Yang, E. Ferrara, and F. Menczer, "Botometer 101: Social Bot Practicum for Computational Social Scientists," *arXiv preprint arXiv:2201.01608*, 2022, doi: 10.48550/arXiv.2201.01608.

[40] C. Grimme, M. Preuss, L. Adam, and H. Trautmann, "Changing Perspectives: Is It Enough to Detect Social Bots?," *Social Media + Society*, vol. 4, no. 3, 2018, doi: 10.1177/2056305118796858.

[41] D. M. Beskow and K. M. Carley, "Bot-Hunter: A Tiered Approach to Detecting & Characterizing Automated Activity on Twitter," in *Confer-ence on Applications of Network Analysis and Information (ASONAM)*, 2018. doi: 10.1145/3269206.3272049.

[42] L. Luceri, F. Cardoso, S. Giordano, and E. Ferrara, "Detecting Troll Behavior via Inverse Reinforcement Learning: A Case Study of Russian Trolls in the 2016 US Election," vol. 14, pp. 417–427, 2020, doi: 10.1609/icwsm.v14i1.7311.

[43] T. D. Nguyen, L. Holzinger, A. Binder, and W. Samek, "Federated Learning for Social Bot Detection: Benchmarks and Challenges," *arXiv preprint arXiv:2212.11628*, 2022, doi: 10.48550/arXiv.2212.11628.

[44] S. Cresci, "A Decade of Social Bot Detection," *Communications of the ACM*, vol. 63, no. 10, pp. 72–83, 2020, doi: 10.1145/3409116.