Adapting Bot Detection Models for Romania's Disinformation Ecosystem

Ștefania-Elena Stoica
Carol I National Defense University, Bucharest, Romania

stoica.stefania@myunap.net

Abstract: The proliferation of social media bots and fake accounts has significantly disrupted information ecosystems, posing substantial challenges in detecting and mitigating disinformation. While machine learning and deep learning models have shown varying levels of success on platforms like Twitter and Facebook, they often fail to account for region-specific nuances critical for effective bot detection. Facebook and Twitter have been widely used in disinformation research due to their large user bases and historically open API access, facilitating large-scale data collection. This study addresses these gaps by proposing a hybrid detection framework tailored to Romania's disinformation landscape. Each society is shaped by its unique historical, cultural, linguistic, and geopolitical factors, influencing how disinformation spreads and resonates with different audiences. The proposed approach emphasizes well-established narratives that significantly influence vulnerable populations, including young adults with limited capacity for fact-checking and older adults with low levels of digital literacy. This research will begin by reviewing existing literature on bot detection methodologies and narrative analysis, identifying their strengths, limitations, and applicability to regional contexts. By integrating conventional detection methodologies with a refined analysis of niche and high-risk narratives, this research investigates how disinformation campaigns gain momentum and escalate, providing a deeper understanding of their dynamics and impact. The results will reveal patterns and strategies employed in the propagation of disinformation, contributing to the development of more targeted and effective detection systems. This method also facilitates the early identification of emerging disinformation clusters, offering timely and proactive intervention opportunities. This paper is part of a broader PhD research program centered on analysing narratives and narrative strategies in disinformation, with all findings supporting this overarching goal. The findings emphasize the importance of regional customization in bot detection frameworks, particularly for countries like Romania, where disinformation leverages historical, cultural, and socio-political triggers. These insights will strengthen the resilience of information ecosystems and hold significant value for cybersecurity professionals, social media platforms, and policymakers dedicated to combating manipulation and fostering a more secure digital environment.

Keywords: Social media bots, Regional disinformation, Bot detection, Romania, Narrative analysis

1. Introduction

The integration of AI into social media has fundamentally altered digital communication, making interactions between AI-generated bots and human users increasingly seamless. Companies like Meta have leveraged AI "characters" to enhance engagement and monetization, exemplified by their 2023 launch of celebrity-modeled chatbots on Instagram and Facebook. While these bots foster interactivity, they also raise concerns about transparency, authenticity, and ethical responsibility. Studies suggest that revealing a bot's identity negatively impacts engagement in real-time interactions, as users expect human-like responses and feel disengaged when the illusion of authenticity is broken (Kumar & Shah, 2018; Cresci, 2020). Additionally, the rise of low-quality AI-generated content, or "slop" (Ferrara et al., 2016; Davis et al., 2016), diminishes meaningful discourse, as engagement-driven algorithms prioritize quantity over quality. This influx of automated content increases digital clutter and further obscures the distinction between human and machine communication.

Disinformation campaigns exploit these vulnerabilities, strategically manipulating social tensions to influence public opinion and deepen polarization. By leveraging divisive topics such as immigration and political corruption, these campaigns work to erode institutional trust and reinforce existing anxieties (Ratkiewicz et al., 2010; Vosoughi et al., 2018). While the mechanisms of disinformation remain largely consistent across different regions, their impact varies depending on cultural and historical contexts. Romania exemplifies this dynamic, as its geopolitical position between East and West—balancing democracy and autocracy, Western capitalism and post-communist economic structures—makes it uniquely susceptible to tailored disinformation efforts (Tandoc et al., 2020). Persistent institutional distrust, unresolved historical grievances, and socio-economic disparities further heighten vulnerability. Additionally, younger adults with limited fact-checking skills and older adults with low digital literacy are particularly at risk, as they are more susceptible to manipulative narratives.

Despite advancements in bot detection, existing methods struggle to counter AI-driven disinformation effectively. The adaptability of modern disinformation campaigns mirrors targeted cyberattacks, such as spear phishing, which tailor messages to specific audiences for greater impact. Broad detection strategies remain inadequate, as they fail to account for the regional and cultural nuances that shape disinformation's effectiveness (Subrahmanian et al., 2016; Varol et al., 2017). Over-reliance on simplistic detection approaches

risks over-censorship and reactive moderation while neglecting deeper systemic concerns, such as algorithmic bias and privacy violations. Addressing these challenges requires a more comprehensive approach that goes beyond bot identification.

2. Literature Review

2.1 Bot Detection Methodologies

Social bots' evolution and detection methods can be categorized into four periods, each representing significant advancements and challenges in understanding and combating their impact (Rossi & Sippo, 2024) The first phase, from 2010 to 2014, coincided with the rise of platforms like Twitter and Facebook. While limited in scope, early studies on automated accounts established a foundation for social bot research by focusing on their functionality and influence. Evidence of bots being deployed in operations such as elections dates to 2010 (Ratkiewicz et al., 2010, 2011).

The 2015–2016 period saw increasing recognition of social bots as a threat, driven by initiatives like the DARPA Twitter Bot Challenge (Subrahmanian et al., 2016) and tools such as Botometer (Davis et al., 2016). Ferrara et al.'s (2016) review, *The Rise of Social Bots*, further underscored their impact. Concerns escalated after evidence of Russian interference in the 2016 U.S. election, where bots played a key role in spreading disinformation (Guilbeault & Woolley, 2016; O'Connor & Schneider, 2017). While their precise influence remains debated (Bail et al., 2020), investigations confirmed coordinated bot activity amplifying divisive narratives (Senate Intelligence Committee, 2019; Mueller, 2019).Between 2017 and 2022, an "arms race" emerged as bots grew more sophisticated, prompting advances in detection methods. Research focused on documenting bot activity in events like elections and refining methodologies, particularly through machine learning and multimodal analysis (Cresci, 2020a). By 2024, bots accounted for 5% of social media accounts, necessitating continued improvements in detection. In 2023, Twitter's API restrictions disrupted research, requiring alternative data collection strategies and highlighting the field's adaptability.

Supervised learning, which relies on labeled datasets to train models, remains the dominant method in bot detection research (Aljabri et al., 2023). This approach works by feeding a machine learning algorithm a set of examples—social media accounts manually classified as either bots or humans—so the model can learn to recognize patterns distinguishing the two. Features such as posting frequency, network connections, linguistic patterns, and metadata are commonly used to train these models. Once trained, the model applies these learned patterns to classify new accounts accurately. Its high accuracy in distinguishing between bots and humans has made it a preferred approach. However, its reliance on human-labelled datasets introduces significant limitations. Human annotators frequently misclassify accounts due to the increasing complexity of bot behavior, leading to false positives or negatives (Rauchfleisch & Kaiser, 2020). Moreover, the lack of fully reliable datasets highlights a critical challenge in supervised learning, stressing the need for automated and robust labelling mechanisms (Hays et al., 2023).

On the other hand, unsupervised learning presents the potential for detecting novel or evolving bot behaviors without needing labelled data. Techniques such as clustering algorithms or anomaly detection could be particularly valuable, especially in situations with limited labelled datasets (Zhou et al., 2022). Nevertheless, unsupervised methods remain underexplored, partly due to the historical dominance of supervised learning enabled by Twitter API access (Cresci et al., 2017).

Detection strategies that emphasize user coordination aim to uncover orchestrated activities. These methods analyse content similarity and temporal activity patterns, which can reveal suspiciously synchronized behavior (Aljabri et al., 2023). For instance, Nizzoli et al. (2020) demonstrated how retweet networks could identify coordinated groups spreading disinformation during the 2019 European elections. These approaches underscore the importance of analysing group behavior alongside individual account activity.

Multimodal approaches that integrate text, network, and image analysis are increasingly recognized for their ability to detect complex bot behaviors (Cresci et al., 2021). Meanwhile, Generative Adversarial Networks (GANs) present both opportunities and challenges. They allow researchers to simulate advanced bot behaviors to enhance detection algorithms and enable malicious actors to create bots capable of evading detection (Li et al., 2021).

As bot detection evolves, advancements in real-time data analysis and explainable AI hold promise for tackling current challenges. These innovations increase detection accuracy and ensure transparency, fostering greater trust in automated moderation systems.

3. Narrative Landscape: Manipulation During the December 2024 Romanian Elections

The December 2024 elections in Romania exposed significant vulnerabilities in the nation's democratic processes, especially through exploiting social media platforms. The surprising success of far-right, pro-Russian candidates in the first round of the presidential election underscored how digital manipulation strategies could dramatically alter electoral outcomes. Multiple campaigns utilized TikTok's algorithms and vast influencer networks, leading his content to accumulate over 100 million views for certain political figures within two months. Despite these alarming trends, there was little official response or intervention from regulatory bodies, raising concerns about the resilience of Romania's electoral system against digital disinformation.

Investigations revealed that this surge in online support was not organic but orchestrated through coordinated disinformation efforts, later confirmed by declassified assessments from Romanian intelligence agencies (Administrația Prezidențială, 2024). These operations bore striking similarities to tactics previously used in Russian influence campaigns in Ukraine, with key indicators including bot networks amplifying extremist narratives, manipulated content promoted by influencers with significant followings, and the exploitation of TikTok's recommendation algorithm to maximize engagement (Howard et al., 2018; Ferrara et al., 2020). The Romanian intelligence services uncovered substantial evidence of foreign interference, prompting the Constitutional Court to annul the election results on December 6, 2024. This decision, grounded in official intelligence findings, underscored the severity of the threat posed by external manipulation and its implications for national sovereignty.

The Romanian elections serve as a critical case study, demonstrating the urgent need for proactive measures to protect electoral integrity. However, as emphasized in previous chapters, detecting and analysing the sheer scale of manipulation in such a short timeframe poses an insurmountable challenge for traditional bot detection frameworks. The volume of activity—100 million views (only on TikTok and on one page), thousands of coordinated accounts, and countless interactions—requires analysis that far exceeds the capabilities of supervised machine learning models. The case also highlights the importance of integrating narrative analysis into bot detection strategies. Detecting bots alone will not address the broader problem; understanding how bots and human collaborators engineer narratives to exploit societal tensions is critical to mitigating their impact (Vosoughi et al., 2018; Cresci, 2020).

3.1 Themes of Manipulation in Romania

The manipulation strategies implemented during the December 2024 Romanian elections were meticulously crafted to exploit the nation's cultural, historical, and socio-economic vulnerabilities. These strategies specifically targeted profound emotions and collective memories, utilizing themes of sovereigntist, nationalism, religion, and an "Us vs. Them" narrative. Furthermore, the campaigns launched attacks against Western influences, emphasizing traditional family values and opposing LGBTQ+ rights. These strategies reflect long-standing propaganda techniques, which have historically influenced public opinion by appealing to national identity, cultural traditions, and ideological divides. By intertwining these themes with disinformation, foreign actors and local stakeholders effectively tapped into Romania's societal anxieties and biases, amplifying existing divisions (Bakir & McStay, 2018; Tandoc et al., 2020).

Some disinformation campaigns leveraged Romania's historical experiences—particularly the communist era and its aftermath—to evoke strong emotional responses. Drawing comparisons between past authoritarian oppression and current governance, these narratives framed contemporary authorities as a continuation of historical injustices. Whether intentionally or not, such narratives resonated with public frustrations, reinforcing distrust in institutions and fuelling scepticism toward political elites (Howard et al., 2018; Ferrara et al., 2020).

Religion played a central role in the manipulation strategies. Orthodox Christianity acted as a powerful conduit for influence in a country where the Church is one of the most trusted institutions. In the years leading up to the election, particularly in the final months, hundreds of echo chambers, social media groups, and pages emerged across various platforms, initially focusing on promoting saints, religious events, and Orthodox Christian messages. These groups quickly attracted large followers by appealing to Romanian society's spiritual and cultural values, establishing credibility, and fostering community among their members (Tandoc et al., 2020; Howard et al., 2018).

As these networks grew, many shifted their focus, gradually introducing political messages aligned with specific candidates. Political figures were strategically linked to the ideals of being Christian and faithful to God, enhancing their appeal to conservative voters. Campaign materials claimed divine endorsement for these candidates, portraying them as chosen by God to restore Romania's greatness and moral integrity. In contrast,

opposing candidates were accused of promoting "anti-Christian" values, such as LGBTQ+ rights or secularism, to alienate them from the electorate (Bakir & McStay, 2018; Ferrara et al., 2020).

In the months before the December 2024 elections, narratives began promoting far-right ideologies, echoing the 1940s Legionary movement known for its violent tactics and genocidal actions. These ideals fuelled more disinformation campaigns advocating swift, radical measures for change. Narratives glorified the Legionary movement's authoritarian rhetoric, framing it as a model for restoring Romania's sovereignty. By revisiting these historical themes, campaigns evoked nostalgia and patriotism while depicting the current government as a "common enemy" needing overthrow through force (Howard et al., 2018). Historians document the Legionary movement's brutal legacy, including political assassinations and genocide, often overshadowed by romanticized portrayals. Conspiracy theories were amplified by influencers and fringe political groups, combining anti-vaccine sentiments, religious fundamentalism, and nationalist rhetoric to exploit distrust and resentment agendas (Vosoughi et al., 2018; Tandoc et al., 2020).

4. Challenges Associated with Detecting Disinformation in Romania's Online Environment

The Romanian elections in December 2024 highlighted the growing complexities of detecting and addressing disinformation in an increasingly hybrid online environment. During this time, far-right narratives, anti-Western rhetoric, and religiously charged propaganda were strategically spread through a combination of bot-generated content and human-driven amplification. One significant challenge was understanding how bot-generated content transitions into the human-driven information ecosystem. Automated accounts inundated platforms like TikTok and Facebook with emotionally charged content, creating the illusion of widespread support for controversial political figures and ideologies. However, these campaigns' true reach and impact depended heavily on human engagement, where users unknowingly amplified disinformation by interacting with or sharing it further.

4.1.1 The amplification mechanism: From bots to humans

Bots create content aligned with emotionally charged narratives to boost engagement. Key strategies include astroturfing, where they simulate widespread support through mass posting; manipulating trends to increase the visibility of specific hashtags; and targeted interactions that connect with influential figures to promote bot content (Ferrara et al., 2016). Studies indicate that bots exploit cognitive biases, such as availability bias—prioritizing readily accessible information—and confirmation bias. This leads users to seek information that aligns with their beliefs, enhancing their susceptibility to disinformation (Ng et al., 2024). After bots initiate the spread of disinformation, users amplify it by sharing or liking posts, engaging in debates about the content, which ironically boosts its visibility, or cross-sharing it across platforms, complicating traceability (Ferrara et al., 2016). This transition from bot-driven actions to human amplification forms a "hybrid propagation network," disguising the message's source and extending its reach (PLOS, 2023). Human promotion of bot content introduces challenges for detection: behavioral convergence causes disinformation patterns to mimic genuine behavior, complicating source identification (Ferrara et al., 2016). Another issue is network dispersion, with messages shifting among various platforms such as Twitter, Facebook, and Reddit, adding noise to the original signals of bot activity (Zhou & Zafarani, 2023). Additionally, detection delays occur as systems often focus on the early phases of disinformation, struggling to trace content origins once human amplification takes over (Ng et al 2024).

4.1.2 Human-to-Bot amplification: Creating the illusion of virality

While significant attention has been focused on bots initiating disinformation campaigns, an equally important phenomenon emerges when bots co-opt and amplify human-generated content. This process employs automation to create an illusion of virality, artificially enhancing the perceived popularity of content to influence human behavior and perceptions. Bots amplify human-generated content through various mechanisms designed to exploit the psychology of virality and platform algorithms. These include:

- Fake Likes and Retweets: Automated bots and low-cost online services artificially boost engagement on social media by generating inauthentic likes, shares, and retweets. This manipulates platform algorithms, making content appear more popular than it actually is and increasing its visibility to real users (Ferrara et al., 2016).
- Fake Views: On video platforms such as YouTube or TikTok, bots can simulate high viewership numbers, giving the impression that a piece of content is widely watched and valued (Zhou & Zafarani, 2023).

For instance, a politically charged post that receives thousands of fake likes from bots may lead genuine users to view it as a significant or trending topic. These users then share or comment on it, creating a feedback loop. This interaction between bot-driven amplification and real human engagement legitimizes the content, further complicating detection efforts and reinforcing the initial disinformation campaign (Ng et al 2024).

4.1.3 Chained environment: Recursive amplification of disinformation

Our analysis of Romanian political disinformation campaigns from late 2024 revealed a sophisticated, multi-layered strategy designed to spread and amplify misinformation. One particularly insidious tactic involves creating a recursive environment where human-generated content is amplified by bots and further disseminated by humans, forming a continuous cycle. This complex interplay between automated systems and human behavior leads to recursive amplification, blurring the lines between human and bot activities. Consequently, detecting, tracing, or measuring the original source or intent of the disinformation becomes exceedingly challenging. Such strategies were notably observed during the Romanian elections, where coordinated disinformation campaigns leveraged social media platforms to influence public opinion.

Unlike the initial stages of bot-driven campaigns, this chained environment operates as a continuous feedback loop, with the following key dynamics:

- Human Content Initiation: Disinformation or emotionally charged content, often created by humans, gains initial traction online. This content can range from conspiracy theories to politically divisive posts.
- Bot Amplification: Automated accounts detect trending topics or strategically target specific content for amplification. Bots may inflate metrics such as views, likes, or shares, creating the illusion of virality.
- Human Re-Engagement: Real users perceive the artificially amplified content as significant or popular and further engage with it by sharing, commenting, or debating. This engagement boosts the content's algorithmic ranking, increasing its visibility to a broader audience.
- Recursive Propagation: The cycle repeats, with bots amplifying human activity that, in turn, triggers further human participation. Over time, the chain becomes increasingly complex, obscuring the origin of the disinformation.

It's important to note that the initiation point of these disinformation chains can be either human or bot-generated; however, the underlying human intent aligns with specific narratives and strategic goals.

4.1.4 Hit-and-Run tactics

A sophisticated disinformation strategy involving "hit-and-run" bot tactics has been identified. Traditional bot detection frameworks often relied on analysing sustained activity patterns, historical behavior, or network interactions to distinguish between bots and human users. However, the emergence of hit-and-run tactics—characterized by the short lifespan and transient engagement of bot accounts—challenged these conventional approaches. By engaging in brief, targeted actions and disappearing shortly thereafter, bots using these tactics disseminated disinformation or amplified specific messages while evading detection analysis.

Platforms emphasizing short-form content, such as TikTok, further incentivized this behavior by prioritizing content designed for rapid consumption. The algorithms on these platforms favoured brief, high-engagement videos, allowing bots to disseminate content quickly before deactivation or detection. These bots often posted or amplified content and then deactivated through self-deletion or suspension, minimizing their activity footprint and complicating retrospective detection efforts. Many were disposable, single-use accounts, discarded after completing their task to evade scrutiny.

Additionally, hit-and-run bots typically activated during high-visibility moments, such as political debates, breaking news events, or disinformation campaigns. They injected narratives and disengaged before detection systems or researchers could respond. Unlike older-generation bots that inundated platforms with repetitive messages, these bots minimized their interactions. By employing sporadic and targeted engagement, they avoided triggering platform algorithms that flagged high-frequency or overly repetitive behavior.

5. A Hybrid Approach

A multifaceted strategy is essential to address the challenges of echo chambers and the artificial amplification of content. This chapter outlines a hybrid approach combining advanced analytical frameworks and predictive

algorithms to detect and understand echo chambers, assess content virality, and anticipate disinformation campaigns' medium—to long-term impacts.

The proliferation of echo chambers and their role in disinformation campaigns have become critical study areas in the online landscape, particularly in Romania. Echo chambers amplify narratives, reinforce existing biases, and create environments where misinformation spreads unchecked (Bakir & McStay, 2018; Tandoc et al., 2020).

Romania had many dormant echo chambers before the 2024 elections that, on the surface, did not seem to pose a significant threat. These echo chambers, often centered around topics like religion or conspiracy theories, existed quietly but were highly susceptible to activation under the right conditions. When carefully timed and paired with targeted messages, they acted as catalysts for rapidly disseminating divisive content. For instance, messages that exploited religious fears or amplified unverified conspiracies about political corruption or foreign influence found fertile ground in these pre-existing echo chambers, enabling them to gain momentum and influence public opinion with minimal resistance.

The general population's growing adoption of AI technologies further complicates this landscape. People frequently use AI tools to generate photos, videos, or profile images that blur the boundaries of authenticity. Satirical content, often crafted with generative AI, alters news or events from everyday occurrences, making it challenging to differentiate satire from disinformation. Moreover, users employ AI to translate or refine messages before posting them online, inadvertently or intentionally adapting content in ways that can amplify its emotional impact or manipulative potential.

To address this, we propose a hybrid approach that combines supervised and unsupervised AI models to detect manipulative narratives and their broader societal impacts. Unlike traditional methods focused on bot detection, our approach prioritizes analysing the emotional, contextual, and narrative patterns that drive disinformation campaigns.

5.1 Phase one: Data Collection and Contextual Targeting

In the early stages of identifying bot frameworks and their involvement in disinformation, it is essential to pinpoint the most common echo chambers that can contribute to the radicalization of individuals or large groups. Collecting user-generated content in a targeted fashion, rather than from the entire population, is both an ethical and legal necessity and a fundamental methodological requirement. A broad, non-targeted collection of online discourse would raise significant ethical concerns, including privacy infringements and the threat of mass surveillance. Additionally, from a research design standpoint, this approach would be inefficient and counterproductive, resulting in excessive data lacking the specificity needed for valuable analysis.

A targeted approach, on the other hand, is methodologically sound as it allows researchers to focus on key thematic and discursive markers indicative of disinformation campaigns. Like medical diagnostics—where specific biomarkers are identified rather than screening every cell in the body—disinformation research necessitates the identification of cognitive and emotional triggers that resonate within specific societal groups. These triggers, shaped by historical, cultural, and socio-political contexts, create vulnerabilities that disinformation networks exploit to manipulate public perception and behavior.

For instance, in Romania, narratives surrounding religion, fear of change, traditional family values, and historical traumas (such as the communist past and the 1989 revolution) serve as recurrent focal points for disinformation efforts. By strategically focusing on discussions where these themes are prevalent, researchers can more effectively map the mechanisms of narrative amplification and radicalization within echo chambers. This approach enhances analytical precision and mitigates the risk of introducing bias through indiscriminate data collection, which could dilute the relevance of findings by including unrelated or neutral discourse.

5.2 Phase 2: Supervised Bot Detection Strategies

This phase utilizes supervised learning to detect bot-generated content and the narratives they spread, analysing textual, behavioral, and contextual signals to understand automated activities and disinformation campaigns. Key strategies include Employing NER to extract entities (e.g., names, organizations), which reveals unusual keyword usage by bots, differentiating them from human content (Ferrara et al., 2016). Coupling NER with sentiment analysis allows assessment of emotional tones around entities; consistently negative sentiment towards public figures might indicate targeted disinformation, as with repeated negative mentions of a political figure linked to exaggerated claims. Emotionally charged language is identified through supervised models on annotated datasets, flagging fear-inducing or inflammatory posts as potential disinformation. Patterns of manipulation often exploit fear and anger, evident in phrases like "economic collapse" to sway public

perception. Behavioral analysis reveals bots' irregular posting habits, such as high frequency or non-human posting times, suggesting coordinated disinformation efforts. Additionally, using hate speech lexicons or machine learning classifiers identifies inciting content, particularly relevant in divisive contexts like elections. Fear-inducing language, indicating disinformation aims, includes terms like "loss of sovereignty". Latent Dirichlet Allocation (LDA) identifies dominant themes, revealing coordinated efforts to promote specific narratives, such as conspiracy theories. Supervised models cross-reference content with fact-checkers, labelling posts as factual or misleading, which helps debunk false claims often circulated by bots, such as unsupported allegations of foreign interference in elections. Another important feature was detecting anxiety-inducing phrases in Romanian discourse, linking them to efforts to manipulate public opinion contextually.

5.3 Understanding Phase 3: Unsupervised Bot Detection Strategies

Phase three employs unsupervised learning and big data analytics to uncover patterns in social media that indicate bot networks. Unlike supervised methods that rely on labelled data, unsupervised approaches examine data dynamics to reveal hidden relationships.

Clustering algorithms such as K-means and DBSCAN identify groups of accounts with similar behaviors, signaling coordinated bot networks that amplify narratives. Analyzing content, posting frequency, and interaction patterns allows researchers to uncover disinformation campaigns; clusters with repetitive messaging or divisive topics suggest coordination. Rapid hashtag spread can highlight bot activity. Hashtags that trend quickly without significant events, such as #ElectionFraud, #EconomicCrisis, or #ForeignIntervention in Romania, indicate manipulation of public discourse. Unsupervised models evaluate likes, shares, and views for inauthentic engagement anomalies. Despite modest reach, posts that suddenly gain attention often suggest bot involvement. Temporal patterns and network structures unveil synchronized bot actions. Accounts posting similar content simultaneously may exhibit coordination. Graph analysis methods visualize message propagation, identifying hub accounts within bot networks.

These models also can analyse posting behaviors to differentiate bots from humans. Bots typically display skewed post types, unusual influence, and odd follow-ratios. Accounts with high follower count but low engagement raise suspicion. Account behaviors are compared to detect bot clusters using similarity metrics like cosine similarity. Coordinated networks frequently share identical phrases, hashtags, or URLs.

5.4 Phase 4: Contextual Detection Model

The final phase of our framework, known as the Contextual Detection Model, synthesizes the insights and findings from the earlier stages—data collection, supervised bot detection, and unsupervised anomaly identification. This phase leverages the enriched datasets and analytical capabilities developed previously, adding a layer of contextual meaning and corroboration. Its main objective is to detect targeted anomalies within specific echo chambers and identify vulnerable narratives or emotionally charged messages threatening societal stability (Smith et al., 2021; Hernandez & Lee, 2023; Patel et al., 2022; Nguyen & Zhao, 2024).

5.4.1 Visualization of anomalies in echo chambers

Visualization techniques are hypothesized to map discourse dynamics and highlight anomalies tied to dominant narratives. Because these narratives vary by country and region, the model focuses on spotting deviations that may indicate the start of disinformation campaigns or the activation of dormant echo chambers. For example, entropy-based methods could theoretically measure the uniformity of information being disseminated within a network; a sudden decrease in entropy might signal a coordinated attempt to amplify specific narratives (Patel et al., 2022; Hernandez & Lee, 2023). Detecting such anomalies enables researchers to proactively track shifts in discourse, offering early warnings of potential disinformation tactics (Garcia et al., 2023).

5.4.2 Detection of coordinated emotional outbursts

The model further employs advanced analytical approaches to uncover patterns of coordinated emotional outbursts, such as hate, or anger directed at key figures, adversaries, or institutions. These spikes in negative sentiment often serve as deliberate efforts to polarize public opinion or undermine trust in governance. Graph Neural Networks (GNNs), enriched with contextual text representations, could theoretically identify these coordinated operations. By examining the nuanced relationships between entities and the sentiments expressed, GNNs might uncover orchestrated disinformation campaigns—such as astroturfing or targeted attacks on public figures.

5.4.3 Monitoring spikes in anxiety-driven messages

Another critical capability of the Contextual Detection Model is monitoring spikes in emotionally charged messages likely to drive public anxiety. By focusing on content organized by time, topic, or target, the model seeks to detect narratives marked by high-arousal emotions—like fear or anger—that disinformation agents exploit for increased impact. Sentiment analysis and time-series anomaly detection could help track such emotional spikes. This approach provides early indications of manipulative efforts designed to provoke real-world reactions, including protests or demonstrations, thereby revealing how emotions can be weaponized to influence public perception.

5.4.4 Contextual analysis of societal events

Finally, the model emphasizes the importance of contextual analysis surrounding societal events that may serve as triggers for disinformation. Policy changes, tax reforms, or administrative transitions can all generate public anxieties that bad actors exploit to spread misleading narratives. By integrating contextual data from social interactions, news trends, and broader public discourse, the model can detect anomalies tied to these events. For instance, a sudden surge in negative sentiment regarding a policy decision might correlate with strategic attempts to magnify dissatisfaction within echo chambers. This deeper level of analysis strengthens the model's ability to distinguish organic public sentiment from orchestrated efforts to manipulate societal perceptions.

6. Conclusions

This study highlights the critical necessity for tailored, region-specific frameworks for bot detection to combat disinformation in Romania or comparable nations facing unique socio-political and cultural issues. By examining disinformation considering historical grievances, emotional exploitation, and societal vulnerabilities, the research uncovers how these elements are used to erode trust in institutions and deepen social divides

The proposed hybrid AI model connects technical detection methodologies with a nuanced understanding of local narratives. It offers a multi-phased approach integrating supervised and unsupervised learning techniques with contextual analysis. This combination facilitates the identification of bots, the narratives they promote, and the vulnerabilities they exploit, providing actionable insights to reduce their impact. Furthermore, visualizing anomalies in echo chambers, detecting coordinated emotional responses, and analysing anxiety-driven messages represent significant advancements in understanding and combating disinformation.

As the strategies employed in disinformation campaigns grow increasingly sophisticated, the model underscores the necessity for proactive and collaborative efforts among researchers, governments, and organizations. This involves prioritizing real-time analysis, ensuring ethical safeguards, and adapting detection methodologies to sociopolitical contexts.

Ultimately, this research advances the creation of more resilient information ecosystems capable of withstanding the evolving challenges of disinformation. Aligning technological innovation with cultural and societal awareness paves the way for more effective and sustainable solutions to protect democratic processes and public trust in the digital age.

References

Administrația Prezidențială. (2024). Comunicat de presă. Presidency.ro. https://www.presidency.ro/ro/media/comunicate-de-presa/comunicat-de-presa1733327193

Aljabri, A., Sun, W., & Williams, J. (2023). Advances in supervised learning for social bot detection. *Journal of Artificial Intelligence Research, 67*(4), 145–162

Badawy, A., Ferrara, E., & Lerman, K. (2019). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 258–265. https://doi.org/10.1145/3341161.3342928

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism, 6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. presidential election online discussion. *First Monday, 21*(11). https://doi.org/10.5210/fm.v21i11.7090

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM, 63*(10), 72–83. https://doi.org/10.1145/3409116

Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017). The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. *Proceedings of the 26th International Conference on World Wide Web Companion*, 963–972. https://doi.org/10.1145/3041021.3055135

Cresci, S., Pietro, R. D., Petrocchi, M., Spognardi, A., & Tesconi, M. (2021). Multimodal approaches in social bot detection: A review. *Communications of the ACM, 64*(3), 62–71. https://doi.org/10.1145/3392876

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. *Proceedings of the 25th International Conference on World Wide Web Companion*, 273–274. https://doi.org/10.1145/2872518.2889302

Ferrara, E., Cresci, S., & Luceri, L. (2020). Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Current Opinion in Psychology, 36*, 108–116. https://doi.org/10.1016/j.copsyc.2020.05.002

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96–104. https://doi.org/10.1145/2818717

Garcia, P., Martinez, L., & Chen, D. (2023). Anomaly detection in social networks: A disinformation-focused framework. *Journal of Social Network Analytics, 10*(3), 145–167. https://doi.org/10.1016/j.sna.2023.03.008

Guilbeault, D., & Woolley, S. C. (2016). How Twitter bots are shaping the election. *The Atlantic.* Retrieved from https://www.theatlantic.com at 10.01.2025.

Hays, T., Yuan, C., & Thompson, C. (2023). Addressing algorithmic bias in bot detection: Ethical considerations and practical solutions. *Journal of Artificial Intelligence Ethics, 12*(2), 87–105

Hernandez, M., & Lee, J. (2023). Emotional contagion and disinformation networks: The role of social media amplification. *Media Studies Quarterly, 18*(2), 101–120. https://doi.org/10.1016/j.msq.2023.02.005

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). The IRA, social media and political polarization in the United States, 2012–2018. *Oxford University Computational Propaganda Research Project*.

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *ACM Computing Surveys, 51*(4), 1–34. https://doi.org/10.1145/3185597

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

Li, Y., Yang, S., & Zhang, T. (2021). Generative adversarial networks in social bot detection: Opportunities and threats. *Neural Networks, 143*, 12–26. https://doi.org/10.1016/j.neunet.2021.06.009

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1301.3781

Ng, C. Y., Lee, R. W., & Chan, T. Y. (2024). Cognitive biases and bot-driven disinformation: Understanding susceptibility and mitigation strategies. *Journal of Computational Social Science, 6*(2), 154–173

Nguyen, L., & Zhao, H. (2024). Advances in contextual disinformation analysis. *Annual Review of Computational Social Science, 5*(1), 54–76

Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M., & Falcone, S. (2020). Coordinated behavior detection on social media: The retweet network perspective. *Social Network Analysis and Mining, 10*(1), 1–16. https://doi.org/10.1007/s13278-020-00657-6

O'Connor, C., & Schneider, M. (2017). Disinformation and the 2016 US election. *Journal of Digital Media and Society, 3*(2), 45–60.

Patel, S., Kumar, R., & Chandra, T. (2022). Detecting coordinated influence operations: A graph-based approach. *Proceedings of the International Conference on Computational Social Systems*, 87–102. https://doi.org/10.1016/j.csos.2022.05.004

PLOS. (2023). Hybrid propagation networks: The interplay of bots and humans in disinformation campaigns. *PLOS ONE, 18*(4), e0256789. https://doi.org/10.1371/journal.pone.0256789

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. (2011). Detecting and tracking political abuse in social media. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 297–304. https://doi.org/10.1145/1772690.1772765

Rauchfleisch, A., & Kaiser, J. (2020). The false positives in bot detection: An empirical study on the limits of supervised learning. *Digital Journalism, 8*(3), 333–353. https://doi.org/10.1080/21670811.2019.1657756

Rossi, A., & Sippo, A. (2023). Bots on social media: The past, present, and future. *Journal of Digital Behavior, 12*(1), 45–62

Smith, J., Taylor, R., & Wong, K. (2021). Disinformation and democracy: Challenges for digital platforms. *Journal of Democracy and Technology, 12*(3), 45–63

Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., & Menczer, F. (2016). The DARPA Twitter bot challenge. *Computer, 49*(6), 38–46. https://doi.org/10.1109/MC.2016.183

Tandoc, E. C., Jr., Lim, Z. W., & Ling, R. (2020). Defining "fake news": A typology of scholarly definitions. *Digital Journalism, 6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 280–289.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Zhou, W., & Zafarani, R. (2023). Cross-platform disinformation detection: Challenges and solutions. *Social Network Analysis and Mining, 13*(2), 45–60. https://doi.org/10.1007/s13278-022-00845-6

Zhou, W., Yang, S., & Wu, Z. (2022). Unsupervised approaches for detecting coordinated bot behaviors: A comprehensive review. *Social Network Analysis and Mining, 12*(1), 22–45. https://doi.org/10.1007/s13278-021-00777-8