The Weaponization of the Marketing Funnel: Adapting Commercial Targeting Strategies for Disinformation Campaigns and Troll Detection

Petr Gallus[1], Dominik Staněk[1], Tomáš Ráčil[1], Willi Lazarov[2] and Ivo Klaban[1]
[1]Department of Informatics and Cyber Operations, University of Defence, Brno, Czech Republic
[2]Department of Telecommunications, FEEC, Brno University of Technology, Brno, Czech Republic

petr.gallus@unob.cz

Abstract: In the era of hybrid warfare, the operational mechanics of troll farms have shifted from random disruption to highly organized campaigns that mirror commercial digital marketing. This paper, grounded in the HYBTRINT project research, investigates how standard marketing strategies—specifically the "marketing funnel"—are weaponized to manipulate the information environment and proposes a technical framework for their detection. The study demonstrates how influence operations guide users from initial sensation to radicalization by exploiting societal "pain points" and creating "dopamine loops" for artificial community validation. To counter these sophisticated strategies, we define a detection framework based on a multimodal scoring system involving linguistic toxicity, reaction latency, and behavioural parameters such as "Activity Consistency". The findings, validated within a controlled experimental "Playground" environment, offer a model for distinguishing between organic engagement and coordinated inauthentic behaviour. The research is supported by an analysis of an experimental asset pool with a cumulative monthly reach exceeding 10 million impressions, providing a statistically significant basis for behavior profiling. This research serves as a basis for developing adaptive monitoring modules for defence and security institutions.

Keywords: Troll account, Troll farm, Troll detection, Disinformation, Marketing, Marketing strategy

1. Introduction

The contemporary information environment has evolved into a sophisticated domain of strategic competition, where cognitive integrity is constantly challenged by coordinated influence operations. In the context of hybrid warfare, the tactics employed by adversarial actors – eferred to as "troll farms" – have undergone a significant paradigm shift. These entities no longer rely solely on primitive, high-volume spamming; instead, they have adopted advanced methodologies pioneered by the digital marketing and advertising industries. Information has become a weapon capable of destabilizing society and undermining trust in democratic institutions faster and more cheaply than conventional military means.

In the current digital landscape, manual monitoring of the information space has become unsustainable due to the sheer volume of data. Human analysts lack the capacity to filter millions of comments in real-time, detect coordinated bot networks, or uncover sophisticated psychological operations without advanced automated aid. The Hybrid Threat Intelligence (HYBTRINT) project responds to this reality by moving beyond manual oversight toward a modular system designed for large-scale data mining and threat identification.

This convergence of propaganda and commercial strategy is most evident in the adaptation of the "marketing funnel". By applying professional audience segmentation, lead generation techniques, and data-driven targeting, influence operations can effectively guide users from initial exposure to fringe narratives toward deep-seated radicalization. This process exploits societal "pain points" such as economic instability, migration, or cultural grievances, transforming them into entry points for psychological manipulation.

While technical capabilities for automated content generation have increased with the advent of Large Language Models (LLMs), the underlying strategic framework remains rooted in human-centric marketing principles. The primary objective of this research is twofold:

1. To provide a theoretical deconstruction of how marketing funnels are re-engineered for disinformation and radicalization purposes.
2. To propose a multidimensional detection framework designed to identify coordinated inauthentic behavior through weighted linguistic and behavioral metrics.

By shifting the focus from individual content pieces to the strategic patterns of a campaign, this paper aims to provide a more robust defense mechanism for security institutions, enabling them to distinguish between organic engagement and professional influence operations.

2. Actors Typology

To effectively detect hybrid threats, it is crucial to distinguish between different types of actors operating within the information space. While the resulting effect – such as the spread of disinformation or societal polarization – may appear identical, their motivations, behavioural patterns, and technical signatures differ fundamentally. The HYBTRINT project defines a triad of actors: the Professional, the Citizen, and the Machine.

Table 1 provides a comparative overview of the identified actors, contrasting the high coordination of "Professional Trolls" and "Bots" with the organic nature of the "Frustrated Citizen." This differentiation is critical for detection, as identifying a professional actor requires sophisticated behavioral analysis (such as monitoring "Managed Attribution") rather than just flagging the linguistic toxicity typical of a frustrated citizen.

A campaign typically begins with a "Professional" seeding a narrative that is artificially inflated by "Bots" to create an illusion of consensus. The ultimate goal is to trigger the "Frustrated Citizen," whose authentic engagement provides a "living shield" of legitimacy that complicates moderation efforts and invites accusations of censorship.

Table 1: Actors Typology Features Comparison

Feature	Professional Troll	Frustrated Citizen	Bots & AI
Primary Motivation	Financial / Ideological (External)	Emotional / Identity-based (Internal)	Algorithmic (Pre-programmed)
Activity Pattern	Regular (Working hours, 8–17)	Irregular (Peak during evenings/weekends)	24/7 or Event-triggered
Content Quality	High (Strategic, Narrative-driven)	Variable (High toxicity, Low grammar)	Repetitive or LLM-generated
Technical Footprint	Managed Attribution (VPN, Proxies)	Single Authenticated Account	Cluster creation, API signatures
Role in Funnel	Seeder / Operator	Amplifier / "Living Shield"	Booster / Metric Inflator
Detection Complexity	High (Requires behavioral analysis)	Low (Linguistic & Toxic analysis)	Medium (Pattern recognition)

2.1 The Professional Troll

This actor operates based on external motivation, typically financial or ideological, and follows a clear hierarchy and strategic assignment. They employ sophisticated "Managed Attribution" infrastructures to manage dozens of fake identities (“sockpuppets”) simultaneously. Their activity mimics standard working hours (e.g., 08:00–17:00) and they cold-bloodedly simulate emotions to provoke a reaction.

2.2 The Frustrated Citizen

Quantitatively the most significant group, these users are not the initiators but rather the "unwitting agents" or amplifiers of disinformation. Their motivation is authentic emotion – fear, anger, or a sense of injustice.

They are characterized by the Dunning-Kruger effect (see Figure 1), where low competence in a field leads to an overestimation of their own knowledge, resulting in aggressive defence of false narratives.

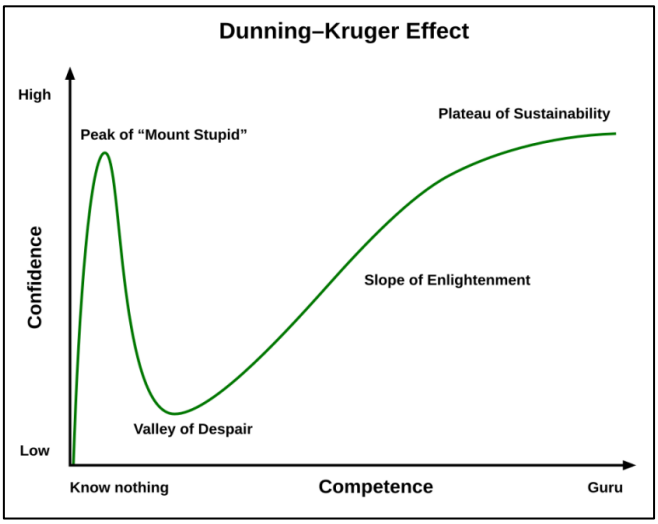

Figure 1:The Dunning-Kruger Effect Graph

2.3 Bots and Automation

These accounts are controlled by software scripts rather than humans. While traditional "Dumb Bots" focus on inflating metrics, modern "Smart Bots" utilize Large Language Models (LLMs) to generate unique content. Their primary detection markers include "superhuman" speed – responding within milliseconds – and 24/7 activity without biological sleep cycles.

3. The Weaponized Marketing Funnel

Modern influence operations have evolved beyond random disinformation; they now represent a sophisticated application of digital marketing and PR procedures. Adversarial actors have "weaponized" methods originally developed for commercial brand building, targeting a citizen's cognitive integrity instead of their consumer behavior. This section explores how the traditional commercial "Customer Journey" is transformed into a systematic process of gradual radicalization.

3.1 Asset Building and Targeting Strategies

In the initial phase, troll farms mirror the activities of marketing agencies by creating "Personas" and profiling vulnerable population segments. To reach these audiences efficiently, attackers utilize low-cost acquisition strategies designed to bypass standard security filters:

- **Bait-and-Switch**: This involves establishing a neutral community (e.g., focused on pets, local history, or nature) to acquire followers at a very low cost-per-result. Once a substantial audience is reached, the page "pivots" its editorial line toward political disinformation.
- **Page Merging and Infrastructure**: Attackers manage multiple smaller thematic pages and subsequently use platform tools to merge them into a single massive entity. This creates an illusion of a large, spontaneous community.
- **Algorithmic Populism**: Content is optimized for conflict. Because social media algorithms prioritize high-emotion engagement (specifically anger), this provides the campaign with massive viral reach without the need for a paid advertising budget.

3.2 Structural Phases of the Radicalization Funnel

The adaptation of the marketing funnel follows a logical progression, guiding a passive reader toward becoming an active, radicalized amplifier. The progression from passive exposure to active amplification is visualized in Figure 2. This process is divided into four key stages:

- **Attract (Awareness)**: The user is exposed to sensationalist headlines or memes designed to trigger an immediate emotional response and "stop the scroll."

- **Inform (Consideration)**: In this phase of "emotional hooking," the user consumes content that offers overly simplistic explanations for complex societal problems, leading to a gradual erosion of trust in mainstream institutions.
- **Convert (Conversion)**: The critical turning point where the user internally adopts the narrative as their own. In this context, conversion is not a purchase, but a shift in attitude or joining a closed, radicalized group.
- **Engage (Loyalty)**: The user becomes an active "Amplifier." They create original content, aggressively defend the narrative against dissent, and become immune to factchecking within a self-reinforcing echo chamber.

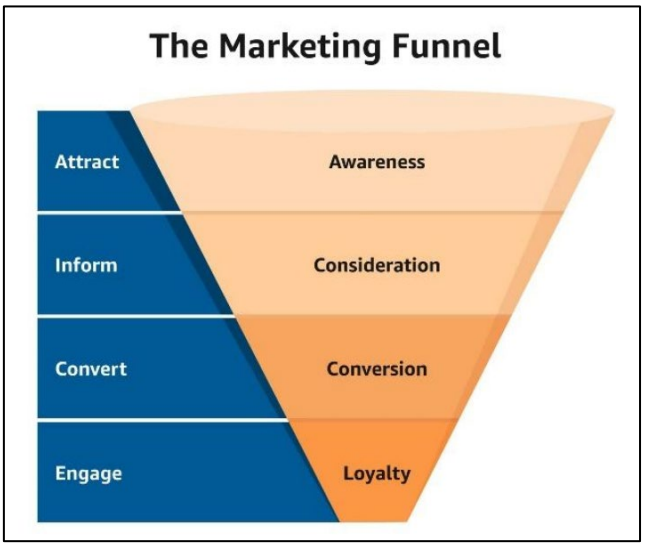

Figure 2: The Marketing Funnel and its Usage for Active Amplification

3.3 Psychological Mechanics: The Dopamine Loop

To ensure the long-term sustainability of the funnel, influence operations leverage the "dopamine loop" of social validation. Professional troll accounts are instructed to provide artificial feedback (likes and supportive comments) to real users who have reached the "Engage" phase. This digital validation provides a sense of belonging and importance. Consequently, the "trained" amplifiers become dependent on this community feedback, eventually forming a self-sustaining ecosystem that requires minimal further intervention from the original attacker.

4. Detection Framework

To counter the sophisticated "weaponized marketing" strategies, the HYBTRINT project defines a multimodal scoring system. Instead of a binary classification, the framework calculates a probability-based risk value, known as the **Hybrid Threat Score (*S*)**. This score is derived from a weighted sum of parameters across three distinct layers: linguistic, behavioral, and platform-specific.

4.1 The Mathematical Model

Each detected parameter (p_i) is assigned a normalized value between 0 and 1. These values are then multiplied by a weight (w_i) reflecting the indicator's severity:

- **Critical (1.0)**: Indicators almost certainly signifying machine behavior or coordinated campaigns (automatic "Red Flag")
- **High (0.7)**: Strong indicators of manipulation requiring cross-verification with other parameters.
- **Medium (0.4)**: Supporting indicators that complete a troll profile when combined with others

The final score is calculated using the following formula:

$$S = \frac{\sum_{i=1}^{n}(w_i \cdot p_i)}{\sum_{i=1}^{n} w_i}$$

Where:

- S: Represents the final **Hybrid Threat Score** within the interval <0, 1>.
- n: The total number of **active detection parameters** evaluated for the specific platform or account.
- w_i: The **weight** assigned to the i-th parameter, reflecting its significance (1.0, 0.7, or 0.4).
- p_i: The **normalized value** of the i-th parameter (<0, 1>), where 0 indicates organic behavior and 1 indicates behavior fully consistent with a threat pattern.

4.2 Detection Layers and Parameters

The detection framework is structured into three distinct analytical layers, each providing a unique perspective on the actor's intent, authenticity, and level of coordination. By combining these layers, the system can identify "Cyborg" accounts—hybrid entities that blend human-like linguistic nuances with machine-like distribution speeds. Given layers, their parameters and weights are summarized in Table 2.

- **Linguistic Layer (NLP Layer):** This layer focuses on the semantic and syntactic properties of the content. Beyond simple keyword matching, it utilizes Natural Language Processing to detect **Toxicity and Sentiment** levels, identifying attempts to exploit societal "pain points" through ad-hominem attacks. The most critical indicator in this layer is **Semantic Duplicity**, which measures the similarity between messages across different accounts. A high similarity score (over 80%) across geographically or socially distant accounts serves as a "Critical" indicator of coordinated "copy-paste" strategies common in professional troll campaigns.
- **Behavioral Layer (Profiling Layer):** This layer shifts the focus from "what" is being said to "how" the actor interacts with the digital environment. It is specifically designed to distinguish between human spontaneity and algorithmic precision. Parameters such as **Reaction Latency**—the time elapsed between a target post's publication and the account's response—are vital. Responses occurring within seconds suggest the use of automated "gaming" of platform algorithms to gain early traction. Furthermore, **Activity Consistency** tracks temporal patterns; a strict adherence to a "working hour" schedule (e.g., 08:00–17:00) reveals a professional operator rather than an organic user.
- **Platform-Specific Layer (Metadata Layer):** This layer utilizes technical footprints unique to individual social networks to uncover hidden infrastructures. For example, on Meta platforms (Facebook/Instagram), a **Follower Spike** that does not correlate with the virality of any specific post is a "Critical" indicator of technical asset building through page merging. Similarly, **Page Renaming** history can reveal a "Bait-and-Switch" operation, where an account originally dedicated to apolitical topics (e.g., "Nature Lovers") is suddenly repurposed for political disinformation.

Table 2: Marketing Techniques and their Technical Vectors

Layer	Parameter	Strategic Link (Marketing/Threat)	Weight
Linguistic	Semantic Duplicity	Test Marketing (Narrative testing/Copy-paste)	Critical
	Toxicity & Sentiment	"Pain Point" exploitation (Rage-baiting)	High
Behavioral	Reaction Latency	Algorithmic Populism (Gaming the algorithm)	Critical
	Activity Consistency	Managed Infrastructure (Scheduling patterns)	High

Layer	Parameter	Strategic Link (Marketing/Threat)	Weight
Platform	Follower Spike	Asset Building (Page merging)	Critical
	Page Renaming	Bait-and-Switch (Content pivot)	High

4.3 Interpretation of Results

The output of the detection mechanism is the Hybrid Threat Score (***S***), a probabilistic value that allows for the efficient prioritization of defensive resources. The weighted approach ensures that a single anomaly (e.g., a user using a single vulgar word) does not trigger a false alarm unless supported by additional behavioral evidence.

- **Safe ($S \in \langle 0.0, 0.3 \rangle$)**: This category includes mostly the organic users. While they may occasionally exhibit high sentiment or frustration, they lack the technical signatures of coordination or automation.
- **Suspicious ($S \in (0.3, 0.7 \rangle$)**: This "Grey Zone" identifies accounts that exhibit significant signs of toxicity or repetitive narrative patterns but lack definitive proof of machine-controlled behavior. These cases are of particular interest for human-AI collaboration; the system flags these profiles for manual validation by security analysts, who can further investigate the nuances of the narrative within the context of the HYBTRINT project.
- **Malicious ($S \in (0.7, 1.0 \rangle$)**: This category is reserved for actors with an overwhelming weighted probability of being professional trolls or automated bots. Profiles in this range exhibit multiple "Critical" indicators, such as near-instant reaction times and high semantic duplicity across known botnet clusters. These results trigger automated reporting and adaptive monitoring modules to mitigate the threat in real-time.

4.4 Illustrative Example of Score Calculation

To demonstrate the practical application of the Hybrid Threat Score (***S***), we present a model calculation for a account labeled **#User-492**, as detected on a microblogging platform. Input data for the illustrative calculation are present in the Table 3.This profile exhibited a mix of high linguistic toxicity and coordination patterns, while maintaining some organic technical characteristics.

Table 3: Input Data for the Illustrative Example of Score Calculation

Parameter (Vector)	Measured Value (p_i)	Weight (w_i)	System Comment
Toxicity (NLP)	0.9	0.7 (High)	Extreme aggression vulgarism
Semantic Duplicity	0.8	1.0 (Critical)	Text matches 15 other accounts
Frequency (Flooding)	0.4	1.0 (Critical)	20 tweets/day (within human limits)
Creation Cluster	0.1	1.0 (Critical)	Registration date is unique
Lack of Solution	1.0	0.4 (Medium)	Purely critical, no constructivity

Final Score Calculation:

$$S = \frac{(0.7 \cdot 0.9) + (1.0 \cdot 0.8) + (1.0 \cdot 0.4) + (1.0 \cdot 0.1) + (0.4 \cdot 1.0)}{0.7 + 1.0 + 1.0 + 1.0 + 0.4}$$

$$S = \frac{0.63 + 0.80 + 0.40 + 0.10 + 0.40}{4.1}$$

$$\boldsymbol{S = \frac{2.33}{4.1} \approx 0.568}$$

Result and Classification: With a final score of **S = 0.57,** the actor is placed in the interval (0.3, 0.7) and classified as **SUSPICIOUS**.

Interpretation: The system did not detect a fully automated bot (due to low frequency and a unique registration date). However, given the high semantic duplicity and toxicity, there is a high probability of a coordinated human actor (Troll) or a "Frustrated Citizen" participating in chain-message campaigns. The profile is sent for manual validation by an analyst.

5. Preliminary Validation: The "Playground" Environment

To validate the theoretical models and the proposed detection framework, the research utilized a controlled testing environment known as the **"Playground"**. This infrastructure serves as a living laboratory for analyzing coordinated behavior, asset accumulation, and the algorithmic response to various narrative types.

5.1 Methodology and Scope

The validation was conducted using a managed portfolio of social media assets specifically curated to simulate the full lifecycle of an influence operation. The primary objective was to empirically verify the **"Asset Building"** and **"Bait-and-Switch"** strategies described in Section III.

- Infrastructure Scale: The experiment monitored a network of pages and profiles that achieved a cumulative monthly reach exceeding 10 million impressions.
- Narrative Testing: The "Playground" utilized a diverse content mix, ranging from low-arousal apolitical topics (nostalgia, local interest) to high-arousal polarizing political narratives.
- Data Collection: Technical footprints, including reaction latency, semantic duplication across clusters, and follower acquisition costs (CPA), were continuously logged to provide a baseline for the detection framework's calibration.

5.2 Algorithmic Amplification and Reach Dynamics

One of the key findings from the Playground experiment is the quantifiable disparity in organic reach between different content categories. This data provides the empirical foundation for **Figure 3** (which can be generated using relative reach metrics).

- **Audience Acquisition (The Bait):** Neutral narratives (e.g., nature photography, historical nostalgia) demonstrated the highest efficiency for low-cost follower acquisition. These "assets" allow attackers to build large, legitimate-looking communities with minimal algorithmic friction or user reports.
- **Narrative Pivoting (The Switch):** Once a critical audience mass was reached, the "pivot" to polarizing content revealed that social media algorithms disproportionately favor high-emotion narratives.
- **Viral Reach Efficiency:** Experimental data shows that **Outrage-based content** (fear and anger) achieves significantly higher organic distribution compared to neutral baseline posts. This algorithmic preference acts as a "force multiplier" for coordinated campaigns, allowing them to achieve massive penetration without advertising spend.

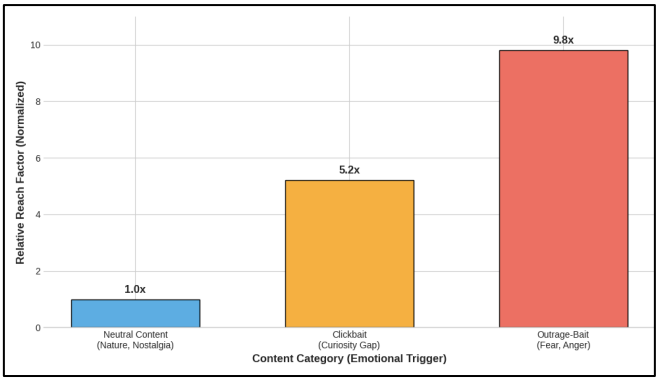

Figure 3: Algorithmic Amplification of Narratives

5.3 Summary of Technical Findings

The preliminary testing within the Playground environment confirmed several critical hypotheses:

- **Low Entry Barrier:** Building a massive, receptive audience is financially and technically accessible, requiring only basic marketing automation tools.
- **Technical Footprinting:** Coordinated clusters exhibited a measurable **"Semantic Duplicity"** and **"Activity Consistency"** that aligned with the Working Hour (8–17) profile of a professional operator.
- **Vulnerability:** Specific demographic segments showed a marked susceptibility to narratives that blend nostalgic sentiment with modern "pain point" exploitation.

6. Conclusion

This paper has demonstrated that modern influence operations have transitioned from sporadic disinformation into a professionalized domain of **weaponized marketing**. By re-engineering the commercial "customer journey" into a systematic radicalization funnel, adversarial actors can efficiently guide users from initial passive exposure to active, self-sustaining amplification.

The empirical data gathered within the **"Playground" environment** confirms the structural vulnerability of current social media landscapes. Our validation shows that while neutral narratives are optimal for low-cost asset building, **outrage-based content achieves nearly 10x (9.8x) the reach efficiency** of baseline information. This confirms that platform algorithms act as unintended force multipliers for coordinated campaigns, providing them with massive, free-of-charge distribution channels.

To counter this, the proposed **Hybrid Threat Score (S)** framework, developed under the HYBTRINT project, shifts the defensive focus from reactive content moderation to proactive behavioral profiling. By integrating linguistic toxicity, reaction latency, and platform-specific metadata into a weighted multimodal system, we provide security institutions with a scalable tool to distinguish between organic engagement and professional manipulation.

The goal of this framework is to provide a baseline for adaptive monitoring modules that can identify and neutralize adversarial infrastructures before they reach the critical phase of societal radicalization.

The research described in this paper was financially supported by the Technology Agency of the Czech Republic (TACR), project No. OZ0102006.

References

Bail, C., 2021. Breaking the Social Media Prism: How to Make Our Platforms Less Polarized. Princeton University Press. ISBN 978-0-691-20342-3.

BEZFAULU.NET, 2024. Argumentační fauly: Přehled a kategorizace [online]. Available at: https://bezfaulu.net/argumentacni-fauly.

Bobba, G. And Nicolas, H., 2021. Populism and the Politicization of the COVID-19 Crisis in Europe. Palgrave Macmillan. ISBN 978-3-030-66010-9.

Bradshaw, S. And Howard, P. N., 2019. The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation. Oxford Internet Institute.

Ferrara, E. et al., 2016. The Rise of Social Bots. Communications of the ACM. vol. 59, no. 7, pp. 96–104.

HYBTRINT, 2025. Studie k analýze trollích farem a vymezení metod pro vytváření cíleného obsahu. Project No. OZ01020006. University of Defence.

Kotler, P., Kartajaya, H. And Setiawan, I., 2017. Marketing 4.0: Moving from Traditional to Digital. Wiley. ISBN 978-1-119-34120-8.

Pomerantsev, P., 2019. This Is Not Propaganda: Adventures in the War Against Reality. Faber & Faber.

Sockpuppet.IO, 2024. Managed Attribution Infrastructure: How It Works [online]. Available at: https://www.sockpuppet.io/about-us/.

Tavris, C. And Aronson, E., 2020. Mistakes Were Made (but Not by Me): Why We Justify Foolish Beliefs, Bad Decisions, and Hurtful Acts. HarperCollins.

Tindale, C. W., 2007. Fallacies and Argument Appraisal. Cambridge University Press.