*informatics*

MDPI

*Article*

# Digital Polarization Index (DPI): An Operational Framework for Detecting and Characterizing Coordinated Operations in Latin America

**Jorge Sánchez-Erazo [1,*] and Jorge Cruz-Cárdenas [2]**

[1] Sánchez Erazo Consulting, Quito 170901, Ecuador
[2] School of Economics, Administration and Business, ESTec Research Center, Universidad Tecnológica Indoamérica, Quito 170103, Ecuador; jorgecruz@uti.edu.ec
* Correspondence: jorge@sanchezerazo.com

**Abstract**

Since X-Twitter suppressed public access to engagement metrics in February 2023, classical models for detecting coordinated operations lost viability. Astroturfing is one modality among several, alongside narrative saturation and targeted amplification, requiring frameworks that use the structural and behavioral signals still available. This article proposes the Digital Polarization Index (DPI; in Spanish, Índice de Polarización Digital), a six-stage operational framework classifying each unique account into four categories (Trolls, Activists, Public opinion, Neutral), reporting four inauthenticity dimensions (D1–D4) and assigning typological labels with scale and intensity qualifiers. The framework applies to political conversation directed at an identifiable actor that has reached trending status on X. Validation covers ten political cases in Ecuador, the Dominican Republic and Mexico (28,007 accounts; 56,623 mentions), with two external contrasts: Botometer-Lite, kappa = 0.368, and Platform Sentiment, kappa = −0.015, the latter supporting a Stance variable directional toward the target actor. The three modalities are differentially activated, though not separable beyond chance. Stance agreement with independent human annotation reaches kappa = 0.411 against kappa = 0.752 between annotators. Four prespecified non-political trending corpora fall outside that scope and receive at least one operational label in three of four, bounding the attribution the framework supports.

**Keywords:** astroturfing; narrative saturation; targeted amplification; social network analysis; X-Twitter; digital political communication; coordinated operation detection; trans-event observatories; Ecuador; Dominican Republic; Mexico

Check for updates

Academic Editor: Matthew K.O. Lee

Received: 23 June 2026
Revised: 27 August 2026
Accepted: 9 September 2026
Published: 30 September 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Digital political conversation today adopts scales and modalities that no methodological framework prior to 2023 fully anticipated. Digital platforms have expanded the voices that can participate in public deliberation, but they have also enabled technical mechanisms that allow this participation to be simulated. The practice of fabricating the appearance of popular consensus when it is being fabricated by a centralized actor is called astroturfing [1] and has been documented in at least 81 countries [2].

The literature has discussed less frequently that astroturfing is not a single phenomenon. There are at least three operationally distinct modalities of astroturfing: mass textual replication operations, in which a single message is copied by thousands of accounts; narrative saturation operations, in which distinct accounts attack the same actor

within a short time window without replicating the text; and focused targeted amplification operations, in which a small core of accounts questions a specific figure in a concentrated manner. Each modality requires distinct metrics to be detected [3,4]. Identifying these modalities precisely matters for society as well as for political actors. It matters for society because the confusion between organic conversation and fabricated operations erodes the legitimacy of public debate, contaminates the markers citizens use to interpret the political climate, and facilitates electoral or public-policy decisions based on artificial signals [5]. It matters for political actors because the inability to distinguish genuine criticism from coordinated attack leads to erroneous strategic responses: overreacting to a manufactured operation as if it were adverse public opinion, or ignoring a genuine wave by interpreting it as automated noise. Rigorous and typified detection therefore constitutes both a tool of democratic hygiene and an instrument of informed reputation management.

To this methodological complexity, a major technical restriction was added in February 2023: X (formerly Twitter) removed public access to the "like" counts at all levels of its API (Application Programming Interface). The decision invalidated a large part of the models published between 2018 and 2022 [6,7]. The recent literature [5,8,9] has responded by proposing a new generation of detectors relying exclusively on the structural and behavioral signals that remain available.

This article fits into this line of work and makes three contributions. The first is the Digital Polarization Index (DPI), an operational framework that classifies each unique account into four conceptual categories and typifies the modality of coordinated behavior present in a corpus, over a class of input it states explicitly: political conversation directed at an identifiable actor that has reached trending status on X. The second is its validation over ten political cases of distinct nature distributed across three Latin American countries (Ecuador, the Dominican Republic, and Mexico), demonstrating that the same methodology operates consistently over corpora of widely varying magnitude and over heterogeneous political contexts, complemented by a matched within-corpus control and by four pre-specified corpora that fall outside that class and bound what the framework can be used to claim. The third is a quantitative comparison against two external references: a local reimplementation of Botometer [10], a historical reference of the field now in disuse due to API restrictions, and the Sentiment variable that commercial platforms automatically export for each mention. This dual comparison isolates what the framework adds to each reference: a behavioral troll profile that a structural detector does not flag, and a direction toward the target actor that an affect variable does not carry.

The article is organized into six sections. This introduction states the problem, justifies the importance of typified detection for both society and political actors, and synthesizes the three contributions of the work. Section 2 reviews the related work, articulating the evolution of detectors before and after the API closure, coordination beyond textual astroturfing, the industrial persistence of operators, and the conceptual distinctions the framework depends on. Section 3 details the methodology, presenting the workflow followed to develop the DPI (Section 3.1) and the validation design, with its entry condition and its three prespecified hypotheses (Section 3.2). Section 4 applies the framework to ten empirical corpora distributed across Ecuador, the Dominican Republic, and Mexico: it documents the DPI development on the cases (Section 4.1) and validates the hypotheses case by case and transversally, including the contrast with Botometer-Lite, with the Platform Sentiment variable and with a matched within-corpus control (Section 4.2). Section 5 discusses the contributions, strengths and implications of the study, and reports the four control corpora that fall outside the entry condition together with the limits they establish. Section 6 closes with the conclusions. Eleven Supplementary Material Files document the glossary and nomenclature, the operational thresholds, the specification of the external

reference, the sensitivity analysis of the thresholds, the lexical pattern load per case, the organic control corpora, the robustness check with respect to the calibration corpus, the per-country dimensional panels, the divergence between the two stance variables, the matched within-corpus control, and the specification of the stance classifier.

## 2. Related Work

This section reviews the state of the art in detecting coordinated operations on digital platforms and is organized into four sections. Section 2.1 presents the transition from bot detection to operation detection following the closure of public access to engagement metrics by X in February 2023, includes a characterization of the historical Botometer baseline and its recent deterioration, and states where the present proposal departs from the post-API research line it belongs to. Section 2.2 characterizes the modalities of astroturfing beyond textual replication and the methodological challenges each one raises. Section 2.3 discusses the literature on the industrial persistence of operators, the regional evidence on which this framework was calibrated, and the emerging proposal of trans-event observatories as a longitudinal monitoring device. Section 2.4 establishes the conceptual distinctions the framework depends on: between bots and trolls, between stance and sentiment, and between measuring polarization and measuring the operations that act within it.

### *2.1. From Bot Detection to Operation Detection*

The pioneering models for detecting coordination on social networks come from social informatics and computational sociology. Ratkiewicz et al. [1] proposed the Truthy system, which detected "political memes" through propagation speed and network patterns, and Vosoughi, Roy, and Aral [11] demonstrated in an analysis of 126,000 cascades that false narratives propagate with greater depth and speed than true ones, establishing temporal thresholds that have become a central reference for subsequent detectors. To this base were added Ferrara et al. [12], who discuss temporal bursts—abnormal concentrations of replication in short windows, and Pacheco et al. [13], who systematized the problem through a typology of coordinated inauthentic behaviors. These foundational works remain the conceptual base of the field and provide the operational thresholds that this article recovers in the formulation of the DPI.

This methodological corpus was profoundly altered by a major technical restriction. In February 2023, X removed public access to "like" counts at all levels of its API, invalidating a large part of the detection models that predominated between 2018 and 2022 [6,7]. The literature has responded by shifting toward methods that rely exclusively on structural and behavioral signals still accessible. Cresci [8] argued that the supervised, account-by-account detection paradigm had reached its limits and proposed guidelines for a new generation based on aggregate behavior and account patterns; Yang and Menczer [9] presented a characterization of botnets driven by large language models; and Goldstein et al. [5] documented post-LLM generative influence operations. The DPI explicitly fits into this line.

The programmatic shift that Cresci [8] proposes, from account-by-account supervised classification to the detection of coordinated groups through aggregate behavior, has since consolidated into a research line with its own survey [14]. It is therefore necessary to state where the present proposal sits inside that line and where it departs from it, because the two are not the same instrument. Three differences are substantive. The first concerns the object of the output. The dominant post-Cresci work detects whether a group is coordinated and returns a binary decision or a coordination score; the framework proposed here takes a corpus already under suspicion and asks a different question, namely which modality of operation is present, returning a typological label with explicit scale and intensity

qualifiers rather than a verdict on the existence of coordination. The second concerns the assumed agent. Group-level detectors inherit from bot detection the premise that the coordinating unit is automated, and consequently privilege signals of automation: posting regularity, creation bursts, and synchrony below human reaction time. The documented structure of troll farms describes a different operator, salaried and human, whose individual output is not separable from ordinary posting at the level of a single account [15,16]. Rule R3 of the framework is designed for that operator: it isolates accounts that sustain aggressive vocabulary with unidirectional stance and without verifiable presence across at least two mentions, a behavioral signature that automation-oriented features do not capture, and it does so without requiring any inference about whether a human or a script produced the text. The third difference is temporal scope. Coordination detection is normally posed within a single event, whereas this framework accumulates classifications across successive events of the same country, which is what allows an operator to be characterized as persistent rather than as opportunistic.

This evolutionary line is closed by the obligatory reference to the historical baseline of the field. Varol et al. [10] introduced Botometer, a supervised learning model that classifies accounts as bots or humans based on approximately one thousand one hundred fifty attributes, and which for years was the operational reference of the field. Its precision began to deteriorate from 2020 onward. Rauchfleisch and Kaiser [17] demonstrated that it overestimates bot presence in political contexts, presents temporal instability, and depends critically on access to attributes the API currently restricts. Cresci [8] consolidated this critique by describing Botometer as a transitional method, historically important today but methodologically obsolete for the post-API era. For this reason, the present article preserves Botometer as an obligatory historical reference and contrasts it quantitatively with the DPI through a local reimplementation, called Botometer-Lite.

### *2.2. Astroturfing Modalities and Coordination Beyond Textual Replication*

In parallel to the methodological migration forced by the API closure, contemporary literature has recognized that classical textual astroturfing is only one modality among several. Ng and Carley [3] proposed BotBuster, a multi-platform detection system based on a mixture of experts, and Schoch et al. [4] showed that coordination patterns reveal political astroturfing across countries—a manifestation of coordinated inauthentic behavior (CIB)—arguing that future detectors must go beyond the strict textual criterion. In the same line, Pacheco et al. [13] showed that the temporal synchrony of coordinated action—distinct accounts acting on the same target in the same time window without literally replicating the text—is a robust marker of coordination independent of textual replication, while Mosleh and Rand [18] documented that exposure to manipulated political signals is structured by the network of followed elites rather than by content alone.

On this basis, this article identifies three operationally distinct modalities of coordinated operation: mass textual replication, in which a single message is copied by thousands of accounts; narrative saturation, in which distinct accounts attack the same actor within a short time window without replicating the text; and focused targeted amplification, in which a small core of accounts questions a specific figure in a concentrated manner. The distinction is relevant because each modality requires distinct metrics to be detected and leaves distinct structural traces in the corpus. The DPI recovers this typological classification as a central piece of its operational output and formalizes it through explicit rules over a panel of four dimensions, presented in Section 3.

### *2.3. Operator Persistence and Trans-Event Observatories*

The evidence on industrial persistence of operators has turned continuous monitoring into a methodologically justified practice, not an accessory operational choice. Bradshaw, Bailey, and Howard [2] documented through systematic review the existence of a global disinformation industry in at least eighty-one countries, organized through contracts among governments, political parties, and private companies that maintain pools of reusable accounts across successive events. Suárez-Serrato et al. [19] presented regional evidence of automated account networks in Latin America, and DiResta et al. [20] showed that accounts appearing as trolls in one operation tend to reappear in operations of a distinct nature weeks or months later.

These findings justify the shift from case-by-case analysis to longitudinal monitoring through observatories that accumulate classifications across successive events. A trans-event observatory allows identifying accounts that persist over time fulfilling the same operational role, distinguishing recurrent patterns from patterns of opportunity, and building profiles whose validity does not depend exclusively on a single corpus. This article incorporates that proposal in the Sapiens Observatory, described operationally in Section 3.1 and applied in Section 4 to the ten study cases. On the preceding landscape, the specific contribution of the proposed DPI fits within three converging lines: the new generation of post-API detectors based on structural and behavioral signals; the explicit typification of operation modalities beyond textual replication; and the incorporation of trans-event monitoring as a device methodologically articulated with the analysis of each case.

The regional evidence deserves to be stated on its own terms, because the framework was designed for Latin American political conversation and calibrated on it, and a regional focus is a choice that requires justification rather than assertion. Three features of the setting bear directly on the design. The first is the density of the empirical record: coordinated manipulation in the region is documented but studied less extensively than its North Atlantic counterpart, and the available work concentrates on a small number of national cases. Suárez-Serrato et al. [19] documented automated account networks in Mexico, and Recuero, Soares, and Gruzd [21] analyzed more than eight million tweets from the 2018 Brazilian presidential election, finding that as hyperpartisan outlets gained centrality the traditional media outlets lost it, and that disinformation circulated within both partisan clusters rather than within one alone. The second is linguistic. The instruments the field relies on were developed for English, and the Spanish-language resources that do exist are fragmented across datasets built to different specifications: a recent effort to fine-tune transformer models for misinformation detection in Spanish had first to unify those sources into a single corpus of sixty-one thousand six hundred seventy-four articles, and found a Spanish-specific encoder to outperform the multilingual architectures evaluated on the same task [22]. That asymmetry is the reason the framework declares a case-specific lexicon rather than a model transferred across targets: where no labeled corpus exists per actor, an auditable lexicon is deployable and a supervised classifier is not. The third concerns the receiving public. A survey of four hundred and four respondents aged eighteen to thirty-four in Medellín, analyzed through exploratory factor analysis and structural equation modeling, associates the immediate resharing of false content with the wish to alert one's immediate circle, with the consistency of the message with prior beliefs, and with the absence of time to verify it [23]. Those are properties of the audience rather than of the message or of its sender, which supports treating the audience composition of a corpus as part of the object of measurement and not as noise around it. The three features converge on one design consequence: an instrument for this region must not depend on a trained classifier per target actor, must not assume that the operator is automated, and must be reproducible by a third party who holds the lexicon.

### *2.4. Bots, Trolls, and the Polarization the Operations Act Within*

The framework classifies accounts into categories that carry contested names, and the contest over those names cannot be settled by an operational definition alone. Two of those names, bot and troll, are routinely used as if they were points on one scale of inauthenticity. They are not; they classify on different axes. Bot is a claim about the mechanism, that the posting is automated. Troll is a claim about conduct and about role. The interactional lineage defines trolling as a deliberate provocation whose purpose is to disrupt a conversation, and it did so long before political operations were the object of study [24], with subsequent work locating a dispositional correlate in the individual [25]. The operational lineage is separate and more recent: it describes the troll not as a personality but as a role inside an organized structure, assigned to one of five handle categories whose daily behavior is consistent within each category and differs across them; the authors characterize the effort as industrial, mass-produced from a system of interchangeable parts in which each class fulfills a specialized function [15], documented in state-sponsored operations and in their downstream effects [16,26,27]. Nothing in either lineage implies automation, and nothing in automation implies either.

The distinction matters because the attribution of automation is empirically unstable in exactly the setting where this framework operates. The category of social bot covers heterogeneous objects that a typology has to separate before any measurement is meaningful [28], and the assumption that automation is detectable from platform-visible behavior has been contested on its own terms [29]. The strongest evidence is comparative: applying three established detection methods to the same five political discourses, over one hundred twenty-two thousand accounts and two hundred sixty-three thousand tweets in five Western democracies, yields bot populations with hardly any overlap between them, so the answer to whether an account is automated depends on which instrument is asked [30]. This framework therefore does not claim automation for any account. Its categories are defined over observable conduct within the corpus, which is the level at which the evidence is stable, and its contrast with Botometer-Lite in Section 4.2.5 is presented as a comparison between two different questions rather than as agreement between two estimates of one quantity.

The term troll is nonetheless imperfect for what the category actually captures here, and it is better to state the imprecision than to let an operational definition absorb it silently. In this framework an account is classified as Troll when it satisfies one of the six rules of Section 3.1, all of which describe corpus-internal conduct: name patterns typical of bulk registration, sustained aggressive vocabulary with unidirectional stance and without verifiable presence, systematic textual replication, or pure amplification of coordinated content. The label therefore asserts that an account behaves as an operational instrument within the event under analysis. It does not assert that the account is automated, that it is paid, that it is coordinated with any other specific account, or that the person behind it holds the positions the account expresses. Alternatives were considered and rejected: an inauthentic account presumes the attribution the framework declines to make, and an operational account is unfamiliar enough that it would require the same explanation while losing comparability with the literature the results are read against. The term is retained because it is the one the field uses for this role, and the operational definition, not the connotation, is what the results should be read through. Readers who prefer a neutral reading may substitute operationally instrumental accounts throughout without any change to the measurements.

The informational input the framework consumes is stance, not sentiment, and that choice places it in a literature with its own history. Stance detection was established as a shared task on tweets [31] and has been surveyed for its feature-based and classical

machine learning generation [32] and for its neural and transformer generation [33]. The two constructs are routinely conflated and are not interchangeable: sentiment is a property of the text considered in isolation, whereas stance is directional toward a declared target, so a text can be hostile in tone while favoring the target actor, and favorable in tone while attacking it. The consequence for an auditing instrument is that a sentiment score cannot substitute for a stance label without changing what is being measured, which is why Section 4.2.5 reports the divergence between the two rather than treating either as a validation of the other. Where this framework departs from the surveyed tradition is in the classifier itself, which is lexical and deterministic rather than trained, for reasons of traceability, deployment cost and exact reproducibility that Section 3.1 states in full together with what the choice forgoes.

Finally, the framework carries polarization in its name and must be explicit about the claim that name does and does not make. Political science distinguishes ideological distance from affective polarization, the mutual hostility between partisan groups that intensified over recent decades and that the literature treats as a division distinct from issue-based disagreement, rooted in partisanship as a social identity rather than in policy distance [34]. Coordinated operations are legible against that background, but the direction of the relation is not something this framework establishes. The question of whether operational accounts produce polarization or exploit a polarization already present is causal, and none of the corpora analyzed here can answer it: they are event-bounded observational samples, selected for public relevance, without pre-event baselines or any exogenous variation to exploit. What the framework does measure is the operational footprint left inside an already polarized conversation, in four dimensions and with a typological label, on the premise that the footprint is a necessary object of study whichever way the causal arrow runs. This position is stated here rather than in the discussion because it bounds the interpretation of every result reported below, and Section 5 returns to what would be required to test the causal question rather than describe its traces.

## 3. Methodology

Section 3.1 describes the workflow from the event mention corpus to the assignment of typological labels, organized into independent operational stages and supported by a synthetic figure of the full pipeline. Section 3.2 presents the validation design applied to the ten study cases, states the three prespecified hypotheses with their operational success criteria, and refers to Section 4 for the quantitative evaluation, including the contrast with two external references: a local reimplementation of Botometer and the Sentiment variable that commercial platforms automatically export.

### *3.1. Methodology for Developing the Index*

To build a DPI capable of operating on the signals still available after the closure of public access to engagement metrics on X, a six-stage workflow was followed that transforms the event mention corpus into a typified classification of the detected operation. Figure 1 synthesizes this flow. The design fits within the guidelines proposed by Cresci [8] for the new generation of post-API detectors, takes up the notion of burst from Ferrara et al. [12], and the temporal thresholds documented by Ratkiewicz et al. [1] and Vosoughi, Roy, and Aral [11] are empirical references for calibration.

The first stage consists of corpus acquisition. Each analyzed case starts from a set of mentions extracted from platform X through Brandwatch (Brandwatch, Brighton, UK; https://www.brandwatch.com, accessed between June 2025 and August 2026). The pipeline is agnostic regarding the extraction tool, provided that the exported corpus preserves the required canonical fields: mention identifier, author, datetime, text, verification

status, follower count, and mention URL. The second stage assigns to each mention a proprietary DPI variable, called Stance, that codes the directional position of the text toward the target actor of the case at three levels: "Attacks", "Defends", or "No stance". Stance is operationally distinguished from the Sentiment that some commercial platforms export with the corpus: Sentiment is automatic emotional polarity assigned to the text in abstract and does not look at the target actor of the case, whereas Stance is directional toward the actor and constitutes the informational input of the DPI.

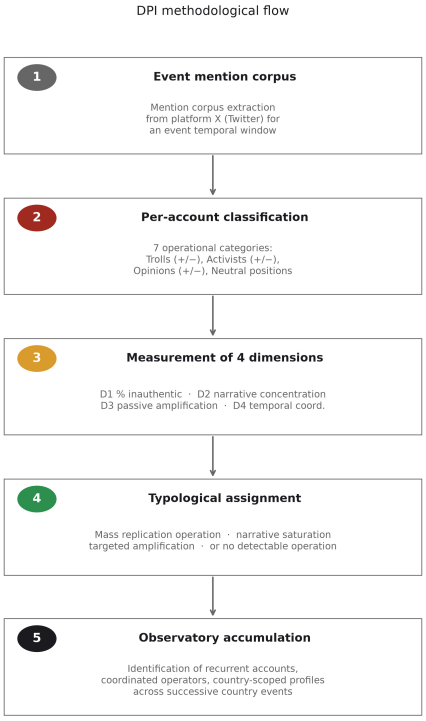



**Figure 1.** Methodological flow of the proposed Digital Polarization Index (DPI). Processing starts from the event mention corpus and successively produces the following: (1) classification of each unique account into one of seven operational categories; (2) measurement of four independent dimensions; (3) assignment of one or more typological labels; (4) accumulation in the Sapiens Observatory to detect recurrent accounts and coordinated operators across successive events from the same country.

Stance is assigned by a deterministic lexical classifier and not by a trained model. Each case declares a curated lexicon with two lists of surface patterns, attack patterns, which express a position against the target actor, and defense patterns, which express a position in its favor. Every unique text of the corpus is compared against both lists by literal substring containment. Writing n_anti(t) and n_pro(t) for the number of distinct patterns of each list occurring in the text, the rule assigns "Attacks" when n_anti(t) > n_pro(t), "Defends" when n_pro(t) > n_anti(t), and "No stance" when the two counts are equal. The tie branch is deliberately absorbing: it covers both the text that activates no pattern at all and the text that activates as many patterns on each side, so a text is coded directionally only when one side strictly dominates. Ties with hits on both sides are rare: 21 unique texts and 177 mentions over the ten study corpora. No probability, threshold or random component enters the rule, so successive runs over the same corpus and the same lexicon reproduce the labels exactly. The decision is taken once per unique text, 8377 decisions over the 56,623 mentions of the ten cases, and propagated to every mention carrying the same text identifier: a retweet inherits the position of the text it replicates, and the pipeline never resolves the amplification graph in order to code position. The lexicon is curated per case from the most replicated texts of the corpus and from the hashtags and slogans that organize the event,

and is not transferred between cases. Table S22 in Supplementary Material File S11 reports its composition and its yield case by case.

The classifier has no supervised training stage and no internal validation split. There are no labeled training examples, no fitted parameters and no held-out partition, because the rule is fully specified by the lexicon and the counting comparison; its external check is the independent manual annotation reported in Section 5.3. What the design does expose to measurement is lexical coverage, the share of mentions receiving a directional code rather than falling into "No stance": 78.8% over the ten cases in the aggregate, with a per-case median of 92.4% and a range from 33.0% to 100.0%. Coverage is near total in corpora dominated by mass replication of a few hostile molds and drops where the conversation around the actor is diffuse and largely informational. Two failure modes follow directly from the mechanism and bound what the variable measures. Matching against the case lexicon is substring-based and without tokenization, so a pattern can fire inside a longer word; it differs in this from the aggressive lexicon of rule R3, whose matching is whole-token and is specified in Supplementary Material File S2. The second failure mode is insensitivity to negation and to quotation, so a text that reports or denies an accusation is coded as if it made it. Both are documented limitations of surface lexical stance detection [32,33], and their aggregate effect is bounded by the annotation contrast rather than assumed away. The choice buys three properties a trained classifier does not offer, and they are the ones Section 2.4 defers to this point: full traceability, since every Stance label reduces to the patterns present in the text and can be contested pattern by a third party holding the lexicon; absence of annotation cost at deployment, since a new case requires a lexicon and not a labeled corpus, which is what makes per-event auditing feasible outside a research laboratory; and exact reproducibility across reruns.

The third stage identifies the coordinated texts of the corpus. Let T be the set of unique texts and, for each text, let vol be the number of distinct accounts that replicated it, vel the rate at which those distinct accounts accumulate per minute over its activity window, and burst the maximum concentration of replications observed in a consecutive fifteen-minute window. A text is classified as coordinated if it meets the condition summarized in Equation (1):

$$\text{vol} \geq 100 \wedge (\text{vel} \geq 1.0 \vee \text{burst} \geq 30) \quad (1)$$

The fourth stage assigns to each mention a narrative, understood as a dominant interpretive frame, through a case-specific lexicon complemented by an aggressive reference lexicon for Spanish (HurtLex ES). The fifth stage aggregates information by account and computes the panel of four DPI dimensions, expressed as percentages. D1 (inauthentic) measures the proportion of the corpus classified as Troll. D2 (narrative concentration) measures the proportion of mentions absorbed by the dominant narrative. D3 (passive amplification) measures the proportion of mentions coming from accounts that act exclusively as amplifiers. D4 (temporal coordination) measures the proportion of mentions falling within fifteen-minute windows with statistically abnormal synchrony of attacks without textual replication. The sixth stage translates the D1–D4 panel into an operational classification. Each account receives one of seven categories (positive or negative Trolls, positive or negative Activists, positive or negative Opinions, and Neutral position). The operation detected in the case receives one or more of three non-exclusive typological labels (Mass replication, Narrative saturation, Targeted amplification), accompanied by two descriptive qualifiers: scale (defined by the size of the corpus) and intensity (defined by D1).

Verifiable presence, used in rules R2, R3 and R6, denotes the institutional verification status declared by the platform for the account, and nothing else. It is a binary attribute read directly from the export field. Follower count is deliberately excluded from it: follower

counts are purchasable and inheritable, and admitting them into the criterion would let an operational account with acquired visibility pass as an authentic presence.

The structure of the D1–D4 panel deliberately articulates two complementary detection modalities. Textual detection, supported by Equation (1) and dimensions D2 and D3, captures operations in which coordination materializes in the literal replication of one or several texts by thousands of accounts. Non-textual temporal detection, supported by dimension D4, captures operations in which distinct accounts attack the same actor in a short window without replicating the text, the characteristic mechanism of the narrative saturation modality [13]. The simultaneous inclusion of both modalities is what allows the DPI to discriminate between typological labels that the classical textual detector collapses into a single phenomenon.

The scale and intensity qualifiers serve the function of qualifying the magnitude and the harshness of the detected operation without altering its typification. Scale is defined in four ranges by the size of the mentioned corpus (low, standard, high, massive), allowing distinction between bounded operations and operations of magnitude comparable to nationwide campaigns. Intensity is defined in three ranges by dimension D1 (low, medium, high) and operates as an aggregate indicator of the proportion of the corpus classified as Trolls. Both qualifiers are reported alongside the typological label to avoid dichotomous readings of the DPI.

DPI computation was implemented in Python 3.12.10 (pandas 3.0.3, numpy 2.4.6, scikit-learn 1.9.0, networkx 3.6.1) over the corpora exported from Brandwatch.

The aggressive reference lexicon operates over the account aggregation stage and derives from HurtLex ES [35], a multilingual lexicon of words to hurt organized in seventeen semantic categories and two inclusion levels, conservative and inclusive. The Spanish distribution employed contains 5006 entries over 3354 distinct lemmas. The DPI does not apply the resource as distributed. A filter cascade retains fourteen of the seventeen categories, restricts the selection to the conservative level and to entries flagged as explicit stereotypes, and discards multiword expressions and forms shorter than four characters; a manual curation step then removes entries whose Spanish political usage is descriptive rather than offensive and adds regional forms observed in the corpora. The derived lexicon actually used by the pipeline contains 470 unigrams, 417 of them inherited from HurtLex after filtering and 53 added by hand. Table S3 documents the cascade step by step, and Table S4 reports the categorical composition; Supplementary Material File S2 also states the tokenization and the known direction of its error.

The 0.30 cut-off of rule R3 is not inherited from a published threshold. It results from empirical calibration over the study corpora and is documented as such. Its empirical basis is the shape of the pct_aggressive distribution, which is close to degenerate at its extremes rather than continuous. Over the 28,007 unique accounts of the ten study cases, 67.4% score exactly zero and 16.7% score exactly one, and only 4.9% fall strictly inside the open interval between zero and 0.30; the cut-off therefore sits at the 72nd percentile of the pooled study distribution and at the 87th percentile of the control corpus, in a region of the support where almost no probability mass is located. The practical consequence is that the exact position of the cut-off is close to immaterial: displacing it across the whole range from 0.20 to 0.60 moves the aggregate troll share of the study set within a span of 0.71 percentage points, from 76.03% to 76.74%, and the corresponding share of the control corpus within a span of 0.24 points. Table S5 reports the distribution per case, and Table S6 the sensitivity.

### *3.2. Validation Methodology and Hypothesis Statement*

**Scope condition and case selection.** The DPI operates on a defined class of input. A corpus enters the framework when the conversation it captures satisfies three conditions

jointly: it is political, it is directed at an identifiable target actor, and it has reached trending status on X. This condition defines the population to which the framework applies rather than a sample drawn from a wider one.

The third condition is not established by the researchers. Trending status is determined by the platform, on criteria external to this study, and it is what makes the conversation directional: a trending topic organized around an actor establishes, prior to any analysis, that a body of accounts is addressing that actor. In six of the ten study cases the conversation is articulated by a central hashtag; in the remaining four it is articulated by the target actor's own name in the trending list. Whether the trending conversation is favorable or hostile to the target actor does not enter the criterion, and the ten cases include both.

A constructed hashtag that reaches trending status is itself preliminary evidence of coordination, in the sense in which Nizzoli et al. [36] and Giglietto et al. [37] use the term: someone assembled the slogan and promoted it. It is not evidence of inauthenticity. The framework is therefore not proposed as a device for discovering that coordination exists in a conversation, which the entry condition already indicates, but for characterizing its modality and reporting the inauthenticity markers that accompany it.

The empirical validation of the DPI is applied to a set of ten political cases of distinct nature distributed across three Latin American countries: five in Ecuador, two in the Dominican Republic, and three in Mexico. The cases were selected for their public political relevance in the period analyzed and for the availability of consolidated corpora with the canonical fields required by the pipeline. Each case was processed with the same methodology, without modification of criteria between cases or between countries, which allows attributing the differences observed in the outputs to empirical properties of the analyzed operations and not to heterogeneous calibration decisions. The selection of the ten cases responded to criteria of public political relevance and corpus availability in the study period, without claim of exhaustiveness. Events with potential operational signatures occur with daily frequency in the three countries considered, and the Sapiens Observatory incorporates them incrementally. The trans-event findings reported in Section 4.2.3—persistent cores and operational sub-cores by country—reflect the current state of the accumulated corpus and will be refined as new events are documented and processed with the same homogeneous methodology.

On this design, three validation hypotheses are prespecified. Each hypothesis is formulated as a claim that the DPI must satisfy if it functions as a valid detector of coordinated operations, and is accompanied by an operational success criterion verified in Section 4.

H1—discrimination between organic conversation and coordinated operation. The DPI assigns operationally distinct typological labels to cases where structural evidence of coordination exists and to cases where such evidence does not exist, without resorting to per-case threshold adjustment. The operational success criterion states that the four panel labels (Mass replication, Narrative saturation, Targeted amplification, and absence of detectable operation) are differentially activated across the ten study cases. The entry condition has a direct consequence for the first hypothesis. Public conversation without prior evidence of coordination is incompatible with it by construction, so H1 cannot be evaluated on the study set alone. Section 5.3 reports what the framework returns when it is applied to four corpora that fall outside the entry condition.

H2—empirical differentiation of operational modalities. The three non-exclusive typological modalities of the DPI give rise to panel profiles that are separable in the D1–D4 space. The operational success criterion states that corpora carrying different label combinations occupy regions of that space whose separation exceeds what random assignment of the same labels to the same panels would produce, evaluated by permutation over the observed set. The correspondence between a panel and the label derived from it

is definitional rather than empirical, since the assignment rules of the typological labels fix it; the hypothesis therefore does not test that correspondence but the prior question of whether the panel positions themselves are structured by modality.

H3—industrial persistence of trans-event operators. When successive events from the same country are analyzed with the same methodology, a subpopulation of accounts emerges that reappears with a Troll profile in more than one event, evidencing reused digital infrastructure in the sense of Bradshaw, Bailey, and Howard [2] and DiResta et al. [20]. The operational success criterion states that the Sapiens Observatory identifies at least one non-trivial set of accounts with the COORDINATED_OPERATOR profile (Troll in at least two events from the same country) or PERSISTENT (Troll in a sustained majority of consecutive events from the same country) upon closure of the set of analyzed cases.

The validation scheme is completed with two quantitative contrasts external to the DPI reported in detail in Section 4.2.5. The first confronts the DPI account classification with the output of Botometer-Lite, a local reimplementation of the public Botometer logic [9,10]. The second confronts the DPI Stance with the Sentiment variable that commercial platforms automatically export for each mention. Both contrasts are designed to evidence whether the DPI captures modalities that the external references miss or whether, on the contrary, both methods converge in their outputs.

**Use of AI-assisted tools.** Generative AI-assisted tools were used in a strictly supportive capacity during this research; the authors reviewed and validated all outputs and take full responsibility for the content:

(i) Brandwatch IRIS was used to support the exploratory identification of narratives within the mention corpus. It assisted in clustering and summarizing large volumes of mentions to surface candidate themes, which the authors then reviewed and curated manually. IRIS did not assign stance, operational categories, or DPI scores; its output was an analytical input only and never a decision factor in the account classification, which is produced by the deterministic rules of the DPI pipeline.

(ii) GitHub Copilot (https://github.com/features/copilot, accessed on 25 August 2026) was used to assist in debugging, validating, and optimizing the Python code that implements the methodology. It provided code completions and suggestions that helped detect and correct errors and improve computational efficiency in the data-processing scripts. Every suggestion was reviewed and tested by the authors against expected outputs; the algorithmic logic, the rules (R1–R6), the dimensions (D1–D4), and the thresholds were designed and implemented by the authors, not by the tool.

(iii) Google Translate was used to assist the English rendering of specific highly technical passages originally written in Spanish, in order to preserve their precise meaning; all translated text was reviewed and verified by the authors.

## 4. Analysis and Validation

Section 4.1 presents the DPI as a finished operational index: it states the troll detection rules, the seven resulting account categories, the four aggregated corpus dimensions, and the four typological labels with their qualifiers. Section 4.2, developed in subsequent sections, applies the index to the ten empirical corpora of the study and evaluates the three hypotheses prespecified in Section 3.2, also incorporating the two external quantitative contrasts against Botometer-Lite and against the Platform Sentiment variable.

### *4.1. DPI Development*

The DPI comprises four successive operational layers. The first layer is a set of six behavioral rules (R1–R6) that classify an account as a Troll if it satisfies at least one of them. The second layer combines the troll classification with the account-dominant

stance toward the case actor, producing seven operational categories grouped into four conceptual sides. The third layer aggregates that classification over the corpus through four independent dimensions (D1–D4), each with a tricolor chromatic code. The fourth layer translates the dimensional panel into one or more of four non-exclusive typological labels accompanied by two descriptive qualifiers. Each layer is described below.

A corpus account is classified as a Troll if its behavioral footprint within the analyzed event satisfies at least one of six prespecified structural or behavioral rules. The rules are applied in hierarchical order (R1 before R2, and so on): the account is assigned to the side corresponding to the first rule it satisfies. Each rule captures an operational signature documented in the field literature, complemented by empirical calibration on the study cases. Table 1 synthesizes the six rules with their formal condition and bibliographic support. The operational terms used throughout are defined in File S1.

**Table 1.** Operational rules for troll detection (R1–R6).

| Rule | Operational Name | Condition | Bibliographic Support |
|---|---|---|---|
| R1 | Strong alphanumeric pattern | username ends in $\geq$6 consecutive digits | Cresci et al. (2017), [38] |
| R2 | Moderate alphanumeric pattern without presence | name ends in 4–5 consecutive digits $\wedge$ $\neg$verified $\wedge$ clear stance | Cresci et al. (2017), [38] |
| R3 | Lexical aggressiveness with a clear stance | pct_aggressive $\geq 0.30 \wedge$ clear stance $\wedge$ $\neg$verified $\wedge$ $n \geq 2$ mentions | Bassignana, Basile, and Patti (2018), [35] on The HurtLex lexicon |
| R4 | Repeated amplifier | $n \geq 3$ mentions $\wedge$ pct_rt $\geq 0.95$ | Yang and Menczer (2024), [9] |
| R5 | High textual replication | rep_textual $\geq 0.70 \wedge n \geq 2$ mentions | Ratkiewicz et al. (2011), [1]; Pacheco et al. (2021), [13] |
| R6 | Pure amplifier | $n \geq 1$ mention $\wedge$ pct_rt $\geq 0.95 \wedge$ $\neg$verified $\wedge$ clear stance | Ng and Carley (2023), [3] |

The six rules are logically disjunctive: it suffices for one to be met for the account to be classified as a Troll, without requiring simultaneous satisfaction of several. Their union covers the operational modalities documented in Section 2. Rules R1 and R2 capture the structural profile of mass account generation through alphanumeric patterns in the username [38]. Rule R3 incorporates the aggressive lexical dimension with directional stance without verifiable institutional presence, grounded in the HurtLex lexicon for Spanish [35]. Rules R4 and R5 capture the classical profile of textual astroturfing (accounts that amplify through repeated retweets or literally replicate text shared among distinct accounts). Rule R6 captures the pure burner-type amplifier modality, single-use accounts that retweet coordinated content without their own content, which the classical recidivist textual detector (R4) does not catch due to the recidivism threshold. The DPI classifies behavior within the analyzed corpus, not the technical nature of the account: the DPI Troll category is not synonymous with bot nor implies automation; it describes an observable behavioral pattern and is reinterpreted event by event under the principle of situational reading.

Once detection rules are applied, each corpus account receives a final classification into one of seven operational categories. The assignment combines the result of rules R1–R6 with the account of the dominant stance toward the case actor, computed as a simple majority over the total author mentions showing a clear stance. The seven categories are grouped into four conceptual sides (Trolls, Activists, Public opinion, and Neutral) that synthesize the corpus composition. Table 2 presents the seven categories with their side and operational definition.

**Table 2.** Seven operational categories of the DPI grouped into four sides.

| Operational Category | Side | Definition |
|---|---|---|
| Negative troll | Trolls | Satisfies at least one of rules R1–R6 with a dominant attacking stance toward the actor. |
| Positive troll | Trolls | Satisfies at least one of rules R1–R6 with a dominant defending stance toward the actor. |
| Negative activist | Activists | Clear attacking stance, verifiable presence, low aggressiveness, and at least three mentions. |
| Positive activist | Activists | Clear defending stance, verifiable presence, low aggressiveness, and at least three mentions. |
| Negative opinion | Public opinion | Clear attacking stance without meeting Troll or Activist criteria. |
| Positive opinion | Public opinion | Clear defending stance without meeting Troll or Activist criteria. |
| Neutral position | Neutral | Accounts without a dominant stance toward the case actor. |

The classification is interpreted under the principle of situational reading: the categories assigned to an account describe its behavior within a given event corpus; they do not constitute permanent or intrinsic attributes of the account. The same account can receive distinct classifications in distinct events according to its specific behavioral footprint. This separation between observed behavior and the technical nature of the account differentiates the DPI from detectors producing permanent per-account labels [9,10] and allows keeping the index reading coupled to the analyzed event.

Account classifications are aggregated over the corpus through four independent dimensions, each expressed as a percentage and accompanied by a tricolor chromatic code (green, amber, red) whose thresholds are documented below. Table 3 fixes the notation of Equations (1)–(5). The equations range over three universes that are kept apart: T, the set of unique texts, over which the coordinated text detector of Equation (1) operates; A, the set of unique accounts, the universe of D1; and M, the set of mentions, the universe of D2, D3, and D4.

Dimension D1 measures the inauthentic percentage of the corpus: the share of unique accounts classified as Troll over all unique accounts of the corpus, as in Equation (2).

$$D_1 = 100 \times \frac{|\mathcal{A}_{\text{troll}}|}{|\mathcal{A}|} \tag{2}$$

The unit of D1 is the account and not the mention: an account contributes once to the numerator and denominator whatever its mention volume, so D1 is insensitive to the asymmetry between an operator that posts fifty times and one that posts once. The chromatic thresholds are green 0–33%, amber 33–66%, and red 66–100%. The lower red threshold (66%) marks the point where trolls cease to be a detectable minority and become the dominant composition of the account universe.

Dimension D2 measures narrative concentration: the share of corpus mentions absorbed by the largest substantive narrative, as in Equation (3).

$$D_2 = 100 \times \frac{\max_{k \in \mathcal{K}_*} |\mathcal{M}_k|}{|\mathcal{M}|} \tag{3}$$

**Table 3.** Notation used in Equations (1)–(5).

| Symbol | Definition |
|---|---|
| M | Set of mentions of the event corpus. |
| A | Set of unique accounts (authors) appearing in M. |
| T | Set of unique texts of the corpus, identified by text hash; t denotes one such text. |
| vol(t) | Number of distinct accounts that replicated t. |
| dur(t) | Minutes elapsed between the first and the last mention of t, floored at one minute. |
| vel(t) | vol(t) divided by dur(t): distinct replicating accounts per minute. |
| burst(t) | The largest number of mentions of t is contained in a fifteen-minute interval closed at both ends, so two mentions exactly fifteen minutes apart fall in the same interval. The sliding windows of W used by Equation (5) are half-open instead. |
| Coord | Subset of T whose texts satisfy Equation (1). |
| M(a) | Mentions are authored by account a; $n_a = \lvert M(a) \rvert$. |
| M(A′) | Mentions authored by the accounts of a subset A′ of A. |
| A_troll | Accounts of A satisfying at least one of the rules R1 to R6. |
| r_a | Share of M(a) that are retweets. |
| v_a | Platform verification status of account a. |
| s_a | Dominant stance of a: “Attacks” if more than half of M(a) attacks, “Defends” if more than half defends, “No stance” otherwise. |
| K, K* | Narrative categories of the case lexicon; K* is K without the residual category “Others/general”. |
| M_k | Mentions assigned to narrative k. |
| A_amp | Single-use pure amplifiers, defined in Equation (4). |
| W | Family of sliding windows: the window of a mention with a valid timestamp is the half-open interval that begins at that timestamp and lasts fifteen minutes. |
| A_w, c_w, u_w | Distinct accounts in window w; share of the mentions of w carrying the majority stance direction; share of the mentions of w authored by non-verified accounts. |
| W_coord | Subset of W whose windows meet the three criteria of Equation (5). |
| M_coord | Union of the windows of W_coord. |

The maximum runs over K*, the narrative categories of the case lexicon after removal of the residual category “Others/general”, which collects the mentions matching no operational frame and is a lexical remainder rather than a narrative. The denominator, by contrast, is the complete corpus M. The exclusion is therefore asymmetric: the residue leaves the numerator but stays in the denominator, which bounds D2 above by the framed share of the corpus. Where that remainder is large, the ceiling is restrictive: the framed share is 29.2% in the Banco Pichincha case and 36.5% in the Gloria Reyes case, and the D2 values observed there are 10.5% and 15.3%. A low D2 is accordingly read as the absence of a single dominant frame across the whole conversation, not across its framed part alone. The chromatic thresholds are green 0–50%, amber 50–80%, and red 80–100%. High narrative concentration does not per se imply inauthenticity: a genuine crisis may narratively concentrate a conversation. D2 gains interpretive weight in combination with D1, D3, and D4.

Dimension D3 measures passive amplification: the share of corpus mentions produced by single-use pure amplifier accounts, as in Equation (4).

$$D_3 = 100 \times \frac{|\mathcal{M}(\mathcal{A}_{\text{amp}})|}{|\mathcal{M}|} \tag{4}$$

A_amp collects the accounts whose entire footprint in the event is one mention, at least 95% retweet, without platform verification, and with a clear stance toward the target actor [3]. Rule R6 states the pure amplifier condition with the weaker requirement n_a >= 1; D3 tightens it to n_a = 1, so A_amp is a strict subset of the accounts meeting the R6 condition. Across the ten study cases A_amp holds 13,762 of the 19,820 accounts meeting that condition, that is 69.4%, and the gap widens in mention terms because multi-mention amplifiers contribute several mentions each: in the Noboa–Uribe case the accounts of A_amp produce 39.4% of the corpus while the full set of accounts meeting the R6 condition produces 81.7%. Since the six rules are applied hierarchically, meeting the R6 condition and being labeled by R6 are distinct properties: an account that also matches an earlier rule is labeled by that earlier rule. A_amp is contained in A_troll nonetheless, because only R1 and R2 can precede R6 for a single-mention account and both return a Troll label when the stance is clear. D3 is thus a mention share computed over an account subpopulation of the D1 numerator, and the two dimensions do not decompose additively. The thresholds are green 0–25%, amber 25–50%, and red 50–100%. High D3 is the characteristic indicator of the Narrative saturation modality; combined with high D1 it signals saturation operations articulated on single-use account networks.

Dimension D4 measures non-textual temporal coordination: the share of the corpus falling inside at least one coordinated fifteen-minute window, as in Equation (5).

$$D_4 = 100 \times \frac{|\mathcal{M}_{\text{coord}}|}{|\mathcal{M}|} \tag{5}$$

The windows of W slide: each mention anchors a half-open interval that begins at its own timestamp and lasts fifteen minutes, so windows overlap and are not aligned to clock quarters, and an operation straddling a quarter hour is not split across two counts. A window is coordinated when it holds at least thirty distinct accounts, when the larger of its "Attacks" and "Defends" counts reaches 80% of its mentions, and when at least half of its mentions come from non-verified accounts. Both shares are computed over all the mentions of the window, those with "No stance" included, and not over its distinct accounts. M_coord is the union of the coordinated windows, so a mention contained in several of them is counted once. The window counts reported per case in Section 4.2 are not |W_coord| but the number of maximal contiguous runs of flagged mentions: coordinated windows that overlap or abut are reported as a single episode. The distinction leaves the typological condition of at least one coordinated window unaffected, since the two counts vanish together. Mentions lacking a valid timestamp cannot be flagged and remain in the denominator; they amount to 11 of the 56,623 mentions of the ten study cases. The thresholds are green 0–10%, amber 10–30%, and red 30–100%. D4 is the only dimension independent of textual replication; it detects convergent attacks over time where each account writes something different [13].

The correspondence between the equations as printed here and the operations the pipeline performs was verified rather than asserted. Each of the five was reimplemented from its published statement alone, without consulting the production code, and the two outputs were compared over the whole corpus set: the four inputs and the outcome of Equation (1) for the 15,950 unique texts of the fourteen corpora, and the four dimensions of Equations (2)–(5) for the fourteen panels, 79,806 values in all, with no discrepancy. No

figure reported in this article changes as a result. The formulations that the submitted version printed were incorrect as stated and not as implemented.

The specific values of every threshold in the four layers were subjected to a systematic sensitivity analysis over the whole corpus set rather than over a single case. Each of the twenty-four thresholds of the framework was varied to −50%, −25%, +25% and +50% of its nominal value, one at a time and with the remaining twenty-three held at nominal, and the complete panel was recomputed on the eleven corpora of the study. Across the eighty non-chromatic configurations, 96.5% of the corpus-configuration cells retain the typological label produced at nominal thresholds, each of the three active modalities remains instantiated in at least one of the ten study cases in all eighty configurations, and the ordering of the ten cases by D1 preserves a Spearman correlation with the nominal ordering of at least 0.806. Six thresholds are critical in the sense that some variation within the tested envelope reassigns a label; the remaining eighteen reassign none. The full procedure, the threshold-by-threshold stability table, the documented rationale of each value, the quantified effect of corpus size on the only absolute threshold of the framework, and the limitations of the exercise are reported in Supplementary Material File S4.

The four-dimensional panel automatically combines into a typological label of the operation detected in the corpus. Labels are non-exclusive: the same case can receive one, two, or three labels simultaneously, according to satisfaction of the operational conditions described in Table 4. If none of the three main labels is activated, the corpus is explicitly reported as conversation without detectable operation.

**Table 4.** Assignment rules for DPI typological labels.

| Typological Label | Operational Condition | Characteristic Mechanism |
|---|---|---|
| Mass replication operation | $\lvert \text{Coord} \rvert \geq 1$ (at least one coordinated text in the sense of Equation (1)) | A single text is replicated by hundreds or thousands of accounts in a short window. |
| Narrative saturation operation | $(\text{D2} > 80\% \wedge \text{D3} > 50\%) \vee (\lvert \text{W_coord} \rvert \geq 1)$ | Synchronous convergence of distinct accounts on a common frame without textual replication, or sustained viral amplification. |
| Targeted amplification operation | $\text{D1} > 10\% \wedge \lvert \text{recurrent_operators} \rvert > 0$ | A non-trivial proportion of corpus troll accounts belong to the persistent core of the Sapiens Observatory in previous events from the same country. |
| Conversation without detectable operation | none of the three conditions above is met | The aggregate behavior is consistent with organic conversation. |

A terminological clarification is owed here, because the framework carries the word index in its name and does not return a scalar. Composite indicator practice distinguishes an aggregated index, which collapses several dimensions into a single number, from a dashboard, which reports the dimensions side by side and leaves the aggregation to the reader [39]. The DPI is a dashboard in that sense: D1 to D4 are reported jointly and are never summed or weighted into a composite score, precisely because the combination of dimensions, and not their sum, is what identifies a modality. The name is retained for continuity with the Spanish denomination under which the framework has been deployed, and the absence of a scalar output is stated here so that no reader expects one.

The three active labels—Mass replication, Narrative saturation, and Targeted amplification—are independent conditions on the dimensional panel; a corpus may satisfy one, two, or all three simultaneously, according to the combinations of D1–D4 activated. The "No detectable operation" label is exclusive and applies only when none of the other three is met.

Each typological label is accompanied by two descriptive qualifiers. Scale qualifies the corpus magnitude in four ranges by total mention size: low (<500), standard (500–5000), high (5000–20,000), and massive (≥20,000). Intensity qualifies the corpus harshness in three ranges derived from D1: low (D1 < 33%), medium (33% ≤ D1 < 66%), and high (D1 ≥ 66%). Both qualifiers are reported alongside the label to avoid dichotomous readings of the DPI: a narrative saturation operation may have low scale or massive scale, and the difference is informatively relevant even if the typology is the same. The four dimensions operate over proportions and maxima of the observed corpus: the framework does not perform population inference nor assume a minimum sample size. The detection of operational modalities is by construction robust against case scale, in the sense that a small operation with an extreme narrative signature and a massive operation with the same signature receive identical typological labels; the scale qualifier records the public weight of the operation, and complete reading of a case requires both descriptors.

The DPI typological label classifies the observable behavioral conduct in the corpus, not the political nature of the event to which that corpus belongs. A politically legitimate event may receive a detectable operation label if the corpus meets the operational conditions indicated; reciprocally, an event interpretable as a coordinated attack by its political context may not receive a label if the extracted corpus does not satisfy the conditions of D1, D2, D3, or textual coordination. This separation between behavioral reading of the corpus and contextual reading of the event is deliberate and is discussed in Section 5.

### *4.2. Empirical Validation*

The empirical validation of the DPI is applied to the set of ten study cases, processed with the homogeneous methodology described in Section 3.1 and with the canonical rules R1–R6 documented in Section 4.1. The cases are distributed across three countries: five in Ecuador, two in the Dominican Republic, and three in Mexico. Section 4.2.1 presents a synoptic panel of the ten cases in their quantitative profile (dimensions D1–D4, typological label, and qualifiers). Section 4.2.2 details the individual analysis by country. Section 4.2.3 synthesizes the cross-country transversal findings. Section 4.2.4 contrasts the three hypotheses prespecified in Section 3.2. Section 4.2.5 reports the external cross-validation against Botometer-Lite and against the Platform Sentiment variable.

#### 4.2.1. Comparative Panel

The synoptic panel of the ten study cases is presented in Table 5. For each case, the country code, the internal Sapiens Observatory identifier, the target actor, the corpus size in mentions, the numerical values of the four dimensions D1–D4 expressed as percentages, the resulting typological label, and the two scale and intensity qualifiers are reported. The figures were computed by applying the methodological flow of Section 3.1 to each corpus independently and without modification of thresholds between cases. The individual analysis of each case, with its dominant narrative and temporal dynamics, is developed in Section 4.2.2.

The panel exhibits broad coverage of magnitudes: corpus size varies between 31 mentions (DO/EV001) and 32,381 mentions (EC/EV005), three orders of magnitude of difference. The heterogeneity of qualifiers accompanies this variability: scale is distributed across the four DPI ranges (low, standard, high, and massive) and intensity across the three ranges (low, medium, and high). The three DPI operational modalities are activated in at least one of the ten analyzed cases: the Targeted amplification label appears in nine cases, Narrative saturation in seven, and Mass replication in four. The fourth typological category of Table 4 (conversation without detectable operation) is not activated in any of the ten cases, which is consistent with the entry condition stated in Section 3.2, since a trending

conversation built around an actor already carries preliminary evidence of coordination. The combinations of labels reported in Table 5 (between one and three simultaneous labels per case) fit the combinatorial scheme documented in Section 4.1. The case-by-case analysis of Section 4.2.2 examines the recurrent operators, dominant narratives, and temporal dynamics that the synoptic panel does not capture.

**Table 5.** Comparative panel of the ten study cases. Dimensions D1–D4 are reported after the elimination of non-canonical rules; the typological labels are assigned according to the operational thresholds documented in Table 4. When a case receives more than one label, the modalities are reported separated by the + operator. Own elaboration on the Sapiens Observatory corpora.

| Country | ID | Case | *n* | D1 (%) | D2 (%) | D3 (%) | D4 (%) | Typological Label | Qualifiers |
|---|---|---|---|---|---|---|---|---|---|
| EC | EV001 | Noboa–Uribe | 9708 | 86.6 | 97.6 | 39.4 | 42.0 | Mass replication + Saturation + Amplification | high · high |
| EC | EV002 | J. J. Neira | 197 | 35.3 | 67.0 | 28.4 | 0.0 | Amplification | low · medium |
| EC | EV003 | I. Vélez | 4062 | 90.6 | 60.3 | 33.9 | 33.6 | Mass replication + Saturation + Amplification | standard · high |
| EC | EV004 | C. A. Vera | 1031 | 73.4 | 100.0 | 44.6 | 0.0 | Amplification | standard · high |
| EC | EV005 | Banco Pichincha | 32,381 | 63.3 | 10.5 | 11.4 | 27.9 | Mass replication + Saturation + Amplification | massive · medium |
| DO | EV001 | Yayo Sanz | 31 | 6.5 | 100.0 | 0.0 | 100.0 | Saturation | low · low |
| DO | EV002 | Gloria Reyes | 919 | 29.4 | 15.3 | 9.6 | 10.0 | Saturation + Amplification | standard · low |
| MX | EV001 | A. Armenta | 5765 | 94.8 | 100.0 | 48.3 | 10.3 | Mass replication + Saturation + Amplification | high · high |
| MX | EV002 | F. Noroña | 130 | 23.9 | 100.0 | 0.0 | 0.0 | Amplification | low · low |
| MX | EV003 | Clara Brugada | 2399 | 89.6 | 76.8 | 61.4 | 22.1 | Saturation + Amplification | standard · high |

#### 4.2.2. Country Analysis

Section Ecuador analyzes the five Ecuador cases; Section Dominican Republic the two Dominican Republic cases; and Section Mexico the three Mexico cases. Table 6 reports, for the ten cases at once, the composition of the account population classified as Troll, disaggregated by the rule that admits each account, together with the two aggregates that the typological labels read; Table 7 reports the narrative and temporal structure of each corpus. Both tables close the block of each country with a subtotal row, so that they can be read case by case and country by country without splitting the comparison into six tables. The per-case text that follows therefore carries the context of the episode, the reading of its panel and, where applicable, the singular operational dynamic that the corpus illustrates, and does not restate the figures collected in those two tables or the dimensional values already reported in Table 5. The case-by-case comparisons and transversal findings are reported in Section 4.2.3. Figure 2 synthesizes the operational composition of each corpus into the four aggregated DPI categories and Figure 3 represents the temporal propagation dynamics of mentions with shading of the coordination windows detected by Equation (5).

**Table 6.** Composition of the account population classified as Troll, by study case. Columns R1 to R6 count the accounts that each rule admits when the six rules are applied in the hierarchical order documented in Section 4.1, so every account is counted once, under the first rule it satisfies, and the six columns partition the Troll total. Pure amplifiers is the set that defines dimension D3, namely accounts with a single mention, a retweet ratio of at least 0.95, no verifiable presence and a clear stance; it is not the same set as column R6, because an account meeting that description may be admitted earlier by a structural rule. Recurrent operators count the accounts of the corpus whose country-scoped profile in the Sapiens Observatory is COORDINATED_OPERATOR or PERSISTENT. A subtotal row closes the block of each country. The six rule columns and the Trolll total add up over the cases of the country; the pure amplifier column likewise. The recurrent operator column does not, because one account may appear in several events of the same country, so its subtotal reports instead the distinct identities that hold a COORDINATED_OPERATOR or PERSISTENT profile in the national record. Own elaboration on the Sapiens Observatory corpora.

| Case | Trolls | R1 | R2 | R3 | R4 | R5 | R6 | Pure Amplifiers | Recurrent Operators |
|---|---|---|---|---|---|---|---|---|---|
| EC/EV001 Noboa–Uribe | 5563 | 624 | 572 | 1002 | 181 | 3 | 3181 | 3822 | 721 |
| EC/EV002 J. J. Neira | 65 | 15 | 5 | 0 | 0 | 1 | 44 | 56 | 119 |
| EC/EV003 I. Vélez | 2072 | 148 | 221 | 15 | 261 | 0 | 1427 | 1377 | 918 |
| EC/EV004 C. A. Vera | 603 | 74 | 101 | 78 | 4 | 2 | 344 | 460 | 459 |
| EC/EV005 Banco Pichincha | 7541 | 940 | 879 | 566 | 1727 | 0 | 3429 | 3702 | 1249 |
| Ecuador, five cases | 15,844 | 1801 | 1778 | 1661 | 2173 | 6 | 8425 | 9417 | 1393 |
| DO/EV001 Yayo Sanz | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| DO/EV002 Gloria Reyes | 117 | 19 | 14 | 0 | 4 | 0 | 80 | 88 | 1 |
| Dominican Republic, two cases | 119 | 20 | 15 | 0 | 4 | 0 | 80 | 88 | 1 |
| MX/EV001 A. Armenta | 3696 | 642 | 566 | 390 | 54 | 1 | 2043 | 2785 | 27 |
| MX/EV002 F. Noroña | 27 | 12 | 13 | 2 | 0 | 0 | 0 | 0 | 11 |
| MX/EV003 Clara Brugada | 1745 | 79 | 101 | 3 | 69 | 3 | 1490 | 1472 | 17 |
| Mexico, three cases | 5468 | 733 | 680 | 395 | 123 | 4 | 3533 | 4257 | 27 |

**Table 7.** Narrative and temporal structure, by study case. Narrative frames count the frames of the case lexicon, excluding the residual bucket that groups mentions with no identified frame. The dominant frame is the share of the whole corpus absorbed by the largest substantive frame and is the quantity reported as dimension D2. Coordination windows count the fifteen-minute intervals retained by the D4 detector. Coordinated texts count the unique texts satisfying Equation (1), and the last column gives the mentions carrying at least one of those texts, with their share of the corpus. A subtotal row closes the block of each country. Frames, windows, coordinated texts and mentions add up over the cases of the country, and the share in the last column is taken over the mentions of that country. The dominant frame column is a maximum per case and therefore does not add up; the subtotal reports its range. Own elaboration on the Sapiens Observatory corpora.

| Case | Narrative Frames | Dominant Frame (%) | Coordination Windows | Coordinated Texts | Mentions in Coordinated Texts |
|---|---|---|---|---|---|
| EC/EV001 Noboa–Uribe | 5 | 97.6 | 8 | 3 | 5987 (61.7%) |
| EC/EV002 J. J. Neira | 2 | 67.0 | 0 | 0 | 0 |
| EC/EV003 I. Vélez | 2 | 60.3 | 13 | 1 | 755 (18.6%) |
| EC/EV004 C. A. Vera | 1 | 100.0 | 0 | 0 | 0 |

**Table 7.** *Cont.*

| Case | Narrative Frames | Dominant Frame (%) | Coordination Windows | Coordinated Texts | Mentions in Coordinated Texts |
|---|---|---|---|---|---|
| EC/EV005 Banco Pichincha | 6 | 10.5 | 30 | 16 | 7196 (22.2%) |
| Ecuador, five cases | 16 | 10.5 to 100.0 | 51 | 20 | 13,938 (29.4%) |
| DO/EV001 Yayo Sanz | 1 | 100.0 | 1 | 0 | 0 |
| DO/EV002 Gloria Reyes | 4 | 15.3 | 3 | 0 | 0 |
| Dominican Republic, two cases | 5 | 15.3 to 100.0 | 4 | 0 | 0 |
| MX/EV001 A. Armenta | 1 | 100.0 | 7 | 2 | 3181 (55.2%) |
| MX/EV002 F. Noroña | 1 | 100.0 | 0 | 0 | 0 |
| MX/EV003 Clara Brugada | 4 | 76.8 | 8 | 0 | 0 |
| Mexico, three cases | 6 | 76.8 to 100.0 | 15 | 2 | 3181 (38.4%) |

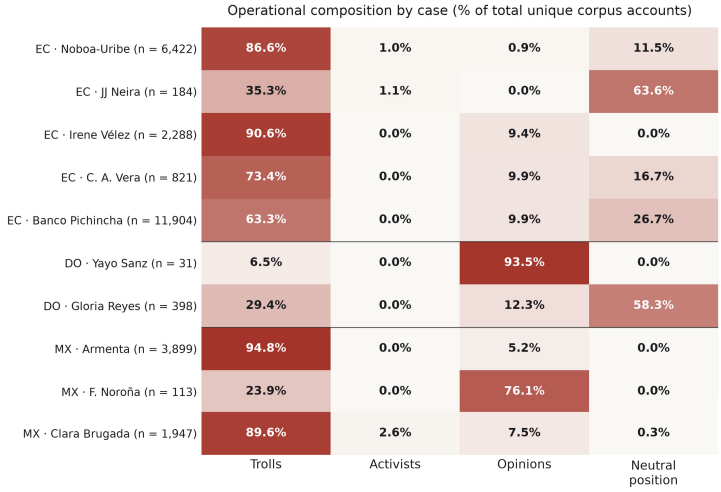



**Figure 2.** Operational composition of the ten cases according to the four aggregated DPI categories (Trolls, Activists, Opinions, and Neutral position), represented as a chromatic matrix. Each cell reports the percentage of total unique accounts of the corpus of each case (n_users in each label, distinct from n_mentions reported in Table 5) belonging to the corresponding category. The chromatic intensity of the accent red is proportional to the percentage. Horizontal lines separate the three study countries. The chromatic heterogeneity of the panel evidences that cases with active typological labels concentrate the highest proportion of Troll accounts, while small corpora present differentiated profiles according to the nature of the episode.

## Ecuador

The five Ecuadorian cases of the study cover political events of distinct nature between June 2025 and April 2026: three controversies against government figures (Noboa–Uribe, J. J. Neira, and Irene Vélez), one case targeting a political analyst (C. A. Vera), and one institutional case of high public exposure (Banco Pichincha). The set covers the four operational scales of the DPI (low, standard, high, and massive) and activates the three typological modalities in distinct combinations. Figure S1 of Supplementary Material File S8 compares the dimensional panels of the country cases.

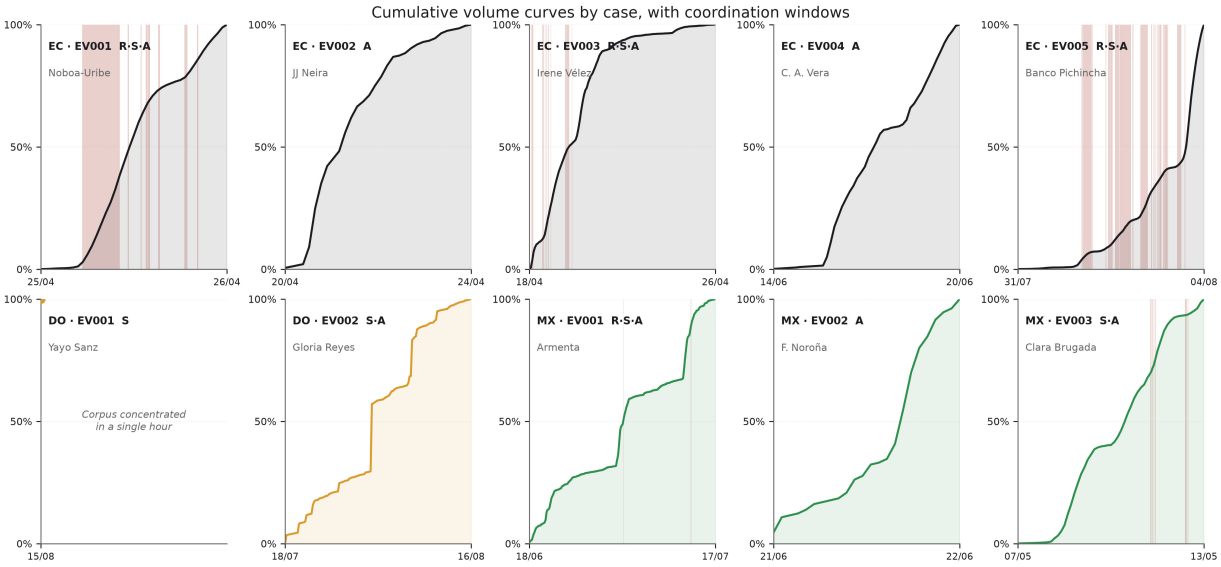



**Figure 3.** Cumulative mention volume curves over time for the ten study cases, normalized to 100% of each case total. Translucent red shaded segments identify temporal coordination windows detected by Equation (5). The heterogeneity of the curves evidences contrasting temporal signatures: concentrated bursts in mass replication operations, extended propagation in cumulative virality, and flat curves in corpora without operational signatures. The typological abbreviations on each panel (R: Mass replication, S: Narrative saturation, A: Targeted amplification) summarize the DPI label of the case.

Case EV001—Noboa–Uribe

Case EV001 documents the political controversy raised on 25 and 26 April 2026 following the public appearance of photographs evidencing a meeting between Ecuadorian president Daniel Noboa and former Colombian president Álvaro Uribe, a meeting that had been previously officially denied. The corpus was extracted from platform X during the 40 h window immediately after the photographs were released. The target actor of the analysis is Daniel Noboa, and the corpus gathers 9708 mentions from 6422 unique accounts.

The quadruple panel signature is consistent with a textual replication operation supported by a pure amplifier core, with temporal concentration in short windows of the night of 25 to 26 April. The synchrony component of that signature does not survive the null model reported in Section 4.2.4: the observed D4 of this case falls below its null mean, so the evidence of coordination here rests on textual replication rather than on temporal alignment between accounts.

The predominant profile is the pure amplifier, which is the main driver of dimension D3, and a second contingent satisfies the lexical-aggressiveness rule R3 and the sustained-replication rule R5; Table 6 reports the composition of the troll population together with the number of corpus accounts already registered as recurrent operators in the Sapiens Observatory, that is, accounts that had appeared with a troll profile in previous events of the same country. The dominant narrative refers to the initial denial of the meeting and to the photographs as evidence, and absorbs most of the corpus, while four minority frames cover the international intersection with the Colombian presidency, the tariff repercussion, personal qualifiers toward Uribe, and the mention of the target actor's family environment. The coordination windows that dimension D4 detects are distributed in fifteen-minute intervals during the main event peak, as Table 7 reports.

The textual coordination detector identifies three coordinated texts, jointly replicated across 5987 corpus mentions, or 61.7% of the corpus, each of them satisfying the conditions of Equation (1). The most replicated of the three reaches 2249 distinct accounts, the

highest single-text volume of the study set, which makes this case the canonical instance of coordinated mass replication in the study.

Case EV002—J. J. Neira

Case EV002 documents the public conversation generated between 20 and 24 April 2026, following the designation of José Julio Neira as Minister of Environment and Energy in Ecuador. The target actor of the analysis is the designated official. The corpus gathers 197 mentions from 184 unique accounts, placing it at the lowest scale of the study.

The moderate-high narrative concentration and the medium proportion of Trolll accounts contrast with the total absence of temporal coordination, which is explained by the limited corpus size: 30 convergent attacks within the same fifteen-minute window are not reached.

The dominant operational profile combines a moderate alphanumeric pattern in the username, under rules R1 and R2, with a pure amplifier group, and the structural signature rests on the recurrent operators of the Sapiens Observatory reported in Table 6. The majority frame refers to the ministerial designation and the appointment, and the second, directed at the target actor, questions his technical capacity and prior experience. The corpus temporality is dispersed across four days, with no identifiable peak and no textual replication burst: the operational signature comes from repeated amplification of a small core, not from mass replication or synchronous saturation.

The case illustrates that the Targeted amplification label can be activated from low-scale corpora, provided the intersection with the Observatory recurrent operators is non-trivial. Detection does not require corpus magnitude to be read: the operational signature is sustained in the overlap of the corpus with the historical record of country troll accounts.

Case EV003—Irene Vélez

Case EV003 documents the conversation originating between 18 and 26 April 2026, following a video posted by Communications Secretary Irene Vélez in an exclusive city sector, a video that generated controversy due to its contrast with the social and sanitary situation of the country. The target actor of the analysis is the official in the video. The corpus gathers 4062 mentions from 2288 unique accounts.

The operational signature combines passive amplification with temporal synchrony and a minimal component of coordinated textual replication.

The dominant operational core is the pure amplifier, which constitutes most of dimension D3, while rules R1 and R2 contribute an additional non-negligible fraction; Table 6 reports the composition and the recurrent-operator intersection, the second largest of the study set. The dominant frame refers to the moral contrast between the public conduct of the official and the social situation of the country, and the complementary one refers to the video itself and its aesthetic component. The coordination windows concentrate at the event peak, which is what places D4 at its red level.

The case constitutes the first record of the study where the post-revision methodology activates the Mass replication label on a corpus that in the previous version of the analysis did not present detectable coordinated text. The increased sensitivity of the textual detector identifies a text replicated above operational thresholds in this case, enriching the typology without modifying the substantive findings about dimensions D1, D2, D3, and D4.

Case EV004—C. A. Vera

Case EV004 documents the public conversation around the hashtag "#HijoBobo" directed at political analyst Carlos Andrés Vera, with a peak on 19 June 2025 and an observation window between 14 and 20 June. The target actor of the analysis is the analyst. The corpus gathers 1031 mentions from 821 unique accounts.

The combination of saturated D2 and null D4 is the characteristic signature of an operation articulated around a single lexical frame that is amplified in a sustained manner but without literal textual replication or abnormal temporal synchrony.

The dominant profile is again the pure amplifier, and the structural signature combines a majority presence of burner accounts with persistence documented in prior Ecuadorian events; Table 6 reports the composition. The single narrative frame is articulated around the hashtag "#HijoBobo" and the family dimension of the attack on the target actor, and absorbs the whole framed corpus, which produces the maximum value of D2. The temporal distribution is extended without a concentrated peak: the attack operates by cumulative virality across several days, not by synchronous mass replication.

Case EV005—Banco Pichincha

Case EV005 documents the public conversation between 31 July and 4 August 2025, around an institutional incident of high exposure associated with Banco Pichincha (the circulation of a customer service episode in which the president of the financial institution intervened, compounded by repeated interruptions of the bank digital service during the same period). The target actor of the analysis is the financial institution. The corpus gathers 32,381 mentions from 11,904 unique accounts, placing it at the massive scale of the DPI.

The significant proportion of Troll accounts coexists with moderate-low values in narrative concentration and passive amplification, and with an amber level of temporal coordination. The combination is atypical compared to the other cases of the Ecuadorian set, where high D2 usually accompanies the triple label.

The operational composition is heterogeneous: pure amplifiers form the largest group, a second significant set satisfies rules R3 and R5, and the corpus carries the largest recurrent-operator intersection of the study set; jointly, pure amplifiers and recurrent operators constitute the major part of the aggregate operational signature, and Table 6 reports the counts. The dominant narrative refers to the failures of the digital service, and the complementary ones include the denunciation of the customer service incident, the anti-bank narrative with political charge and the institutional defense, none of which individually absorbs more than eleven percent of the framed corpus. The coordination windows are distributed in clusters of a few hours linked to the circulation of the incident news.

The case constitutes the only institutional record of the study, in which the target actor is a financial entity and not a political figure. The corpus exhibits a bipolar stance distribution, with a significant volume of defense mentions coexisting with the majority attack; this coexistence operates on disjoint narrative frames and produces the heterogeneous profile of the panel. The case additionally validates the operational concept of the reach operator, a structural profile combining absence of institutional verification with a high number of followers and pure retweet conduct of coordinated text.

Dominican Republic

The two Dominican cases of the study correspond to events occurring in July and August 2025, with peaks separated by six days and partial overlap of observation windows. Both target actors are State officials of the Dominican Republic. The incremental nature of the Sapiens Observatory implies that the accumulated corpus for the Dominican Republic is structurally smaller than the Ecuadorian one (two events versus five). This asymmetry remains present when reading the trans-event comparisons reported in Section 4.2.3 and will be refined as new Dominican events are documented and processed with the same homogeneous methodology. Figure S2 of Supplementary Material File S8 compares the dimensional panels of the country cases.

Case EV001—Yayo Sanz Lovatón

Dominican case EV001 documents the public conversation of 15 August 2025 directed at Yayo Sanz Lovatón, head of the General Customs Directorate in the Dominican Republic. The event trigger is the public circulation of signals linking political campaign activities with the use of institutional infrastructure under his command. The corpus gathers 31 mentions from 31 unique accounts, all concentrated in a single day, which places it at the minimum scale of the study.

The absolute narrative concentration and absolute temporal coordination coexist with a very low proportion of Troll accounts, since only two of the 31 unique accounts satisfy structural or behavioral troll criteria. The framework operates on proportions and maxima of the observed corpus, so the signature is readable at this size. It is also the weakest evidence of the set: Section 5.3 grades the typological label of corpora below one thousand mentions as exploratory and states that the absolute density criterion of D4 is what places a temporally compact corpus of thirty-one mentions at the maximum of that dimension.

The operational composition is minimal: two Troll accounts with attacking stances, one of them registered as a recurrent operator of the Sapiens Observatory and present with a troll profile in the other Dominican event of the set. No pure amplifiers are detected, consistent with the absence of retweets in a corpus where every mention comes from a distinct account, and the rest of the corpus is distributed between accounts with clear stance and no troll signature in the negative opinion category and accounts without a dominant stance. The single narrative, articulated around the use of the public institution as a campaign platform, absorbs the whole framed corpus and produces the maximum value of D2, and the one coordination window holds the thirty-one unique accounts attacking the same target within a fifteen-minute interval, which is where the maximum value of D4 comes from.

The case illustrates that narrative saturation detection is readable even when the corpus is limited to tens of mentions. The label reads the structure of the D1–D4 panel rather than the corpus size directly, but it is not independent of it in either direction: the density criterion of D4 is an absolute count, which is why this corpus reaches the maximum of that dimension, and Section S.4.6 of Supplementary Material File S4 shows that Equation (1) cannot fire in a corpus of this size under any configuration tested. The scale qualifier simultaneously records the limited public weight of the operation.

Case EV002—Gloria Reyes

Dominican case EV002 documents the public conversation between 18 July and 16 August 2025, with a peak on 9 August, around official Gloria Reyes in the context of a public debate over her tenure in office and presidential endorsement. The target actor of the analysis is the official. The corpus gathers 919 mentions from 398 unique accounts, distributed across an approximately thirty-day window.

The four dimensions are placed in the green-yellow chromatic zone, far from red. The Saturation signature is activated by the presence of three temporal coordination windows: the amplification one, by the intersection between the corpus and the population of recurrent operators of the Sapiens Observatory.

Troll status is distributed asymmetrically with respect to stance toward the target actor: fifty accounts carry a negative sign, with attacking stance, and sixty-seven a positive sign, with defending stance. The defense numerically exceeds the attack, which inverts the pattern observed across the Ecuadorian set; Table 6 reports the composition. The dominant frame refers to the presidential endorsement in the debate over the official tenure, and the complementary ones include an officialist counter-narrative that disqualifies critics, whose presence in the corpus is informatively relevant, and further frames of lesser weight. The frame distribution is among the most dispersed of the study set, and the coordination windows are spread across the one-month observation window.

The case constitutes a paradigmatic record of bipolar polarity with coordinated defense. The presence of an explicit counter-narrative operating with its own signature of inauthentic accounts—an officialist lexicon that disqualifies those who criticize the target actor—sustains a proportion of defending Troll accounts greater than that of attacking Troll accounts. The DPI typological label captures both sides as Trolls correctly, and the internal distinction between positive and negative signs reveals the bipolar structure of the operation. The operational thread connecting this case with Dominican case EV001 is developed in Section 4.2.3.

Mexico

The three Mexican cases of the study occurred between June 2025 and May 2026 and share a contextual trait: the three target actors are Morena figures with current public responsibility in the study period (a state governor, a senator, and a head of government). This coincidence allows reading the three cases as windows on the digital dynamics around the Mexican ruling party during the period. With three accumulated events, the Mexican trans-event analysis reported in Section 4.2.3 has greater power than the Dominican one but less than the Ecuadorian one. As in the other countries, the findings will be refined as new events are incorporated into the Sapiens Observatory. Figure S3 of Supplementary Material File S8 compares the dimensional panels of the country cases.

Case EV001—A. Armenta

Mexican case EV001 documents the public conversation between 18 June and 17 July 2025, with a peak on 3 July, around Alejandro Armenta, governor of the state of Puebla. The event trigger is the public discussion of a state initiative regulating social media content that the opposition characterized as a censorship mechanism. The target actor of the analysis is the governor. The corpus gathers 5765 mentions from 3899 unique accounts, placing it at a high scale.

The proportion of Troll accounts reaches the highest value in the set of ten study cases, and the narrative concentration is absolute. The combination is structurally equivalent to that of case EC/EV001 Noboa–Uribe, but the Mexican corpus reaches a higher aggregate intensity.

The pure amplifier core is the main driver of dimension D3, and accounts with strong and moderate alphanumeric pattern, under rules R1 and R2, contribute another significant contingent. The totality of the Sapiens Observatory coordinated operators for Mexico, twenty-seven accounts with troll profiles registered in at least two Mexican events, participates in this corpus, which is what sustains the Targeted amplification signature; Table 6 reports the composition. A single central narrative articulated around the insult hashtag against the target actor absorbs the whole framed corpus and produces the maximum value of D2. The textual detector and the temporal detector are activated simultaneously, as Table 7 shows, which gives rise to the triple panel signature.

The case illustrates the upper extreme of the DPI in terms of aggregate troll intensity: when D1 exceeds 90%, practically the entire composition of the account corpus satisfies structural or behavioral troll criteria, which is the upper extreme of the observed range and lies well above every control corpus reported in Section 5.3. That subsection also states why the framework returns no positive verdict on the authenticity of a conversation. The participation of the totality of the country's coordinated operators introduces the trans-event operational thread developed in Section 4.2.3.

Case EV002—F. Noroña

Mexican case EV002 documents the public conversation of 21 and 22 June 2025 around Gerardo Fernández Noroña, senator for Morena. The event trigger is statements by the

target actor in the national public debate over content regulation. The corpus gathers 130 mentions from 113 unique accounts, which places the case at a low scale of the study.

The narrative concentration is absolute, but passive amplification and temporal coordination are not activated. The operational signature is sustained exclusively on the intersection of the corpus with the population of recurrent operators of the Sapiens Observatory.

No pure amplifiers are detected in the corpus, consistent with a small corpus without sustained retweet bursts, and the operational substrate of the case is the set of accounts registered as recurrent operators of the Sapiens Observatory, reported in Table 6. A central narrative articulated around a campaign hashtag that questions the target actor absorbs the whole framed corpus. The temporality concentrates on two days with no identifiable intermediate peak and no coordination window that activates dimension D4.

Case EV003—Clara Brugada

Mexican case EV003 documents the public conversation between 7 and 13 May 2026, with a peak on 12 May, around Clara Brugada Molina, head of government of Mexico City. The event trigger is atypical floods in the capital that gave rise to a viral conversation with a strong satirical and meme-type component articulated around the term "ajolotización". The corpus gathers 2399 mentions from 1947 unique accounts.

Coordinated text detection is not activated: the case illustrates a saturation operation articulated on pure amplifiers without massive literal textual replication, characteristic of the cumulative meme frame.

The dominant core is the pure amplifier, which constitutes practically the totality of dimension D3; the proportion of corpus produced by burner accounts, and a significant fraction of the Mexican record accumulated in previous events participate in the corpus; Table 6 reports the composition. A dominant satirical-tone frame absorbs most of the framed corpus, complemented by several frames of lesser weight. The coordination windows concentrate on one day of delay, on the second day of virality rather than on the day of the triggering event. This signature, a displaced peak with meme-type mechanics and high amplification without coordinated textual replication, is the characteristic signature of cumulative virality, understood as sustained diffusion that narratively saturates the corpus without triggering the coordinated mass replication detector of Equation (1).

The case illustrates that the Narrative saturation label can be activated through meme-type virality mechanics without literal textual coordination. The high D3 reflects an operation sustained by pure amplifiers that retweet satirical content without contributing their own; the high-not-maximum D2 reflects the presence of a dominant frame with complementary frames; the moderate D4 captures the temporal convergence windows that the meme produced during the virality peak.

#### 4.2.3. Transversal Findings

The following patterns emerge only when considering the ten study cases together and are not visible from the individual analysis of Section 4.2.2. The findings are organized in six observations: the synoptic reading of the D1–D4 panel across the ten cases, the structure of persistent cores intra-country and the operational sub-cores by narrative frame, the polarity pattern of each case and the appearance of bipolar operations with coordinated defense, the cross-event intra-country and trans-country continuity of operational accounts, the distribution of typological modalities across the three countries, and the composition of the operational core of the Sapiens Observatory at the close of the study period. Figure 4 reports the synoptic panel of the ten cases, and Figure 5 disaggregates the internal composition of the Troll universe by canonical rule R1–R6.




**Figure 4.** Synoptic matrix of the four DPI dimensions across the ten study cases. Cells display the numerical value of each dimension with chromatic coding according to the operational thresholds. The right-annotated column summarizes the typology activated in each case (R = mass replication, S = narrative saturation, A = targeted amplification). Horizontal lines separate the three study countries. Color encodes the threshold band of each dimension—green for low, amber for intermediate and red for high—with dimension-specific cut-offs: D1 green 0–33%, amber 33–66%, red 66–100%; D2 green 0–50%, amber 50–80%, red 80–100%; D3 green 0–25%, amber 25–50%, red 50–100%; D4 green 0–10%, amber 10–30%, red 30–100%.

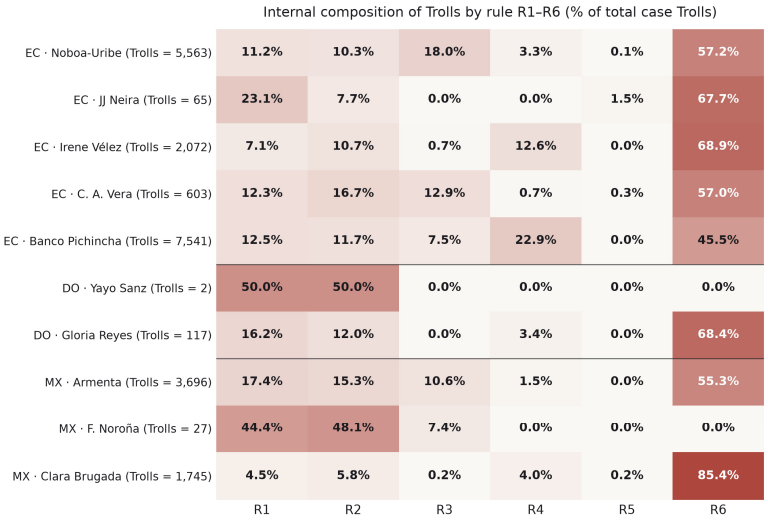



**Figure 5.** Internal composition of accounts with Troll status by canonical R1–R6 DPI rule, disaggregated by case and represented as a chromatic matrix. Each cell reports the percentage of the case total Trolls assigned to that rule under the operational priority order of the code (R1 > R2 > R3 > R4 > R5 > R6, first-match) to avoid double-counting. The accent red intensity is proportional to the percentage. Horizontal lines separate the three study countries. The chromatic heterogeneity of the panel evidences that operations differ not only in scale or typological modality but also in the dominant operational paradigm of the accounts composing them.

Figure 4 synthesizes the numerical values of the four dimensions D1–D4 across the ten study cases, with chromatic coding according to the DPI operational thresholds. Three cases exhibit a synoptic signature near the maximum on the four dimensions: EC/EV001 Noboa–Uribe, EC/EV003 I. Vélez, and MX/EV001 A. Armenta. The three receive the triple typology

Mass replication plus Narrative saturation, plus Targeted amplification, and their panels are visually analogous despite scale differences. EC/EV005 Banco Pichincha receives the same triple typology but with a distinct synoptic profile: D2 and D3 in the green zone and D1 and D4 in amber; the operational signature comes from the corpus magnitude combined with temporal coordination windows, not from monolithic narrative concentration. DO/EV001 Yayo Sanz constitutes the only case with D2 and D4 at absolute maximum on a minimal corpus, which illustrates the robustness against scale property documented in Section 4.1. The remaining cases exhibit partially green panels with a single modality activated.

The Sapiens Observatory allows distinguishing, within each country's accumulated corpus, a class of accounts satisfying troll criteria in multiple successive events. Under the DPI operational definition, profile assignment is a mutually exclusive cascade: an account receives the COORDINATED_OPERATOR profile if it presents troll classification in at least two country events, the PERSISTENT profile if it appears in at least three country events without meeting that first criterion, the RECURRENT profile if it appears in exactly two, and OCCASIONAL if it appears in one. With five accumulated events, Ecuador is the only study country whose dataset satisfies the operational condition to detect the PERSISTENT profile: forty-eight Ecuadorian accounts appear in at least three of the five country events without meeting the coordinated operator criterion, while three hundred sixty-five accounts present troll classification in three or more of those events. The strict intersection of accounts present in the five Ecuadorian corpora is fourteen accounts. This set constitutes the most stable persistent core identifiable with the available evidence.

For the other two countries, the sample size of events does not reach the structural threshold of the definition. The Dominican Republic, with two events, is by construction outside the PERSISTENT criterion; the Observatory registers a single account with the COORDINATED_OPERATOR profile. Mexico, with three events, has the minimum theoretical magnitude for PERSISTENT but an empty set in that category; the Observatory registers twenty-seven accounts with the COORDINATED_OPERATOR profile. The asymmetries among per-country counts reflect the incremental scale of the Sapiens Observatory more than substantive differences in the digital structure of the three countries; Section 3.2 states that the figures will be refined as additional events are documented.

Internal examination of the Mexican operational core reveals a structure of disjoint sub-cores by narrative frame. The totality of the twenty-seven accounts with the COORDINATED_OPERATOR profile from Mexico participates in EV001 A. Armenta, around the anti-Morena ideological frame articulating the case. Ten of the twenty-seven reappear in EV002 F. Noroña, an event with an analogous frame. Seventeen reappear in EV003 Clara Brugada, an event with a satirical frame distinct from the previous one; no account appears in the three events simultaneously. The separation is not between active and dormant accounts, but between accounts mobilized for direct ideological frames and accounts mobilized for meme-type satirical frames (Figure 6).

On the considered pairs, the temporal overlap of windows (EV002 falls within the EV001 window; EV003 is separated by ten months) does not explain the operational signature: the COORDINATED_OPERATOR criterion requires troll classification in each of the events, not mere temporal concurrence.

The polarity pattern varies qualitatively across the ten study cases. Nine of the ten cases exhibit unilateral hostile polarity: most of the Troll account corpus carries an attacking stance toward the target actor, and the defending stance is marginal or null. Case DO/EV002 Gloria Reyes constitutes the only structural exception: the composition of the Troll corpus is distributed between fifty accounts with attacking stance and sixty-seven accounts with defending stance, so that 57.3% of its one hundred seventeen Troll accounts defend the target actor, which inverts the dominant pattern. The defense of the target actor operates

with its own signature of inauthentic accounts, articulated around an explicit counter-narrative that disqualifies those who criticize the official. The simultaneous presence of positive and negative trolls defines the bipolar signature of the case, already operationally documented in Section Dominican Republic.

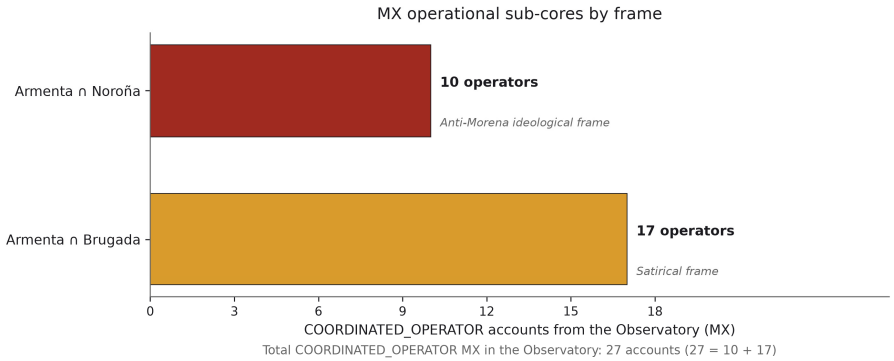



**Figure 6.** Distribution of the twenty-seven accounts with the COORDINATED_OPERATOR profile from the Sapiens Observatory in Mexico across the three country cases. Ten operators specialize in the Armenta-Noroña sub-core with an anti-Morena ideological frame; seventeen operators specialize in the Armenta-Brugada sub-core with a satirical frame. No account operates across the three cases simultaneously. The organization into disjoint sub-cores by frame operationalizes the notion of specialized digital infrastructure [20].

Case EC/EV005 Banco Pichincha presents an informative tail intermediate on the spectrum between unilateral and bipolar polarity. The composition of the case Troll corpus includes six thousand nine hundred twenty-seven accounts with an attacking stance and six hundred fourteen accounts with a defending stance, a proportion dominated by attack but with a non-trivial defense presence. The minority institutional defense narrative is registered in the case lexicon. The difference between the Ecuadorian case and the paradigmatic Dominican case is one of magnitude: eight percent of positive Trolls in EV005 versus fifty-seven percent in Gloria Reyes. The tail of the Ecuadorian institutional case suggests that the bipolar pattern is not a binary but a gradual property of the corpus.

Intra-country operational continuity is quantifiable through the intersection of unique accounts appearing in distinct events of the same country. In Ecuador, the strict intersection of the five Ecuadorian corpora contains fourteen accounts: fourteen distinct identities appear in the five analyzed events uninterruptedly. Paired continuity exhibits intermediate values: the overlap between EV001 Noboa–Uribe and EV003 I. Vélez, both April 2026 events with overlapping temporal windows, registers several dozen accounts; pairs with temporally separated events (EV004 June 2025 and EV005 July–August 2025 versus the April 2026 ones) register smaller overlaps, congruent with the presence of the persistent core. Intra-country continuity in Mexico and the Dominican Republic is documented in the Observatory findings reported in the preceding paragraphs: twenty-seven accounts with the COORDINATED_OPERATOR profile distributed in disjoint sub-cores by frame in Mexico, and a single account with the same profile in the Dominican Republic.

The deduplication arithmetic is stated explicitly, since the two totals reported in this article measure different objects. The gross total of twenty-eight thousand and seven accounts is the sum of the unique accounts of each corpus taken separately, so an account participating in two events of the same country is counted twice. The deduplicated total of twenty-five thousand two hundred seventy-four accounts is the number of distinct identities in the Sapiens Observatory once those repetitions are collapsed within each country. The two thousand seven hundred thirty-three accounts that separate the two figures are distributed as two thousand six hundred fifty-one in Ecuador, fifty-one in Mexico and

thirty-one in the Dominican Republic, and that distribution is itself the trans-event persistence the framework reports: Ecuador accumulates five events and concentrates the collapse, whereas the Dominican and Mexican sets, with two and three events, barely produce it.

Trans-country continuity constitutes an informatively relevant negative observation. The intersection between the set of unique accounts appearing in Ecuadorian corpora and the analogous set of Dominican accounts presents an empty intersection. The intersection between Ecuadorian corpora and Mexican corpora contains a single account. The intersection between Dominican and Mexican corpora presents an empty intersection. Over twenty-five thousand two hundred seventy-four unique identities distributed across the ten corpora (after trans-event deduplication; the gross total summing unique accounts case by case is twenty-eight thousand seven), a single account appears in events from two distinct countries. The result is consistent with the hypothesis of operations anchored to each country's political geography and with digital ecosystems that maintain predominantly local operational account populations.

Figure 7 synthesizes the distribution of the three active typological modalities among the ten study cases, grouped by country. Ecuador activates the three modalities in at least one case: Mass replication in three out of five cases, Narrative saturation in three out of five, and Targeted amplification in all five. Mexico also activates the three modalities, but with Mass replication concentrated in a single case (EV001 A. Armenta), the highest-magnitude one in the country. The Dominican Republic does not activate Mass replication in any of its two cases; Narrative saturation activates in both and Targeted amplification in one.

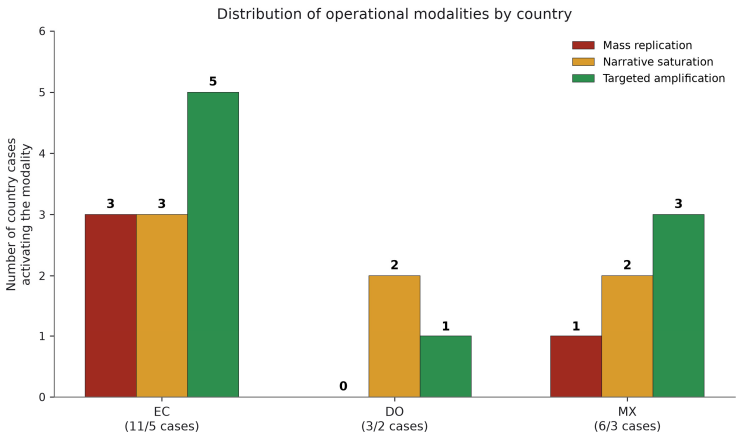



**Figure 7.** Distribution of the three DPI operational modalities across the ten study cases, grouped by country. Bar height indicates the number of country cases activating the corresponding modality. Below each country, the total activated modalities over the total analyzed cases are reported. Targeted amplification is the only modality activated in at least one case from each of the three countries.

Targeted amplification is the only one of the three labels activated in at least one case of each of the three countries, which is consistent with it being the label with the lowest activation threshold. That coverage is not a property of the modality alone: the label requires a non-empty intersection with the accumulated record of the country, and the size of that record differs by three orders of magnitude between Ecuador and the Dominican Republic, so the label is not equally reachable in the three. No case in the study activates it through operators identified outside the closed set of ten corpora, because the Sapiens Observatory is built from those same corpora, and no external register of coordinated operators exists for these countries that could be used instead. Section 5.3 states the limit this places on the label, and Section 5.3 grades the trans-event results by country accordingly. Mass replication, in contrast, concentrates empirically in the highest-

scale cases, above four thousand mentions in each of the four cases activating it, which is congruent with its operational definition: Equation (1) requires a minimum of one hundred accounts replicating the same text, a threshold incompatible with small corpora.

Figure 8 disaggregates the accumulated record of the Sapiens Observatory at the close of the study, separated by country and by profile. The record groups the three profiles above the occasional threshold: COORDINATED_OPERATOR, PERSISTENT and RECURRENT. The total cross-country core groups two thousand fifty-two accounts, distributed asymmetrically among Ecuador (one thousand nine hundred seventy-one accounts), Mexico (fifty accounts), and the Dominican Republic (thirty-one accounts). The Ecuadorian proportion simultaneously reflects the highest number of accumulated events and the documented presence of the persistent core.

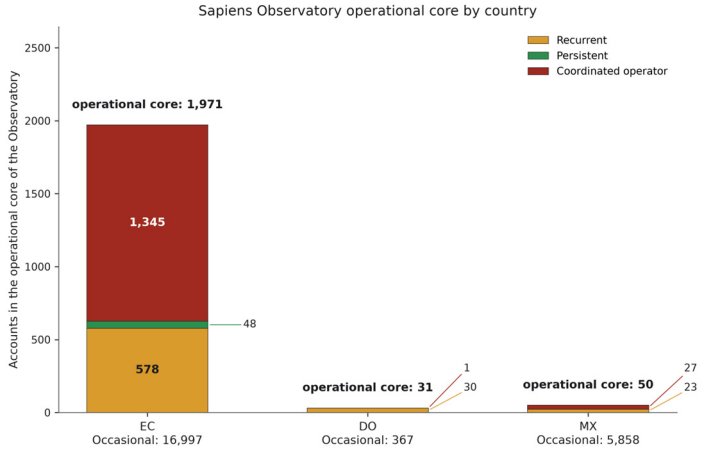



**Figure 8.** Operational core of the Sapiens Observatory disaggregated by profile and by country. The figure deliberately excludes the OCCASIONAL profile (accounts present in a single event), whose count is reported below the horizontal axis as context. The stacked bars show the three operational profiles. The total operational core is annotated above each bar.

Internal composition of the accumulated record varies qualitatively by country, and only the COORDINATED_OPERATOR profile carries operational classification: the PERSISTENT and RECURRENT profiles record sustained presence across events without meeting the troll criterion in two or more of them. In Ecuador, the one thousand three hundred forty-five COORDINATED_OPERATOR accounts represent sixty-eight percent of the Ecuadorian record, alongside forty-eight PERSISTENT and five hundred seventy-eight RECURRENT accounts. In Mexico, the twenty-seven COORDINATED_OPERATOR accounts represent fifty-four percent of the Mexican record. In the Dominican Republic, the sample is informatively limited by construction. The figures will be refined as new events are documented and processed with the same homogeneous methodology.

#### 4.2.4. Hypothesis Results

The contrast of the three hypotheses prespecified in Section 3.2 is organized against the empirical evidence reported case by case in Section 4.2.2 and transversally in Section 4.2.3. For each hypothesis, the original statement is reproduced, the associated operational success criterion is specified, and the evidence supporting, limiting, or introducing methodological caveats on its scope is synthesized. The evaluation of the three hypotheses with their qualitative conclusive state is summarized at the end of the subsection.

Hypothesis H1, stated in Section 3.2, requires the four panel labels to be differentially activated across the ten study cases.

The empirical evidence supports the differentiation among the three operational labels. The ten study cases are distributed among the modalities non-homogeneously: Mass repli-

cation is activated in four cases (EC/EV001 Noboa–Uribe, EC/EV003 I. Vélez, EC/EV005 Banco Pichincha, and MX/EV001 A. Armenta), Narrative saturation in seven cases, and targeted amplification in nine. The per-case combinations vary between one and three simultaneous labels (Table 5 and Figure 4). Dimensions D1–D4 present qualitative amplitude across cases: D1 ranges between 6.5% (DO/EV001 Yayo Sanz) and 94.8% (MX/EV001 A. Armenta); D2 between 10.5% and 100.0%. The pipeline assigns these labels without threshold modification between countries or cases.

The operational criterion also states the differentiated activation of the fourth label, absence of detectable operation. This label is not activated in any of the ten study cases, and the entry condition stated in Section 3.2 explains why. A conversation enters the framework only once it is political, directed at an identifiable target actor and trending on X, and a constructed hashtag that reaches trending status already carries preliminary evidence of coordination. Public conversation without such evidence is therefore incompatible with the entry condition by construction, and no selection made within the study set could have produced it.

Two readings of H1 must be kept apart. The operational success criterion as stated is met: the three active labels are differentially activated across the ten study cases, in the combinations reported in Table 5. The hypothesis statement, which asserts discrimination between organic conversation and coordinated operation, is not evaluable on the study set alone, since the entry condition guarantees the second term for every case in it. To bound that gap, four control corpora were prespecified before extraction and processed with the same rules and the same unmodified thresholds. All four are trending conversations directed at an identifiable actor and none of them is political, so they satisfy two of the three entry conditions and fail the third. The framework returns at least one operational label in three of the four, and Section 5.3 reports that outcome together with the mechanism behind each activation. Hypothesis H1 is therefore not established. What the control corpora measure is the behavior of the framework beyond its declared scope rather than its specificity within it, and an organic control that is political, trending and directed at an actor remains the test the framework has not undergone.

Hypothesis H2 requires, as prespecified in Section 3.2, that corpora carrying different label combinations occupy regions of the D1–D4 space whose separation exceeds what random assignment of the same labels to the same panels would produce. The correspondence between a panel and the label derived from it is not part of the test, since the assignment rules of Table 4 fix it by definition.

Each of the three modalities is activated in at least one study case, as reported under H1. Figure 4 displays the panel of each case with the resulting typological label annotated, as a reading aid and not as a result. Cases with Mass replication all exhibit at least one coordinated text detected by Equation (1) and high D1 and D2 dimensions; cases with Narrative saturation exhibit combinations of high D2 with high D3 or non-textual temporal coordination windows; cases with Targeted amplification exhibit significant presence of recurrent operators of the Sapiens Observatory and D1 above the operational threshold.

The observed empirical combinations are consistent with the combinatorial scheme documented in Section 4.1: four cases activate the three labels simultaneously, two cases activate two labels, and four cases activate a single label. That correspondence is not evidence for H2, because the assignment rules derive the label from the panel; the hypothesis asks the prior question of whether the panel positions themselves are structured by modality.

That question was tested by permutation. Each of the three labels defines two groups over the ten panels, presence and absence, and the separation between them is measured in the D1–D4 space by the pseudo-F statistic of a two-group PERMANOVA on Euclidean distances; the four dimensions are percentages on a common scale and are therefore not

standardized. Because the number of distinct assignments of the observed labels to the ten panels is small, the permutation is exhaustive rather than sampled and the resulting $p$-value is exact. Table 8 reports the outcome. No label separates the panel space beyond chance, and a second statistic computed as a control agrees: the leave-one-out accuracy of a nearest-neighbor rule in the same space reaches 0.40 for Mass replication against a base rate of 0.60, 0.70 for Narrative saturation and 0.90 for Targeted amplification, with permutation $p$-values of 0.719, 0.242 and 0.300. Recomputed over the nine non-calibrating corpora, the outcome is the same, with $p$-values of 0.405, 0.321 and 0.111.

**Table 8.** Exhaustive permutation test of the separability criterion of hypothesis H2 over the D1–D4 space. Each label defines two groups over the ten study panels, and pseudo-F is the two-group PERMANOVA statistic on Euclidean distances. Assignments is the number of distinct ways the observed number of labels can be distributed over the ten panels, so the permutation is exhaustive and the $p$-value exact. The minimum attainable $p$-value is the reciprocal of that number and bounds the significance the design can reach, irrespective of the data.

| Label | Cases Carrying It | Pseudo-F | Assignments | $p$ (Exact) | Minimum Attainable $p$ |
|---|---|---|---|---|---|
| Mass replication | 4 | 1.265 | 210 | 0.295 | 0.005 |
| Narrative saturation | 7 | 1.343 | 120 | 0.283 | 0.008 |
| Targeted amplification | 9 | 4.042 | 10 | 0.100 | 0.100 |

Two properties of the design bound what that result can be read to mean. Targeted amplification is present in nine of the ten cases, so only ten distinct assignments exist and the smallest attainable $p$-value is 0.100; the observed configuration is in fact the most extreme of those ten, and significance is unreachable by construction rather than by absence of signal. For Mass replication and Narrative saturation the attainable minimum is 0.005 and 0.008, so there the negative result is informative. The success criterion of H2 as prespecified is therefore not met, and the claim of empirical separability is withdrawn. What the set does establish is the differential activation of the three labels, which is a property of the sample rather than of the framework. What the framework retains is the operational content of those labels: two of the three require evidence that the panel does not contain, a coordinated text satisfying Equation (1) for Mass replication and a non-empty intersection with the Sapiens Observatory record for Targeted amplification, so the correspondence between panel and label is definitional without being tautological. Whether panel positions cluster by modality remains open, and ten cases are not enough to decide it.

The coordination that dimension D4 measures requires a statistical baseline of its own, and one was computed. A null model of temporal synchrony tests whether the alignment measured by D4 exceeds what the activity of the same accounts would produce by chance. Two nulls were used, each with five hundred permutations per case and a fixed seed. The first applies an independent random circular shift to every account timeline, which preserves the volume and the posting rhythm of each account while destroying the alignment between accounts. The second preserves the observed instants and permutes which account owns each of them. The first is the null that isolates synchrony; the second saturates in corpora of near-total polarity and is reported for completeness. Table 9 reports the outcome case by case.

**Table 9.** Null model of temporal synchrony over the ten study cases. Null A applies an independent random circular shift per account; null B permutes account ownership over the observed instants. Five hundred permutations per case and null, seed 20260820. A $p$-value of 0.002 is the minimum attainable with five hundred permutations and is reported as $p < 0.01$. Cases with D4 equal to zero admit no test and are marked with a dash. Column $n$ counts the mentions carrying a valid timestamp, which for MX/EV003 is eleven fewer than the corpus size reported in Table 5.

| Case | $n$ | Global Stance (%) | D4 Observed (%) | Null A Mean (%) | $p$ (A) | Null B Mean (%) | $p$ (B) |
|---|---|---|---|---|---|---|---|
| DO/EV001 Yayo Sanz | 31 | 100.0 | 100.0 | 100.0 | 1.000 | 100.0 | 1.000 |
| DO/EV002 Gloria Reyes | 919 | 27.7 | 10.0 | 0.0 | <0.01 | 0.0 | <0.01 |
| EC/EV001 Noboa–Uribe | 9708 | 76.4 | 42.0 | 67.4 | 1.000 | 57.7 | 1.000 |
| EC/EV002 J. J. Neira | 197 | 33.0 | 0.0 | 0.0 | — | 0.0 | — |
| EC/EV003 I. Vélez | 4062 | 99.9 | 33.6 | 0.0 | <0.01 | 43.2 | 1.000 |
| EC/EV004 C. A. Vera | 1031 | 85.5 | 0.0 | 0.0 | — | 1.9 | — |
| EC/EV005 Banco Pichincha | 32,381 | 64.0 | 27.9 | 3.0 | <0.01 | 1.1 | <0.01 |
| MX/EV001 A. Armenta | 5765 | 100.0 | 10.3 | 0.0 | <0.01 | 12.0 | 1.000 |
| MX/EV002 F. Noroña | 130 | 100.0 | 0.0 | 0.0 | — | 0.0 | — |
| MX/EV003 Clara Brugada | 2388 | 99.7 | 22.2 | 0.0 | <0.01 | 26.5 | 1.000 |

Under the first null, five of the seven cases with D4 above zero exceed the chance at $p < 0.01$: DO/EV002, EC/EV003, EC/EV005, MX/EV001 and MX/EV003. In four of them the null mean is exactly zero, so once deliberate alignment is removed the detector finds no qualifying window at all. This is the statistical evidence the submitted version did not report.

Two scope conditions follow, and both are stated as limitations. The first is that in corpora whose global stance concentration already exceeds the 0.80 threshold, the stance condition of D4 is satisfied by construction: six of the ten corpora are in that situation, and there the discriminating power of D4 comes entirely from the density criterion of thirty distinct accounts within fifteen minutes rather than from the stance criterion. The second is that EC/EV001, the largest and most cited case of the study, does not exceed the synchrony null: its observed D4 of 42.0 falls below a null mean of 67.4, because nine thousand seven hundred eight mentions concentrated around a single peak over forty hours produce more qualifying windows than the observed ones even after alignment between accounts is destroyed. Its Mass replication label does not change, since Equation (1) sustains it with three coordinated texts covering 61.7% of the corpus, but its Narrative saturation label rests on the coordination windows alone, because the alternative branch of Table 4 requires D3 above 50% and the case reaches 39.4%. The evidence of coordination in it is therefore textual replication and not temporal synchrony. The same reading applies to DO/EV001, whose only label is Narrative saturation and whose D4 exceeds neither null.

Section 4.1 reports that the specific thresholds of each dimension were verified through sensitivity analysis on one Ecuadorian corpus, EC/EV001 Noboa–Uribe, which also enters the validation set. Verifying thresholds and confirming the framework on shared evidence can inflate the apparent internal coherence of a design, so every aggregate result of the study was recomputed over the nine corpora that took no part in threshold verification. Rules, thresholds, and code paths are held at their nominal values; only the composition of the evidence set changes. Supplementary Material File S7 reports the full comparison, the mechanism analysis of D4, and the threshold sensitivity measured on corpora that

did not calibrate. The reduced set retains 82.9% of the mentions and 77.1% of the unique accounts, every aggregate result of Section 4 survives on it, and the check also identifies two boundaries of the framework: the stance-concentration condition of D4 governs the dimension only in high-volume corpora, and the typological label is no longer guaranteed to be invariant to threshold perturbation below one thousand mentions.

Hypothesis H3 requires the Sapiens Observatory to identify, upon closure of the analyzed set, at least one non-trivial set of accounts with the COORDINATED_OPERATOR or PERSISTENT profiles.

The empirical evidence supports the hypothesis, with magnitudes varying according to the size of the per-country accumulated corpus. In Ecuador, with five analyzed events, the Observatory identifies one thousand three hundred forty-five accounts with the COORDINATED_OPERATOR profile, that is, classified as Troll in at least two of the five country events. Of those, three hundred sixty-five are classified as Troll in three or more events, eighty-nine in four or more, and nine in all five; the strict intersection of the five Ecuadorian corpora contains fourteen accounts, which constitutes the most stable persistent core identifiable with the available evidence. In Mexico, with three analyzed events, the Observatory identifies twenty-seven accounts with the COORDINATED_OPERATOR profile; these are organized in disjoint sub-cores by narrative frame (ten accounts present in the ideological frames of Armenta and Noroña, seventeen accounts present in the frames of Armenta and Brugada, and no account present in the three events simultaneously). In the Dominican Republic, with two analyzed events, the Observatory identifies one account with the COORDINATED_OPERATOR profile present in both corpora (Figure 8).

The operational success criterion—at least one non-trivial set of accounts with COORDINATED_OPERATOR or PERSISTENT profile—is met in the three study countries. The power of the trans-event finding depends on the magnitude of the accumulated corpus: the Ecuadorian persistent core rests on five events and is robust; the Mexican and Dominican cores are sustained on three and two events, respectively, and are less robust by sampling construction. Hypothesis H3 is supported: in the three countries where successive events were accumulated, a subpopulation of reusable operational accounts emerges; the specific figures will be refined as new events are documented and processed with the same homogeneous methodology, according to the principle stated in Section 3.2. Table 10 summarizes the empirical contrast of the three hypotheses.

**Table 10.** Summary of the empirical contrast of the three hypotheses stated in Section 3.2. The Evidence column synthesizes the individual findings of Section 4.2.2 and the transversal findings of Section 4.2.3. The State column reports the qualitative conclusion of the subsection; the methodological caveats are developed in the body.

| Hypothesis | Brief Statement | Operational Success Criterion | Empirical Evidence | State |
|---|---|---|---|---|
| H1 | Discrimination between organic conversation and coordinated operation. | The four panel labels are differentially activated across the ten cases. | The three operational modalities are differentially activated (4 Mass replication, 7 Narrative saturation, 9 Targeted amplification); the fourth label, no detectable operation, is not activated because the entry condition of Section 3.2 excludes conversation without prior evidence of coordination. | Not established: not evaluable within scope. |

**Table 10.** *Cont.*

| Hypothesis | Brief Statement | Operational Success Criterion | Empirical Evidence | State |
|---|---|---|---|---|
| H2 | Empirical differentiation of operational modalities. | The separation between label groups in the D1–D4 space exceeds what random assignment of the same labels to the same panels would produce by exhaustive permutation. | Each of the three modalities is activated in at least one case, but no label separates the panel space beyond chance: $p = 0.295$, 0.283 and 0.100 (Table 8). | Not supported. |
| H3 | Industrial persistence of trans-event operators. | The Observatory identifies at least one non-trivial set of COORDINATED_OPERATOR or PERSISTENT accounts. | EC: 1345 COORDINATED_OPERATOR over 5 events, of which 365 are Troll in 3 or more events. MX: 27 COORDINATED_OPERATOR over 3 events. DO: 1 COORDINATED_OPERATOR over 2 events. | Supported. |

The balance of the empirical contrast is the following: hypothesis H2 is not supported, since the three modalities are differentially activated but their panel positions are not separated beyond chance; hypothesis H3 is supported in the three countries with magnitudes differentiated by the incremental coverage of the Sapiens Observatory; hypothesis H1 is not established, since the entry condition makes it not evaluable on the study set alone. Section 4.2.5 reports two additional tests on the framework through quantitative contrasts external to the DPI: a local reimplementation of Botometer and the Sentiment variable that commercial platforms automatically export, both described in Section 3.2.

#### 4.2.5. External Cross-Validation

Two external validity tests on the DPI framework complement the hypothesis contrast of Section 4.2.4. Both tests compare the pipeline output against independent references formulated prior to the DPI design. Section DPI vs. Botometer-Lite confronts the DPI Troll classification with the automated bot detection of Botometer-Lite, a local reimplementation of the public signals documented by the literature [10]. Section DPI Stance vs. Platform Sentiment contrasts the DPI Stance, a proprietary variable directional with respect to the target actor, with the Platform Sentiment, a variable exported by commercial monitoring tools without directionality with respect to the actor.

##### DPI vs. Botometer-Lite

For the ten study cases, a Botometer-Lite score was computed per unique account following the canonical pipeline formulation, calibrated with the recent literature on automated bot detection [8–10]. The score combines five structural signals available in any commercial platform extraction: alphanumeric pattern of the username, absence of verifiable social presence, retweet ratio over own content, abnormally high or low mention volume, and textual replication with other corpus accounts. The score produces a value in the zero-to-one range; following the canonical Botometer threshold, an account with a score equal to or above 0.5 is labeled as bot. The account-by-account comparison with the DPI Troll classification produces four disjoint and exhaustive subpopulations per case.

Over the aggregate set of the ten corpora, the two classifications coincide bot-to-bot on sixteen thousand three hundred twenty-one unique accounts and coincide non-bot-to-non-bot on four thousand three hundred fifty-nine accounts, yielding a raw concordance of 73.8% over $n = 28{,}007$ unique accounts and an aggregate Cohen kappa of 0.368. The standard

interpretation of this magnitude places the agreement between the two frameworks in a fair range [40]. The reciprocal coverage figures are asymmetric: 76.2% of accounts classified as Troll by the DPI also receive the bot label by Botometer-Lite, and 88.0% of accounts classified as bot by Botometer-Lite are simultaneously DPI Trolls.

The discordance concentrates in two quadrants that deserve separate reading. Five thousand one hundred ten accounts receive the DPI Troll label but not the bot label by Botometer-Lite. Inspection of this subpopulation reveals a consistent profile: accounts without verifiable social presence, with a retweet ratio below the structural threshold but with a concentrated aggressive lexicon and a clear directional stance toward the target actor, that is, the behavioral human troll profile captured by rule R3 of the framework. Two thousand two hundred seventeen accounts receive the bot label by Botometer-Lite but not the DPI Troll label. Their profile is complementary: usernames with extreme alphanumeric patterns, high retweet ratios, and null social presence, but without directional stance toward the target actor within the mention corpus. They are automated accounts belonging to another conversation or amplifying non-thematic content with respect to the actor.

The agreement varies among cases, with kappa values ranging between 0.18 and 0.69. Case-by-case inspection shows that agreement is maximum where the detected operation is predominantly sustained on structural signatures—rules R1, R2, R4, and R6 of the framework—and decreases in cases where the operation is sustained on behavioral signatures documented by rules R3 and R5. Table 11 disaggregates this distribution by case. One reading of the contrast holds that the DPI does not replicate Botometer but complements it: it incorporates behaviorally coordinated accounts that escape structural-signal automated detection and, simultaneously, discards automated accounts whose activity is not directionally oriented to the corpus actor.

**Table 11.** Account-by-account comparison between the DPI Troll classification and the Botometer-Lite Bot detection, disaggregated by case. The four categories are disjoint and exhaustive: Both (DPI Troll ∧ Botometer Bot), Only DPI (behavioral human troll without extreme structural signature), Only Botometer (automated bot without directional stance), and None (accounts that no framework flags). The per-case κ reports the Cohen agreement; the cross-country aggregate is fair (κ = 0.368).

| Case | *n* | Both (*n* %) | Only DPI (*n* %) | Only Botometer (*n* %) | None (*n* %) | κ |
|---|---|---|---|---|---|---|
| EC · Noboa–Uribe | 6422 | 4356 (67.8%) | 1207 (18.8%) | 456 (7.1%) | 403 (6.3%) | 0.184 |
| EC · JJ Neira | 184 | 63 (34.2%) | 2 (1.1%) | 86 (46.7%) | 33 (17.9%) | 0.191 |
| EC · Irene Vélez | 2288 | 1546 (67.6%) | 526 (23.0%) | 2 (0.1%) | 214 (9.4%) | 0.353 |
| EC · C. A. Vera | 821 | 496 (60.4%) | 107 (13.0%) | 130 (15.8%) | 88 (10.7%) | 0.234 |
| EC · Banco Pichincha | 11,904 | 5148 (43.2%) | 2393 (20.1%) | 1498 (12.6%) | 2865 (24.1%) | 0.325 |
| DO · Yayo Sanz | 31 | 1 (3.2%) | 1 (3.2%) | 0 (0.0%) | 29 (93.5%) | 0.652 |
| DO · Gloria Reyes | 398 | 103 (25.9%) | 14 (3.5%) | 41 (10.3%) | 240 (60.3%) | 0.688 |
| MX · Armenta | 3899 | 3116 (79.9%) | 580 (14.9%) | 0 (0.0%) | 203 (5.2%) | 0.359 |
| MX · F. Noroña | 113 | 11 (9.7%) | 16 (14.2%) | 0 (0.0%) | 86 (76.1%) | 0.511 |
| MX · Clara Brugada | 1947 | 1481 (76.1%) | 264 (13.6%) | 4 (0.2%) | 198 (10.2%) | 0.528 |
| Aggregated cross-country | 28,007 | 16,321 (58.3%) | 5110 (18.2%) | 2217 (7.9%) | 4359 (15.6%) | 0.368 |

Botometer-Lite is not the Botometer model. It is a five-feature heuristic reimplementation whose weights and cut points were fixed from the published description of the structural markers the original detector exploits, without supervised fitting on labeled accounts and without a performance evaluation of its own. Supplementary Material File S3 specifies the five features, their weights, their cut points and the decision threshold, and

documents three properties of the reference that bear on the reading of the agreement: the score is discrete and places 34.2% of the ten-corpus account set exactly on the decision boundary, the presence feature applies a follower-count criterion that the DPI itself rejects, and the reference has no measured accuracy against which the disagreement could be apportioned. A kappa of 0.368 obtained against a reference never validated on labeled accounts admits more than one reading. The complementarity reading stated above is consistent with the quadrant profiles and with the case-by-case gradient between structural and behavioral rules; the alternative reading, under which part of the disagreement records the arbitrariness of the reference rather than a substantive difference in what the two frameworks detect, is equally available on the present evidence. Separating them requires ground truth over labeled accounts, which this study does not have. The contrast is therefore reported as a descriptive characterization of two outputs, and none of the three hypotheses of Section 3.2 rests on it.

DPI Stance vs. Platform Sentiment

The second external test contrasts the Stance classified by the DPI—"Attacks" the actor, "Defends" the actor, "No stance"—against the Platform Sentiment exported by the monitoring tool, which takes one of the three values: positive, negative, or neutral. The comparison is performed mention by mention and aggregated cross-country over fifty-six thousand five hundred seventy-nine mentions. The null hypothesis tested is the operational equivalence between both variables, formalized as the diagonal correspondence "Attacks"–negative, "Defends"–positive, and "No stance"–neutral. The methodological caveat recalled in Section 3.1 states that Platform Sentiment is assigned to the text in abstract, without reference to the target actor of the analysis.

The aggregated contingency matrix (Figure 9) reports a diagonal concordance of 45.2% and a discordance of 54.8% over the total. The aggregate Cohen kappa is $-0.015$, a magnitude indistinguishable from the null value of equivalence under the hypothesis of chance. Of the ten individual cases, six present kappa less than or equal to zero—three of them exactly zero because the corpus contains a single Stance level, which makes the coefficient null by construction—and three present kappa between zero and 0.1; none of these nine exceeds 0.083. Case EC/EV002 JJ Neira ($n$ = 197 mentions) is the single exception, with kappa = 0.9886. The exception is not statistical degeneracy: the matrix occupies three cells, two of them diagonal, and the modal cell holds 66.5% of the mentions. It is a replication artifact. The corpus contains twenty-five unique texts, and two of them concentrate 82.2% of the mentions: a neutral administrative announcement that the DPI classifies as "No stance" and a critical text with negative sentiment that the DPI classifies as "Attacks". When a corpus reduces to two massively replicated texts whose Stance and Platform Sentiment happen to align, the agreement measured over mentions is near perfect and rests on what are effectively two independent observations.

The disagreement is not distributed uniformly over the text of the corpora. The thirty-one thousand thirty-two discordant mentions were profiled on eleven measurements computed directly on the mention: length in characters and in words, retweet condition, number of hashtags, number of third-party account handles, presence of a link, presence of the calibrated aggressive lexicon in Spanish that rule R3 uses [35], presence of praise lexicon, presence of laughter or scare-quote markers, presence of greeting, congratulation or gratitude formulas, presence of reportorial register markers, and the counts of actor-directed anti- and pro-patterns of the case lexicon that determine the Stance. On these profiles the discordance separates into eight types. Each mention receives exactly one type through a fixed precedence over the measurements, so that the eight types are mutually exclusive and exhaustive over the thirty-one thousand thirty-two mentions. Table S20

reports each type with its volume, its share of the discordance and of the corpus, and the measurements that distinguish it. The eight resulting types are mutually exclusive and exhaustive, and two families organize seven of them: divergences of anchoring, in which the affect of the text is not directed at the target actor, and divergences of attack vehicle, in which both variables point at the same object but the aggression does not travel in the lexical affect. Supplementary Material File S9 reports the full typology, worked examples of each type and the structural properties that qualify the proportions.

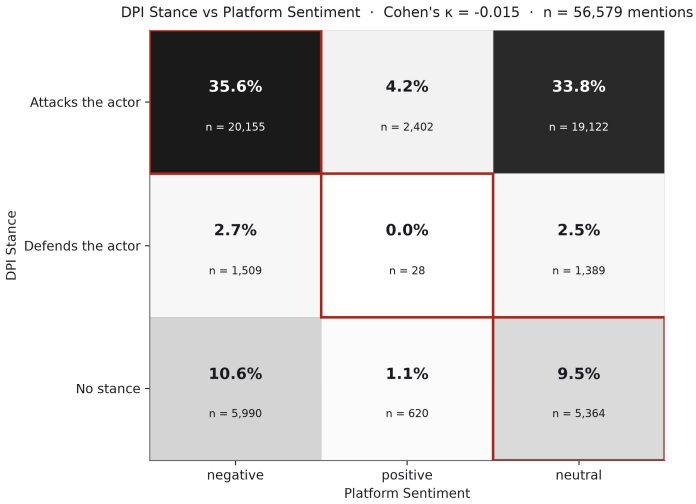



**Figure 9.** Aggregated cross-country contingency matrix between the DPI Stance (proprietary variable, directional toward the target actor) and the Platform Sentiment (variable exported by the commercial monitoring tool, without directionality with respect to the actor). Each cell reports the percentage over the total mentions and the absolute count. Cell shading encodes the percentage of mentions, from light for the lowest values to dark for the highest. The red diagonal marks the expected concordance under the null hypothesis of equivalence between both variables. The aggregate agreement is null (Cohen $\kappa = -0.015$ over $n = 56{,}579$ mentions), which documents that the Platform Sentiment is not a valid proxy for the author's Stance with respect to the actor when the analysis objective is to identify targeted political polarization.

The mechanism of ironic courtesy framing admits a direct empirical test within a single corpus. In case MX/EV001, the DPI assigns "Attacks" to all five thousand seven hundred sixty-five mentions, so the Stance is constant and every variation in Platform Sentiment is attributable to the text. One thousand six hundred sixty-nine of those mentions contain a greeting, congratulation, or gratitude formula, and 97.6% of them receive positive sentiment, against 6.8% of the four thousand ninety-six mentions without such a formula, an odds ratio of 555.0. The corpus is a single insult-hashtag campaign against one actor; the formulas that precede the insult—a morning greeting, an invitation to celebrate the day—are what the classifier scores. Because the corpus is massively replicated, the same contrast was recomputed over unique texts: fifty of the nine hundred forty-three unique texts carry a courtesy formula, and 74.0% of them have most of their mentions scored positive, against 7.7% of the texts without a formula, an odds ratio of 34.0. The effect shrinks by an order of magnitude when replication is removed, and it remains large and in the same direction. The finding isolates the mechanism: the platform variable measures the affective register of the surface formula, not the direction of the act.

The central asymmetry of the result is that Platform Sentiment is assigned to the text in abstract, without reference to the target actor of the analysis, whereas the DPI Stance is assigned by explicitly interpreting the role of the actor within the mention. This methodological caveat, described in Section 3.1 and documented in the glossary, translates empirically into the divergence reported by Figure 9. The operational consequence is

that using Platform Sentiment as a proxy for targeted political polarization produces systematic errors of the order of 55% of mentions, errors that are not randomly distributed but concentrate in non-diagonal cells with magnitude comparable to the diagonal cells.

The operational consequence sharpens the balance stated above. Using Platform Sentiment as a proxy for targeted political polarization does not merely produce errors on the order of 55% of mentions: it produces errors of at least three distinct kinds that a single correction cannot repair. Slightly more than seven tenths of the divergence would require a classifier able to score aggression carried by hashtag tokens, by coined labels, by images, by reportorial framing, and by ironic courtesy—five vehicles that dominate the attack repertoire of the corpora analyzed here and that lexical polarity scoring does not reach. A further 8.4% is definitional, and no improvement of the sentiment classifier would remove it, because the platform variable has no target actor and cannot distinguish affect directed at the actor from affect directed at anything else in the same text: this covers adversary redirection and the strictly lateral mentions of the first family. The remaining 19.5% measures the opposite thing, the coverage of the case lexicon on which the DPI Stance rests, and is the quantified form of the classifier caveat declared in Section 5.3. The finding supports the methodological decision of the DPI to operate with a proprietary variable directional toward the actor, and it locates the reason for that decision in the structure of the discordance rather than in its magnitude alone; it also locates, with the same instrument, the direction in which the DPI Stance itself must be strengthened.

#### 4.2.6. Matched Within-Corpus Control

The control condition reported here is internal to the conversation: it contrasts the accounts the rules classify as Trolls against matched accounts of the same corpus, and it therefore operates entirely within the entry condition stated in Section 3.2. A second control condition applies the framework to four corpora that fall outside that entry condition. Because those corpora are not the object the framework is defined over, their outcome is not a validation of it and is reported in Section 5.3, with its tables in Supplementary Material File S6.

Rather than contrasting one corpus against another, a comparison that confounds coordination with topic, target, language and collection window, the accounts that rules R1 to R6 classify as Trolls are contrasted against an equally sized random sample of the accounts of the same corpus that the rules do not so classify. Everything that separates one conversation from another is held fixed by construction, and the only surviving contrast is the classification itself. Supplementary Material File S10 reports the sampling design, the circularity audit and the full outcome in Table S21.

The hierarchy itself was audited in the same spirit, since applying the rules in priority order can suppress a classification that a later rule would have made. It has a single such point: rule R2 sends an account with a moderate alphanumeric handle and no verifiable presence to the Neutral category when its stance is undefined, and rules R3 to R6 are then never evaluated on it. Over the ten study corpora, 549 accounts are cut at that point, and 90 of them satisfy one of the later rules. Restoring all ninety raises D1 by at most 0.7 percentage points in EC/EV005 Banco Pichincha, and by exactly zero in eight of the ten corpora, and it alters no typological label. Under non-exclusive assignment, in which every rule an account satisfies is recorded, the Troll population grows by those same ninety accounts and accounts satisfy between 1.00 and 1.54 rules on average, with a maximum of four. The hierarchy therefore costs little coverage and buys the property the panel depends on, namely that the six rule counts partition the Troll population and can be read as a composition without double counting, which is how Table 6 reports them.

Two conclusions follow and should not be conflated. What survives the circularity audit is narrow and real: inside a single conversation, holding topic and target constant, the accounts the rules label as Trolls are also less followed and more often attached to text replicated across accounts, on variables no rule reads, and the follower deficit holds even for the subset defined by handle morphology with no behavioral rule satisfied. What does not survive is any claim that the rules isolate a discrete population. The distributions overlap heavily, most of the contrast on rule inputs is definitional, and the split behaves in an organic sports conversation much as it does in a suspected operation. The Troll label must therefore be read as a behavioral stratum with an elevated prior probability of inauthenticity, not as an account-level determination of intent, and the case-level inference must rest on the four-dimension panel rather than on the classification alone. Whether that panel in turn separates the control conversation from the case corpora is the question the event-level comparison takes up.

## 5. Discussion and Limitations

This section synthesizes the findings reported in Section 4 in light of the previous literature, discusses the contribution that the DPI framework makes against the external references with which it was contrasted, and enumerates the limitations identified in the current study design. The synthesis introduces no new evidence: it groups the results of Section 4.2 in terms of methodological contribution, places them in dialog with the notion of reused digital infrastructure documented by the literature, and characterizes them with respect to automated bot detectors and Platform Sentiment. The implications for future research and professional practice are developed in Section 5.4 as the closing of the discussion and do not anticipate Section 6 of conclusions.

### *5.1. Synthesis of Findings and Framework Contribution*

The hypothesis contrast presented in Section 4.2.4 establishes that the DPI differentially activates the three non-exclusive typological modalities—Mass replication, Narrative saturation, and Targeted amplification—over the ten study cases (the separability criterion of H2 is not met, and that claim is withdrawn), with per-case combinations varying between one and three simultaneous labels according to the combinatorial scheme documented in Section 4.1 (Table 5 and Figure 4). Hypothesis H3 on the industrial persistence of trans-event operators is supported in the three countries where the Sapiens Observatory accumulates at least two successive events, with the Ecuadorian coverage documenting the most stable persistent core identifiable with the available evidence. Hypothesis H1 on the discrimination between organic conversation and coordinated operation is not established: the entry condition guarantees the coordinated term for every case of the study set, so the hypothesis is not evaluable on it.

The external validity tests of Section 4.2.5 characterize the framework contribution against pre-established references. The account-by-account contrast with Botometer-Lite over $n$ = 28,007 unique accounts yields an aggregate Cohen kappa of 0.368, a fair magnitude on the Landis and Koch scale [40]. The coverage asymmetry is informative: eighty-eight percent of the automatic detector bots are simultaneously DPI Trolls, while seventy-six point two percent of the DPI Trolls also receive the bot label. The five thousand one hundred ten accounts that the DPI flags but the automatic detector does not, correspond to the behavioral human troll profile captured by rule R3 of the framework; the two thousand two hundred seventeen accounts that the automatic detector flags but the DPI does not, correspond to bots without directional stance toward the target actor (Table 11). What the contrast establishes is that the DPI output is not a relabeling of a structural bot detector. The reading under which it complements automated detection in the direction

of targeted conversational behavior, a direction the literature signals as the central void of traditional bot detectors [8,9], is consistent with the quadrant profiles, and Section DPI vs. Botometer-Lite states why the present evidence does not establish it over the alternative.

The contrast between DPI Stance and Platform Sentiment over $n$ = 56,579 mentions produces an aggregate Cohen kappa of $-0.015$, a magnitude indistinguishable from the null agreement and consistent across the ten study cases. Nineteen thousand one hundred twenty-two mentions with "Attacks" stance receive neutral sentiment from the textual classifier, two thousand four hundred two receive positive sentiment, and one thousand five hundred nine mentions with "Defends" stance receive negative sentiment (Figure 9). The balance of the contrast empirically validates the methodological decision of the DPI to operate with a proprietary variable directional toward the actor instead of inheriting Platform Sentiment. The framework contribution is synthesized in four elements: a stable operational typology of non-exclusive modalities, a panel of four independent dimensions reported jointly and never aggregated into a scalar, a directional Stance variable that Platform Sentiment does not capture, and a trans-event accumulation structure that documents the notion of reused digital infrastructure in country key.

### *5.2. Dialog with the Literature*

The transversal findings of the Sapiens Observatory (Section 4.2.3 and Figure 8) operationalize the notion of reused digital infrastructure that Bradshaw, Bailey, and Howard [2] and DiResta et al. [20] document as a central trait of contemporary coordinated operations. The PERSISTENT and COORDINATED_OPERATOR profiles defined by the framework are the operational formalization of that notion on direct empirical evidence: the set of accounts reidentifiable across successive country events constitutes the statistical footprint of a preexisting infrastructure that is being reused, not the random residue of independent interactions. The quantified magnitude of the Ecuadorian persistent core (fourteen accounts present in the five corpora, three hundred sixty-five accounts classified as Troll in at least three) is consistent with the order of magnitude reported by the industrial studies of the field [5,19].

The set of rules R1 to R6 reorganizes previous detection paradigms. Rule R1 on extreme alphanumeric pattern and rule R2 on moderate pattern without verifiable institutional presence operate on the most rudimentary structural signature documented by Cresci et al. [38], which is the footprint of first-generation social spambots; the paradigm shift described by those authors toward behaviorally sophisticated operators is covered by rules R3, R5, and R6 on directional aggressive lexicon, coordinated textual replication, and pure amplifiers with clear stance. Rule R5 picks up the operationalization of coordinated textual replication of Ratkiewicz et al. [1] and Pacheco et al. [13]. Rule R6 captures the narrative-saturation profile of pure amplifiers that operate once and disappear. Rule R3 incorporates the aggressive lexicon of the HurtLex resource in Spanish [35] as a behavioral signature when it concurs with a clear directional stance and absence of verifiable social presence.

The debate on the distinction between bots and trolls [12] receives in the DPI an operational resolution: the framework does not classify the permanent nature of the account but the behavior observed within the corpus of an event. An account can behave as a troll in one event and not in another, which is consistent with the event-bounded nature of the coordinated operations studied by Pacheco et al. [13]. This operational choice is consistent with the critique of Rauchfleisch and Kaiser [17] on the over-attribution of botness derived from the indiscriminate use of automated detectors on polarized human accounts. The local reimplementation of Botometer-Lite used in Section DPI vs. Botometer-Lite follows the public Botometer formulation of Varol et al. [10] on the structural signals available for reproducible detectors, and the contrast documented in Table 11 isolates five thousand one

hundred ten accounts that the DPI labels and the structural detector do not, a quadrant whose profile and whose two possible readings Section DPI vs. Botometer-Lite reports.

The interpretation of the aggregate Cohen kappa of the external cross-validation follows the canonical conventions of Landis and Koch [40] on agreement levels (negative, slight, fair, moderate, substantial, almost perfect). On this scale, the DPI ↔ Botometer-Lite agreement is fair, and the Stance ↔ Sentiment agreement is indistinguishable from null agreement. Both results are consistent with the methodological notion of complementarity between dimensions: distinct frameworks detect distinct subpopulations. The reading of complementarity between frameworks explicitly responds to the void signaled by the recent literature on the need for integral frameworks that combine structural and behavioral signatures without aggregating the two planes into a single metric [5,8].

### *5.3. Scope, Control Corpora and Limitations*

Four organic control corpora were prespecified before extraction, each with its selection rule, extraction query, time window and expected outcome registered in advance, and each processed with the same pipeline, the same rules R1 to R6 and the same unmodified thresholds as the ten study cases. All four trended on X and are directed at an identifiable actor, and none of them is political: the platform classifies trending topics in sport and entertainment as well as in politics, and these four trended in the former. They therefore meet two of the three entry conditions and fail the third, which is why their outcome is reported here as a bound on the applicability of the framework rather than in Section 4 as a validation of it. They were not written into the Sapiens Observatory, so the accumulated record against which the Targeted amplification condition is evaluated is identical to the one used for the study cases. Tables S15–S17 of Supplementary Material File S6 report the figures.

Three of the four control corpora receive at least one operational label, and one of them receives the same three-label combination as the four strongest study cases. None of the four dimensions separates the two groups: for D1, D2, D3 and D4 alike the control range lies inside the study range, and two study cases return a lower D1 than three of the four controls. The single control that receives no label does so because the accumulated national record against which the Targeted amplification condition is evaluated contains one account, which is a property of the country rather than of the conversation analyzed.

The replicated texts that trigger Equation (1) in the two affected controls are, in order of volume, the target actor's own post, a widely followed sports commentator, an established sports broadcaster, and the official communiqué of a football club. Equation (1) reads volume velocity and burst and never reads who wrote the text that is being replicated. Mass replication of a message issued by the protagonist of the event therefore satisfies it.

The morphology the framework measures is a single actor concentrating emission and a periphery of low-presence accounts replicating it, and it is present in all fourteen corpora. Six of the seven measures reported in Table S17 overlap between the two groups, and the overlap holds when the two study corpora too small for the measures to be stable are excluded. What separates the groups is position rather than form: the study corpora sit consistently at the extreme of the same distribution. This is the continuum that Nizzoli et al. [36] describe, observed within a single instrument, and it has a direct consequence for the framework: where two populations differ in degree along overlapping distributions, any fixed threshold produces false positives near the cut. Two of the results reported here sit at that boundary, one control failing the volume condition of Equation (1) by two mentions and another exceeding the stance concentration threshold by four hundredths.

The evidence just reported establishes four limits on what the DPI can be used to claim, and they are stated here without qualification.

**Attribution.** The framework has no positive verdict for organic conversation. Its only negative outcome, absence of detectable operation, is a residual state reached when no label condition is met, and it is not evidence of authenticity. Applied outside its entry condition, the framework returns at least one operational label in three of four corpora in which no operation exists. The panel and the typological labels therefore describe the structure of a conversation and the inauthenticity markers it carries; they do not establish that an operator exists behind it, and any operational deployment intended to make that attribution requires evidence the framework does not supply.

**Origin.** Equation (1) reads volume, velocity and burst, and imposes no condition on the origin of the replicated text. Mass replication of a post issued by the target actor, by an official institutional account or by an established media outlet satisfies it. The condition is therefore a measure of replication intensity and not of the provenance of what is replicated.

**Lexical coverage.** The Narrative saturation condition requires directional stance to reach eighty percent of all mentions within a fifteen-minute window, so mentions carrying no coded stance enter the denominator. The label is bounded by the lexical coverage of the case lexicon, which is curated by the analyst. On the cycling control, three lexicons of increasing coverage over an identical corpus, processed with identical code and identical thresholds, return D4 values of zero, five point six and eighty-eight point two percent, while the number of windows meeting the density condition remains unchanged at eight thousand six hundred fifty-four. Two consequences follow. The label reports a property of the lexicon at least as much as a property of the conversation, and the earlier observation that no corpus fell between the observed stance concentration values reflects limited lexical coverage in the control corpora rather than threshold robustness. For the same reason, the stance variable and the entry condition share signal in corpora defined by a hostile trending term: Supplementary Material File S5 reports, for each study case, the share of mentions whose stance is decided by the most load-bearing lexical pattern alone, which ranges from zero to eighty-eight percent and reaches forty-five percent in the largest corpus of the set.

**Absolute thresholds on corpora of unequal size.** The density condition of the Narrative saturation label is an absolute count of thirty distinct accounts within a fifteen-minute window rather than a proportion of the corpus. A temporally compact corpus of thirty-one mentions satisfies it, and the smallest corpus in the study set returns the maximum value of D4 for that reason.

The present study has further limitations. The set of analyzed cases is temporal: the ten corpora documented in Section 4.2.2 cover three countries, and new cases will be accumulated as the Sapiens Observatory work progresses, so the aggregate figures presented in Section 4 are a snapshot at the close of the study and not a census of the phenomenon. The cases also meet the entry condition stated in Section 3.2, that is, political conversation directed at an identifiable target actor and trending on X, and among the conversations meeting it, the selection favored events already under suspicion. Both facts bound the population the reported figures describe.

A third limitation is that the operational thresholds of the coordination window were verified on a corpus that also belongs to the validation set. The leave-one-out check reported in Section 4.2.4 and Supplementary Material File S7 sustains the differential activation of the three modalities, the amplitude of the four dimensions, and both external contrasts on the nine non-calibrating corpora, and identifies two boundaries of the framework: the stance-concentration condition of D4 governs the dimension only in high-volume corpora, and the typological label, although not the numerical panel, is no longer guaranteed to be invariant to threshold perturbation in corpora below one thousand mentions.

The cross-country coverage is asymmetric by sampling construction: five events in Ecuador, three in Mexico, and two in the Dominican Republic. The trans-event findings

of Section 4.2.3 are, consequently, more robust for the first country than for the other two, where the magnitude of the identified operational cores is sustained on fewer events by construction. The incremental accumulation of the Observatory in the three countries—and eventually in others—will reduce this asymmetry with the incorporation of new corpora. Additionally, two of the cases present small corpora (DO/EV001 Yayo Sanz with thirty-one unique accounts and MX/EV002 F. Noroña with one hundred thirteen unique accounts), where the external cross-validation statistics of Section 4.2.5 have greater variability due to reduced sample coverage. This variability is reported case by case and does not compromise the stability of the aggregate figures, but it is worth keeping in mind when examining the individual figures.

That asymmetry makes it necessary to state explicitly how much weight each reported result can carry, and Table 12 grades them. A result is called robust when it does not depend on the number of accumulated events and survives the leave-one-out check of Supplementary Material File S7; intermediate when the evidence reported here supports it but it rests on a national record or on an instrument that has not been independently validated; and exploratory when it is decided by a corpus or a registry too small for the quantity to be stable. The grade is a property of the evidence and not of the framework, and a result may move between grades as the Sapiens Observatory accumulates events.

The DPI Stance variable is inferred by automated text classification against the target actor and retains the error margin of any inferred variable. Its agreement with human annotation was measured directly. Two annotators independently coded a stratified sample drawn from the study corpora, and agreement between them reaches a Cohen kappa of 0.7524, with a ninety-five percent confidence interval of 0.650 to 0.842, which is substantial. The agreement between the classifier and each annotator is lower, moderate at 0.4708 against the first and fair at 0.3711 against the second, and moderate at 0.4108 once reweighted. The weakest class is “No stance”, where agreement falls to between 0.25 and 0.33: the framework attributes a direction to texts that a human reader takes as neutral. Since the rules that require a clear stance decide 77.3% of the Troll labels of the study set, that error propagates to the panel. The task is therefore well defined for human readers, while the instrument does not follow them closely, which bounds the confidence any single account-level label can carry. The HurtLex resource in Spanish used by rule R3 is likewise calibrated but not exhaustive.

The design of that measurement is declared in full, because the coefficient should be read against it. The sample was drawn from the 8377 unique texts of the study corpora and stratified by predicted class rather than in proportion to the corpus, which is why the balanced coefficient and the reweighted one are both reported. The two annotators coded a block of 120 texts, of which 119 entered the computation after one text that both marked undecidable. Neither annotator is an author of this study and neither had access to the case lexicons. The confidence intervals are bootstrap intervals over two thousand resamples of the annotated pairs. Two facts are recorded against the exercise: the two annotators flagged 2 and 21 texts, respectively, as doubtful, an asymmetry that is itself evidence that they applied independent criteria, and both received the sample in the same order, so the design does not protect against correlated fatigue.

The Platform Sentiment variable used in Section DPI Stance vs. Platform Sentiment comes from the commercial monitoring tool with which the corpus was extracted, a tool that is not identified by name in the present work. This option has an explicit methodological consequence: exact reproducibility of the contrast of Figure 9 with a different tool will require inspecting the particular definition of positive, negative, and neutral that each provider implements. The qualitative finding—that Platform Sentiment is not a valid proxy for directional Stance—holds robustly regardless of the underlying classifier, due to the

conceptual asymmetry between a variable that looks at the text in abstract and a variable that looks at the role of the target actor within the mention, but the exact quantitative figures are specific to the corpus extracted with the tool used in this study.

**Table 12.** Evidential status of the results reported in this article, graded by the coverage that sustains each one. Robust, intermediate and exploratory are defined in the paragraph above. The last column states the evidence and the bound in the same line, so that no grade is asserted without the quantity behind it.

| Result | Status | Evidence and Bound |
|---|---|---|
| Numerical panel D1 to D4 per corpus | Robust | Computed from proportions and maxima of the observed corpus, so it does not depend on the number of accumulated events; the independent reimplementation reproduces it exactly at nominal thresholds over the eleven corpora of the sweep (Table S8). |
| Typological label, corpora above one thousand mentions | Robust | Invariant across both perturbation grids of the D4 thresholds over the nine non-calibrating corpora (Supplementary Material File S7). |
| Typological label, corpora below one thousand mentions | Exploratory | Two of the four corpora of that size change label when the unique-account threshold moves to thirty-two, and the other two are invariant only because D4 is null in every configuration tested. |
| Separability of the panel positions by modality, H2 | Not supported | Exhaustive permutation over the ten panels returns $p = 0.295$, 0.283 and 0.100 for the three labels, and the same outcome over the nine non-calibrating corpora; the differential activation of the labels does hold on both sets. |
| Discrimination between organic and coordinated conversation, H1 | Not established | The entry condition guarantees coordination for every study case, and the four control corpora fall outside that condition, so the false-positive rate over political trending conversation remains unmeasured. |
| Trans-event persistence of operators, Ecuador | Robust | Five accumulated events and an operational core of 1971 accounts, of which 1414 retain the profile when the calibration corpus is withdrawn. |
| Trans-event persistence of operators, Mexico | Intermediate | Three accumulated events and twenty-seven coordinated operators, enough to identify sub-cores by narrative frame but not to estimate their stability over time. |
| Trans-event persistence of operators, Dominican Republic | Exploratory | Two accumulated events, and a single account reaches the coordinated-operator criterion, so the Targeted amplification condition is decided by a registry of size one. |
| Targeted amplification label | Intermediate | Activation requires a non-empty intersection with the national accumulated record, whose size differs by three orders of magnitude between Ecuador and the Dominican Republic. |
| Trans-country continuity of operators | Exploratory | One account appears in events of two countries in the full set and none in the reduced set, so the negative observation rests on a single case. |
| Complementarity with Botometer-Lite | Intermediate | Fair aggregate agreement, Cohen kappa 0.368, against a local reimplementation that was neither trained nor statistically calibrated (Supplementary Material File S3). |
| Null agreement between DPI Stance and Platform Sentiment | Robust | Cohen's kappa of $-0.015$ over 56,579 mentions of the three countries, and $-0.027$ over the nine non-calibrating corpora. |
| Agreement of the DPI Stance with human annotation | Intermediate | Cohen's kappa of 0.7524 between two independent annotators and 0.4108 reweighted for the classifier, over one stratified sample; the No stance class falls to between 0.25 and 0.33. |
| Per-case cross-validation statistics for DO/EV001 and MX/EV002 | Exploratory | Thirty-one and one hundred thirteen unique accounts, respectively, so the per-case statistics carry high variability and only the aggregate figures are stable. |

Finally, the DPI framework is formulated on the signals available on platform X (Twitter). Rules R1 to R6 presuppose variables such as username, retweet ratio over own content, verifiable institutional presence, and textual replication among accounts, which have direct analogs on that platform. The portability of the framework to other platforms—Facebook,

TikTok, Instagram, and messaging platforms with public space—requires adaptation of the rules to the specific features of each platform and to the coordination detection mechanisms, without the general framework architecture being compromised by the change in medium. That adaptation belongs to the extension agenda developed in Section 5.4.

### *5.4. Future Work*

The immediate extension of the research program follows from the limits stated in Section 5.3. The first line is an organic control that satisfies the entry condition, that is, a political conversation directed at an actor and trending on X but without prior evidence of coordination; the four control corpora reported in Section 5.3 fall outside the entry condition on the political attribute and therefore bound the applicability of the framework rather than its specificity within scope.

Three extensions follow directly from the limits above and are stated as the agenda this evidence opens. The first is an origin condition on Equation (1), restricting the Mass replication label to replicated texts that do not originate from the target actor, from an institutional account or from an established media outlet. The second is the systematic reporting of lexical coverage alongside every stance derived dimension, with a coverage floor below which the Narrative saturation label is withheld rather than returned as zero. The third is a density criterion expressed as a proportion of the corpus rather than as an absolute count of accounts. None of the three was applied to the results reported here, which use the thresholds as originally specified.

A fifth line answers the question left open in Section 2.4, on whether operational accounts produce polarization or exploit a polarization already present. Nothing in the present design can decide it: the ten corpora are event-bounded observational samples without pre-event baselines and without exogenous variation to exploit. Testing it would require a different design, in which the conversation about a target actor is observed before, during and after the event, and in which some source of variation independent of the operation itself separates the two directions. That is a longitudinal design rather than an extension of this instrument, and the framework contributes to it the measurement of the operational footprint rather than the identification of its effect.

The implications for the professional practice of digital operations analysis are immediate. The DPI delivers a per-event replicable operational panel that distinguishes between the three framework modalities and reports the country's persistent core on trans-event evidence, which allows those engaged in political consulting, electoral monitoring, or communication audits to have a common framework with public operational definitions and thresholds calibrated on empirical evidence. The complementarity with automated bot detectors documented in Section DPI vs. Botometer-Lite suggests a combined workflow where the DPI provides the directional behavioral reading and structural detectors provide coverage on automated accounts without directional stance toward the target actor. The integration of a complementary Public Opinion Analysis module on the same corpora—documented in the project glossary and methodological reasoning—is part of the framework extension agenda and will be reported in subsequent program publications.

## 6. Conclusions

This article proposes the Digital Polarization Index (DPI), a non-binary, multidimensional, and behavioral classification framework to audit polarization operations on the X-Twitter platform. The framework distinguishes itself from automated bot detectors—explicitly contrasted with a local reimplementation of Botometer-Lite—and from the Platform Sentiment exported by commercial monitoring tools. The central difference is operational: the DPI does not classify the permanent nature of an account but its behavior within

the corpus of an event, and operationalizes the troll category through six public behavioral rules (R1–R6) that combine structural signatures of the username with behavioral markers of replication, amplification, and aggressive lexicon directional toward the target actor of the analysis.

The central findings establish that the three operational modalities of the framework are differentially activated within its stated scope. Mass replication, Narrative saturation, and Targeted amplification occur across the ten study cases in combinations that the assignment rules derive from the panel. The stronger claim that the panel positions themselves separate by modality is not sustained: an exhaustive permutation test over the ten cases returns no separation beyond chance, and that claim is withdrawn. What is established over conversation that meets the entry condition of Section 3.2 is over four trending but non-political corpora; the labels activate in three of four cases, so the panel and the labels describe the structure of a conversation and the inauthenticity markers it carries without establishing on their own that an operator stands behind it.

The two external contrasts delineate the specific contribution of the framework against the pre-established references of the field, and Section 5.1 reports their magnitudes. The comparison with Botometer-Lite is asymmetric in an informative direction: the accounts the DPI labels without the automatic structural signature detecting them correspond to the behavioral human troll profile captured by rule R3. The comparison with Platform Sentiment returns agreement indistinguishable from null, which documents empirically that an affect variable assigned to the text in abstract is not a valid proxy for polarization directed at a target actor and sustains the decision to operate with a directional variable of the framework.

The implications of the framework are three. For academic research, the DPI offers an operationalizable framework on corpora from platform X, with public rules, public thresholds and a deterministic pipeline, capable of functioning in the post-API ecosystem that invalidated the previous engagement-metric-based models. Its reproducibility is qualified in two respects stated in Section 5.3: The pipeline code is available from the authors on request and is not yet deposited in a public repository, and the Stance variable on which the rules depend agrees only moderately with independent human annotation. For the platforms, the distinction between situational operational behavior and permanent automated nature enables moderation policies that act on the operation and not on the account as a stable entity. For public policy and the regulation of digital campaigns, the trans-event documentation of reused digital infrastructure—operationalized in the PERSISTENT and COORDINATED_OPERATOR profiles of the Sapiens Observatory—constitutes a tool to identify and characterize coordinated campaigns beyond the individual inspection of accounts.

Five lines of future work complete the program and are stated in Section 5.4.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/informatics13100156/s1, File S1 Glossary; File S2. Operational thresholds [41]; File S3. Botometer-Lite specification; File S4. Sensitivity analysis of the operational thresholds; File S5. Lexical pattern load per study case; File S6. Organic control corpora; File S7. Robustness with respect to the calibration corpus; File S8. Per-country dimensional panels; File S9. Divergence between the DPI Stance and the Platform Sentiment; File S10. Matched within-corpus control [42,43]; File S11. Stance classifier specification.

**Author Contributions:** Conceptualization, J.S.-E. and J.C.-C.; methodology, J.S.-E.; software, J.S.-E.; validation, J.S.-E. and J.C.-C.; formal analysis, J.S.-E.; investigation, J.S.-E.; resources, J.S.-E.; data curation, J.S.-E.; writing—original draft, J.S.-E.; writing—review and editing, J.S.-E. and J.C.-C.; visualization, J.S.-E.; supervision, J.C.-C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable. The study analyzed publicly available social media data and did not involve interventions with human participants or animals.

**Informed Consent Statement:** Not applicable. No human participants were recruited; only publicly available social media content was analyzed.

**Data Availability Statement:** The raw data contain partial account information and are subject to platform X policies. The processed corpora (per-account classifications and aggregated D1–D4 dimensions) can be shared under reasonable request to the corresponding author. The pipeline code that implements the rules, the dimensions and the thresholds is likewise available from the corresponding author on request. The classification resources on which the reported analyses depend are available on the same terms: the per-case stance lexicons, the derivation of the aggressive lexicon from its published source, the annotation package with the two returned codings, and the lexical lists and the precedence order that define the divergence types of Supplementary Material File S9.

**Acknowledgments:** The authors thank the constructive comments of colleagues and anonymous reviewers who contributed to the development of the present work. During the preparation of this manuscript, the authors used Brandwatch IRIS (Brandwatch, Brighton, UK) for exploratory narrative identification, GitHub Copilot (GitHub, Inc., San Francisco, CA, USA) for code debugging and optimization, and Google Translate (Google LLC, Mountain View, CA, USA) for translation assistance with technical passages. The authors have reviewed and edited all resulting outputs and take full responsibility for the content of this publication.

**Conflicts of Interest:** Jorge Sánchez-Erazo is the founder and director of Sánchez Erazo Consulting, a political communication consultancy, and conducted this research in that capacity. The consultancy received no funding for this study, and no client of the consultancy commissioned it or had any role in the design of the study; in the collection, analysis or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results. None of the ten cases analysed was selected by the authors: all correspond to conversations that had reached trending status on X, which is the entry condition stated in Section 3.1. Jorge Cruz-Cárdenas declares no conflict of interest.

## References

1. Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; Menczer, F. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April 2011*; Association for Computing Machinery: New York, NY, USA, 2011; pp. 249–252. [CrossRef]
2. Bradshaw, S.; Bailey, H.; Howard, P.N. *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*; Computational Propaganda Research Project; Oxford Internet Institute: Oxford, UK, 2021.
3. Ng, L.H.X.; Carley, K.M. BotBuster: Multi-platform bot detection using a mixture of experts. *Proc. Int. AAAI Conf. Web Soc. Media* **2023**, *17*, 686–697. [CrossRef]
4. Schoch, D.; Keller, F.B.; Stier, S.; Yang, J. Coordination patterns reveal online political astroturfing across the world. *Sci. Rep.* **2022**, *12*, 4572. [CrossRef] [PubMed]
5. Goldstein, J.A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; Sedova, K. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*; Stanford Internet Observatory: Stanford, CA, USA, 2023.
6. Stella, M.; Ferrara, E.; De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 12435–12440. [CrossRef] [PubMed]
7. Alizadeh, M.; Shapiro, J.N.; Buntain, C.; Tucker, J.A. Content-based features predict social media influence operations. *Sci. Adv.* **2020**, *6*, eabb5824. [CrossRef] [PubMed]
8. Cresci, S. A decade of social bot detection. *Commun. ACM* **2020**, *63*, 72–83. [CrossRef] [PubMed]
9. Yang, K.-C.; Menczer, F. Anatomy of an AI-powered malicious social botnet. *J. Quant. Descr. Digit. Media* **2024**, *4*, 1–36. [CrossRef]
10. Varol, O.; Ferrara, E.; Davis, C.A.; Menczer, F.; Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. *Proc. Int. AAAI Conf. Web Soc. Media* **2017**, *11*, 280–289. [CrossRef]
11. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. [CrossRef] [PubMed]
12. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The rise of social bots. *Commun. ACM* **2016**, *59*, 96–104. [CrossRef]
13. Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B.T.; Flammini, A.; Menczer, F. Uncovering coordinated networks on social media: Methods and case studies. *Proc. Int. AAAI Conf. Web Soc. Media* **2021**, *15*, 455–466. [CrossRef]

14. Mannocci, L.; Mazza, M.; Monreale, A.; Tesconi, M.; Cresci, S. Detection and characterization of coordinated online behavior: A survey. *ACM Comput. Surv.* **2026**, *Advance online publication*. [CrossRef]
15. Linvill, D.L.; Warren, P.L. Troll factories: Manufacturing specialized disinformation on Twitter. *Political Commun.* **2020**, *37*, 447–467. [CrossRef]
16. Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; Blackburn, J. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion Proceedings of the 2019 World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 218–226. [CrossRef]
17. Rauchfleisch, A.; Kaiser, J. The false positive problem of automatic bot detection in social science research. *PLoS ONE* **2020**, *15*, e0241045. [CrossRef] [PubMed]
18. Mosleh, M.; Rand, D.G. Measuring exposure to misinformation from political elites on Twitter. *Nat. Commun.* **2022**, *13*, 7144. [CrossRef] [PubMed]
19. Suárez-Serrato, P.; Roberts, M.E.; Davis, C.; Menczer, F. On the influence of social bots in online protests: Preliminary findings of a Mexican case study. In *Social Informatics: SocInfo 2016*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2016; Volume 10047, pp. 269–278. [CrossRef]
20. DiResta, R.; Goldstein, J.A.; Grossman, S. *Middle East Influence Operations: Observations Across Social Media Takedowns*; Project on Middle East Political Science (POMEPS); Stanford Internet Observatory: Stanford, CA, USA, 2021. Available online: https://pomeps.org/middle-east-influence-operations-observations-across-social-media-takedowns (accessed on 25 August 2026).
21. Recuero, R.; Soares, F.B.; Gruzd, A. Hyperpartisanship, disinformation and political conversations on Twitter: The Brazilian presidential election of 2018. *Proc. Int. AAAI Conf. Web Soc. Media* **2020**, *14*, 569–578. [CrossRef]
22. Avilés, G.H.; Reyes-Ortiz, J.A.; Mora-Gutiérrez, R.A.; Cuevas, J.P.; Alcántara, Ó.H. Systematic fine-tuning of transformer models for domain-specific misinformation detection in Spanish social media text. *Informatics* **2026**, *13*, 83. [CrossRef]
23. Valencia-Arias, A.; Arango-Botero, D.M.; Cardona-Acevedo, S.; Paredes Delgado, S.S.; Gallegos, A. Understanding the spread of fake news: An approach from the perspective of young people. *Informatics* **2023**, *10*, 38. [CrossRef]
24. Hardaker, C. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *J. Politeness Res.* **2010**, *6*, 215–242. [CrossRef]
25. Buckels, E.E.; Trapnell, P.D.; Paulhus, D.L. Trolls just want to have fun. *Personal. Individ. Differ.* **2014**, *67*, 97–102. [CrossRef]
26. Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; Gilbert, E. Still out there: Modeling and identifying Russian troll accounts on Twitter. In *Proceedings of the 12th ACM Conference on Web Science, Southampton, UK, 6–10 July 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–10. [CrossRef]
27. Badawy, A.; Ferrara, E.; Lerman, K. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Barcelona, Spain, 28–31 August 2018*; IEEE: New York, NY, USA, 2018; pp. 258–265. [CrossRef]
28. Gorwa, R.; Guilbeault, D. Unpacking the social media bot: A typology to guide research and policy. *Policy Internet* **2020**, *12*, 225–248. [CrossRef]
29. Assenmacher, D.; Clever, L.; Frischlich, L.; Quandt, T.; Trautmann, H.; Grimme, C. Demystifying social bots: On the intelligence of automated social media actors. *Soc. Media Soc.* **2020**, *6*, 2056305120939264. [CrossRef]
30. Martini, F.; Samula, P.; Keller, T.R.; Klinger, U. Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data Soc.* **2021**, *8*, 20539517211033566. [CrossRef]
31. Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; Cherry, C. SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*; Association for Computational Linguistics: San Diego, CA, USA, 2016; pp. 31–41. [CrossRef]
32. Küçük, D.; Can, F. Stance detection: A survey. *ACM Comput. Surv.* **2020**, *53*, 1–37. [CrossRef]
33. ALDayel, A.; Magdy, W. Stance detection on social media: State of the art and trends. *Inf. Process. Manag.* **2021**, *58*, 102597. [CrossRef]
34. Iyengar, S.; Lelkes, Y.; Levendusky, M.; Malhotra, N.; Westwood, S.J. The origins and consequences of affective polarization in the United States. *Annu. Rev. Political Sci.* **2019**, *22*, 129–146. [CrossRef]
35. Bassignana, E.; Basile, V.; Patti, V. HurtLex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). CEUR Workshop Proceedings*; Cabrio, E., Mazzei, A., Tamburini, F., Eds.; CEUR Workshop Proceedings: Turin, Italy, 2018; Volume 2253.
36. Nizzoli, L.; Tardelli, S.; Avvenuti, M.; Cresci, S.; Tesconi, M. Coordinated behavior on social media in 2019 UK general election. *Proc. Int. AAAI Conf. Web Soc. Media* **2021**, *15*, 443–454. [CrossRef]
37. Giglietto, F.; Righetti, N.; Rossi, L.; Marino, G. It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections. *Inf. Commun. Soc.* **2020**, *23*, 867–891. [CrossRef]

38. Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; Tesconi, M. 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, Perth, Australia, 3–7 April 2017*; International World Wide Web Conferences Steering Committee: Republic and Canton of Geneva, Switzerland, 2017; pp. 963–972. [CrossRef]
39. Greco, S.; Ishizaka, A.; Tasiou, M.; Torrisi, G. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Soc. Indic. Res.* **2019**, *141*, 61–94. [CrossRef]
40. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
41. Talpur, B.A.; O'Sullivan, D. Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter. *Informatics* **2020**, *7*, 52. [CrossRef]
42. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]
43. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* **1993**, *114*, 494–509. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.